ΠΡΑΚΤΙΚΑ



37ου Πανελλήνιου και 3ου Διεθνούς Συνεδρίου Στατιστικής

Λάρισα | 24-27 Απριλίου 2025

ΕΛΛΗΝΙΚΟ ΣΤΑΤΙΣΤΙΚΟ ΙΝΣΤΙΤΟΥΤΟ

ΗΡΑΚΛΕΙΤΟΣ

Ε. Σ. Ι.

ΕΛΛΗΝΙΚΟ
ΣΤΑΤΙΣΤΙΚΟ
ΙΝΣΤΙΤΟΥΤΟ

Ελληνικό Στατιστικό Ινστιτούτο
Greek Statistical Institute

ΠΡΑΚΤΙΚΑ

37ου Πανελληνίου & 3ου Διεθνούς Συνεδρίου Στατιστικής

Λάρισα, 24-27 Απριλίου 2025

**ΚΑΙΝΟΤΟΜΙΕΣ ΣΤΗ ΣΤΑΤΙΣΤΙΚΗ ΚΑΙ ΤΗΝ ΑΝΑΛΥΤΙΚΗ ΔΕΔΟΜΕΝΩΝ:
ΔΙΑΜΟΡΦΩΝΟΝΤΑΣ ΤΟ ΜΕΛΛΟΝ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ**

Οργάνωση

Ελληνικό Στατιστικό Ινστιτούτο

Πανεπιστήμιο Θεσσαλίας

ISSN 1792-2461

Περιεχόμενα

Πρόλογος (Preface in Greek)……7
Επιτροπές (Commitees in Greek)……9
Πρόγραμμα Συνεδρίου / Conference Program……...11

Εργασίες στα Ελληνικά – Papers in Greek

Αλεξανδρή Ε-Μ., Νάκας Χ.Θ.: Εξωτερική επαλήθευση μοντέλων πρόγνωσης καρδιαγγειακού επεισοδίου στη βιοτράπεζα του Ηνωμένου Βασιλείου ……..19
Καπετανοπούλου Π.: Διερεύνηση αξιοποίησης ανανεώσιμων πηγών ενέργειας από τα κράτη μέλη του ευρωπαικού νότου…….....37
Καρανικόλας Π., Φουσέκης Π., Τσακλίδη Γ.: Συνδεσιμότητα ανάμεσα στις τιμές εμπορευμάτων και τους δείκτες μεταβλητότητας τους. Εμπειρική διερεύνηση για τον χρυσό και το πετρέλαιο……..53
Λυμπερόπουλος Δ.: Μια επισκόπηση των δεικτών τιμών ακινήτων στις επίσημες στατιστικές……...65
Μαυρίδης Π., Μπάγκαβος Δ.: Έλεγχος καλής προσαρμογής για τη συνάρτηση κινδύνου……...85
Μηλιώνης Α., Βαρλάγκας Β-Θ.: Προγνωστική ικανότητα τεχνικών συναλλακτικών κανόνων στα χρηματιστήρια Αθηνών και Νέας-Υόρκης και αποτελεσματικότητα κεφαλαιαγορών ……103
Μηλιώνης Α., Γαλανόπουλος Ν.: Εγκυρότητα και επιπτώσεις των οικονομετρικών ελέγχων για την υπόθεση των αποτελεσματικών αγορών…….123
Μπουκουβάλα Α., Τσακλίδης Γ..: Μοντελοποίηση ενός ομογενούς μαρκοβιανού συστήματος ως ιξωδοελαστικού μέσου ……..143
Ξανθόπουλος Θ., Τσιλίκα Κ.: Εμπειρική διερεύνηση μέτρων συσχέτισης με τεχνικές και εργαλεία αναλυτικής δεδομένων ……..155
Παπαδοπούλου Ε., Κουγιουμτζής Δ.: Διερεύνηση υπερπαραμέτρων σε μοντέλα reservoir computing……...177
Σιόλου Ε., Συκιανάκης Ν., Χαλικιάς Μ.: Κατανομή Poisson: Ένα εναλλακτικό στατιστικό μοντέλο για την πρόβλεψη των ποδοσφαιρικών αποτελεσμάτων……..199
Τσαφής Δ., Αυλογιάρης Γ.: Οπτικόποιηση κύριων δεικτών απόδοσης (KPIs) στην καλαθοσφαίριση…….215

Εργασίες στα Αγγλικά – Papers in English

Azzalini A.: On the numerical representation of ordered factors used as explanatory variables: A proposal……..239
Chasiotis V., Karlis D.: On the estimation of complex statistics combining different surveys……...257
Fountoukidis K.G., Antzoulakos D.L., Rakitzis A.C.: …….277
Georgakis A., Papageorgiou V.E., Stamatellos G.: A multivariate fay-herriot model for small area estimation of tree diameter distribution…….297
Georgiou K., Mittas N., Angelis L.: Skill competitiveness and concentration in the software engineering market using prioritization approaches……..313
Ismyrlis V.: Ethical consideration for the use of data in statistics ……333

Kavargyris D-C., Gkorezis P., Mittas N., Angelis L.: Predicting skills in digital publication policies: A linear discriminant analysis and machine learning classification approach……..347
Kyrana Z., Papafilippou N., Pratsinakis E., Michaelidou A.M., Dordas C., Markos A., Menexes G.: Dimensionality reduction method for four-dimensional data cubes based on variability decomposition……369
Liotiri E., Ganopoulou M., Kavargyris D-C., Angelis L.: Skills in Policies: Understanding digital competencies in European policies through causal discovery……...383
Petridis P., Kugiumtzis D.: Comparison of regression and sparse regression models on multivariate time series of stochastic and deterministic systems ……..401

Πρόλογος

Το Πανελλήνιο Συνέδριο Στατιστικής αποτελεί μία από τις σημαντικότερες και μακροβιότερες επιστημονικές διοργανώσεις στον χώρο της Στατιστικής και των Πιθανοτήτων στην Ελλάδα. Διοργανώνεται σε ετήσια βάση, από το 1985, από το Ελληνικό Στατιστικό Ινστιτούτο (ΕΣΙ), μέλος της Ομοσπονδίας Ευρωπαϊκών Εθνικών Στατιστικών Εταιρειών (FENStat), σε συνεργασία με Ελληνικό ή Κυπριακό Ακαδημαϊκό Ίδρυμα. Η μακρόχρονη και συνεπής παρουσία του Συνεδρίου αντανακλά τη διαρκή ανάπτυξη της Στατιστικής Επιστήμης στην Ελλάδα και τη σημαντική συμβολή της στην εξέλιξη της σύγχρονης επιστημονικής έρευνας.

Το **37ο Πανελλήνιο & 3ο Διεθνές Συνέδριο Στατιστικής** πραγματοποιήθηκε στη Λάρισα, από τις **24 έως τις 27 Απριλίου 2025**, με κεντρικό θεματικό άξονα τις **«Καινοτομίες στη Στατιστική και την Αναλυτική Δεδομένων: Διαμορφώνοντας το Μέλλον της Πληροφορίας»**. Η θεματική του Συνεδρίου αντανακλά τον καθοριστικό ρόλο που διαδραματίζουν η Στατιστική και η Αναλυτική Δεδομένων στη σύγχρονη επιστημονική και τεχνολογική πραγματικότητα, σε μια εποχή κατά την οποία η εκθετική αύξηση της διαθέσιμης πληροφορίας δημιουργεί νέες δυνατότητες, αλλά και σύνθετες προκλήσεις για την εξαγωγή αξιόπιστης και χρήσιμης γνώσης.

Το Συνέδριο αποτέλεσε ένα διεπιστημονικό πεδίο συνάντησης ερευνητών και ερευνητριών, ακαδημαϊκών, επαγγελματιών και νέων επιστημόνων από ένα ευρύ φάσμα γνωστικών αντικειμένων. Στο πλαίσιό του παρουσιάστηκαν σύγχρονες θεωρητικές και μεθοδολογικές προσεγγίσεις, καθώς και εφαρμογές της Στατιστικής και της Αναλυτικής Δεδομένων σε διαφορετικούς επιστημονικούς και επαγγελματικούς χώρους. Η ανταλλαγή ιδεών, η παρουσίαση νέων ερευνητικών αποτελεσμάτων και η ανάπτυξη επιστημονικών συνεργασιών αποτέλεσαν βασικούς άξονες της διοργάνωσης.

Η διαχρονική πορεία του Συνεδρίου είναι στενά συνδεδεμένη με την εξέλιξη της Στατιστικής Επιστήμης στην Ελλάδα. Κάθε διοργάνωση λειτουργεί ως ένα δυναμικό σημείο αναφοράς για την αποτύπωση των σύγχρονων τάσεων στη θεωρία, στη μεθοδολογία και στις εφαρμογές της Στατιστικής και των Πιθανοτήτων. Παράλληλα, αναδεικνύει τη διαρκώς αυξανόμενη αλληλεπίδραση της Στατιστικής με άλλες επιστημονικές περιοχές, στις οποίες η αξιοποίηση δεδομένων και η ποσοτική ανάλυση αποτελούν βασικά εργαλεία για τη μελέτη σύνθετων προβλημάτων και τη λήψη τεκμηριωμένων αποφάσεων.

Ιδιαίτερη θέση στην ιστορία του Συνεδρίου κατέχει τα τελευταία χρόνια η ενίσχυση της διεθνούς του διάστασης. Η διεύρυνση της συμμετοχής ερευνητών και ερευνητριών από διαφορετικές χώρες και επιστημονικές κοινότητες ενισχύει τον διεθνή χαρακτήρα της διοργάνωσης και δημιουργεί πρόσθετες ευκαιρίες για ανταλλαγή γνώσης, εμπειριών και ερευνητικών προσεγγίσεων. Το 37ο Πανελλήνιο & 3ο Διεθνές Συνέδριο Στατιστικής εντάσσεται σε αυτή τη συνεχή προσπάθεια εξωστρέφειας και διεθνοποίησης, διευρύνοντας τον επιστημονικό διάλογο πέρα από τα εθνικά όρια.

Παράλληλα, το Συνέδριο αποτελεί έναν σημαντικό χώρο συνάντησης διαφορετικών γενεών της επιστημονικής κοινότητας. Καταξιωμένοι ερευνητές και ερευνήτριες και ακαδημαϊκοί δάσκαλοι έχουν τη δυνατότητα να παρουσιάσουν το ερευνητικό τους έργο και να μοιραστούν την εμπειρία και τη γνώση τους, ενώ νέοι και νέες ερευνητές και ερευνήτριες βρίσκουν ένα πρόσφορο περιβάλλον για την παρουσίαση των πρώτων ερευνητικών τους αποτελεσμάτων και την ένταξή τους στην ευρύτερη επιστημονική κοινότητα. Η συνύπαρξη διαφορετικών ερευνητικών προσεγγίσεων και επιπέδων εμπειρίας συμβάλλει ουσιαστικά στη δημιουργία ενός γόνιμου και δημιουργικού επιστημονικού διαλόγου.

Στο πλαίσιο του 37ου Πανελλήνιου & 3ου Διεθνούς Συνεδρίου Στατιστικής πραγματοποιήθηκε ένα πλούσιο επιστημονικό πρόγραμμα, το οποίο περιλάμβανε κεντρική και προσκεκλημένες ομιλίες, ειδικές συνεδρίες και παρουσιάσεις ερευνητικών εργασιών, καλύπτοντας ένα ευρύ φάσμα θεματικών περιοχών της σύγχρονης Στατιστικής και της Αναλυτικής Δεδομένων. Ιδιαίτερη έμφαση δόθηκε στις σύγχρονες τεχνολογικές και μεθοδολογικές εξελίξεις, καθώς και στις νέες δυνατότητες που προσφέρουν η υπολογιστική στατιστική, η τεχνητή νοημοσύνη, η μηχανική μάθηση και οι σύγχρονες τεχνικές ανάλυσης δεδομένων.

Στο πλαίσιο της διοργάνωσης πραγματοποιήθηκε, επίσης, Εκπαιδευτικό–Επιμορφωτικό Σεμινάριο με θέμα «**Εφαρμογές της Μηχανικής Προτροπών (Prompt Engineering) στη Στατιστική Ανάλυση Δεδομένων»**, το οποίο συντονίστηκε από τον συνάδελφο Δημήτρη Πανάρετο. Η ενσωμάτωση θεματικών που συνδέονται με τις σύγχρονες τεχνολογίες και την αξιοποίηση νέων εργαλείων στην ανάλυση δεδομένων αναδεικνύει τη σημασία της συνεχούς επικαιροποίησης των γνώσεων και των δεξιοτήτων των επιστημόνων του πεδίου.

Στα χέρια σας κρατάτε τα **Πρακτικά του 37ου Πανελλήνιου & 3ου Διεθνούς Συνεδρίου Στατιστικής**, τα οποία αποτελούν την επιστημονική αποτύπωση μέρους της ερευνητικής δραστηριότητας που παρουσιάστηκε στο πλαίσιο της διοργάνωσης. Στον παρόντα τόμο περιλαμβάνονται **23 επιστημονικές εργασίες**, οι οποίες υποβλήθηκαν και αξιολογήθηκαν στο πλαίσιο της προβλεπόμενης διαδικασίας επιστημονικής κρίσης.

Η Επιτροπή Έκδοσης των Πρακτικών του ΕΣΙ εκφράζει τις θερμές ευχαριστίες της προς τα μέλη της Επιτροπής Κρίσης, κ.κ. **Ε. Ανδουλάκη, Ι. Βόντα, Δ. Δημητρίου, Β. Καρυώτη, Π. Μαραβελάκη, Δ. Μπάγκαβο, Α. Μποζίκα, Π. Μπομποτά, Ι. Νίκα, Κ. Ντότση, Γ. Παπαγεωργίου, Π. Παπασταμούλη, Κ. Πολίτη, Α. Ρακιτζή, Α. Σαχλά και Β. Χασιώτη**, για τον χρόνο που αφιέρωσαν και για την επιμελημένη και προσεκτική αξιολόγηση των υποβληθεισών εργασιών. Η συμβολή τους υπήρξε ιδιαίτερα σημαντική για τη διασφάλιση της επιστημονικής ποιότητας του παρόντος τόμου.

Η σειρά παρουσίασης των εργασιών στον παρόντα τόμο είναι αλφαβητική, με βάση το επώνυμο του πρώτου συγγραφέα.

Κλείνοντας, εκφράζουμε την πεποίθηση ότι τα Πρακτικά του 37ου Πανελλήνιου & 3ου Διεθνούς Συνεδρίου Στατιστικής θα αποτελέσουν ένα χρήσιμο επιστημονικό τεκμήριο, συμβάλλοντας στη διάδοση σύγχρονων ερευνητικών αποτελεσμάτων και στην περαιτέρω ανάπτυξη του επιστημονικού διαλόγου γύρω από τη Στατιστική, τις Πιθανότητες και την Αναλυτική Δεδομένων. Παράλληλα, ευχόμαστε η παρούσα έκδοση να αποτελέσει αφετηρία για νέες ερευνητικές αναζητήσεις, επιστημονικές συνεργασίες και δημιουργικές προσεγγίσεις στις προκλήσεις που θέτει η διαρκώς εξελισσόμενη εποχή της πληροφορίας.

Η σειρά παρουσίασης των εργασιών στον παρόντα τόμο είναι αλφαβητική με βάση το επώνυμο του πρώτου συγγραφέα.

Οι Συντονιστές των Πρακτικών

Γεώργιος Ηλιόπουλος, Αλέξανδρος Καραγρηγορίου, Σωτηρία Μαλεφάκη,
Απόστολος Μπατσίδης, Πολυχρόνης Οικονόμου, Μιλτιάδης Χαλικιάς &
Γεώργιος Ψαρράκος

Επιτροπές

Τοπική Οργανωτική Επιτροπή

Αγγελής Ε., Καθηγητής, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
Γεωργακόπουλος Σ., Επίκουρος Καθηγητής, Πανεπιστήμιο Θεσσαλίας
Γεωργίου Α., Καθηγητής, Πανεπιστήμιο Μακεδονίας
Δρακόπουλος Β., Αναπληρωτής Καθηγητής,Πανεπιστήμιο Θεσσαλίας
Καραγρηγορίου Α., Καθηγητής, Πανεπιστήμιο Πειραιώς
Καστέλλη Ι., Επίκουρη Καθηγήτρια, Πανεπιστήμιο Θεσσαλίας
Κοντού Π., Επίκουρη Καθηγήτρια, Πανεπιστήμιο Θεσσαλίας
Μαλεφάκη Σ., Αναπληρώτρια Καθηγήτρια, Πανεπιστήμιο Πατρών
Μπατσίδης Α., Αναπληρωτής Καθηγητής, Πανεπιστήμιο Ιωαννίνων
Μπομποτάς Π., Επίκουρος Καθηγητής, Πανεπιστήμιο Πατρών
Μπούσια Α., Επίκουρη Καθηγήτρια, Πανεπιστήμιο Θεσσαλίας
Μωυσιάδης Χ., Ομότιμος Καθηγητής, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης.
Νάκας Χ., Καθηγητής, Πανεπιστήμιο Θεσσαλίας
Σαχλάς Α., Επίκουρος Καθηγητής, Πανεπιστήμιο Θεσσαλίας
Συρακούλης Κ., Καθηγητής, Πανεπιστήμιο Θεσσαλίας
Τσακλίδης Γ., Καθηγητής, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
Τσέλιος Δ., Αναπληρωτής Καθηγητής, Πανεπιστήμιο Θεσσαλίας

Επιστημονική Επιτροπή

Αγγελής Ε., Καθηγητής, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
Αυλογιάρης Γ., Επίκουρος Καθηγητής, Πανεπιστήμιο Δυτικής Μακεδονίας
Βαμβακάρη Μ., Καθηγήτρια, Χαροκόπειο Πανεπιστήμιο
Δαφνής Σ., Αναπληρωτής Καθηγητής, Πανεπιστήμιο Αιγαίου
Δρακόπουλος Β., Αναπληρωτής Καθηγητής,Πανεπιστήμιο Θεσσαλίας
Ευαγγελάρας Χ., Αναπληρωτής Καθηγητής, Πανεπιστήμιο Πειραιώς
Ηλιόπουλος Γ., Καθηγητής, Πανεπιστήμιο Πειραιώς
Καραγρηγορίου Α., Καθηγητής, Πανεπιστήμιο Πειραιώς
Καρλής Δ., Καθηγητής, Οικονομικό Πανεπιστήμιο Αθηνών
Κουνιάς Σ., Ομότιμος Καθηγητής, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών, Επίτιμος Πρόεδρος του ΕΣΙ
Κούτρας Μ., Καθηγητής, Πανεπιστήμιο Πειραιώς
Μαλεφάκη Σ., Αναπληρώτρια Καθηγήτρια, Πανεπιστήμιο Πατρών
Μηλιένος Φ., Αναπληρωτής Καθηγητής, Πάντειο Πανεπιστήμιο Κοινωνικών και Πολιτικών Επιστημών
Μπατσίδης Α., Αναπληρωτής Καθηγητής, Πανεπιστήμιο Ιωαννίνων
Μπουρνέτας Α., Καθηγητής, Πανεπιστήμιο Αθηνών
Μπούσια Α., Επίκουρη Καθηγήτρια, Πανεπιστήμιο Θεσσαλίας
Μωυσιάδης Χ., Ομότιμος Καθηγητής, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
Ντζούφρας Ι., Καθηγητής, Οικονομικό Πανεπιστήμιο Αθηνών
Οικονόμου Π., Αναπληρωτής Καθηγητής, Πανεπιστήμιο Πατρών
Παπαδόπουλος Γ., Καθηγητής, Γεωπονικό Πανεπιστήμιο Αθηνών
Παπαϊωάννου Τ., Ομότιμος Καθηγητής, Πανεπιστήμιο Πειραιώς και Πανεπιστήμιο Ιωαννίνων και Επίτιμος Πρόεδρος του ΕΣΙ

Ρακιτζής Α., Επίκουρος Καθηγητής, Πανεπιστήμιο Πειραιώς
Συρακούλης Κ., Καθηγητής, Πανεπιστήμιο Θεσσαλίας
Τασιάς Κ., Επίκουρος Καθηγητής, Πανεπιστήμιο Δυτικής Μακεδονίας
Τζαβελάς Γ., Αναπληρωτής Καθηγητής, Πανεπιστήμιο Πειραιώς
Τριανταφύλλου Ι., Αναπληρωτής Καθηγητής, Πανεπιστήμιο Πειραιώς
Τσαγρής, Μ., Αναπληρωτής Καθηγητής,Πανεπιστήμιο Κρήτης
Τσαϊρίδης, Χ., Επίκουρος Καθηγητής, Πανεπιστήμιο Θράκης
Τσακλίδης Γ., Καθηγητής, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
Τσέλιος Δ., Αναπληρωτής Καθηγητής, Πανεπιστήμιο Θεσσαλίας
Τσιλίκα Κ., Αναπληρώτρια Καθηγήτρια, Πανεπιστήμιο Θεσσαλίας
Φουσκάκης Δ., Καθηγητής, Εθνικό Μετσόβειο Πολυτεχνείο
Χαλικιάς Μ., Καθηγητής, Πανεπιστήμιο Δυτικής Αττικής
Χατζηπαντελής Θ., Καθηγητής, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
Ψαρράκος Γ., Αναπληρωτής Καθηγητής, Πανεπιστήμιο Πειραιώς

Διαλέξεις - Presentations

Κεντρικοί Ομιλητές – Plenary Talks

Adelchi Azzalini, Univ. of Padua, Italy
Dimitrios Karlis, Athens Univ. of Economics & Business, Greece
Vasileios Maroulas, Univ. of Tennessee, Knoxville,USA
Konstantinos Tasias, Univ. of Western Macedonia, Greece
Ciprian Tudor, University of Lille, France
Vlad S. Barbu, University of Rouen – Normandy, France

Προσκεκλημένοι Ομιλητές – Invited Talks

2025 Eleneio Award for the Best PhD Thesis for 2023-2024

Alex Pasiouras National and Kapodistrian Univ. of Athens, Greece

Πρόγραμμα Συνεδρίου

Πέμπτη – Thursday 24/4/25
ΑΜΦΙΘΕΑΤΡΟ 1 – AUDITORIUM 1

14:00 - 15:00	Εγγραφή Συνέδρων – Registration of the Participants
15:00 – 15:30	**Έναρξη του Συνεδρίου – Conference Opening** **Προεδρεύων/Chair: Κ. Συρακούλης/K. Sirakoulis**
	Χαιρετισμοί – Welcome Addresses
15:30 - 17:00	**Plenary Talks I.0** **Chair: A. Batsidis**
15:30 - 16:00	**A. Azzalini** ***Cacoullos Memorial Lecture 2025:*** *Two formulations for the construction of multivariate families of distributions*
16:00 - :16:30	**V. Maroulas** *The shape of uncertainty*
16:30 – 17:00	**D. Karlis** *On modelling high dimensional time series*
17:00 - 17:30	***ΔΙΑΛΕΙΜΜΑ – ΚΑΦΕΣ – COFFEE BREAK***
17:30 - 19:10	**Session I.1 Statistical Learning and Applications** **Chair: D. Karlis**
17:30-17:50	**D. Karatzios** *Machine learning approaches for monitoring and optimizing water treatment processes*
17:50-18:10	**G. Papageorgiou** *Text mining algorithm from scratch using deep learning*
18:10-18:30	**E. Liotiri** *Skills in policies: Understanding digital competencies in European policies through causal discovery*
18:30-18:50	**D. C. Kavargyris** *Predicting skills in digital publication policies: A linear discriminant analysis and machine learning classification approach*
18:50-19:10	**K. Bourazas** *Network-informed Bayesian anomaly detection by using Gaussian processes*
19:15	***WELCOME COCKTAIL***

Παρασκευή - Friday 25/4/25

Αίθουσα /Room 1 09:00-10:20	**Session / Συνεδρία II.1 Actuarial Science** **Chair: G. Psarrakos**
09:00-09:20	**S. Zampelis** *Bayesian variable selection methods for insurance lapsation modelling*
09:20-09:40	**G. Pitselis** *Exponential regression models for estimating mortality rates*
09:40-10:00	**A. Bozikas** *From survival beliefs to life insurance premiums: A parametric approach*
10:00-10:20	**S. Tzaninis** *Simulation for ruin probabilities in a renewal risk model*
10:20-10:40	**J.D. Economides** On a generalized Gerber-Shiu function assuming dependent claim sizes and inter-claim times
Αίθουσα /Room 2 09:00-10:20	**Session II.2 Applied Statistics** **Chair: A. Georgiou**
09:00-09:20	**K. M. Kontouli** *Bridging single arm studies with individual participant data in network meta-analysis of randomized controlled trials: A simulation study*
09:20-09:40	**T. Daras** *SARIMA models in weather parameters' forecasting and climate change: A case study four Greek cities*
09:40-10:00	**E. N. Kalligeris** *Gaussian processes for uncertainty quantification of complex atmospheric models*
10:00-10:20	**K. Georgiou** Skill competitiveness and concentration in the software engineering market using prioritization approaches
10:20-10:40	**D. Dawidowicz, G.S. Piorico** *Assessment of gender inequality in EU countries using the Gender Financial Equality Index*
10:45-11:15	***ΔΙΑΛΕΙΜΜΑ – ΚΑΦΕΣ – COFFEE BREAK***
Αίθουσα /Room 1 11:15-12:55	**Session II.3 Applied Probability** **Chair: V. Maroulas**
11:15-11:35	**G. Psarrakos** *On the stochastic variability order and delta method*
11:35-11:55	**A. Karagrigoriou** *Multicomponent stress-strength reliability: Theory and practice*
11:55-12:15	**S. Dafnis** *On a new family of r-modified reliability systems*
12:15-12:35	**I. Triantafyllou** *On the age replacement policy of consecutive-type structures with non-identical components*
12:35-12:55	**C. Papadopoulos** *A new approximation in renewal function of Weibull distribution*
Αίθουσα /Room 2 11:15-12:55	**Συνεδρία II.4 Εφαρμοσμένη Στατιστική** **Προεδρεύουσα: Κ. Τσιλίκα**
11:15-11:35	**Π. Καπετανοπούλου** *Διερεύνηση αξιοποίησης ανανεώσιμων πηγών ενέργειας από τα κράτη-μέλη του ευρωπαϊκού νότου*
11:35-11:55	**Ε. Μ. Αλεξανδρή** *Εξωτερική επαλήθευση μοντέλων πρόγνωσης καρδιαγγειακού επεισοδίου στα δεδομένα της UK Biobank*
11:55-12:15	***Ε. Στεργίου*** *Ανάλυση συστάδων και ποιοτική συγκριτική ανάλυση (QCA): Πλεονεκτήματα και μειονεκτήματα*
12:15-12:35	**Π. Οικονόμου** *Στοχαστική βελτιστοποίηση δυναμικής τιμολόγησης σε συστήματα κρατήσεων*
12:35-12:55	**Θ. Ξανθόπουλος**

	Εμπειρική διερεύνηση μέτρων συσχέτισης με τεχνικές και εργαλεία αναλυτικής δεδομένων
13:00 – 15:00	***ΜΕΣΗΜΒΡΙΝΗ ΔΙΑΚΟΠΗ-LUNCH BREAK***
Αίθουσα/ Room 1 15:00-16:20	**Session II.5 Statistics** **Chair: S. Dafnis**
15:00-15:20	**V. Ismyrlis** *Ethical considerations for the use of data in Statistics*
15:20-15:40	**V. Chasiotis** *On the estimation of complex statistics combining different surveys*
15:40-16:00	**A. Rakitzis** *On the performance of a control chart for rates and proportions when process parameters are estimated*
16:00-16:20	**K. Fountoukidis** *Two modified CUSUM-type control charts for joint monitoring origin and scale of two-parameter exponential distribution*
Αίθουσα/ Room 2 15:00-16:20	**Ειδική Συνεδρία ΙΙ.6 Στατιστική για Χρηματοοικονομικά** **Προεδρεύων: Α. Μηλιώνης**
15:00-15:20	**A. Tsioutsios** *Dynamic connectedness analysis of the balance of payments: A disaggregated perspective*
15:20-15:40	**Β. Βαρλάγκας** *Προγνωστική ικανότητα τεχνικών συναλλακτικών κανόνων στο Χρηματιστήριο Αθηνών, έλεγχος data-snooping και αποτελεσματικότητα κεφαλαιαγορών*
15:40-16:00	**Ν. Γαλανόπουλος** *Εγκυρότητα και επιπτώσεις των οικονομετρικών ελέγχων για την υπόθεση των αποτελεσματικών αγορών*
16:00-16:20	**Α. Μηλιώνης** *Μικροδομή μοντέλου αγοράς στο Χρηματιστήριο Αθηνών την περίοδο της οικονομικής κρίσης. Απρόσμενα αποτελέσματα και ερμηνεία τους*
16:30 – 17:00	***ΔΙΑΛΕΙΜΜΑ – ΚΑΦΕΣ – COFFEE BREAK***
Auditorium 1 17:00 - 18:00	**Invited Talks** **Chair: G. Iliopoulos**
17:00 - 17:30 **17:30 - 18:00**	**V.S. Barbu** *Goodness-of-fit tests based on Tsallis and Kaniadakis entropies* **K. Tasias** *Optimization approaches for aerial firefighting resource allocation*
Auditorium 1 18:00 – 18:30	**Ελένειο Βραβείο / Eleneio Award** **Chair: I. Vonta**
18:00 – 18:30	**A. Pasiouras** *Bayesian inversion techniques for stochastic differential equations*
20:30	***ΔΕΙΠΝΟ – CONFERENCE DINNER***

Σάββατο-Saturday 26/4/25

Αίθουσα/ Room 1 10:00-11:20	**Ειδική Συνεδρία ΙΙΙ.1 Ανάλυση Πολυμεταβλητών Χρονοσειρών (Μέρος Α) Προεδρεύων: Δ. Κουγιουμτζής**
10:00-10:20	**Γ. Μυλλής** *Παράλληλη υπολογιστική ροή εργασίας (HPC pipeline) για την επιτάχυνση της εκπαίδευσης μοντέλων και την πρόβλεψη χρονοσειρών με Apache Spark, MPI και GPU acceleration*
10:20-10:40	**Α. Τράσιας** *Γραμμικές και μη γραμμικές αιτιακές σχέσεις σε πολυμεταβλητές χρονοσειρές: Μελέτη περίπτωσης της σχέσης μεταξύ ειδήσεων και χρηματοοικονομικών δεικτών*
10:40-11:00	**Α. Φωτιάδης** *Αραιή μοντελοποίηση χρονοσειρών υψηλής διάστασης*
11:00-11:20	**Ε. Παπαδοπούλου** *Διεύρυνση υπερπαραμέτρων σε μοντέλα reservoir computing*
Αίθουσα/Room 2 10:00-11:20	**Συνεδρία ΙΙΙ.2 Στατιστική Προεδρεύων: Ι. Τριανταφύλλου**
10:00-10:20	**Χ. Ευαγγελάρας** *Παραγοντικοί σχεδιασμοί δύο επιπέδων και J–χαρακτηριστικά*
10:20-10:40	**Μ. Χαλικιάς** *Σχεδιασμοί επαναλαμβανόμενων μετρήσεων: Η ύπαρξη ομάδας ελέγχου/αγωγής ελέγχου (control treatment) σε μοντέλο δύο αγωγών*
10:40-11:00	**Μ. Παπαγεωργίου** *Δειγματοληψία σε συμμετρικά δίκτυα: Αμεροληψία εκτιμητριών*
11:00-11:20	**Α. Γεωργάκης** *Επιλογή μεγέθους δείγματος και δειγματοληπτικών επιφανειών για την εκτίμηση του ξυλαποθέματος*
11:30 - 12:00	***ΔΙΑΛΕΙΜΜΑ – ΚΑΦΕΣ – COFFEE BREAK***
Αίθουσα /Room 1 12:00– 13:00	**Ειδική Συνεδρία ΙΙΙ.3 Ανάλυση Πολυμεταβλητών Χρονοσειρών (Μέρος Β) Προεδρεύων: Δ. Κουγιουμτζής**
12:00-12:20	**D. Kugiumtzis** *Information measures for balancing redundancy and relevance in causality analysis of multivariate time series*
12:20-12:40	**P. Petridis** *Comparative simulation analysis on stochastic and deterministic systems of multivariate time series*
12:40-13:00	**G. Gkyzis** *Instantaneous causality and structural changes in multivariate times series systems: Detection and quantification of instantaneous relationships*
Αίθουσα /Room 1 13:00 -13:30	**Invited Talks Chair: A. Karagrigoriou**
13:00 – 13:30	**C. Tudor** *Inference for the nonlinear stochastic heat equation based on observations in time or in space*
13:30 – 15:30	***ΜΕΣΗΜΒΡΙΝΗ ΔΙΑΚΟΠΗ-LUNCH BREAK***
Αίθουσα/ Room 1 15:30 – 17:10	**Session III.4: Statistics Chair: H. Evangelaras**
15:30-15:50	**C. Meselidis** *Censored data analysis via generalized divergence measures*
15:50-16:10	**T. Gkelsinis** *Validating credit risk transition models: A semi-Markov approach using divergence measures*
16:10-16:30	**G. Iliopoulos** *A direct proof of the monotonicity of power of classical t- and F-tests with respect to sample size*
16:30-16:50	**C. Kitsos** *Sampling from the γ-order generalized normal Nγ(μ,σ2) using the ziggurat algorithm*
16:50-17:10	**I. Stamatiou** *On the γ-order generalized distributions*

Αίθουσα/ Room 2 15:30-17:10	**Συνεδρία ΙΙΙ.5 Στατιστική** **Προεδρεύων: Κ. Συρακούλης**
15:30-15:50	**Δ. Λυμπερόπουλος** *Μια επισκόπηση των δεικτών τιμών ακινήτων στις Επίσημες Στατιστικές*
15:50-16:10	**Θ. Χατζηπαντελής** *Ο ενημερωμένος και κριτικά σκεπτόμενος πολίτης*
16:10-16:30	**Σ. Καφφές** *Πειραματική αξιολόγηση σοβαρού παιχνιδιού για τη διδασκαλία της μεθόδου CPM στη διαχείριση έργων*
16:30-16:50	**Π. Καρανικόλας** *Συνδεσιμότητα ανάμεσα στις τιμές εμπορευμάτων και τους δείκτες μεταβλητότητας τους. Εμπειρική διερεύνηση για τον χρυσό και το πετρέλαιο*
16:50 – 17:10	**Δ. Τσαφής** *Μέθοδοι οπτικοποίησης κύριων δεικτών απόδοσης (KPIs) στην καλαθοσφαίριση*
17:15 - 17:45	***ΔΙΑΛΕΙΜΜΑ – ΚΑΦΕΣ – COFFEE BREAK***
17:30 -19:15	***POSTER SESSION***
1	**P. Bobotas:** *A unified model for sampling schemes from exponential distributions with common location parameter*
2	**A. Bousia:** *Competitive K-nearest neighbors algorithm for electric vehicles energy dempand forecasting*
3	**K. Exarchou-Kouveli, P. Kontou, P. Bagos:** *Bayesian t-tests in STATA*
4	**P. Kontou, G. Tsionkis, E. Andronidou, G. Braliou, P. Bagos**: *Meta-analysis using IC50 values: Approaches and applications*
5	**Z. Kyrana, N. Papafilippou, E. Pratsinakis, A.-M. Michaelidou, C. Dordas, A. Markos, G. Menexes:** *Dimensionality reduction method for four-dimensional data cubes*
6	**E. Nikolitsa, P. Kontou, P. Bagos:** *metacp: A versatile software package for combining dependent or independent p-values*
7	**T. Ntoumani:** *Data analysis methods for cancer genomics*
8	**E. Paraskevadaki, A. Dania:** *A Catastrophe Theory model for research training in higher education: A complex systems approach to educational change*
9	**E. Paraskevadaki:** *Limited statistical thinking in students and professionals: Challenges and implications*
10	**Ε. Σιόλου, Ν. Συκιανάκης, Μ. Χαλικιάς**: *Κατανομή Poisson: Ένα εναλλακτικό στατιστικό μοντέλο για την πρόβλεψη των ποδοσφαιρικών αποτελεσμάτων*

Κυριακή-Sunday 27/4/25

Αίθουσα/ Room 1 09:30-11:10	**Special Session IV.1 Generalized Markovian Mixed Models** **Chair: A. Papadopoulou**
09:30-09:50	**V. Dimitriou** *On certain aspects of a multilevel manpower planning model in continuous time*
09:50-10:10	**A. Georgiou** *Efficiency evaluation of normative Markovian models in hierarchical systems*
10:10-10:30	**Z. Kyritsis** *A review of Markov and semi-Markov reward processes: Theoretical foundations, methodologies, and applications*
10:30-10:50	**A. Makrides** *Modelling earthquake risk using stochastic hidden Markov processes*
10:50-11:10	**A. Papadopoulou** *A semi-Markov reward model for cost evaluation and capacity planning in healthcare systems*
Αίθουσα/ Room 2 09:30-11:10	**Συνεδρία IV.2 Στατιστική και Εφαρμοσμένες Πιθανότητες** **Προεδρεύων: Π. Οικονόμου**
09:30-09:50	**Α. Τσάτση** *Έλεγχοι πολυδιάστατης κανονικότητας με 2-βηματικά μονότονα ελλιπή δεδομένα*
09:50-10:10	**Π. Μαυρίδης** *Έλεγχος καλής προσαρμογής για τη συνάρτηση ρυθμού επικινδυνότητας*

10:10-10:30	**Α. Μπατσίδης** *Στατιστική συμπερασματολογία υπό μεροληπτική δειγματοληψία*
10:30-10:50	**Λ. Κανελλόπουλος** *Μελέτη ιδιοτήτων γήρανσης μιας οικογένειας συνεχών μονοδιάστατων κατανομών με εφαρμογές στη θεωρία στοχαστικών διατάξεων*
10:50-11:10	**Ά. Μπουκουβάλα** *Μοντελοποίηση ενός ομογενούς μαρκοβιανού συστήματος ως ιξωδοελαστικού μέσου*
11:15 - 11:45	***ΔΙΑΛΕΙΜΜΑ – ΚΑΦΕΣ – COFFEE BREAK***
11:45	**ΛΗΞΗ ΣΥΝΕΔΡΙΟΥ – CONFERENCE CLOSING**

Εργασίες στα Ελληνικά

Papers in Greek

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 37ου Πανελληνίου & 3ου Διεθνούς Συνεδρίου Στατιστικής (2025), σελ.19-35

ΕΞΩΤΕΡΙΚΗ ΕΠΑΛΗΘΕΥΣΗ ΜΟΝΤΕΛΩΝ ΠΡΟΓΝΩΣΗΣ ΚΑΡΔΙΑΓΓΕΙΑΚΟΥ ΕΠΕΙΣΟΔΙΟΥ ΣΤΗ ΒΙΟΤΡΑΠΕΖΑ ΤΟΥ ΗΝΩΜΕΝΟΥ ΒΑΣΙΛΕΙΟΥ

Ελευθερία Μαρία Αλεξανδρή και Χρήστος Θ Νάκας
Εργαστήριο Βιομετρίας, Τμήμα Γεωπονίας Φυτικής Παραγωγής & Αγροτικού Περιβάλλοντος, Πανεπιστήμιο Θεσσαλίας
elefalexandri@uth.gr & cnakas@uth.gr

ΠΕΡΙΛΗΨΗ

Τα μοντέλα πρόβλεψης κινδύνου καρδιαγγειακού επεισοδίου (CVD risk prediction models) αποτελούν χρήσιμα εργαλεία στην κλινική πρακτική. Ωστόσο, ο ορισμός του συμβάντος CVD διαφέρει ευρέως στη βιβλιογραφία. Στόχος της ανάλυσης είναι να υπογραμμίσει τις προκλήσεις της χρήσης τυποποιημένων μοντέλων κινδύνου σε διάφορους πληθυσμούς και τη σημασία της μεταβλητότητας του ορισμού των συμβάντων κατά την ερμηνεία των προβλέψεων κινδύνου CVD σε πραγματικές εφαρμογές. Στην παρούσα μελέτη, τέσσερα ευρέως χρησιμοποιούμενα μοντέλα εφαρμόστηκαν ανά φύλο στο εκτεταμένο σύνολο δεδομένων της Βιοτράπεζας HB και αξιολογήθηκαν ως προς την ικανότητα διάκρισης και αξιοπιστίας των εκτιμήσεων κινδύνου, χρησιμοποιώντας και τους τέσσερις ορισμούς συμβάντων. Προέκυψαν αξιοσημείωτες διαφορές μεταξύ των διαφορετικών ορισμών του συμβάντος και συστηματικά καλύτερες προβλέψεις στις γυναίκες. Συμπεραίνουμε ότι η χρήση στατικών μοντέλων που δεν αξιοποιούν δεδομένα σε πραγματικό χρόνο έχουν περιορισμένη χρησιμότητα στην πράξη.

Λέξεις Κλειδιά: Αξιοπιστία, Διάκριση, Επαλήθευση, Στατιστικά Μοντέλα Πρόβλεψης

1. ΕΙΣΑΓΩΓΗ

Η καρδιαγγειακή νόσος παραμένει ένας καθοριστικός παράγοντας θνησιμότητας και νοσηρότητας σε παγκόσμιο επίπεδο επιβάλλοντας έναν σημαντικό περιορισμό σε άτομα και συστήματα υγειονομικής περίθαλψης. Καθώς η καρδιαγγειακή νόσος αναπτύσσεται αθόρυβα με την πάροδο των ετών, η αποτελεσματική πρόβλεψη του κινδύνου είναι ζωτικής σημασίας για τον έγκαιρο εντοπισμό ατόμων με υψηλό κίνδυνο και την εφαρμογή έγκαιρων παρεμβάσεων πρόληψης. Τα μοντέλα πρόβλεψης κινδύνου καρδιαγγειακής νόσου έχουν αναδειχθεί ως χρήσιμα εργαλεία στην κλινική πρακτική προσφέροντας μια δομημένη και τεκμηριωμένη προσέγγιση για την αξιολόγηση της πιθανότητας ενός ατόμου να εμφανίσει καρδιαγγειακό επεισόδιο σε συγκεκριμένο χρονικό διάστημα. Τα μοντέλα αυτά ενσωματώνουν πολλαπλούς παράγοντες κινδύνου, επιτρέποντας στους γιατρούς να λαμβάνουν αποφάσεις προσαρμοσμένες στο μοναδικό προφίλ υγείας του ασθενούς.

Μία από τις κύριες χρησιμότητες των μοντέλων πρόβλεψης κινδύνου καρδιαγγειακής νόσου έγκειται στην ικανότητα τους να ταξινομούν τα άτομα με βάση το συνολικό τους κίνδυνο. Μοντέλα όπως το Framingham Risk Score, το QRISK και το ASCVD Risk Calculator ενοποιούν δεδομένα όπως η ηλικία, το φύλο, το κάπνισμα, τα επίπεδα χοληστερόλης και η αρτηριακή πίεση, για να δημιουργήσουν εξατομικευμένες εκτιμήσεις κινδύνου. Αυτή η ταξινόμηση βοηθά τους γιατρούς στον καθορισμό των απαιτούμενων παρεμβάσεων, από την τροποποίηση του τρόπου ζωής και τη φαρμακολογική θεραπεία έως τη συχνότερη παρακολούθηση. Με την εστίαση των πόρων σε άτομα υψηλού κινδύνου, τα μοντέλα αυτά ενισχύουν την αποτελεσματικότητα της παροχής υγειονομικής περίθαλψης ενώ ελαχιστοποιούν τις περιττές παρεμβάσεις σε πληθυσμούς χαμηλού κινδύνου.

Επιπλέον, τα μοντέλα αυτά παρέχουν ένα ισχυρό πλαίσιο για τη λήψη κοινών αποφάσεων μεταξύ γιατρών και ασθενών. Η ποσοτικοποίηση του κινδύνου μπορεί να βοηθήσει τους ασθενείς να κατανοήσουν καλύτερα τα δυνητικά οφέλη των προληπτικών μέτρων προωθώντας τη μεγαλύτερη συμμόρφωση στα θεραπευτικά σχέδια. Για παράδειγμα, ένας ασθενής με 15% 10-ετή κίνδυνο εμφάνισης καρδιαγγειακής νόσου μπορεί να έχει περισσότερα κίνητρα να υιοθετήσει πιο υγιεινό τρόπο ζωής ή να ξεκινήσει θεραπεία με στατίνες. Επίσης, τα μοντέλα πρόβλεψης κινδύνου καρδιαγγειακού επεισοδίου χρησιμεύουν στην καθοδήγηση πρωτοβουλιών δημόσιας υγείας. Σε επίπεδο πληθυσμού, βοηθούν τους υπεύθυνους χάραξης πολιτικής στον εντοπισμό ομάδων υψηλού επιπολασμού και στην ιεράρχηση της κατανομής πόρων για προληπτικά προγράμματα. Στην έρευνα, παρέχουν πολύτιμες γνώσεις σχετικά με την αλληλεπίδραση διαφόρων παραγόντων κινδύνου συμβάλλοντας στην ανάπτυξη νέων θεραπευτικών στρατηγικών και κατευθυντήριων γραμμών πρόληψης.

Παρά τη χρησιμότητά τους, τα μοντέλα πρόβλεψης κινδύνου καρδιαγγειακής νόσου φέρουν περιορισμούς όπως μεταβλητότητα στην ακρίβεια των εκτιμήσεων του κινδύνου μεταξύ διάφορων πληθυσμών και το ενδεχόμενο να παραβλέπουν αναδυόμενους παράγοντες κινδύνου. Ωστόσο, πρόοδοι στην επιστήμη των δεδομένων και τη μηχανική μάθηση είναι έτοιμοι να βελτιώσουν την προβλεπτική τους ικανότητα. Συνολικά, τα μοντέλα πρόβλεψης κινδύνου καρδιαγγειακού επεισοδίου είναι απαραίτητα στην σύγχρονη υγειονομική περίθαλψη καθώς γεφυρώνουν το κενό μεταξύ έρευνας και πρακτικής εφαρμογής. Η μεταφερσιμότητα των μοντέλων αυτών αναφέρεται στην ικανότητά τους να διατηρούν την ακρίβεια και την αξιοπιστία τους όταν εφαρμόζονται σε πληθυσμούς διαφορετικούς από αυτούς που χρησιμοποιήθηκαν για την ανάπτυξή τους. Όλα τα μοντέλα προέρχονται από διαφορετικές κοόρτες και ίσως να μην αποδίδουν βέλτιστα σε πληθυσμούς με διαφορετικά δημογραφικά, γενετικά ή υγειονομικά πλαίσια. Για να εξασφαλιστεί η ευρεία εφαρμογή τους, τα μοντέλα πρέπει να υποβάλλονται σε εξωτερική επαλήθευση σε διάφορους πληθυσμούς με προσαρμογές ή αναβαθμίσεις όπως επαναβαθμονόμηση του μοντέλου ή προσθήκη νέων παραγόντων κινδύνου για να λαμβάνονται υπόψη οι διαφορές στην κατανομή των παραγόντων κινδύνου και τα περιστατικά ασθενειών. Ως εξωτερική επαλήθευση ορίζεται η διαδικασία αξιολόγησης της απόδοσης ενός προγνωστικού μοντέλου σε ανεξάρτητο πληθυσμό διαφορετικό από εκείνον στον οποίο αναπτύχθηκε. Στόχος της είναι να εκτιμηθεί η δυνατότητα γενίκευσης του μοντέλου και η αξιοπιστία της εφαρμογής του σε νέα δεδομένα. Πρόσφατες μελέτες εξωτερικής επαλήθευσης σύγχρονων μοντέλων πρόβλεψης καρδιαγγειακού κινδύνου στη Βιοτράπεζα του Ηνωμένου Βασιλείου (BHB) έδειξαν ότι, παρά την ικανοποιητική τους ικανότητα διάκρισης, η βαθμονόμησή τους παρέμεινε μη βέλτιστη, αναδεικνύοντας την ανάγκη εξωτερικής επαλήθευσης πριν από την κλινική εφαρμογή τους. [1] Ως ικανότητα διάκρισης ορίζεται η ικανότητα του μοντέλου να διακρίνει τα άτομα υψηλού από εκείνα χαμηλού κινδύνου μέσω της περιοχής κάτω από την καμπύλη ROC (Receiver Operating Characteristic) (AUC) και ως βαθμονόμηση, η συμφωνία μεταξύ του εκτιμώμενου και του παρατηρούμενου κινδύνου.

Ο ορισμός του συμβάντος καρδιαγγειακής νόσου διαφέρει ευρέως στη βιβλιογραφία, αντικατοπτρίζοντας διαφορές στους σκοπούς, τους πληθυσμούς και τις κλινικές συνθήκες των μελετών. Ορισμένες μελέτες επικεντρώνονται σε συμβάντα όπως έμφραγμα του μυοκαρδίου, εγκεφαλικό επεισόδιο ή θάνατο, ενώ άλλες υιοθετούν ευρύτερους ορισμούς που περιλαμβάνουν τη στηθάγχη, την καρδιακή ανεπάρκεια ή επεμβάσεις επαναιμάτωσης. Μεταβλητότητα υπάρχει ακόμη στον τρόπο κατηγοριοποίησης αυτών των συμβάντων καθώς μερικοί ορισμοί βασίζονται σε αυτοαναφερόμενα αποτελέσματα, ενώ άλλοι χρησιμοποιούν κλινικές διαγνώσεις ή βιοδείκτες. Οι ασυνέπειες αυτές μπορούν να επηρεάσουν τη συγκρισιμότητα των μελετών και την απόδοση των μοντέλων πρόβλεψης κινδύνου, αναδεικνύοντας την ανάγκη για τυποποιημένους ορισμούς καρδιαγγειακών συμβάντων, προκειμένου να ενισχυθεί η εγκυρότητα και η εφαρμοσιμότητα της έρευνας, δεδομένου ότι οι διάφοροι αλγόριθμοι πρόβλεψης καρδιαγγειακού κινδύνου στη βιβλιογραφία βασίζονται συχνά σε διαφορετικούς ορισμούς των καρδιαγγειακών συμβάντων. Αυτές οι αποκλίσεις στους ορισμούς των συμβάντων εισάγουν επιπλέον προκλήσεις στη μεταφερσιμότητα και τη δυνατότητα γενίκευσης αυτών των μοντέλων όταν εφαρμόζονται σε διαφορετικούς πληθυσμούς και κλινικές συνθήκες. Αυτή η μεταβλητότητα επισημαίνει την εγγενή δυσκολία

της συνήθους χρήσης τέτοιων αλγορίθμων για να αξιολογήσουν με ακρίβεια τον κίνδυνο ενός ατόμου στην πράξη, καθώς οι διαφορές στους ορισμούς μπορεί να οδηγήσουν σε ασυνέπειες στις προβλέψεις και τις κλινικές αποφάσεις. Η αντιμετώπιση αυτών των περιορισμών μέσω τυποποιημένων ορισμών ή προσαρμόσιμων αλγορίθμων θα μπορούσε να βελτιώσει την αξιοπιστία και τη χρησιμότητά τους στην καθημερινή κλινική πρακτική.
Στην παρούσα μελέτη εφαρμόσαμε εξωτερική επαλήθευση τεσσάρων ευρέως χρησιμοποιούμενων αλγορίθμων πρόβλεψης κινδύνου καρδιαγγειακής νόσου, των QR4 [2], PREVENT [3,4], SCORE2 [5,6] και ASCVD [7] στο εκτεταμένο σύνολο δεδομένων της BHB, ξεχωριστά ανά φύλο. Οι αλγόριθμοι εφαρμόστηκαν όπως ακριβώς παρέχονται από τους δημιουργούς τους διασφαλίζοντας ότι όλες οι μεταβλητές πρόβλεψης αντιστοιχούσαν στις απαιτούμενες προδιαγραφές εισόδου. Η ανάλυση πραγματοποιήθηκε ξεχωριστά ανά φύλο, προκειμένου να διασφαλιστεί η αξιοπιστία των αποτελεσμάτων μεταξύ διαφορετικών πληθυσμιακών ομάδων, αξιολογώντας τη διαγνωστική τους ικανότητα μέσω ανάλυσης ROC και της ποιότητας βαθμονόμησης. Η προγνωστική ικανότητα των μοντέλων αξιολογήθηκε βάσει και των τεσσάρων ορισμών καρδιαγγειακού συμβάντος, οδηγώντας συνολικά σε 16 διαφορετικούς συνδυασμούς μοντέλου και ορισμού συμβάντος. Οι συγκεκριμένοι αλγόριθμοι επιλέχθηκαν επειδή αποτελούν σύγχρονα και ευρέως χρησιμοποιούμενα μοντέλα πρόβλεψης καρδιαγγειακού κινδύνου, τα οποία έχουν αναπτυχθεί σε διαφορετικούς πληθυσμούς και βασίζονται σε διαφορετικές στατιστικές προσεγγίσεις, γεγονός που καθιστά ιδιαίτερα ενδιαφέρουσα την αξιολόγηση της απόδοσής τους σε έναν εξωτερικό πληθυσμό.
Με τη συστηματική εφαρμογή αυτών των μοντέλων σε πολλαπλούς ορισμούς καρδιαγγειακών συμβάντων, επιδιώκουμε να διερευνήσουμε σε ποιο βαθμό τα μοντέλα πρόβλεψης καρδιαγγειακού κινδύνου γενικεύονται μεταξύ διαφορετικών ορισμών συμβάντος. Εάν ένα μοντέλο παρουσιάζει καλή απόδοση μόνο όταν εφαρμόζεται στον αρχικό του ορισμό συμβάντος, αλλά χαμηλή απόδοση σε άλλους ορισμούς, αυτό θα αναδείκνυε έναν σημαντικό περιορισμό στην καθολική εφαρμογή αλγορίθμων πρόβλεψης κινδύνου στην κλινική πράξη. Αντιθέτως, εάν ένα μοντέλο επιδεικνύει σταθερή απόδοση σε πολλούς ορισμούς, αυτό θα μπορούσε να υποδηλώνει μεγαλύτερη ευελιξία και δυνατότητα γενίκευσης.
Στόχος της παρούσας ανάλυσης είναι να επισημάνει τις προκλήσεις της χρήσης τυποποιημένων μοντέλων κινδύνου σε διάφορους πληθυσμούς καθώς επίσης και τη σημασία της συνεκτίμησης της μεταβλητότητας του ορισμού συμβάντος κατά την ερμηνεία των προβλέψεων κινδύνου καρδιαγγειακής νόσου σε πραγματικές εφαρμογές.

2. ΜΕΘΟΔΟΙ

Τα δεδομένα για τη μελέτη αυτή προήλθαν από τη BHB, μια μεγάλης κλίμακας βιοϊατρική βάση δεδομένων. Υλοποιήσαμε τους αλγόριθμους πρόβλεψης 10ετούς κινδύνου για καρδιαγγειακή νόσο ακριβώς όπως είχαν δοθεί από τους αρχικούς τους δημιουργούς. Για την αξιολόγηση της απόδοσης των μοντέλων, πραγματοποιήσαμε ανάλυση καμπύλης ROC ώστε να εκτιμήσουμε την ικανότητα διάκρισης, και χρησιμοποιήσαμε τόσο γραφικές μεθόδους όσο και τον δείκτη Brier για την εκτίμηση της ποιότητας βαθμονόμησης. Όλες οι αναλύσεις πραγματοποιήθηκαν με τη χρήση του λογισμικού R, έκδοση 4.4.1 (The R Foundation for Statistical Computing, Βιέννη, Αυστρία), και των πακέτων RiskScorescvd [8] και PooledCohort [9], διασφαλίζοντας τη συνέπεια με τις μεθοδολογίες που είχαν καθοριστεί από τους αρχικούς συγγραφείς.

2.1 Μέθοδος προσαρμογής μοντέλου

Μια βασική πρόκληση στην πρόβλεψη κινδύνου είναι η μεταβλητότητα στον τρόπο με τον οποίο ορίζονται τα συμβάντα καρδιαγγειακού επεισοδίου μεταξύ διαφορετικών μοντέλων. Κάθε ένας από τους τέσσερις αλγορίθμους αναπτύχθηκε αρχικά με βάση έναν συγκεκριμένο ορισμό συμβάντος που ποικίλει από έμφραγμα του μυοκαρδίου και εγκεφαλικό επεισόδιο, έως ευρύτερους ορισμούς που περιλαμβάνουν καρδιαγγειακή ανεπάρκεια, επεμβάσεις επαναγγείωσης ή καρδιακό θάνατο. Για τη συστηματική αξιολόγηση της γενικευσιμότητάς

τους, προσαρμόσαμε κάθε μοντέλο σε όλους τους παραπάνω ορισμούς συμβάντος, δημιουργώντας ένα πλήρες σύνολο συνδυασμών μοντέλου–συμβάντος.
Η ανάλυση για τα μοντέλα QR4 [2], SCORE2 [5,6] και ASCVD [7] βασίστηκε στον αποκλεισμό των συμμετεχόντων που είχαν ελλείπουσες τιμές στις απαιτούμενες μεταβλητές εισόδου. Το μοντέλο PREVENT [3,4] περιλαμβάνει πρόσθετες μεταβλητές, όπως η αναλογία λευκωματίνης προς κρεατινίνη (ACR), οι οποίες χρησιμοποιούνται όταν είναι διαθέσιμες, χωρίς να αποτελούν απαραίτητες μεταβλητές για την εφαρμογή του. Αντίθετα, για τις βασικές μεταβλητές του μοντέλου, όπως η ολική χοληστερόλη, η χοληστερόλη λιποπρωτεϊνών υψηλής πυκνότητας (ΧΥΠ), η συστολική αρτηριακή πίεση (ΣΑΠ), ο δείκτης μάζας σώματος και ο εκτιμώμενος ρυθμός σπειραματικής διήθησης (εΡΣΔ), απαιτούνταν διαθέσιμες τιμές. Για αυτές τις μεταβλητές εφαρμόστηκαν τα προβλεπόμενα όρια τιμών με αποτέλεσμα να αποκλειστούν παρατηρήσεις τόσο με τιμές εκτός των επιτρεπόμενων ορίων όσο και με ελλείπουσες τιμές στις υποχρεωτικές μεταβλητές. Ο αριθμός των παρατηρήσεων που χρησιμοποιήθηκαν για την εφαρμογή καθενός από τα τέσσερα μοντέλα σε καθέναν από τους τέσσερις ορισμούς συμβάντος παρουσιάζεται στον Πίνακα 1.
Κάθε μοντέλο αξιολογείται και στους τέσσερις ορισμούς των συμβάντων χρησιμοποιώντας τον δείκτη Concordance (C-index) ή AUC, τις καμπύλες ROC και τις καμπύλες βαθμονόμησης (calibration), τόσο συνολικά όσο και ανά ηλικία και ΣΑΠ. Η ηλικία και η ΣΑΠ επιλέχθηκαν ως τυπικά κλινικά κριτήρια που χρησιμοποιούνται από τους ιατρούς στην καθημερινή πρακτική, πέρα από το πλαίσιο των μοντέλων. Επιπλέον, και οι δύο παράγοντες περιλαμβάνονται στις εξισώσεις των μοντέλων, παίζοντας σημαντικό ρόλο στον υπολογισμό του εκτιμώμενου κινδύνου. Η ικανότητα διάκρισης εστιάζει στη σωστή ταξινόμηση, ενώ η βαθμονόμηση αξιολογεί την αξιοπιστία των εκτιμώμενων πιθανοτήτων. Ένα μοντέλο μπορεί να αποδίδει πολύ καλά στο ένα αλλά όχι στο άλλο, γι' αυτό και αξιολογούνται και οι δύο για μία ολοκληρωμένη εκτίμηση της απόδοσης του μοντέλου.
Οι συμμετέχοντες που είχαν υποστεί κάποιο συμβάν πριν από την παρακολούθηση αποκλείστηκαν, με αποτέλεσμα τα δεδομένα να είναι «αριστερά κομμένα», ένα μειονέκτημα που είναι εγγενές σε όλα τα μοντέλα, καθώς αυτή είναι η συνήθης πρακτική για την εξαγωγή αυτών των μοντέλων, αν και η εγκυρότητά της έχει αμφισβητηθεί στη βιβλιογραφία [10,11]. Ωστόσο, ο στόχος αυτής της μελέτης είναι η αξιολόγηση των υπαρχόντων μοντέλων με τα πλεονεκτήματα και τα μειονεκτήματά τους.

Πίνακας 1. *Πλήθος των ατόμων που χρησιμοποιήθηκαν για κάθε εφαρμογή, ανά φύλο.*

Ορισμός συμβάντος	Μοντέλο	Πλήθος Γυναικών	Πλήθος Ανδρών
QR4	PREVENT	186,124	155,041
	SCORE2	152,989	114,298
	ASCVD	191,825	153,318
	QR4	179,624	138,531
PREVENT	PREVENT	190,556	190,556
	SCORE2	153,298	114,579
	ASCVD	192,341	153,799
	QR4	179,477	138,242
ASCVD	PREVENT	186,791	155,811
	SCORE2	153,329	114,639
	ASCVD	192,515	154,067
	QR4	179,624	138,531
SCORE 2	PREVENT	143,058	110,399
	SCORE2	153,329	114,639
	ASCVD	147,227	109,360
	QR4	144,260	105,971

2.2 Επισκόπηση μοντέλων πρόγνωσης

Χρησιμοποιήθηκε το προγνωστικό μοντέλο QR4 το οποίο βασίζεται σε στοχευμένης αιτιολογίας μοντέλα ανάλογων κινδύνων κατά Cox (*cause-specific Cox models*), λαμβάνοντας υπόψη τον θάνατο από μη καρδιαγγειακά αίτια ως ανταγωνιστικό κίνδυνο [2]. Η ανάλυση πραγματοποιήθηκε ξεχωριστά για άνδρες και γυναίκες. Το μοντέλο QR4 περιλαμβάνει ευρύ φάσμα δημογραφικών, κλινικών και κοινωνικοοικονομικών μεταβλητών πρόγνωσης. Οι βασικές μεταβλητές που χρησιμοποιήθηκαν είναι: συστολική αρτηριακή πίεση (ΣΑΠ), τυπική απόκλιση ΣΑΠ, λόγος ολικής χοληστερόλης προς χοληστερόλη υψηλής πυκνότητας (ΧΥΠ), δείκτης κοινωνικής αποστέρησης (ανά 5 μονάδες αύξηση), κατηγορία καπνίσματος, εθνικότητα, ύπαρξη ή μη μαθησιακής αναπηρίας, σύνδρομο Down, οικογενειακό ιστορικό στεφανιαίας νόσου, σακχαρώδης διαβήτης τύπου 1 και τύπου 2, υπέρταση υπό φαρμακευτική αγωγή, ρευματοειδής αρθρίτιδα, κολπική μαρμαρυγή, νεφρική ανεπάρκεια, ημικρανία, χρήση κορτικοστεροειδών, συστηματικός ερυθηματώδης λύκος, χρήση άτυπων αντιψυχωσικών, σοβαρή ψυχική νόσος, χρόνια αποφρακτική πνευμονοπάθεια (ΧΑΠ), καθώς και ιστορικό καρκίνου του πνεύμονα, του στόματος, του αίματος ή του εγκεφάλου. Επιπλέον, συμπεριλήφθηκαν φύλο-ειδικοί παράγοντες κινδύνου: στις γυναίκες η επιλόχειος κατάθλιψη και η προεκλαμψία, ενώ στους άνδρες η στυτική δυσλειτουργία.

Επιπλέον, εφαρμόστηκαν οι εξισώσεις PREVENT της Αμερικανικής Καρδιολογικής Εταιρείας που επιτρέπουν την εκτίμηση του 10ετούς και 30ετούς κινδύνου για συνολική καρδιαγγειακή νόσο, η οποία περιλαμβάνει τόσο τη σύνθετη αθηροσκληρωτική καρδιαγγειακή νόσο όσο και την καρδιακή ανεπάρκεια. Οι εξισώσεις αυτές προσαρμόζονται για τον ανταγωνιστικό κίνδυνο μη-καρδιαγγειακού θανάτου. Αφορούν ενήλικες ηλικίας 30 έως 79 ετών [3,4]. Η εφαρμογή τους έγινε μέσω της συνάρτησης predict_10yr_cvd_risk(…, equation_version = "Khan_2023",…) του πακέτου PooledCohort [8]. Οι εξισώσεις που επιτρέπουν την εξατομικευμένη εκτίμηση του καρδιαγγειακού κινδύνου λαμβάνοντας υπόψη παραμέτρους όπως: ηλικία, ολική χοληστερόλη (ΟΧ), ΧΥΠ, ΣΑΠ, εκτιμώμενος ρυθμός σπειραματικής διήθησης (εΡΣΔ), παρουσία διαβήτη, κάπνισμα, και χρήση φαρμακευτικών αγωγών δίνονται στον Πίνακα 2 του Παραρτήματος.

Για την εκτίμηση του καρδιαγγειακού κινδύνου χρησιμοποιήθηκε επίσης το μοντέλο SCORE2, το οποίο βασίζεται σε μοντέλα Fine & Gray [12] για ανταγωνιστικό κίνδυνο, εφαρμόζοντας την προσέγγιση των ανάλογων υποκατανεμημένων κινδύνων. Η μεθοδολογία αυτή επιτρέπει την ακριβέστερη πρόβλεψη του κινδύνου για καρδιαγγειακά συμβάντα, λαμβάνοντας υπόψη τον θάνατο από μη καρδιαγγειακά αίτια ως ανταγωνιστικό κίνδυνο [5,6]. Η εφαρμογή πραγματοποιήθηκε μέσω της συνάρτησης SCORE2_scores του πακέτου RiskScorescvd [9], επιλέγοντας την κατάλληλη παραμετροποίηση για την «χαμηλή» γεωγραφική περιοχή καρδιαγγειακού κινδύνου. Οι εξισώσεις του μοντέλου SCORE2 για την εκτίμηση του 10ετούς καρδιαγγειακού κινδύνου διατίθενται ξεχωριστά ανά φύλο και ηλικιακή κατηγορία. Οι εξισώσεις για τις γυναίκες και τους άνδρες, κάτω των 70 αλλά και των 70 και άνω ετών παρουσιάζονται στους Πίνακες 3 και 4 του Παραρτήματος, αντίστοιχα.

Τέλος, χρησιμοποιήθηκε το μοντέλο ASCVD**,** το οποίο βασίζεται σε μοντέλα ανάλογων κινδύνων κατά Cox, με διαχωρισμό ανά φύλο και φυλή. Οι μεταβλητές που περιλαμβάνονται στο μοντέλο είναι: η ηλικία, η θεραπευμένη ή μη θεραπευμένη ΣΑΠ (mm Hg), η ΟΧ (mg/dL), η ΧΥΠ (HDL-C) (mg/dL), το κάπνισμα (ναι=1/όχι=0) και η παρουσία σακχαρώδους διαβήτη (ναι=1/όχι=0) [7]. Η εφαρμογή του μοντέλου πραγματοποιήθηκε μέσω της συνάρτησης ASCVD_scores από το πακέτο RiskScorescvd [9] της R. Οι αντίστοιχες εξισώσεις ανά φυλή, για γυναίκες και άνδρες δίνονται στους Πίνακες 5 και 6 του Παραρτήματος.

3. ΑΠΟΤΕΛΕΣΜΑΤΑ

3.1 Σύγκριση Απόδοσης των Μοντέλων Πρόβλεψης Κινδύνου Καρδιαγγειακής Νόσου σε Διαφορετικούς Ορισμούς Συμβάντος ως προς την Ικανότητα διάκρισης

Η ανάλυσή μας αποκάλυψε σημαντικές διαφορές στην απόδοση των μοντέλων ανάλογα με τον ορισμό του συμβάντος. Τα αποτελέσματα απεικονίζονται στην Εικόνα 1. Για τις γυναίκες, ο ορισμός συμβάντος της μελέτης PREVENT αποδίδει τις υψηλότερες τιμές της περιοχής κάτω από την καμπύλη (AUC), ενώ για τους άνδρες αυτό ισχύει για τον ορισμό συμβάντος της μελέτης SCORE2. Αξιοσημείωτο είναι ότι, για τις γυναίκες, το μοντέλο PREVENT ήταν σταθερά καλύτερο από τα υπόλοιπα μοντέλα πρόβλεψης κινδύνου, ανεξάρτητα από τον εφαρμοζόμενο ορισμό συμβάντος. Αυτό υποδηλώνει ότι το PREVENT είναι ιδιαίτερα αποτελεσματικό στη διάκριση ατόμων με υψηλό έναντι χαμηλότερου καρδιαγγειακού κινδύνου σε πολλαπλές ταξινομήσεις καρδιαγγειακών συμβάντων, καθιστώντας το ένα ισχυρό εργαλείο πρόβλεψης, ακόμη και όταν εφαρμόζεται εκτός του αρχικού του πλαισίου, ειδικά σε γυναίκες. Αντίθετα, το μοντέλο QR4 εμφάνισε τη χαμηλότερη προβλεπτική ακρίβεια σε όλους τους ορισμούς συμβάντος, με τη χαμηλότερη ικανότητα διάκρισης μεταξύ ατόμων που παρουσίασαν ή δεν παρουσίασαν καρδιαγγειακό συμβάν, τόσο για άνδρες όσο και για γυναίκες. Αυτό αναδεικνύει πιθανές περιορισμένες δυνατότητες μεταφερσιμότητας του μοντέλου QR4 και υποδηλώνει, βάσει της επαλήθευσης στο σύνολο δεδομένων της BHB, ότι ίσως απαιτείται αναπροσαρμογή ή βελτίωση του μοντέλου κατά την εφαρμογή του σε εξωτερικούς πληθυσμούς (Εικόνα 2).

Τα μοντέλα SCORE2 και ASCVD παρουσίασαν ενδιάμεση απόδοση στις γυναίκες, καταλαμβάνοντας σταθερά θέσεις μεταξύ των μοντέλων PREVENT και QR4, τόσο ως προς την ικανότητα διάκρισης (AUC) όσο και ως προς τη βαθμονόμηση. Παρόλο που τα μοντέλα αυτά εμφάνισαν μέτρια ακρίβεια, η απόδοσή τους διέφερε ανάλογα με τον ορισμό του συμβάντος που χρησιμοποιήθηκε, γεγονός που υποδηλώνει ότι η προβλεπτική τους ικανότητα ενδέχεται να είναι πιο ευαίσθητη στις διαφορές στον τρόπο ορισμού των καρδιαγγειακών συμβάντων.

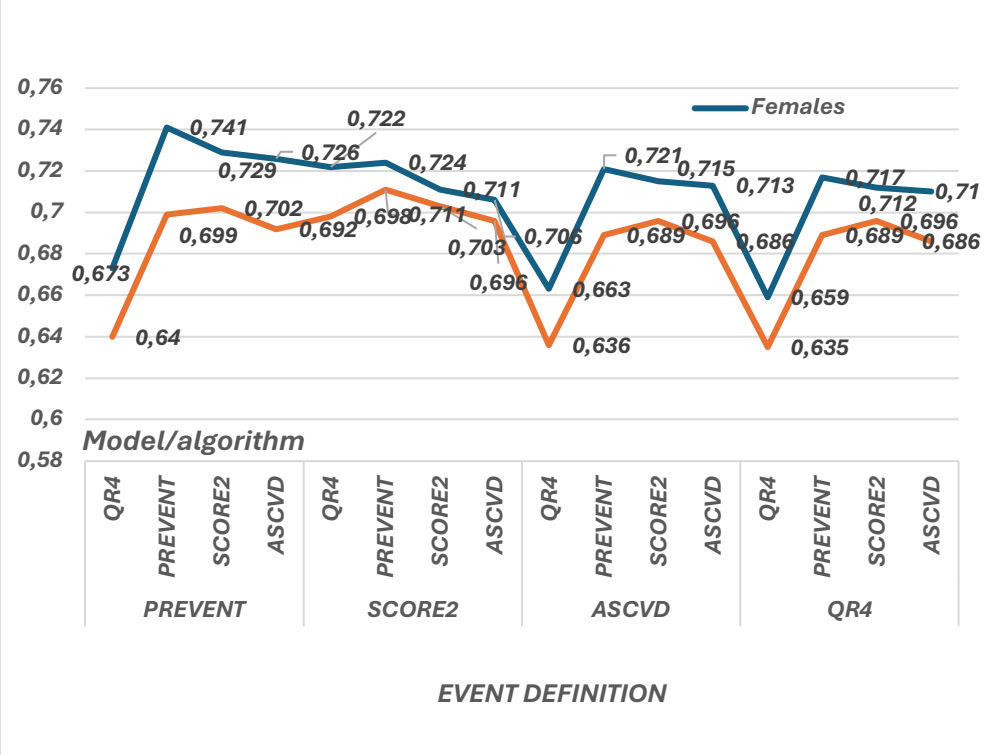
Για τους άνδρες, το μοντέλο με την καλύτερη απόδοση ήταν το PREVENT με βάση τον ορισμό συμβάντος SCORE2, ακολουθούμενο από το μοντέλο SCORE2 με τον ίδιο ορισμό συμβάντος (C=0.711 έναντι 0.703), γεγονός που υποδηλώνει καλή μεταφερσιμότητα του μοντέλου PREVENT στους άνδρες και για τον ορισμό συμβάντος SCORE2. Αξιοσημείωτο είναι ότι, σε όλα τα μοντέλα και για όλους τους ορισμούς συμβάντος, η προβλεπτική απόδοση ήταν συστηματικά καλύτερη για τις γυναίκες σε σύγκριση με τους άνδρες (Εικόνα 2). Αυτό υποδηλώνει ότι τα μοντέλα ενδέχεται να εντοπίζουν τους παράγοντες κινδύνου πιο αποτελεσματικά στις γυναίκες ή ότι οι υποκείμενες διαφορές φύλου στην εμφάνιση και εξέλιξη της καρδιαγγειακής νόσου επηρεάζουν την ακρίβεια των μοντέλων. Το εύρημα αυτό εγείρει επίσης σημαντικά ερωτήματα σχετικά με πιθανές μεροληψίες στα εργαλεία πρόβλεψης κινδύνου, αναδεικνύοντας την ανάγκη για περαιτέρω διερεύνηση σε προσεγγίσεις μοντελοποίησης με βάση το φύλο, ώστε να διασφαλιστεί δίκαιη εκτίμηση κινδύνου τόσο για άνδρες όσο και για γυναίκες.

Συνολικά, τα ευρήματα αυτά αναδεικνύουν τον καθοριστικό ρόλο που παίζει η επιλογή του ορισμού συμβάντος στην απόδοση των μοντέλων και υπογραμμίζουν τη σημασία της επιλογής του κατάλληλου εργαλείου πρόβλεψης κινδύνου, με βάση τον συγκεκριμένο πληθυσμό και το κλινικό πλαίσιο. Η ανώτερη απόδοση του μοντέλου PREVENT σε διαφορετικούς ορισμούς συμβάντος υποδηλώνει ότι μπορεί να είναι πιο προσαρμόσιμο, ενώ η χαμηλότερη ακρίβεια του μοντέλου QR4 καταδεικνύει πιθανούς περιορισμούς στη γενικότερη εφαρμοσιμότητά του. Η κατανόηση αυτών των λεπτών διαφορών είναι κρίσιμη για τη βελτίωση της κλινικής χρησιμότητας των μοντέλων πρόβλεψης καρδιαγγειακού κινδύνου και τη βέλτιστη εφαρμογή τους σε πραγματικές συνθήκες.

3.2 Σύγκριση Απόδοσης των Μοντέλων Πρόβλεψης Κινδύνου Καρδιαγγειακής Νόσου σε Διαφορετικούς Ορισμούς Συμβάντος ως προς την Βαθμονόμηση

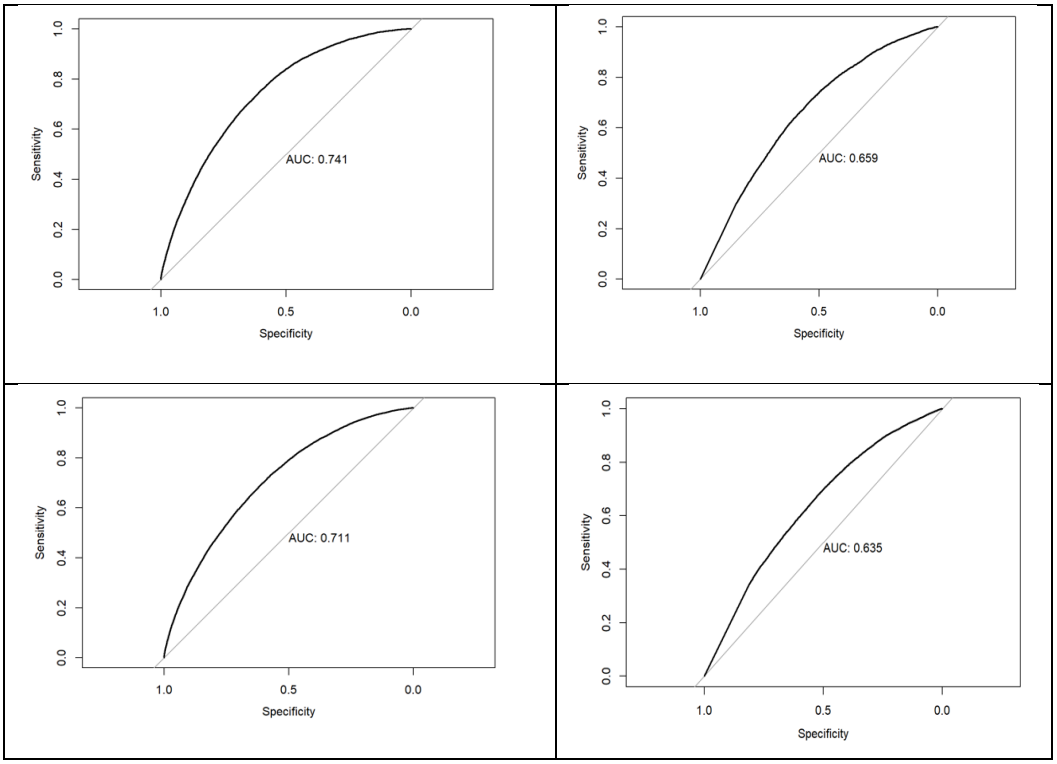
Ο Πίνακας 7 παρουσιάζει τις τιμές Brier για τις υλοποιήσεις των διαφορετικών μοντέλων σε κάθε έναν από τους τέσσερις διαφορετικούς ορισμούς συμβάντος. Ένα καλά βαθμονομημένο μοντέλο εμφανίζει χαμηλότερες τιμές Brier, κάτι που υποδηλώνει ότι οι προβλεπόμενες πιθανότητες συμφωνούν περισσότερο με τα παρατηρούμενα αποτελέσματα. Η βαθμονόμηση των μοντέλων είναι σταθερά καλύτερη για τις γυναίκες σε σχέση με τους άνδρες, και το μοντέλο SCORE2 αποδίδει γενικά καλύτερα από τα υπόλοιπα, αν και χρησιμοποιεί τον ορισμό συμβάντος PREVENT, ακολουθούμενο από τον ορισμό ASCVD. Τα μοντέλα PREVENT και ASCVD είναι συνεπή υπό την έννοια ότι είναι αρκετά καλά βαθμονομημένα για τους αντίστοιχους ορισμούς συμβάντος καρδιαγγειακής νόσου. Η σχετική ασυνέπεια μεταξύ μοντέλων και αντίστοιχων ορισμών συμβάντος υποδηλώνει την ανάγκη για έναν συνεπή και ευρέως αποδεκτό ορισμό συμβάντος.

Εικόνα 1*.Εκτιμήσεις του δείκτη C από την εξωτερική επαλήθευση διαφορετικών μοντέλων πρόβλεψης του κινδύνου ΚΑΝ σύμφωνα με τον ορισμό του συμβάντος ΚΑΝ, ανά φύλο.*



Συνολικά, ένας διαφορετικός συνδυασμός μεταξύ ορισμού συμβάντος και μοντέλου φαίνεται να παρουσιάζει την καλύτερη βαθμονόμηση. Συγκεκριμένα, ο ορισμός συμβάντος PREVENT σε συνδυασμό με το μοντέλο SCORE2 αποδίδει τις χαμηλότερες τιμές Brier. Ωστόσο, όσον αφορά τη συνέπεια μεταξύ ορισμού συμβάντος και χρησιμοποιούμενου μοντέλου, το σύστημα ASCVD υπερτερεί σταθερά έναντι των υπολοίπων, ακολουθούμενο από το PREVENT, ενώ το QR4 φαίνεται να έχει τη χειρότερη απόδοση για τα δεδομένα της ΒΗΒ (Εικόνα 3).

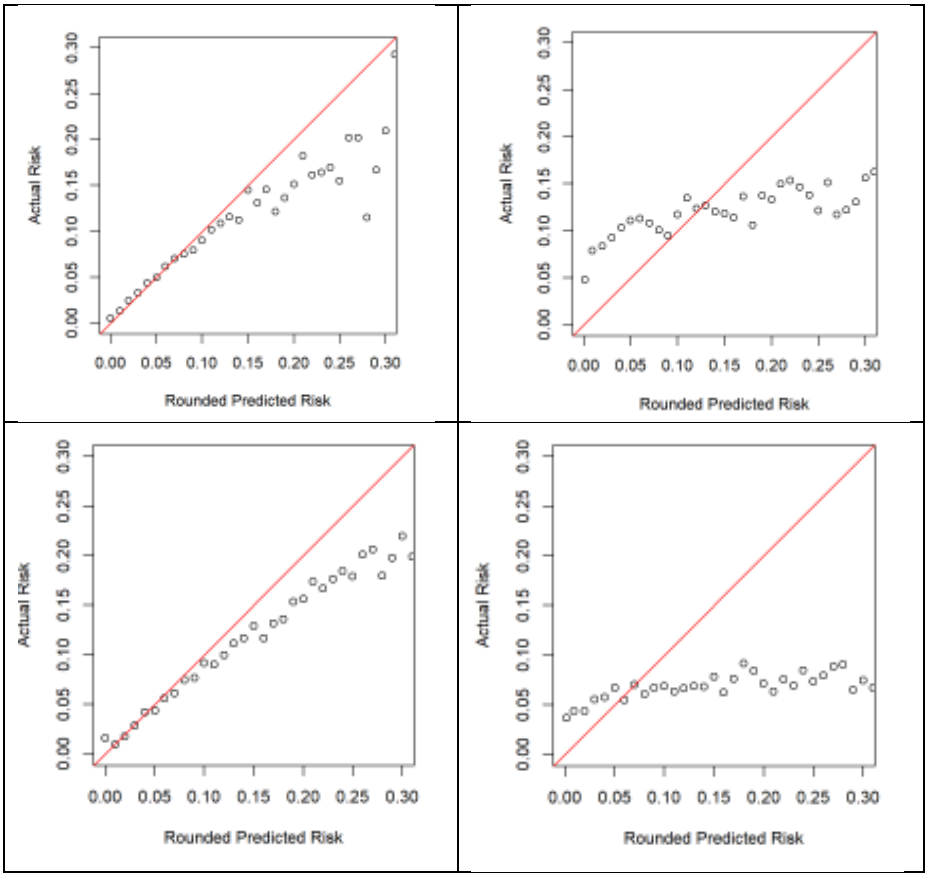
***Εικόνα 2**. Καμπύλες ROC και αντίστοιχες AUC για τα καλύτερα έναντι των χειρότερων μοντέλων με βάση την ακρίβεια (αριστερά: καλύτερο, δεξιά: χειρότερο), ανά φύλο (επάνω: γυναίκες, κάτω: άνδρες).*



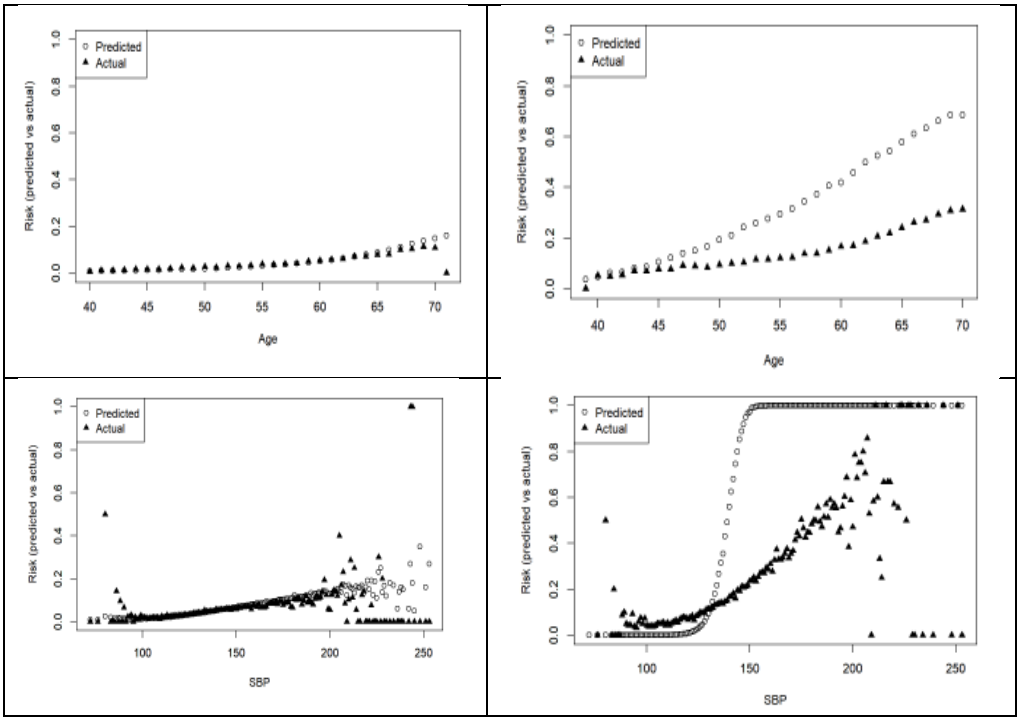
Τα διαγράμματα βαθμονόμησης ανάλογα με την ηλικία (Εικόνα 4) απεικονίζουν την προγνωστική απόδοση των μοντέλων ASCVD και QR4 για τον γυναικείο πληθυσμό. Το μοντέλο ASCVD παρουσιάζει καλή βαθμονόμηση, καθώς οι προβλεπόμενοι κίνδυνοι (ανοιχτοί κύκλοι) συμφωνούν αρκετά με τους πραγματικούς κινδύνους (μαύρα τρίγωνα), υποδηλώνοντας αξιοπιστία σε διάφορες ηλικιακές ομάδες. Αντίθετα, το μοντέλο QR4 εμφανίζει κακή βαθμονόμηση, με τις προβλέψεις να υπερεκτιμούν σταθερά τους πραγματικούς κινδύνους, ιδίως στις μεγαλύτερες ηλικίες. Αυτό υποδεικνύει ότι το QR4 στερείται ακρίβειας και πιθανώς απαιτεί αναβαθμονόμηση για τον συγκεκριμένο πληθυσμό. Τα δύο αυτά μοντέλα επιλέχθηκαν ως το καλύτερα και χειρότερα βαθμονομημένο, αντίστοιχα, χρησιμοποιώντας τον ίδιο συνδυασμό ορισμού συμβάντος και αντίστοιχου μοντέλου.

Το ίδιο μοτίβο παρατηρείται και στον ανδρικό πληθυσμό, όπως απεικονίζεται στην Εικόνα 5. Η ηλικία και η συστολική αρτηριακή πίεση αποτελούν καλούς δείκτες για την εκτίμηση του 10ετούς καρδιαγγειακού κινδύνου. Το μοντέλο ASCVD αποτυπώνει με ακρίβεια τους αναμενόμενους κινδύνους βάσει αυτών των βιοδεικτών, ενώ το QR4 υπερεκτιμά σημαντικά τους αντίστοιχους κινδύνους, γεγονός που ενδέχεται να οδηγήσει σε περιττό φαρμακευτικό φορτίο, με κίνδυνο υπερθεραπείας και αυξημένου κόστους υγειονομικής περίθαλψης.

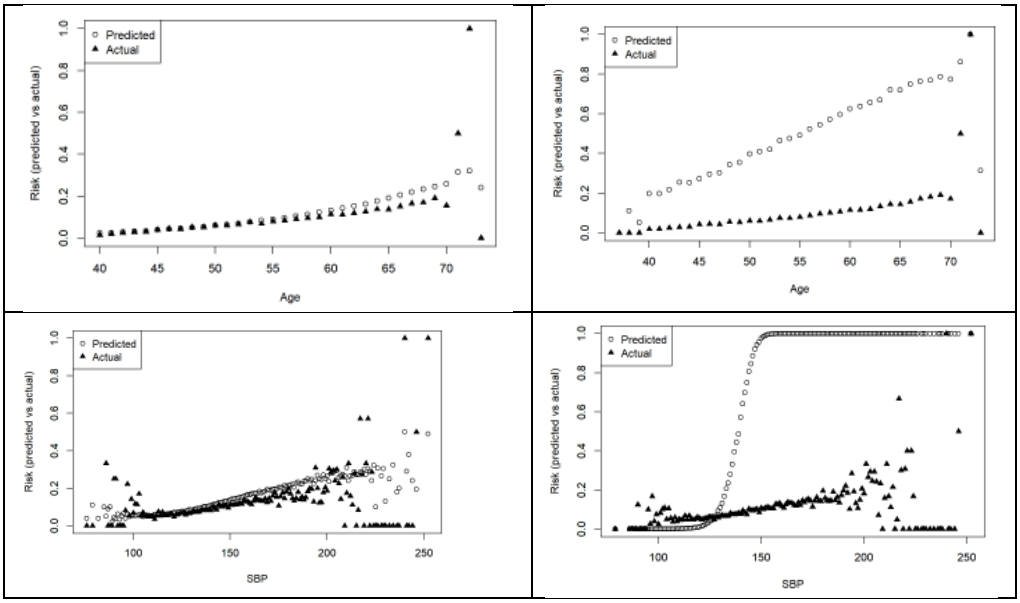
Εικόνα 3. *Διαγράμματα βαθμονόμησης (συνολικά) για ASCVD (αριστερά) & QR4 (δεξιά), ανά φύλο (επάνω: γυναίκες, κάτω: άνδρες). Ίδιο μοντέλο μελέτης & ορισμός συμβάντος.*



Εικόνα 4. *Βαθμονόμηση του ASCVD (αριστερά) και του QR4 (δεξιά) ως προς την ηλικία και την ΣΑΠ για τις γυναίκες. Ίδιο μοντέλο μελέτης και ορισμός συμβάντος.*



Εικόνα 5. *Βαθμονόμηση του ASCVD (αριστερά) και του QR4 (δεξιά) ως προς την ηλικία και την ΣΑΠ για τους άνδρες. Ίδιο μοντέλο μελέτης και ορισμός συμβάντος.*



Πίνακας 7. *Περίληψη των τιμών Brier για την αξιολόγηση της βαθμονόμησης των διαφορετικών μοντέλων πρόβλεψης 10ετούς καρδιαγγειακού κινδύνου, ανά φύλο.*

Ορισμός συμβάντος	Μοντέλο	Γυναίκες	Άνδρες
PREVENT	QR4	0.336	0.4207
	PREVENT	0.0442	0.0862
	SCORE2	0.0308	0.0661
	ASCVD	0.0413	0.0821
SCORE2	QR4	0.2496	0.3104
	PREVENT	0.1297	0.1763
	SCORE2	0.1379	0.187
	ASCVD	0.1341	0.1699
ASCVD	QR4	0.3368	0.4216
	PREVENT	0.0426	0.0819
	SCORE2	0.0329	0.0672
	ASCVD	0.0435	0.0830
QR4	QR4	0.3361	0.4198
	PREVENT	0.0468	0.0864
	SCORE2	0.0369	0.0713
	ASCVD	0.0477	0.087

4. Συζήτηση

Συνολικά, τα αποτελέσματα έδειξαν ότι, ενώ τα συγκεκριμένα μοντέλα παρουσιάζουν μέτρια μεταφερσιμότητα, περιορίζονται από την εξάρτησή τους από μεταβλητές πρόβλεψης που καταγράφονται κατά την έναρξη της παρακολούθησης. Αυτό αναδεικνύει έναν σημαντικό περιορισμό της τρέχουσας κλινικής πρακτικής: τα στατικά μοντέλα, αν και χρήσιμα, ενδέχεται να μην αποτυπώνουν πλήρως τη δυναμική φύση της υγείας ενός ατόμου με την πάροδο του χρόνου. Αυτοί οι περιορισμοί επισημαίνουν την ανάγκη για προσοχή κατά την καθολική εφαρμογή αυτών των εργαλείων σε ετερογενείς πληθυσμούς. Η βελτίωση της μεταφερσιμότητάς τους θα μπορούσε να επιτευχθεί μέσω επαναβαθμονόμησης σε τοπικούς πληθυσμούς ή ενσωμάτωσης πρόσθετων παραγόντων κινδύνου που αντανακλούν καλύτερα τα χαρακτηριστικά του πληθυσμού εφαρμογής.

Η μεταφερσιμότητα των μοντέλων μπορεί να θεωρηθεί αμφίβολη, με συνήθεις περιορισμούς στην επικύρωσή τους, όπως έχει ήδη αναφερθεί στη σχετική βιβλιογραφία [13,14], καθώς οι ορισμοί του συμβάντος ενδέχεται να διαφέρουν σε τοπικό επίπεδο, απουσία ευρέως αποδεκτών κατευθυντήριων οδηγιών. Επιπλέον, τα χαρακτηριστικά των συμμετεχόντων μπορεί επίσης να ποικίλλουν σημαντικά μεταξύ διαφορετικών πληθυσμών. Συνολικά, η ικανότητα διάκρισης των μοντέλων έχει περιοριστεί σε τιμές AUC μικρότερες του 0.75, γεγονός που υπογραμμίζει ότι τα στατικά μοντέλα έχουν περιορισμένη χρησιμότητα στην πράξη, παρόλο που μπορούν να προσφέρουν χρήσιμες πληροφορίες και κατανόηση.

Συνολικά, τα μοντέλα παρουσίασαν καλύτερη ικανότητα διάκρισης στις γυναίκες συγκριτικά με τους άνδρες. Η απόδοσή τους θα μπορούσε να βελτιωθεί περαιτέρω με την ενσωμάτωση μεταβλητών που είναι ειδικές για το γυναικείο φύλο και επηρεάζουν την προγνωστική τους ικανότητα.

Τα μοντέλα επικυρώθηκαν στο σύνολο δεδομένων της BHB. Ωστόσο, διαφορετικοί πληθυσμοί φέρουν διαφορετικά χαρακτηριστικά λόγω κοινωνικών, φαινοτυπικών και γενετικών παραγόντων που μπορεί να ποικίλλουν σημαντικά σε διάφορες περιοχές του κόσμου [14]. Επομένως, τα αποτελέσματα της παρούσας μελέτης δεν μπορούν να γενικευθούν αυτομάτως σε άλλους πληθυσμούς χωρίς προηγούμενη εξωτερική επαλήθευση. Από κλινικής πλευράς, τα ευρήματα υποδεικνύουν ότι η επιλογή ενός μοντέλου πρόβλεψης δεν θα πρέπει να βασίζεται αποκλειστικά στις επιδόσεις που παρουσίασε κατά την ανάπτυξή του, αλλά και στον πληθυσμό εφαρμογής, στις διαθέσιμες μεταβλητές και στον ορισμό του καρδιαγγειακού συμβάντος. Δεδομένου ότι η παρούσα μελέτη σχεδιάστηκε ως ανεξάρτητη εξωτερική επαλήθευση των επιμέρους μοντέλων και όχι ως αυστηρή συγκριτική αξιολόγησή τους, δεν είναι δυνατόν να προταθεί ένα μοντέλο ως συνολικά προτιμότερο έναντι των υπολοίπων.

Τα μοντέλα δοκιμάστηκαν συμπεριλαμβάνοντας θετικά συμβάντα πριν την έναρξη της παρακολούθησης, δηλαδή άτομα που είχαν ήδη εμφανίσει ένα καρδιαγγειακό συμβάν (σύμφωνα με τον εκάστοτε χρησιμοποιούμενο ορισμό) πριν από την έναρξη της μελέτης, ως ένα απλό μέτρο αντιμετώπισης της αριστερής περικοπής. Η απόδοση των μοντέλων παρουσίασε ελαφρώς βελτιωμένα αποτελέσματα, αν και με περιορισμένο όφελος, κοντά σε μια διαφορά AUC της τάξης του 0.01.

Κοιτώντας προς το μέλλον, η ενσωμάτωση δυναμικών, σε πραγματικό χρόνο δεδομένων υγείας, φαίνεται να έχει τη δυνατότητα να φέρει επανάσταση στην πρόβλεψη του καρδιαγγειακού κινδύνου. Φορητές συσκευές συνεχούς παρακολούθησης, παρέχουν πλούσια και αξιοποιήσιμα δεδομένα, που περιλαμβάνουν διακυμάνσεις στον καρδιακό ρυθμό, επίπεδα φυσικής δραστηριότητας, πρότυπα ύπνου και άλλους φυσιολογικούς δείκτες. Μέσω της ενσωμάτωσης αυτής της δυναμικής πληροφορίας, οι διαγνωστικοί αλγόριθμοι θα μπορούσαν να ενισχυθούν σημαντικά, προσφέροντας πιο εξατομικευμένες και προσαρμοστικές εκτιμήσεις κινδύνου.

Η παρούσα μελέτη παρουσιάζει ορισμένους περιορισμούς που πρέπει να ληφθούν υπόψη κατά την ερμηνεία των αποτελεσμάτων. Τα αναλυτικά δείγματα που χρησιμοποιήθηκαν για την αξιολόγηση του κάθε μοντέλου, αν και όλα προέρχονταν από το ίδιο σύνολο δεδομένων της BHB, διέφεραν ελαφρώς λόγω των διαφορετικών απαιτούμενων μεταβλητών και κριτηρίων επιλεξιμότητας κάθε μοντέλου. Ως εκ τούτου, δεν πραγματοποιήθηκε αυστηρή στατιστική σύγκριση της απόδοσής τους και τα αποτελέσματα θα πρέπει να ερμηνεύονται ως ανεξάρτητη

εξωτερική επαλήθευση κάθε μοντέλου. Επιπλέον, όλα τα μοντέλα βασίζονται σε χαρακτηριστικά που καταγράφονται κατά την έναρξη της παρακολούθησης και δεν λαμβάνουν υπόψη τις μεταβολές των παραγόντων κινδύνου κατά τη διάρκεια της παρακολούθησης.

ABSTRACT

Cardiovascular disease (CVD) risk prediction models are valuable tools in clinical practice. However, the definition of CVD events varies widely across the literature. The aim of this analysis is to highlight the challenges associated with applying standardized risk models to diverse populations, and to emphasize the importance of outcome definition variability when interpreting CVD risk predictions in real-world settings. In the present study, four widely used models were applied separately by sex to the extensive UK Biobank dataset and evaluated in terms of their calibration and discrimination using all four outcome definitions. Substantial differences emerged between the various event definitions, with consistently better predictive performance observed in women. We conclude that static models that do not incorporate real-time data have limited practical utility.

Keywords: Reliability, Discrimination, Validation, Statistical Prediction Models

ΠΑΡΑΡΤΗΜΑ

Πίνακας 2. *Βασικές εξισώσεις εκτίμησης 10ετούς κινδύνου PREVENT, ανά φύλο.*

Άνδρες	log-Odds = −3.031168 + 0.7688528 × (ηλικία − 55) /10 + 0.0736174 × (ΟΧ − ΧΥΠ − 3.5) − 0.0954431 × (ΧΥΠ − 1.3) /0.3 − 0.4347345 × (min(ΣΑΠ, 110) − 110) /20 + 0.3362658 × (max(ΣΑΠ, 110) − 130) /20 + 0.7692857 × (αν υπάρχει σακχαρώδης διαβήτης) + 0.4386871 × (αν είναι νυν καπνιστής) + 0.5378979 × (min(εΡΣΔ, 60) − 60) / (−15) + 0.0164827 × (max(εΡΣΔ, 60) − 90) / (−15) + 0.288879 × (αν λαμβάνει αντιυπερτασικά) − 0.1337349 × (αν λαμβάνει στατίνη) − 0.0475924 × (αν λαμβάνει αντιυπερτασικά) × (max(ΣΑΠ, 110) − 130) /20 + 0.150273 × (αν λαμβάνει στατίνη) × (ΟΧ − ΧΥΠ − 3.5) − 0.0517874 × (ηλικία − 55) /10 × (ΟΧ − ΧΥΠ − 3.5) + 0.0191169 × (ηλικία − 55) /10 × (ΧΥΠ − 1.3) /0.3 − 0.1049477 × (ηλικία − 55) /10 × (max(ΣΑΠ, 110) − 130) /20 − 0.2251948 × (ηλικία − 55) /10 × (αν υπάρχει διαβήτης) − 0.0895067 × (ηλικία − 55) /10 × (αν είναι νυν καπνιστής) − 0.1543702 × (ηλικία − 55) /10 × (min(εΡΣΔ, 60) − 60) / (−15) Risk = exp(log-Odds) / (1 + exp(log-Odds))
Γυναίκες	log-Odds = −3.307728 + 0.7939329 × (ηλικία − 55) /10 + 0.0305239 × (ΟΧ − ΧΥΠ − 3.5) − 0.1606857 × (ΧΥΠ − 1.3) /0.3 − 0.2394003 × (min(ΣΑΠ, 110) − 110) /20 + 0.360078 × (max(ΣΑΠ, 110) − 130) /20

	+ 0.8667604 × (αν υπάρχει σακχαρώδης διαβήτης) + 0.5360739 × (αν είναι νυν καπνίστρια) + 0.6045917 × (min(εΡΣΔ, 60) − 60) / (−15) + 0.0433769 × (max(εΡΣΔ, 60) − 90) / (−15) + 0.3151672 × (αν λαμβάνει αντιυπερτασικά) − 0.1477655 × (αν λαμβάνει στατίνη) − 0.0663612 × (αν λαμβάνει αντιυπερτασικά) × (max(ΣΑΠ, 110) − 130) /20 + 0.1197879 × (αν λαμβάνει στατίνη) × (ΟΧ −ΧΥΠ − 3.5) − 0.0819715 × (ηλικία − 55) /10 × (ΟΧ-ΧΥΠ − 3.5) + 0.0306769 × (ηλικία − 55) /10 × (ΧΥΠ − 1.3) /0.3 − 0.0946348 × (ηλικία − 55) /10 × (max(ΣΑΠ, 110) − 130) /20 − 0.27057 × (ηλικία − 55) /10 × (αν υπάρχει διαβήτης) − 0.078715 × (ηλικία − 55) /10 × (αν είναι νυν καπνίστρια) − 0.1637806 × (ηλικία − 55) /10 × (min(εΡΣΔ, 60) − 60) / (−15) Risk = exp(log-Odds) / (1 + exp(log-Odds))

Πίνακας 3. *Εξισώσεις SCORE2 για την εκτίμηση 10ετούς καρδιαγγειακού κινδύνου σε γυναίκες ηλικίας <70 και ≥70 ετών.*

Γυναίκες <70 ετών	xx_female <- c(0.4648*(Ηλικία - 60)/5 +0.7744*(αν είναι νυν καπνίστρια) +0.3131*(ΣΑΠ - 120)/20 + 0.8096*(αν υπάρχει σακχαρώδης διαβήτης) + 0.1002*(ΟΧ - 6)/1 + (-0.2606)*(ΧΥΠ - 1.3)/0.5 + (-0.1088)*(Ηλικία - 60)/5 * (αν είναι νυν καπνίστρια) + (-0.0277)*(Ηλικία - 60)/5 * (ΣΑΠ - 120)/20 + (-0.0226)*(Ηλικία - 60)/5 * (ΟΧ - 6)/1 + 0.0613*(Ηλικία - 60)/5 * (ΧΥΠ - 1.3)/0.5 + (-0.1272)*(Ηλικία - 60)/5 * (αν υπάρχει σακχαρώδης διαβήτης)) xx_female2 <- 1 - 0.9776^exp(xx_female) xx_female3<-1-exp(-exp(0.738+0.7019 * log(-log(1-xx_female2)))) x <- round(xx_female3*100, 1)
Γυναίκες ≥ 70 ετών	xx_female_op <- c(0.0789*(Ηλικία - 73) + 0.6010*(αν υπάρχει σακχαρώδης διαβήτης) + 0.4921*(αν είναι νυν καπνίστρια) + 0.0102*(ΣΑΠ - 150) +0.0605*(ΟΧ - 6) + (-0.3040)*(ΧΥΠ - 1.4) +

	(-0.0107)*(Ηλικία - 73) * (αν υπάρχει σακχαρώδης διαβήτης) + (-0.0255)*(Ηλικία - 73) * (αν είναι νυν καπνίστρια) + (-0.0004)*(Ηλικία - 73) * (ΣΑΠ - 150) + (-0.0009)*(Ηλικία - 73) * (ΧΥΠ - 6) + 0.0154*(Ηλικία - 73) * (ΟΧ - 1.4)) xx_female2_op <- 1 - 0.8082^exp(xx_female_op-0.229) xx_female3_op <- 1-exp(-exp(-0.52+1.01 * log(-log(1-xx_female2_op)))) x <- round(xx_female3_op*100, 1)

Πίνακας 4. *Εξισώσεις SCORE2 για την εκτίμηση 10ετούς καρδιαγγειακού κινδύνου σε άνδρες ηλικίας <70 και ≥70 ετών.*

Άνδρες <70 ετών	xx_male <- c(0.3742*(Ηλικία - 60)/5 + 0.6012*(αν είναι νυν καπνιστής) + 0.2777*(ΣΑΠ - 120)/20 +0.6457*(αν υπάρχει σακχαρώδης διαβήτης) + 0.1458*(ΟΧ - 6)/1 + (-0.2698)*(ΧΥΠ- 1.3)/0.5 + (-0.0755)*(Ηλικία - 60)/5 * (αν είναι νυν καπνιστής) + (-0.0255)*(Ηλικία - 60)/5 * (ΣΑΠ - 120)/20 + (-0.0281)*(Ηλικία - 60)/5 * (ΟΧ - 6)/1 + 0.0426*(Ηλικία - 60)/5 * (ΟΧ - 1.3)/0.5 + (-0.0983)*(Ηλικία - 60)/5 *(αν υπάρχει σακχαρώδης διαβήτης)) xx_male2 <- 1 - 0.9605^exp(xx_male) xx_male3 <- 1-exp(-exp(-0.5699+0.7476 * log(-log(1-xx_male2)))) x <- round(xx_male3*100, 1)
Άνδρες ≥ 70 ετών	xx_male_op <- c(0.0634*(Ηλικία - 73) + 0.4245*(αν υπάρχει σακχαρώδης διαβήτης) + 0.3524*(αν είναι νυν καπνιστής) + 0.0094*(ΣΑΠ - 150) + 0.0850*(ΟΧ- 6) + (-0.3564)*(ΧΥΠ - 1.4) + (-0.0174)*(Ηλικία - 73) * (αν υπάρχει σακχαρώδης διαβήτης) + (-0.0247)*(Ηλικία - 73) *(αν είναι νυν καπνιστής) + (-0.0005)*(Ηλικία - 73) * (ΣΑΠ - 150) + 0.0073*(Ηλικία - 73) * (ΟΧ - 6) + 0.0091*(Ηλικία - 73) * (ΧΥΠ - 1.4)) xx_male2_op <- 1 - 0.7576^exp(xx_male_op - 0.0929)

	xx_male3_op <- 1-exp(-exp(-0.34+1.19 * log(-log(1-xx_male2_op)))) x <- round(xx_male3_op*100, 1)

Πίνακας 5. *Συντελεστές εξισώσεων ASCVD για εκτίμηση 10ετούς κινδύνου καρδιαγγειακής νόσου στις γυναίκες, ανά φυλή.*

Μεταβλητές	Συντελεστές Γυναικών	
	Λευκές	Αφροαμερικανές
Ln(Ηλικία)	-29.799	17.114
Ln(Ηλικία^2)	4.884	-
Ln(ΟΧ)	13.54	0.94
Ln(Ηλικία) × Ln(ΟΧ)	-3.114	-
Ln(ΧΥΠ)	-13.578	-18.92
Ln(Ηλικία) × Ln(ΧΥΠ)	3.149	4.475
Log(θεραπευμένη ΣΑΠ)	2.019	29.291
Log(Ηλικία) × Log(θεραπευμένη ΣΑΠ)	-	-6.432
Log(μη-θεραπευμένη ΣΑΠ)	1.957	27.82
Log(Ηλικία) × Log(μη-θεραπευμένη ΣΑΠ)	-	6.087
Νυν καπνιστής/στρια	7.574	0.691
Log(Ηλικία) × Νυν καπνιστής/στρια	-1.665	-
Σακχαρώδης διαβήτης	0.661	0.874
Εκτιμώμενος 10-ετής κίνδυνος	$1 - 0.9665^{e^{(x+29.18)}}$	$1 - 0.9533^{e^{(x-86.61)}}$
x=Άθροισμα (συντελεστή×τιμή)		

Πίνακας 6. *Συντελεστές εξισώσεων ASCVD για εκτίμηση 10ετούς κινδύνου καρδιαγγειακής νόσου στους άνδρες, ανά φυλή.*

Μεταβλητές	Συντελεστές Ανδρών	
	Λευκοί	Αφροαμερικανοί
Log(Ηλικία)	12.344	2.469
Log(ΟΧ)	11.853	0.302
Log(Ηλικία) × Log(ΟΧ)	-2.664	-
Log(ΧΥΠ)	-7.99	-0.307
Log(Ηλικία) × Log(ΧΥΠ)	1.769	-
Log(θεραπευμένη ΣΑΠ)	1.797	1.916

Log(μη-θεραπευμένη ΣΑΠ)	1.764	1.809
Νυν καπνιστής/στρια	-7.837	0.549
Log(Ηλικία) × Νυν καπνιστής/στρια	-1.795	-
Σακχαρώδης διαβήτης	0.658	0.645
Εκτιμώμενος 10-ετής κίνδυνος	$1 - 0.9144^{e^{(x-61.18)}}$	$1 - 0.8954^{e^{(x-19.54)}}$
x=Άθροισμα (συντελεστή×τιμή)		

ΑΝΑΦΟΡΕΣ

Kastrati L, Alexandri EM, Rupp M, Koch L, Garcia-Tirado J, Nakas C, Maron DJ, Herzig D, Bally L. (2026) Independent external validation and head-to-head comparison of guideline-recommended CVD risk prediction models. Am J Prev Cardiol. 2026;28:101625. doi:10.1016/j.ajpc.2026.101625.

Hippisley-Cox, J. et al. (2024). Development and validation of a new algorithm for improved cardiovascular risk prediction. Nature Medicine, 30(5). Published online April 18, 2024. pp. 1440–1447. doi:10.1038/s41591-024-02905-y.

Scheuermann B, Brown A, Colburn T, Hakeem H, Chow CH, Ade C. External Validation of the American Heart Association PREVENT Cardiovascular Disease Risk Equations. JAMA Netw Open. 2024;7(10):e2438311. doi:10.1001/jamanetworkopen.2024.38311.

Khan, S.S. et al. (2023). Novel prediction equations for absolute risk assessment of total cardiovascular disease incorporating cardiovascular-kidney-metabolic health: A scientific statement from the American Heart Association. Circulation, 148(24), pp. 1982–2004. doi:10.1161/cir.0000000000001191.

Hageman, S. et al. (2021). SCORE2 risk prediction algorithms: New models to estimate 10-year risk of cardiovascular disease in Europe. European Heart Journal, 42(25), pp. 2439–2454. doi:10.1093/eurheartj/ehab309.

de Vries, T.I. et al. (2021) 'SCORE2-OP risk prediction algorithms: Estimating Incident Cardiovascular Event Risk in older persons in four geographical risk regions', European Heart Journal, 42(25), pp. 2455–2467. doi:10.1093/eurheartj/ehab312.

Goff, D.C. et al. (2014) '2013 ACC/AHA guideline on the assessment of cardiovascular risk', Circulation, 129 (25_suppl_2). doi:10.1161/01.cir.0000437741.48606.98.

Jaeger, B. (2024). PooledCohort: Predicted Risk for CVD using Pooled Cohort Equations, PREVENT Equations, and Other Contemporary CVD Risk Calculators. R package version 0.0.2. Available at: https://CRAN.R-project.org/package=PooledCohort. doi:10.32614/CRAN.package.PooledCohort.

Perez-Vicencio, D., Doudesis, D., Thurston, A.J.F. and Selva, J. (2025). RiskScorescvd: Cardiovascular Risk Scores Calculator. R package version 0.3.1. Available at: https://CRAN.R-project.org/package=RiskScorescvd. doi:10.32614/CRAN.package.RiskScorescvd.

Cain KC, Harlow SD, Little RJ, et al. Bias Due to Left Truncation and Left Censoring in Longitudinal Studies of Developmental and Disease Processes. Am J Epidemiol. 2011;173(9):1078-1084. doi:10.1093/aje/kwq481.

Han L. Truncated, Not Forgotten — Handling Left Truncation in Time-to-Event Studies. NEJM Evid. 2024;3(12). doi:10.1056/EVIDe2400384.

Fine, J.P. and R.J. Gray. (1999). A proportional hazards model for the subdistribution of a competing risk. Journal of the American statistical association, 94(446), pp. 496-509. doi: 10.1080/01621459.1999.10474144.

Cook, R. J., & Lawless, J. F. (2024). Selection processes, transportability, and failure time analysis in life history studies. Biostatistics. Published online October 27, 2024.

Van Calster, B., Steverberg, E. W., Wynants, L., & Van Smeden, M. (2023). There is no such thing as a validated prediction model. BMC Medicine, 21(1), 70.

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 37ου Πανελληνίου & 3ου Διεθνούς Συνεδρίου Στατιστικής (2025), σελ.37-52

ΔΙΕΡΕΥΝΗΣΗ ΑΞΙΟΠΟΙΗΣΗΣ ΑΝΑΝΕΩΣΙΜΩΝ ΠΗΓΩΝ ΕΝΕΡΓΕΙΑΣ ΑΠΟ ΤΑ ΚΡΑΤΗ ΜΕΛΗ ΤΟΥ ΕΥΡΩΠΑΪΚΟΥ ΝΟΤΟΥ

Καπετανοπούλου Παρασκευή
Τμήμα Μηχανολόγων Μηχανικών, Πανεπιστήμιο Δυτικής Μακεδονίας
pkapetanopoulou@uowm.gr

ΠΕΡΙΛΗΨΗ

Βάσει της Οδηγίας για την προώθηση της χρήσης ενέργειας από Ανανεώσιμες Πηγές Ενέργειας (Οδηγία 2023/2413, Ευρωπαϊκής Επιτροπής), ο στόχος της ΕΕ για τις ανανεώσιμες πηγές ενέργειας για το 2030 έχει αναθεωρηθεί προς τα πάνω από 32 % σε 42,5 % (με στόχο την αύξησή του σε 45 %). Προς τη κατεύθυνση αυτή οι χώρες της ΕΕ πρέπει i) να εντείνουν τις προσπάθειές τους και να συμμορφωθούν συλλογικά με τον νέο στόχο της Ευρωπαϊκής Ένωσης για το 2030 και ii) να αυξήσουν το μερίδιο των Ανανεώσιμων Πηγών Ενέργειας στην ακαθάριστη τελική κατανάλωση ενέργειας. Η παρούσα εργασία έχει ως στόχο τη διερεύνηση και συγκριτική ανάλυση της επίδοσης των κρατών μελών του Ευρωπαϊκού Νότου (Ελλάδα, Ιταλία, Ισπανία και Πορτογαλία) σχετικά με τη χρήση των Ανανεώσιμων Πηγών Ενέργειας και την παραγωγή ηλεκτρικής ενέργειας από αυτές.

Λέξεις Κλειδιά: Ανανεώσιμες Πηγές Ενέργειας, Στατιστική Ανάλυση, Ευρωπαϊκός Νότος.

1. ΕΙΣΑΓΩΓΗ

Οι Ανανεώσιμες Πηγές Ενέργειας (ΑΠΕ) αποτελούν βασικό πυλώνα της ενεργειακής στρατηγικής της Ευρωπαϊκής Ένωσης (ΕΕ) για την αντιμετώπιση της κλιματικής αλλαγής και της ενεργειακής κρίσης, καθώς επίσης και για την προώθηση της ενεργειακής ασφάλειας (European Commission, 2025, Green transition). Με την αυξανόμενη ανησυχία για τις επιπτώσεις των ορυκτών καυσίμων στο περιβάλλον και την ανάγκη για μείωση των εκπομπών αερίων του θερμοκηπίου, οι ΑΠΕ αποκτούν καθοριστική σημασία στο ενεργειακό μείγμα των κρατών-μελών της ΕΕ (European Parliament, 2024, Cutting EU greenhouse gas emissions: national targets for 2030). Τα τελευταία χρόνια γίνονται σημαντικές επενδύσεις σε ολόκληρη την Ευρωπαϊκή Ένωση για την ανάπτυξη και υιοθέτηση τεχνολογιών ανανεώσιμων πηγών ενέργειας (European Parliament, 2025, How the EU is boosting renewable energy). Ωστόσο, η πρόοδος και η αποτελεσματικότητα της χρήσης των ΑΠΕ διαφέρει σημαντικά μεταξύ των κρατών-μελών της ΕΕ. Οι διαφορές αυτές απορρέουν από πολλούς παράγοντες, όπως η πρόσβαση σε φυσικούς πόρους, η κρατική πολιτική, το ρυθμιστικό πλαίσιο και η οικονομική κατάσταση κάθε κράτους.

Σύμφωνα με την οδηγία 2009/28/ΕΚ του Ευρωπαϊκού Κοινοβουλίου, ως ενέργεια από ανανεώσιμες μη ορυκτές πηγές θεωρείται η αιολική, ηλιακή, αεροθερμική, γεωθερμική, υδροθερμική και ενέργεια των ωκεανών, υδροηλεκτρική, από βιομάζα, από τα εκλυόμενα στους χώρους υγειονομικής ταφής αέρια, από αέρια μονάδων επεξεργασίας λυμάτων και από βιοαέρια. Οι Ανανεώσιμες Πηγές Ενέργειας χαρακτηρίζονται και ως «ήπιες» μορφές ενέργειας καθώς για την εκμετάλλευσή τους δεν απαιτείται κάποια ενεργητική παρέμβαση (εξόρυξη, άντληση ή καύση) και δεν αποδεσμεύουν τοξικά/ραδιενεργά απόβλητα και διοξείδιο του άνθρακα. Αφορμή για τη συγκεκριμένη διερεύνηση αποτέλεσαν πρόσφατα δημοσιεύματα με τίτλους «Ο ευρωπαϊκός Νότος «οδηγεί» την οικονομία της ΕΕ», «Πώς ο ευρωπαϊκός Νότος κρατάει όρθια την Ευρωζώνη;» τα οποία αναφέρονται στα ιδιαιτέρως θετικά βήματα στη μετάβαση στις ΑΠΕ και στην ταχύτερη πρόοδο που έχουν επιδείξει οι τέσσερις μεγαλύτερες οικονομίες της Νότιας Ευρώπης (Ελλάδα, Ισπανία, Ιταλία και Πορτογαλία), συγκριτικά με

κράτη του Βορρά. Παρά την κοινή γεωγραφική τους θέση στη Μεσόγειο η Ελλάδα, η Ισπανία η Ιταλία και η Πορτογαλία έχουν διαφορετικές προσεγγίσεις στην ανάπτυξη των ΑΠΕ, ανάλογα με τις κλιματολογικές συνθήκες και την ενεργειακή τους στρατηγική. Η παρούσα εργασία αποτελεί την πρώτη προσπάθεια συγκριτικής αξιολόγησης των τεσσάρων κρατών μελών του Ευρωπαϊκού Νότου και η διερεύνηση που πραγματοποιείται έχει διττό στόχο. Ειδικότερα, στοχεύει στη διεξαγωγή συγκριτικής ανάλυσης αναφορικά με:

i) το συνολικό μερίδιο ενέργειας από ΑΠΕ (ως % της ακαθάριστης τελικής κατανάλωσης ενέργειας) και την εκτίμησή του κατά το έτος 2030 χρονιά ορόσημο που έχει θέσει η ΕΕ για την επίτευξη ποσοστού τουλάχιστον ίσου με 42.5% και
ii) το μερίδιο χρήσης των ΑΠΕ (για ηλεκτρική ενέργεια, θέρμανση-ψύξη και μεταφορές) και την ένταξη των σχετικών μεριδίων στο ενεργειακό μείγμα των τεσσάρων κρατών-μελών του Ευρωπαϊκού Νότου.

Η διάρθρωση της εργασίας έχει ως εξής: Στην ενότητα 2 παρουσιάζεται το βασικό νομοθετικό πλαίσιο (Οδηγίες, Στόχοι και Δράσεις) της Ευρωπαϊκής Ένωσης που καλούνται να ακολουθήσουν όλα τα κράτη μέλη αναφορικά με τις ΑΠΕ. Το μεθοδολογικό πλαίσιο διερεύνησης περιγράφεται στην ενότητα 3 ενώ στην ενότητα 4 παρουσιάζονται τα αποτελέσματα της ανάλυσης. Τέλος, στην ενότητα 5 δίνονται τα συμπεράσματα της μελέτης καθώς και προτάσεις για μελλοντική έρευνα.

2. ΝΟΜΟΘΕΤΙΚΟ ΠΛΑΙΣΙΟ ΕΥΡΩΠΑΙΚΗΣ ΕΝΩΣΗΣ

2.1 Οδηγίες και Στόχοι

Μετά τις ευρωεκλογές τον Ιούνιο του 2024, η ΕΕ έθεσε προτεραιότητες που διαμορφώνουν τις πολιτικές της έως το 2029. Σε αυτό το στρατηγικό θεματολόγιο της Ευρωπαϊκής Ένωσης βασική προτεραιότητα για την περίοδο 2024-2029 αποτελεί η εξασφάλιση της επιτυχίας της πράσινης μετάβασης και η επένδυση σε ένα βιώσιμο ενεργειακό μέλλον για την Ευρώπη (European Commission, 2025 «Renewable energy - directive, targets and rules»). Στο πλαίσιο της Ευρωπαϊκής Πράσινης Συμφωνίας, οι ανανεώσιμες πηγές ενέργειας αποτελούν πυλώνα της μετάβασης σε καθαρές μορφές ενέργειας καθώς χαρακτηρίζονται ως «εγχώριο» προϊόν χαμηλού κόστους γεγονός που μειώνει την εξάρτηση της Ευρώπης από εξωτερικούς προμηθευτές. Αυτός είναι ο λόγος για τον οποίο το επίπεδο φιλοδοξίας της ΕΕ για την αύξηση του μεριδίου των Ανανεώσιμων Πηγών Ενέργειας στο ενεργειακό της μείγμα επανεξετάζεται τακτικά. Τα μέτρα που απαιτούνται για την προώθηση των ΑΠΕ συμβάλουν καθοριστικά στη μείωση των εκπομπών αερίων του θερμοκηπίου στην Ευρώπη και στην τήρηση των δεσμεύσεων της συμφωνίας του Παρισιού. Η οδηγία του 2009 (2009/28/ΕΚ) σχετικά με την προώθηση της χρήσης ενέργειας από ανανεώσιμες πηγές έθετε στόχο μερίδιο 20% των ΑΠΕ στο ενεργειακό μείγμα των κρατών μελών έως το 2020. Το 2018 η οδηγία 2018/2001ΕΚ αποτέλεσε την αναθεώρηση της οδηγίας του 2009 και έθετε νέο στόχο 32% των ΑΠΕ έως το 2030. Η αναθεωρημένη οδηγία του 2018 τέθηκε σε ισχύ τον Δεκέμβριο του 2018 στο πλαίσιο της δέσμης μέτρων «**Καθαρή ενέργεια για όλους τους Ευρωπαίους**». Η εισβολή της Ρωσίας στην Ουκρανία δημιούργησε την ανάγκη επιτάχυνσης της μετάβασης της ΕΕ σε καθαρές μορφές ενέργειας οπότε και αναθεωρείται η οδηγία του 2018. Στις 20 Νοεμβρίου 2023 τίθεται σε ισχύ νέα (η τρέχουσα) οδηγία (2023/2413/ΕΚ). Σύμφωνα με την τρέχουσα οδηγία ο δεσμευτικός στόχος της ΕΕ, για την προώθηση της χρήσης ενέργειας από ΑΠΕ, αυξάνεται σε τουλάχιστον 42,5%, για το 2030 με την φιλοδοξία να φτάνει το 45% (European Commission, 2025, Renewable energy targets).

Η τρέχουσα οδηγία (2023/2413) εισάγει ισχυρότερα μέτρα για να διασφαλίσει την πλήρη αξιοποίηση όλων των δυνατοτήτων για την περαιτέρω ανάπτυξη και αξιοποίηση των ΑΠΕ και να διευκολύνει τις διαδικασίες αδειοδότησης για τα έργα ανανεώσιμων πηγών ενέργειας (European Commission, 2025, Cooperation mechanisms). Η εσωτερική αγορά ενέργειας της Ευρώπης δημιουργεί την ευκαιρία για τις χώρες να συνεργαστούν για την εκμετάλλευση των ΑΠΕ προκειμένου να είναι σε θέση να επιτύχουν τους εθνικούς τους στόχους. Στο πλαίσιο αυτό έχουν θεσπιστεί οι εξής βασικοί μηχανισμοί συνεργασίας:

- στατιστικές μεταβιβάσεις όπου μια ποσότητα ανανεώσιμης ενέργειας αφαιρείται από την πρόοδο μιας χώρας της και προστίθεται σε μια άλλη. Ο συγκεκριμένος μηχανισμός συνεργασίας αποτελεί σημαντικό κίνητρο στις χώρες της ΕΕ να υπερβούν τους στόχους τους, διότι μπορούν να λάβουν πληρωμή για την ενέργεια που μεταφέρουν σε άλλους και επιπλέον επιτρέπει σε χώρες με λιγότερο αποδοτικές ΑΠΕ στη διάθεσή τους να επιτύχουν τους στόχους τους
- κοινά σχέδια μεταξύ χωρών της ΕΕ ή χωρών της ΕΕ και τρίτων χωρών όπου ένα έργο ΑΠΕ στον τομέα της ηλεκτρικής ενέργειας ή της θέρμανσης και ψύξης μπορεί να συγχρηματοδοτηθεί από δύο ή περισσότερες χώρες της ΕΕ και η ενέργεια να μοιραστεί
- κοινά καθεστώτα στήριξης όπου δύο ή περισσότερες χώρες της ΕΕ μπορούν να συγχρηματοδοτήσουν ένα κοινό καθεστώς στήριξης για την ενθάρρυνση της παραγωγής ενέργειας από ανανεώσιμες πηγές σε μία ή και στις δύο επικράτειές τους. Αυτή η μορφή συνεργασίας μπορεί να περιλαμβάνει μέτρα όπως κοινό δασμολόγιο τροφοδότησης, κοινή πριμοδότηση τροφοδότησης ή κοινό καθεστώς ποσοστώσεων και εμπορίας πιστοποιητικών.

Το 2023 το 24,5% της ακαθάριστης τελικής κατανάλωσης ενέργειας στην ΕΕ προήλθε από ΑΠΕ, σημειώνοντας αύξηση κατά 1,4% σε σύγκριση με το 2022. Η Σουηδία είχε το υψηλότερο μερίδιο ανανεώσιμων πηγών ενέργειας στην κατανάλωσή της (66,4 %), έναντι της Φινλανδίας (50,8 %) και της Δανίας (44,9%). Τα χαμηλότερα μερίδια ανανεώσιμων πηγών ενέργειας καταγράφηκαν στο Λουξεμβούργο (11,6%), το Βέλγιο (14,7%) και τη Μάλτα (15,1%). (Eurostat, 2025).

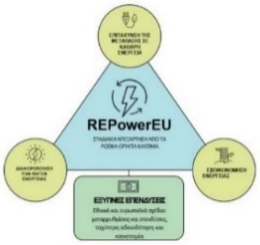
2.2 Πρόγραμμα REPowerEU

Στις 18 Μαΐου 2022 η Ευρωπαϊκή Επιτροπή παρουσίασε το σχέδιο REPowerEU ως απάντηση στις δυσκολίες και τη διαταραχή της παγκόσμιας αγοράς ενέργειας που προκλήθηκε από την εισβολή της Ρωσίας στην Ουκρανία (European Commission, 2025, REPowerEU). Το σχέδιο REPower επιδιώκει να τερματίσει την εξάρτηση της Ευρωπαϊκής Ένωσης (ΕΕ) από τα ρωσικά ορυκτά καύσιμα και να σημειώσει περαιτέρω πρόοδο στην επίτευξη των κλιματικών της στόχων. Οι κύριοι στόχοι του σχεδίου REPowerEU είναι:

- Η μείωση της εξάρτησης από τα Ρωσικά ορυκτά καύσιμα
- Η επιτάχυνση στον τομέα της καθαρής ενέργειας και της πράσινης μετάβασης
- Η ενθάρρυνση μαζικών επενδύσεων σε ΑΠΕ

Το REPowerEU στηρίζεται στην υλοποίηση των προτάσεων της δέσμης Fit for 55 η οποία αποτελεί ένα εκτενές σύνολο νομοθεσίας για το κλίμα, περιλαμβάνει διάφορες νομοθετικές πράξεις και καθορίζει κανόνες και μέτρα για την επίτευξη του στόχου της μείωσης των εκπομπών αερίων του θερμοκηπίου. Ειδικότερα, οι προτάσεις της δέσμης Fit for 55 υποστηρίζουν τον φιλόδοξο στόχο της ΕΕ για την επίτευξη μείωσης κατά τουλάχιστον 55 % των καθαρών εκπομπών αερίων του θερμοκηπίου έως το 2030, και την επίτευξη κλιματικής ουδετερότητας έως το 2050 με μηδενικές καθαρές εκπομπές αερίων του θερμοκηπίου. Με βάση τη δέσμη προτάσεων για την «προσαρμογή στον στόχο του 55%» (Fit for 55) και την ολοκλήρωση των δράσεων που αφορούν την ασφάλεια του ενεργειακού εφοδιασμού και την αποθήκευση ενέργειας, το σχέδιο REPowerEU προτείνει μια πρόσθετη δέσμη δράσεων που αφορούν: i) την εξοικονόμηση ενέργειας, ii) τη διαφοροποίηση του ενεργειακού εφοδιασμού, iii) την ταχεία υποκατάσταση των ορυκτών καυσίμων με την επιτάχυνση της μετάβασης της Ευρώπης σε καθαρές μορφές ενέργειας και iv) τον έξυπνο συνδυασμό επενδύσεων και μεταρρυθμίσεων.

Εικόνα 1. *Σχέδιο REPowerEU (πηγή European Commission)*



Κύριο χρηματοδοτικό μέσο του σχεδίου REPowerEU είναι ο Μηχανισμός Ανάκαμψης και Ανθεκτικότητας (European Council - Council of the European Union, 2025). Όλα τα κράτη μέλη της της Ευρωπαϊκής Ένωσης είναι απαραίτητο να υποβάλλουν δελτία (ένα ανά χώρα της ΕΕ) με τον απολογισμό (λεπτομερή ανάλυση) της προόδου τους. Η πρώτη υποβολή των 27 απογραφικών δελτίων αφορούσε τη διετία 2022-2024 (European Commission, 2025, REPowerEU - 2 years on).

3. ΜΕΘΟΔΟΛΟΓΙΚΟ ΠΛΑΙΣΙΟ ΔΙΕΡΕΥΝΗΣΗΣ

3.1 Συλλογή δεδομένων

Για τη συγκριτική αξιολόγηση της επίδοσης των τεσσάρων κρατών μελών του Ευρωπαϊκού Νότου σχετικά με το μερίδιο ενέργειας από Ανανεώσιμες Πηγές Ενέργειας (ως ποσοστό % της ακαθάριστης τελικής κατανάλωσης ενέργειας) όπως και για τα μερίδια της χρήσης των ΑΠΕ σε τομείς όπως η κατανάλωση ηλεκτρικής ενέργειας, η θέρμανση/ψύξη και οι μεταφορές, ως ποσοστά (%) της ακαθάριστης κατανάλωσης ηλεκτρικής ενέργειας, ενέργειας για θέρμανση/ψύξη και ενέργειας για μεταφορές, αντίστοιχα χρησιμοποιήθηκαν τα πιο πρόσφατα διαθέσιμα δεδομένα από την βάση δεδομένων της Eurostat. Η επικαιροποίηση των μεριδίων/ποσοστών στη βάση δεδομένων της Eurostat πραγματοποιήθηκε στις 14/2/2025. Αντλήθηκαν δεδομένα μεριδίων των τελευταίων 14 ετών για τα τέσσερα κράτη μέλη (Ελλάδα, Ιταλία, Ισπανία και Πορτογαλία).

3.2 Μεθοδολογία ανάλυσης

Η στατιστική ανάλυση των δεδομένων πραγματοποιήθηκε μέσω της ανάλυσης διακύμανσης (Analysis of Variance - ANOVA) η οποία αποτελεί μια από τις πλέον χρήσιμες τεχνικές της στατιστικής, με πολλές εφαρμογές σε προβλήματα όπου εμπλέκονται δύο ή και περισσότερες μεταβλητές. Μια τυπική χρήση της ανάλυσης μεταβλητότητας είναι στον στατιστικό έλεγχο ισότητας μέσων τιμών περισσοτέρων των δύο πληθυσμών, που αποτελεί επέκταση του απλού στατιστικού ελέγχου ισότητας των μέσων τιμών δύο κανονικών πληθυσμών, όπου μέσω της ANOVA διαπιστώνεται αν οι μέσες τιμές της μεταβλητής απόκρισης είναι ίδιες ή διαφορετικές στα διάφορα επίπεδα τιμών του ελεγχόμενου παράγοντα (Ταγαράς, 2001).
Για τη στατιστική ανάλυση στην παρούσα έρευνα ως ανεξάρτητη μεταβλητή (ελεγχόμενος παράγοντας) ορίσθηκε η χώρα η οποία εξετάσθηκε σε τέσσερα (α=4) επίπεδα τιμών (Ελλάδα, Ιταλία, Ισπανία και Πορτογαλία). Ελήφθησαν n=14 παρατηρήσεις (ποσοστά) για κάθε ένα από τα α=4 επίπεδα τιμών του παράγοντα και ορίσθηκε ως y_{ij} η j τιμή της τυχαίας μεταβλητής Y (j = 1, 2, .., n) για το i επίπεδο του παράγοντα (i = 1, 2, 3, 4).
Οι 4 διαφορετικές μεταβλητές απόκρισης (Y) που χρησιμοποιήθηκαν για τη συγκριτική αξιολόγηση των ΑΠΕ στις τέσσερις χώρες του Ευρωπαϊκού Νότου είναι οι ακόλουθες:

- Y_1: Μερίδιο ενέργειας από ΑΠΕ (% της ακαθάριστης τελικής κατανάλωσης ενέργειας)
- Y_2: Μερίδιο ενέργειας από ΑΠΕ για Ηλεκτρική ενέργεια (% της ακαθάριστης κατανάλωσης ηλεκτρικής ενέργειας)
- Y_3: Μερίδιο ενέργειας από ΑΠΕ για Θέρμανση και Ψύξη (% της ακαθάριστης κατανάλωσης ενέργειας για θέρμανση/ψύξη)
- Y_4: Μερίδιο ενέργειας από ΑΠΕ στις Μεταφορές (% της ακαθάριστης κατανάλωσης ενέργειας για μεταφορές).

Τα δεδομένα y_{ij} για κάθε μία από τις τέσσερις πιο πάνω μεταβλητές, θεωρείται ότι ακολουθούν το μαθηματικό πρότυπο

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij} \quad \begin{cases} i = 1,2,\ldots,\alpha \\ j = 1,2,\ldots,n \end{cases} \qquad (1)$$

όπου μ είναι η γενική μέση τιμή (overall mean) της μεταβλητής Y, τ_i είναι η επίδραση του επιπέδου i του παράγοντα (factor effect) και ε_{ij} είναι το τυχαίο σφάλμα (random error), το οποίο ακολουθεί κανονική κατανομή με μέση τιμή 0 και μεταβλητότητα σ^2 (Ταγαράς, 2001). Σύμφωνα με το πρότυπο (1) της ANOVA η (άγνωστη) μέση τιμή της Y στο επίπεδο i είναι $\mu_i = \mu + \tau_i$. Η μεταβλητότητα σ^2 θεωρείται σταθερή σε όλα τα επίπεδα του παράγοντα και προέρχεται από σφάλματα μέτρησης και άλλους παράγοντες που πιθανώς επιδρούν αλλά δεν περιλαμβάνονται στον πειραματικό σχεδιασμό, δηλαδή δεν εξετάζονται άμεσα στον πειραματικό σχεδιασμό. Η ανάλυση πραγματοποιήθηκε σύμφωνα με το πρότυπο ορισμένων επιδράσεων, όπου τα τ_i ορίζονται ως αποκλίσεις από τη γενική μέση τιμή $\sum_{i=1}^{a} \tau_i = 0$ και η στατιστική ανάλυση αποσκοπεί στον έλεγχο της υπόθεσης ισότητας των μέσων τιμών της μεταβλητής απόκρισης Y σε όλα τα α επίπεδα τιμών που επελέγησαν προς εξέταση. Συγκεκριμένα, η μηδενική υπόθεση H_0 και η εναλλακτική υπόθεση H_1 διατυπώνονται μαθηματικά ως εξής:

$$H_0: \tau_1 = \tau_2 = \ldots = \tau_\alpha = 0$$
$$H_1: \exists i: \tau_i \neq 0 \qquad (2)$$

Απόρριψη της μηδενικής υπόθεσης σημαίνει ότι η μέση τιμή της μεταβλητής απόκρισης Y παρουσιάζει στατιστικά σημαντικές διαφορές σε διαφορετικά επίπεδα τιμών του παράγοντα, άρα οι τιμές της μεταβλητής απόκρισης Y επηρεάζονται πράγματι από τις τιμές του ελεγχόμενου παράγοντα. Σε αυτή την περίπτωση, προκειμένου να διαπιστωθεί ποια ζεύγη επίπεδων του παράγοντα διαφέρουν στατιστικώς σημαντικά, σε σχέση με τα υπόλοιπα, χρησιμοποιείται η μέθοδος του Tukey. Η συγκεκριμένη μέθοδος είναι κατάλληλη για την κατασκευή όλων των διαστημάτων εμπιστοσύνης των διαφορών των μέσων τιμών των διαφορετικών επιπέδων ενός σταθερού παράγοντα με ταυτόχρονο επίπεδο σημαντικότητας 100(1-α)% (Κουτρουβέλης, 2002).

Βασική υπόθεση του μαθηματικού προτύπου (1) είναι ότι τα τυχαία σφάλματα ε_{ij} αποτελούν ανεξάρτητες τυχαίες μεταβλητές οι οποίες ακολουθούν κανονική κατανομή με μέση τιμή 0 και μεταβλητότητα σ^2, την ίδια σε όλα τα επίπεδα τιμών του παράγοντα. Εάν οι υποθέσεις αυτές παραβιάζονται, τότε η διαδικασία της ανάλυσης μεταβλητότητας δεν αποτελεί ακριβή έλεγχο των υποθέσεων (2) που εξετάζονται και θα πρέπει ίσως να γίνουν διορθωτικές παρεμβάσεις (Montgomery, 2019).

Αναγκαίο επομένως συμπλήρωμα της ανάλυσης μεταβλητότητας είναι ο έλεγχος της ορθότητας των παραπάνω υποθέσεων. Το κύριο διαγνωστικό εργαλείο είναι τα υπόλοιπα (residuals), οι διαφορές e_{ij} μεταξύ των μετρήσεων y_{ij} και των αντίστοιχων εκτιμητριών $\hat{y}_{ij}=\bar{y}_{i.}$ όπου $e_{ij} = y_{ij} - \hat{y}_{ij} = y_{ij} - \bar{y}_{i.}$. Η κανονικότητα της κατανομής των υπολοίπων ελέγχεται με τον έλεγχο κανονικότητας Anderson Darling. Η σταθερότητα της μεταβλητότητας σε όλα τα επίπεδα τιμών του παράγοντα ελέγχεται μέσω του ελέγχου Bartlett, στην περίπτωση που τα υπόλοιπα ακολουθούν κανονική κατανομή, και μέσω του ελέγχου Levene εάν απορρίπτεται η μηδενική υπόθεση της κανονικότητας των υπολοίπων (Ταγαράς 2001).

Για την εκτίμηση και συγκριτική αξιολόγηση του ποσοστού της ακαθάριστης τελικής κατανάλωσης ενέργειας από ΑΠΕ, ώστε να διερευνηθεί εάν και κατά πόσο το έτος 2030 (το οποίο αποτελεί το ορόσημο που έχει θέσει η ΕΕ) θα έχει επιτευχθεί από τα κράτη μέλη του Ευρωπαϊκού Νότου ο στόχος του 42.5%, αξιοποιήθηκαν τα κλασικά Αυτοπαλίνδρομα Μοντέλα Κινητού Μέσου Όρου που αποτελούν τη συνήθη μέθοδο ανάλυσης μη στάσιμων χρονοσειρών (Hyndman and Athanasopoulos 2021, Hyndman and Khandakar 2008). Εάν $\{X_t\}$, t=1,2,…,T είναι η υπό μελέτη χρονοσειρά τότε το αυτοπαλίνδρομο μοντέλο κινητού μέσου τάξης *p, d, q* δίδεται από την εξίσωση

$$y_t = c + \varphi_1 y_{t-1} + \cdots + \varphi_p y_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t \qquad (3)$$

όπου $y_t = (1-B)^d X_t$, $B^d X_t = X_{t-d}$, t=1,2,..,T, ε_t το σφάλμα τη χρονική στιγμή t και p, d & q οι τάξεις του αυτοπαλίνδρομου τμήματος, της διαφόρησης και του τμήματος του κινητού μέρους της χρονοσειράς. Στα μοντέλα αυτά που συμβολίζονται με ARIMA(p,d,q) θα μπορούσε να ενσωματωθεί και η εποχικότητα που όμως στην περίπτωση των ετήσιων δεδομένων κατανάλωσης ενέργειας δεν παίζει σημαντικό ρόλο.
Η αξιολόγηση της προβλεπτικής ικανότητας ενός μοντέλου μπορεί να βασιστεί σε πολλά κριτήρια αλλά εδώ παρουσιάζονται και αξιοποιούνται τα εξής δύο:

- η ρίζα του Μέσου Τετραγωνικού Σφάλματος (Root Mean Square Error - RMSE) $RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(e_t)^2}$, και
- το Μέσο Απόλυτο Σφάλμα (Mean Absolute Error - MAE)

$$MAE = \frac{1}{T}\sum_{t=1}^{T}(|e_t|)$$

όπου $e_t = y_t - \hat{y}_t$ αντιπροσωπεύουν τα κατάλοιπα ενώ y_t η παρατηρούμενη τιμή και $\hat{y}_t$ είναι η εκτιμώμενη τιμή της y_t, t =1,…,T.
Στην παρούσα έρευνα η ανάλυση των δεδομένων πραγματοποιήθηκε με τη χρήση της γλώσσας προγραμματισμού R και του στατιστικού πακέτου MINITAB, ενώ όλοι οι απαιτούμενοι στατιστικοί έλεγχοι πραγματοποιήθηκαν στα συνήθη επίπεδα σημαντικότητας (α=0,1, α=0,05 και α=0,01).

4. ΣΥΓΚΡΙΤΙΚΗ ΑΞΙΟΛΟΓΗΣΗ ΚΡΑΤΩΝ ΜΕΛΩΝ ΤΟΥ ΕΥΡΩΠΑΙΚΟΥ ΝΟΤΟΥ

4.1 Μερίδιο ενέργειας από ΑΠΕ

Θέτοντας ως μεταβλητή απόκρισης το μερίδιο ενέργειας από ΑΠΕ ως % ποσοστό της ακαθάριστης τελικής κατανάλωσης ενέργειας (μεταβλητή απόκρισης Y_1) η στατιστική ανάλυση των αποτελεσμάτων έδειξε στατιστικά σημαντικές διαφοροποιήσεις μεταξύ των τεσσάρων κρατών μελών του Ευρωπαϊκού Νότου.

Πίνακας 1. *Ανάλυση μεταβλητότητας - Μεταβλητή απόκρισης* Y_1

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Χώρα	3	1644.5	548.17	42.96	0.000
Error	52	663.5	12.76		
Total	55	2308.0			

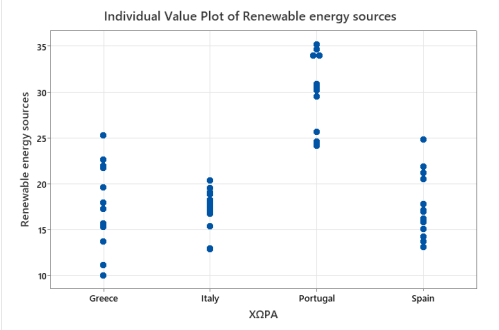
Ειδικότερα, σύμφωνα με τα αποτελέσματα του πίνακα 1, εφόσον p-value=0.000, η H_0 απορρίπτεται για όλα τα συνήθη επίπεδα σημαντικότητας, συνεπώς, ο παράγοντας Χώρα φαίνεται να επηρεάζει στατιστικώς σημαντικά τη μεταβλητή απόκρισης Y_1 (μερίδιο ενέργειας από ΑΠΕ ως % ποσοστό της ακαθάριστης τελικής κατανάλωσης ενέργειας). Οι έλεγχοι επάρκειας του προτύπου της ANOVA έδειξαν ότι οι υποθέσεις του προτύπου δεν παραβιάζονται, συνεπώς δεν υπάρχει λόγος αμφισβήτησης των αποτελεσμάτων. Ειδικότερα, από τον έλεγχο κανονικότητας των υπολοίπων αποδεικνύεται ότι τα υπόλοιπα ακολουθούν κανονική κατανομή σε όλα τα συνήθη επίπεδα σημαντικότητας (p-value=0.567) ενώ και η υπόθεση της σταθερής μεταβλητότητας δεν απορρίπτεται (p-value=0.121).
Εφόσον μέσω της ANOVA διαπιστώθηκε ότι υπάρχει στατιστικά σημαντική διαφοροποίηση μεταξύ των επιπέδων του παράγοντα (Χώρα) και προκειμένου να εντοπιστούν οι διαφοροποιήσεις μεταξύ των τεσσάρων κρατών μελών πραγματοποιήθηκε έλεγχος Tukey. Τα αποτελέσματα των ανά ζεύγη συγκρίσεων παρουσιάζονται στον πίνακα 2.

Πίνακας 2. *Έλεγχος* Tukey

Difference	**T-Value**	**Adjusted p-value**
Italy - Greece	-0.07	1.000
Portugal - Greece	9.28	0.000
Spain - Greece	0.12	0.999
Portugal - Italy	9.36	0.000
Spain - Italy	0.19	0.998
Spain - Portugal	-9.17	0.000

Η σαφής υπεροχή της Πορτογαλίας, η οποία γραφικά αποτυπώνεται στο γράφημα 1, αποδεικνύεται και στατιστικά βάσει των αποτελεσμάτων του ελέγχου Tukey.

Γράφημα 1. *Διάγραμμα κουκίδων μεταβλητής* Y_1



Ειδικότερα, από τις ανά ζεύγη συγκρίσεις (πίνακας 2) των κρατών μελών του Ευρωπαϊκού Νότου οι στατιστικά σημαντικές διαφοροποιήσεις (p-value=0.000) εντοπίζονται μόνο σε τρία (από τα συνολικά έξι) ζεύγη: Πορτογαλία – Ισπανία, Πορτογαλία – Ιταλία και Πορτογαλία – Ελλάδα αναδεικνύοντας την Πορτογαλία πρωτοπόρο στο μερίδιο των ΑΠΕ στο ενεργειακό της μείγμα. Για τα υπόλοιπα κράτη μέλη (Ιταλία - Ισπανία -Ελλάδα) οι διαφοροποιήσεις που «φαίνονται» στο γράφημα 1 δεν αποδεικνύονται στατιστικά σημαντικές σύμφωνα με τα αποτελέσματα του πίνακα 2 (τα σχετικά p-value's είναι μεγαλύτερα από τα συνήθη επίπεδα σημαντικότητας).

Τα αποτελέσματα της ανάλυσης για την εκτίμηση της ετήσιας μέσης μεταβολής και κατ' επέκταση του μεριδίου ενέργειας από ΑΠΕ ως % ποσοστό της ακαθάριστης τελικής κατανάλωσης ενέργειας (μεταβλητή απόκρισης Y_1) για τα τέσσερα κράτη μέλη του Ευρωπαϊκού Νότου παρουσιάζονται στον πίνακα 3.

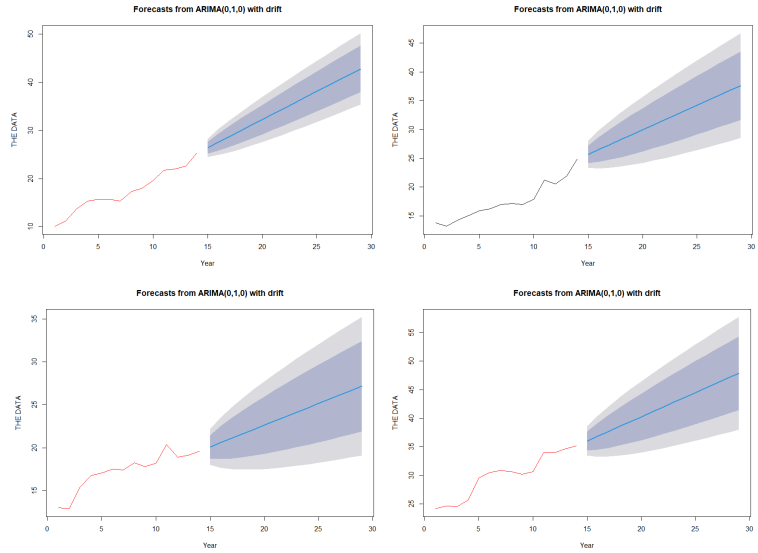
Πίνακας 3. *Τα χαρακτηριστικά του κατάλληλου μοντέλου για το % της ενέργειας στις χώρες του Ευρωπαϊκού Νότου.*

Μοντέλο	**GREECE**	**ITALY**	**PORTUGAL**	**SPAIN**
ARIMA (0,1,0) with drift	1.1686 (0.2594)	0.5055 (0.2833)	0.8472 (0.3457)	0.8515 (0.3191)
AIC	**39.16**	**41.11**	**46.62**	**44.54**
RMSE	0.9013	0.9842	1.201	1.1088
MAE	0.7745374	0.6963447	0.8460051	0.7619016

Οι τιμές των παραμέτρων p, d & q του μοντέλου ARIMA για κάθε χώρα που περιλαμβάνεται στη μελέτη προσδιορίζονται είτε αξιολογώντας τις συναρτήσεις αυτοσυσχέτισης και μερικής αυτοσυσχέτισης είτε αξιοποιώντας τη συνάρτηση auto.arima() της R. Σημειώνεται ότι, για σκοπούς σύγκρισης, σε κάθε περίπτωση πραγματοποιήθηκε επιπλέον ανάλυση και ελήφθησαν τα βασικά χαρακτηριστικά και για το 2ο καλύτερο μοντέλο με βάση το AIC που αποτελεί το κλασικό κριτήριο επιλογής μοντέλων. Βάσει των συγκρίσεων σε όλες τις περιπτώσεις το κατάλληλο μοντέλο είναι ένα μοντέλο ARIMA (0,1,0) με drift, δηλαδή ένας τυχαίος περίπατος (random walk) της μορφής $y_t = c + \varepsilon_t$ όπου η σταθερά c αντιπροσωπεύει τη μέση μεταβολή από χρόνο σε χρόνο.

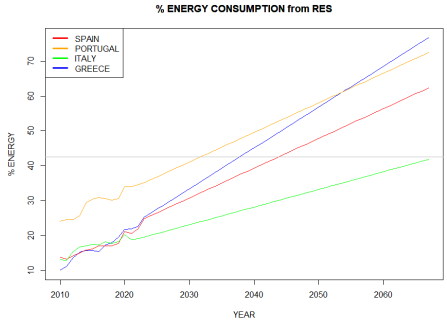
Στην 1η γραμμή του Πίνακα 3 δίδονται οι εκτιμητές των παραμέτρων και τα τυπικά τους σφάλματα (σε παρένθεση) για το μοντέλο ARIMA (0,1,0) τα οποία συνοδεύονται (στη 2η γραμμή του Πίνακα 3) από τις τιμές του κριτηρίου AIC. Στις δύο τελευταίες γραμμές του πίνακα 3 δίδονται οι τιμές των RMSE και MAE.
Με βάση τις τιμές της σταθεράς στον Πίνακα 3, παρατηρούμε ότι η ετήσια μεταβολή του ποσοστού κυμαίνεται από 0.5055 για την Ιταλία έως 1.1686 για την Ελλάδα. Αυτό δείχνει ότι η μέση ετήσια μεταβολή (αύξηση) του ποσοστού για την Ελλάδα είναι ίση με 1.1686% και είναι η μεγαλύτερη που παρατηρείται στον Ευρωπαϊκό Νότο.
Η μέση ετήσια αύξηση για την Ισπανία και την Πορτογαλία είναι σχεδόν ίδια και ισούται με περίπου 0.85% ενώ για την Ιταλία η ετήσια αύξηση είναι πολύ αργή και ίση περίπου με 0.5%. Σημειώνεται ότι για το 2023 (πιο πρόσφατα δεδομένα στο dataset) το ποσοστό κατανάλωσης ενέργειας από ΑΠΕ Ελλάδας και Ισπανίας ήταν σχεδόν 25%, για την Ιταλία ήταν σχεδόν 20% ενώ για την Πορτογαλία ήταν 35%.
Με βάση το μοντέλο που προέκυψε για κάθε χώρα (Πίνακας 3) και την εκτίμηση της σταθεράς ανά χώρα, διαπιστώνεται ότι μόνο η Πορτογαλία εκτιμάται ότι θα επιτύχει το στόχο που έχει θέσει η ΕΕ για το 2030. Πράγματι, η εκτίμηση για την Πορτογαλία για το 2030 είναι σχεδόν ίση με 41.1%. Από την άλλη, για το ίδιο έτος η εκτίμηση για την Ελλάδα είναι 33.5%, για την Ισπανία 31% και για την Ιταλία 23%. Με βάση το ρυθμό αύξησης ανά χώρα, η Ελλάδα πρόκειται να επιτύχει το στόχο το 2037 ενώ η Ισπανία και η Ιταλία θα πρέπει να επιταχύνουν το ρυθμό τους αφού με βάση τα τρέχοντα δεδομένα ο στόχος του 42.5% επιτυγχάνεται το 2042 και το 2067 αντίστοιχα, για τις δύο αυτές χώρες.
Η εξέλιξη του ποσοστού (% της ακαθάριστης τελικής κατανάλωσης ενέργειας) μέχρι το κομβικό χρονικό σημείο του 2030 για κάθε μία από τις 4 χώρες του Ευρωπαϊκού Νότου με τα σχετικά διαστήματα εμπιστοσύνης (σκούρο γκρι=80% - ανοικτό γκρι=95%) φαίνεται στο γράφημα 2.

Γράφημα 2. *Προβλέψεις % κατανάλωσης ενέργειας για Ελλάδα (πάνω αριστερά), Ισπανία (πάνω δεξιά), Ιταλία (κάτω αριστερά), Πορτογαλία (κάτω δεξιά).*



Στο παραπάνω γράφημα οι πρώτες 14 τιμές αφορούν τα διαθέσιμα δεδομένα (2010 - 2023) ενώ οι επόμενες 15 τιμές αφορούν τις προβλέψεις για τα επόμενα 15 χρόνια (2024 - 2038). Με βάση τα διαστήματα εμπιστοσύνης και συγκεκριμένα τα άνω άκρα τους, παρατηρείται ότι το ποσοστό για την Πορτογαλία θα μπορούσε να φτάσει στο τέλος της 15ετίας, μέχρι ακόμα και πάνω από 55%, ενώ τα αντίστοιχα ποσοστά για την Ελλάδα, Ισπανία και Ιταλία είναι περίπου 50%, 45% και 35%.
Το γράφημα 3, όπου ο όρος RES (Renewable Energy Sources) αφορά τις Ανανεώσιμες Πηγές Ενέργειας (ΑΠΕ), είναι ένα πολλαπλό γράφημα που δίνει τις προβλέψεις μαζί με τα διαθέσιμα δεδομένα (2010-2023) για τις 4 χώρες του Ευρωπαϊκού Νότου.

Γράφημα 3. *Προβλέψεις % κατανάλωσης ενέργειας έως το 2067*



Οι προβλέψεις φτάνουν μέχρι το 2067 που είναι το έτος που εκτιμάται ότι και η Ιταλία, που κινείται με τον πιο αργό ρυθμό, θα πετύχει το στόχο της ΕΕ του 42.5%. Λόγω του ταχύτερου ρυθμού ανάπτυξης (με βάση τα σημερινά διαθέσιμα δεδομένα) η Ελλάδα θα έχει προσπεράσει την Πορτογαλία που έχει πιο αργό ρυθμό ανάπτυξης. Το πιο πάνω σχόλιο θα πρέπει να αντιμετωπισθεί με τη δέουσα προσοχή διότι δεν μπορεί να προβλεφθεί ότι ο ρυθμός με βάση τα δεδομένα της περιόδου 2010-2023 θα παραμείνει ίδιος για τα επόμενα 40 και πλέον χρόνια.

4.2 Μερίδιο ενέργειας από ΑΠΕ για Ηλεκτρική ενέργεια

Προκειμένου να διαπιστωθεί εάν τα τέσσερα κράτη μέλη του Ευρωπαϊκού Νότου διαφοροποιούνται στατιστικώς σημαντικά αναφορικά με τη χρήση ΑΠΕ για παραγωγή ηλεκτρικής ενέργειας πραγματοποιήθηκε παρόμοια ανάλυση με αυτή της παραγράφου 4.1. Ειδικότερα, τέθηκε ως μεταβλητή απόκρισης Y_2 το μερίδιο ενέργειας από ΑΠΕ (ως % ποσοστό της ακαθάριστης κατανάλωσης ηλεκτρικής ενέργειας).

Πίνακας 4. *Ανάλυση μεταβλητότητας - Μεταβλητή απόκρισης* Y_2

Source	DF	Adj SS	Adj MS	F-Value	P-Value
ΧΩΡΑ	3	5363	1787.53	30.31	0.000
Error	52	3066	58.97		
Total	55	8429			

Σύμφωνα με τα αποτελέσματα της ANOVA (πίνακας 4) και τα αποτελέσματα του ελέγχου Tukey (πίνακας 5) αποδεικνύεται και πάλι η σαφής υπεροχή της Πορτογαλίας.

Πίνακας 5. *Έλεγχος* Tukey

Difference	T-Value	Adjusted P-Value
Italy - Greece	1.99	0.203
Portugal - Greece	9.05	0.000
Spain - Greece	4.25	0.001
Portugal - Italy	7.06	0.000
Spain - Italy	2.25	0.122
Spain - Portugal	-4.80	0.000

Από τον έλεγχο Tukey και τις ανά ζεύγη συγκρίσεις που παρουσιάζονται στον πίνακα 5 αποδεικνύεται ότι Πορτογαλία παρουσιάζει στατιστικά σημαντικές διαφοροποιήσεις (pvalue's=0.000) συγκριτικά με τα υπόλοιπα τρία κράτη μέλη. Οι συγκεκριμένες διαφοροποιήσεις είναι μη αμφισβητήσιμες καθώς, σύμφωνα με τα αποτελέσματα των ελέγχων επάρκειας του προτύπου της ANOVA, αποδεικνύεται ότι οι υποθέσεις κανονικότητας και ομοσκεδαστικότητας των υπολοίπων δεν απορρίπτονται σε επίπεδο σημαντικότητας 0,01.

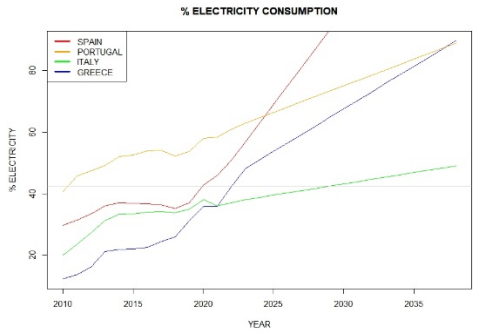
Η μέση τιμή του μεριδίου ΑΠΕ για ηλεκτρική ενέργεια στην Πορτογαλία είναι ίση με 53.018, τιμή πολύ υψηλότερη συγκριτικά με τα υπόλοιπα τρία κράτη μέλη οι μέσες τιμές των οποίων είναι 39.081 για την Ισπανία, 32.538 για την Ιταλία και 26.748 για την Ελλάδα.
Οι διαφοροποιήσεις που προέκυψαν μέσω του ελέγχου Tukey, συνδυαστικά με τις τιμές των μέσων μεριδίων, πέραν της υπεροχής της Πορτογαλίας, αναδεικνύουν ότι μεταξύ Ελλάδας και Ισπανίας η Ισπανία υπερτερεί στο σχετικό μερίδιο ενώ μεταξύ Ελλάδας και Ιταλίας δεν αποδεικνύεται στατιστικά σημαντική διαφοροποίηση. Για την εκτίμηση της ετήσιας μέσης μεταβολής και κατ' επέκταση του μεριδίου ενέργειας από ΑΠΕ για παραγωγή ηλεκτρικής ενέργειας ως % ποσοστό της ακαθάριστης κατανάλωσης ηλεκτρικής ενέργειας (μεταβλητή απόκρισης Y_2) για τα τέσσερα κράτη μέλη του Ευρωπαϊκού Νότου πραγματοποιήθηκε παρόμοια, με αυτή της παραγράφου 4.1, ανάλυση. Τα αποτελέσματα της ανάλυσης παρουσιάζονται στον πίνακα 6.

Πίνακας 6. *Τα χαρακτηριστικά του κατάλληλου μοντέλου για το % της κατανάλωσης ηλεκτρικής ενέργειας στις χώρες του Ευρωπαϊκού Νότου.*

Μοντέλο	**GREECE**	**ITALY**	**PORTUGAL**	**SPAIN**
ARIMA (0,1,0) with drift	2.7629 (0.6201)	ARIMA (0,2,1) -0.5390 (0.2517)	1.7232 (0.4864)	-- (no drift)
AIC	**61.81**	**52.13**	**55.50**	**51.29**
RMSE	2.1544	1.6409	1.6900	1.7469
MAE	1.8952	1.2031	1.1950	1.2584

Στο πολλαπλό γράφημα 4 παρουσιάζονται οι προβλέψεις μαζί με τα διαθέσιμα δεδομένα (2010-2023) για τις 4 χώρες του Ευρωπαϊκού Νότου έως το 2038. Βάσει των αποτελεσμάτων του πίνακα 6, σε αντίθεση με την Ελλάδα και την Πορτογαλία για τις οποίες το καταλληλότερο μοντέλο είναι το ARIMA (0,1,0) με μετατόπιση (drift), η Ισπανία ακολουθεί μεν το ίδιο μοντέλο αλλά χωρίς μετατόπιση ενώ για την Ιταλία το καταλληλότερο μοντέλο είναι το ARIMA (0,2,1) χωρίς μετατόπιση.

Γράφημα 4. *Προβλέψεις % κατανάλωσης ηλεκτρικής ενέργειας έως το 2038*



Η μέση ετήσια (αύξηση) του ποσοστού για την Ελλάδα, βάσει της τιμής της σταθεράς του Πίνακα 6, είναι ίση με 2.7629% και είναι η μεγαλύτερη που παρατηρείται στις χώρες του Ευρωπαϊκού Νότου. Λόγω του ταχύτερου ρυθμού ανάπτυξης της Ελλάδας (με βάση πάντα τα σημερινά διαθέσιμα δεδομένα) περί τα τέλη του ορίζοντα πρόβλεψης (μεταξύ 2035 και 2038) όπως παρουσιάζεται στο γράφημα 4, η Ελλάδα και η Πορτογαλία εκτιμάται ότι θα έχουν την ίδια ανάπτυξη.

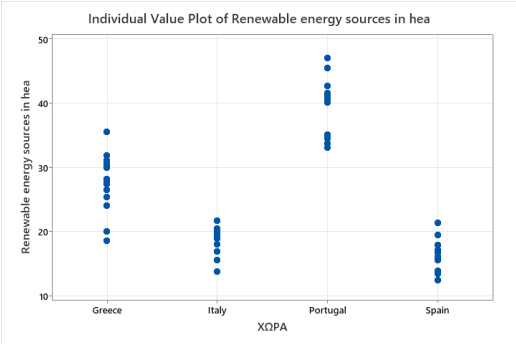
4.3 Μερίδιο ενέργειας από ΑΠΕ για Θέρμανση και Ψύξη

Η ανάλυση για τη μεταβλητή απόκρισης Y_3 (% ποσοστό ΑΠΕ για θέρμανση/ψύξη) διαφοροποιεί το τοπίο στον Ευρωπαϊκό Νότο καθώς μέσω της ανάλυσης μεταβλητότητας (πίνακας 7) και του ελέγχου Tukey (πίνακας 8) σε συνδυασμό με το γράφημα 5 διαπιστώνονται πλήθος στατιστικά σημαντικών διαφορών.

Πίνακας 7. *Ανάλυση μεταβλητότητας - Μεταβλητή απόκρισης* Y_3

Source	DF	Adj SS	Adj MS	F-Value	P-Value
ΧΩΡΑ	3	4830.7	1610.24	131.88	0.000
Error	52	634.9	12.21		
Total	55	5465.6			

Γράφημα 5. *Διάγραμμα κουκίδων μεταβλητής* Y_3



Πρωτίστως αποδεικνύεται ότι η Πορτογαλία υπερέχει και σε αυτή την περίπτωση και διαφοροποιείται στατιστικώς σημαντικά από τα υπόλοιπα κράτη μέλη. Επιπλέον, στατιστικά σημαντικές διαφοροποιήσεις διαπιστώνονται και μεταξύ της Ελλάδας με την Ισπανία και την Ιταλία, με την Ελλάδα να υπερτερεί στατιστικώς σημαντικά στη χρήση ΑΠΕ για θέρμανση/ψύξη.

Πίνακας 8. *Έλεγχος* Tukey

Difference	T-Value	Adjusted P-Value
Italy - Greece	-6.79	0.000
Portugal - Greece	9.28	0.000
Spain - Greece	-8.66	0.000
Portugal - Italy	16.07	0.000
Spain - Italy	-1.87	0.252
Spain - Portugal	-17.95	0.000

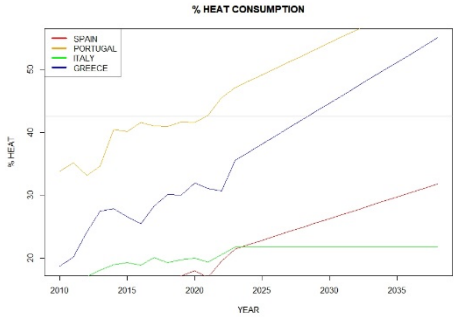
Ειδικότερα, μεταξύ των τεσσάρων κρατών μελών το μέσο μερίδιο ΑΠΕ για θέρμανση/ψύξη στην Ελλάδα είναι το δεύτερο μεγαλύτερο (ίσο με 27.698) με την Πορτογαλία να κατέχει και πάλι την πρώτη θέση με μέση τιμή μεριδίου ίση με 39.959. Οι αντίστοιχες τιμές του μέσου μεριδίου για την Ιταλία είναι 18.73 και για την Ισπανία 16.257. Τα αποτελέσματα του πίνακα 8 συνδυαστικά με τις τιμές των μέσων μεριδίων ΑΠΕ για θέρμανση/ψύξη πέραν της υπεροχής της Πορτογαλίας φανερώνουν ότι η Ελλάδα υπερτερεί στατιστικώς σημαντικά στο σχετικό μερίδιο έναντι και της Ιταλίας και της Ισπανίας. Η μονή διαφοροποίηση που δεν αποδεικνύεται στατιστικά σημαντική, (σύμφωνα με τα αποτελέσματα του πίνακα 8), είναι αυτή μεταξύ της Ισπανίας και της Ιταλίας (γεγονός που αποτυπώνεται και στο γράφημα 5). Από τα αποτελέσματα των ελέγχων επάρκειας του προτύπου της ANOVA διαπιστώθηκε ότι και πάλι δεν απορρίπτονται οι υποθέσεις κανονικότητας και ομοσκεδαστικότητας των υπολοίπων σε

επίπεδο σημαντικότητας 0,01. Τα αποτελέσματα για την εκτίμηση του μεριδίου ενέργειας από ΑΠΕ για θέρμανση/ψύξη ως % ποσοστό ΑΠΕ για θέρμανση/ψύξη (μεταβλητή απόκρισης Y_3) για τα τέσσερα κράτη μέλη του Ευρωπαϊκού Νότου παρουσιάζονται στον πίνακα 9 και στο πολλαπλό γράφημα 6.

Πίνακας 9. *Τα χαρακτηριστικά του κατάλληλου μοντέλου για το % ΑΠΕ για θέρμανση/ψύξη στις χώρες του Ευρωπαϊκού Νότου.*

Μοντέλο	GREECE	ITALY	PORTUGAL	SPAIN
ARIMA (0,1,0) with drift	1.2992 (0.5474)	-- (no drift)	1.0218 (0.5089)	0.6885 (0.2796)
AIC	**58.57**	**45.15**	**56.68**	**41.1**
RMSE	1.9017	1.2257	1.7683	0.9713
MAE	1.5990	0.9506	1.2286	0.7662

Γράφημα 6. *Προβλέψεις % ΑΠΕ για θέρμανση/ψύξη έως το 2038*



Βάσει των αποτελεσμάτων του πίνακα 9 διαπιστώνεται ότι σε όλες τις περιπτώσεις το κατάλληλο μοντέλο είναι ένα μοντέλο ARIMA (0,1,0) όμως στην περίπτωση της Ιταλίας από τη σχετική ανάλυση δεν προέκυψε παράμετρος μετατόπισης προς εκτίμηση. Με βάση τις τιμές των σταθερών στον Πίνακα 9 παρατηρούμε ότι η μέση ετήσια μεταβολή (αύξηση) του ποσοστού για την Ελλάδα (ίση με 1.2992%) είναι η μεγαλύτερη που παρατηρείται στον Ευρωπαϊκό Νότο και ακολουθεί η Πορτογαλία με ποσοστό ίσο με 1.0218%. Με βάση τα διαθέσιμα δεδομένα, από το πολλαπλό γράφημα 6 αναδεικνύεται η υπεροχή της Πορτογαλίας και της Ελλάδας έναντι της Ισπανίας και της Ιταλίας έως το 2038.

4.4 Μερίδιο ενέργειας από ΑΠΕ στις Μεταφορές

Η ανάλυση που πραγματοποιήθηκε για το μερίδιο ενέργειας που χρησιμοποιείται για μεταφορές (μεταβλητή απόκρισης Y_4) ανέδειξε τη χαμηλή θέση της Ελλάδας (συγκριτικά με τα υπόλοιπα κράτη μέλη του Ευρωπαϊκού Νότου).

Πίνακας 10. *Ανάλυση μεταβλητότητας - Μεταβλητή απόκρισης* Y_4

Source	DF	Adj SS	Adj MS	F-Value	P-Value
ΧΩΡΑ	3	174.8	58.265	6.59	0.001
Error	52	460.0	8.846		
Total	55	634.8			

Βάσει των αποτελεσμάτων της ANOVA (πίνακας 10) αποδεικνύεται η ύπαρξη στατιστικά σημαντικών διαφοροποιήσεων (p-value=0.001) μεταξύ των κρατών μελών του Ευρωπαϊκού Νότου ενώ από τις ανά ζεύγη συγκρίσεις (πίνακας 11) οι στατιστικά σημαντικές διαφοροποιήσεις (p-value<0.01) εντοπίζονται μόνο σε δύο (από τα συνολικά έξι) ζεύγη κρατών μελών.

Πίνακας 11. *Έλεγχος* Tukey

Difference	T-Value	Adjusted P-Value
Italy - Greece	4.21	0.001
Portugal - Greece	3.34	0.008
Spain - Greece	2.36	0.097
Portugal - Italy	-0.87	0.822
Spain - Italy	-1.84	0.266
Spain - Portugal	-0.97	0.765

Από την ανάλυση προέκυψε ότι, μεταξύ των τεσσάρων χωρών, το υψηλότερο μέσο μερίδιο ΑΠΕ για μεταφορές κατέχει η Ιταλία (ίσο με 7.4696) ενώ το χαμηλότερο κατέχει η Ελλάδα, μόλις 2.742. Για την Πορτογαλία και την Ισπανία οι τιμές του μέσου μεριδίου είναι 6.498 και 5.40, αντίστοιχα.

Τα αποτελέσματα του πίνακα 11 συνδυαστικά με τις τιμές των μέσων μεριδίων ΑΠΕ για μεταφορές φανερώνουν ότι η Ελλάδα κατέχει την τελευταία θέση και διαφοροποιείται στατιστικώς σημαντικά από όλα τα υπόλοιπα κράτη μέλη του Ευρωπαϊκού Νότου. Μεταξύ των υπολοίπων κρατών μελών (Ιταλία - Ισπανία - Πορτογαλία) η ανάλυση δεν φανέρωσε στατιστικά σημαντικές διαφοροποιήσεις (όλα τα p-value's είναι μεγαλύτερα από τα συνήθη επίπεδα σημαντικότητας). Οι έλεγχοι επάρκειας που πραγματοποιήθηκαν και σε αυτή την περίπτωση δεν θέτουν τα αποτελέσματα υπό αμφισβήτηση καθώς σε επίπεδο σημαντικότητας α=0,01 δεν απορρίπτονται οι υποθέσεις κανονικότητας και ομοσκεδαστικότητας των υπολοίπων, αντίστοιχα. Από την ανάλυση που πραγματοποιήθηκε για την εκτίμηση της ετήσιας μέσης μεταβολής του μεριδίου ενέργειας από ΑΠΕ για μεταφορές (μεταβλητή απόκρισης Y_4) για τα τέσσερα κράτη μέλη του Ευρωπαϊκού Νότου, προέκυψε ότι το κατάλληλο μοντέλο πρόβλεψης είναι σε κάθε περίπτωση ένα απλό μοντέλο κινητού μέσου (MA(1)). Βάσει του συγκεκριμένου μοντέλου οι προβλέψεις για οποιοδήποτε χρονική στιγμή στο μέλλον ταυτίζονται με την τελευταία παρατήρηση που διαθέτουμε, συνεπώς περαιτέρω ανάλυση δεν έχει ουσιαστικά νόημα.

5. ΣΥΜΠΕΡΑΣΜΑΤΑ

Παρά την κοινή γεωγραφική θέση οι τέσσερις μεγαλύτερες οικονομίες του Ευρωπαϊκού Νότου έχουν επιδείξει διαφορετικές προσεγγίσεις στην ανάπτυξη και χρήση των Ανανεώσιμών Πηγών Ενέργειας. Από την ανάλυση που πραγματοποιήθηκε αποδεικνύεται η σαφής υπεροχή της Πορτογαλίας (η οποία αποδεικνύεται στατιστικώς σημαντική συγκριτικά με τα υπόλοιπα κράτη μέλη του Ευρωπαϊκού Νότου) τόσο στο μερίδιο ενέργειας από ΑΠΕ, ως ποσοστό της ακαθάριστης τελικής κατανάλωσης ενέργειας, όσο και στο μερίδιο ενέργειας από ΑΠΕ για παραγωγή ηλεκτρικής ενέργεια και για Θέρμανση και Ψύξη. Η Ελλάδα υπερτερεί έναντι της Ιταλίας και της Ισπανίας στο μερίδιο των ΑΠΕ για θέρμανση / ψύξη καθώς τα τελευταία χρόνια είναι μεταξύ των χωρών που κατέγραψαν τη μεγαλύτερη αύξηση στο σχετικό μερίδιο γεγονός που διαπιστώθηκε από την παρούσα έρευνα. Στο πεδίο της χρήσης ενέργειας από ΑΠΕ για μεταφορές η εικόνα (συνολικά) στην Ευρωπαϊκή Ένωση (ΕΕ) δεν είναι ιδιαιτέρως ενθαρρυντική καθώς το συνολικό μερίδιο της ενέργειας από ΑΠΕ στις μεταφορές είναι μόλις 10,8% στην ΕΕ. Με δεδομένο ότι στην ΕΕ έως το 2030 αναμένονται περισσότερα από $30*10^6$ ηλεκτρικά οχήματα όλα τα κράτη μέλη θα πρέπει να εντείνουν τις προσπάθειές τους ώστε να αξιοποιηθεί η ηλεκτρική ενέργεια από ΑΠΕ στις μεταφορές. Μεταξύ των κρατών μελών του Ευρωπαϊκού Νότου διαπιστώθηκε ότι Ελλάδα παρουσιάζει το χαμηλότερο μερίδιο ενέργειας από ΑΠΕ για μεταφορές και διαφοροποιείται στατιστικώς σημαντικά από τα υπόλοιπα τρία κράτη μέλη τα οποία μεταξύ τους δεν διαφοροποιούνται. Ταυτόχρονα, με βάση την ανάλυση χρονολογικών σειρών που χρησιμοποιήθηκε, διαπιστώθηκε ότι η εξέλιξη του ποσοστού

κατανάλωσης ενέργειας από ΑΠΕ ακολουθεί ένα τυχαίο περίπατο (random walk) με ρυθμό πάνω από 1% ετησίως για την Ελλάδα, περίπου 0.8% για Πορτογαλία και Ισπανία και μόλις 0.50% για την Ιταλία. Με βάση τα διαθέσιμα δεδομένα και τον τρέχοντα ρυθμό η Πορτογαλία το 2030 θα έχει προσεγγίσει (41.1%) το όριο του 42.5% που έχει θέσει η ΕΕ. Η Ελλάδα θα χρειαστεί να περιμένει μέχρι το 2037, η Ισπανία μέχρι το 2042 ενώ η Ιταλία θα πρέπει να περιμένει μέχρι το 2067. Σε οποιαδήποτε περίπτωση η Ελλάδα και η Πορτογαλία θα πρέπει να συνεχίσουν με τουλάχιστον τον ίδιο ρυθμό ενώ Ισπανία και Ιταλία θα πρέπει να αυξήσουν σημαντικά το ρυθμό τους ώστε να βελτιωθούν οι προβλέψεις που έχουν γίνει με βάση τα διαθέσιμα τρέχοντα δεδομένα. Η διερεύνηση που πραγματοποιήθηκε στη συγκεκριμένη εργασία μπορεί να επεκταθεί και στα υπόλοιπα κράτη μέλη της ΕΕ, ή σε περιοχές της ΕΕ που χαρακτηρίζονται από διαφορετικές, σε σχέση με τον Ευρωπαϊκό Νότο, κλιματολογικές συνθήκες. Επιπλέον οι μελέτες στο πεδίο των ΑΠΕ μπορούν να εμπλουτιστούν με μοντέλα προβλέψεων τα οποία θα αποτυπώνουν την εικόνα του μέλλοντος στη χρήση και παραγωγή ενέργειας από ΑΠΕ.

ABSTRACT

Under the Directive on the Use of Energy from Renewable Energy Sources (Directive 2023/2413, European Commission), the EU's 2030 renewable energy target has been revised upwards from 32% to 42.5% (with a view to increasing it to 45%). In this direction, EU countries must i) step up their efforts and collectively comply with the European Union's new 2030 target and ii) increase the share of Renewable Energy Sources in gross final energy consumption. This paper aims to investigate, and comparative analyze the performance of the member states of the European South (Greece, Italy, Spain and Portugal) regarding the use of Renewable Energy Sources and the production of electricity from them.

Ευχαριστίες: Ιδιαίτερες ευχαριστίες στο Πανεπιστήμιο Δυτικής Μακεδονίας για την κάλυψη των δαπανών συμμετοχής στο συνέδριο. Θερμές ευχαριστίες στον κριτή της εργασίας για τα εποικοδομητικά σχόλιά του τα οποία συνέβαλαν σημαντικά στην ολοκλήρωση της εργασίας.

ΑΝΑΦΟΡΕΣ

European Commission, (2025). Green transition. Διαθέσιμο σε https://reform-support.ec.europa.eu/what-we-do/green-transition_en

European Parliament, (2024). Cutting EU greenhouse gas emissions: national targets for 2030. Διαθέσιμο σε https://www.europarl.europa.eu/topics/en/article/20180208STO97442/cutting-eu-greenhouse-gas-emissions-national-targets-for-2030

European Parliament, (2025). How the EU is boosting renewable energy. Διαθέσιμο σε https://www.europarl.europa.eu/topics/en/article/20221128STO58001/how-the-eu-is-boosting-renewable-energy

European Commission, (2025). Renewable energy - directive, targets and rules. Διαθέσιμο σε https://energy.ec.europa.eu/topics/renewable-energy/renewable-energy-directive-targets-and-rules_en?prefLang=el

European Commission, (2025). Renewable energy targets. Διαθέσιμο σε https://energy.ec.europa.eu/topics/renewable-energy/renewable-energy-directive-targets-and-rules/renewable-energy-targets_en?prefLang=el

European Commission, (2025). Cooperation mechanisms. Διαθέσιμο σε https://energy.ec.europa.eu/topics/renewable-energy/renewable-energy-directive-targets-and-rules/cooperation-mechanisms_en?prefLang=el

European Commission, (2025). REPowerEU. Διαθέσιμο σε https://eur-lex.europa.eu/legal-content/EL/TXT/HTML/?uri=CELEX:52022DC0230

European Commission, (2025). REPowerEU - 2 years on. Διαθέσιμο σε https://energy.ec.europa.eu/publications/repowereu-2-years_en

European Council - Council of the European Union, (2025). The REPowerEU plan explained**.** Διαθέσιμο σε https://www.consilium.europa.eu/en/infographics/repowereu/

European Union, (2025). Οδηγία 2009/28/ΕΚ του Ευρωπαϊκού Κοινοβουλίου και του Συμβουλίου, της 23ης Απριλίου 2009. Διαθέσιμο σε https://eur-lex.europa.eu/legal-content/el/TXT/?uri=CELEX%3A32009L0028

European Union, (2025). Οδηγία 2018/2001 του Ευρωπαϊκού Κοινοβουλίου και του Συμβουλίου της 11ης Δεκεμβρίου 2018. Διαθέσιμο σε https://eur-lex.europa.eu/legal-content/EL/TXT/?uri=CELEX:32018L2001

European Union, (2025). Οδηγία 2023/2413 του Ευρωπαϊκού Κοινοβουλίου και του Συμβουλίου, της 18ης Οκτωβρίου 2023. Διαθέσιμο σε https://eur-lex.europa.eu/eli/dir/2023/2413/oj?locale=el

Eurostat, (2025). Renewables account for 24.5% of EU energy use in 2023. Διαθέσιμο σε https://ec.europa.eu/eurostat/en/web/products-eurostat-news/w/ddn-20241219-3

Κουτρουβέλης, Ι. (2002). Σχεδιασμός και ανάλυση πειραμάτων, Τόμος Β, Ελληνικό Ανοιχτό Πανεπιστήμιο.

Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and practice, 3rd edition*. OTexts: Melbourne. OTexts.com/fpp3

Hyndman, R. J., & Khandakar, Y. (2008). Automatic Time Series Forecasting: The forecast Package for *R*. *Journal of Statistical Software*, *27*(3). https://doi.org/10.18637/jss.v027.i03

Montgomery, D.C. (2019). Introduction to Statistical Quality Control, 8th Edition, Wiley,

Ταγαράς, Γ. Ν. (2001). Στατιστικός Έλεγχος Ποιότητας, εκδόσεις Ζήτη, Θεσσαλονίκη.

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 37ου Πανελληνίου & 3ου Διεθνούς Συνεδρίου Στατιστικής (2025), σελ.53-64

ΣΥΝΔΕΣΙΜΟΤΗΤΑ ΑΝΑΜΕΣΑ ΣΤΙΣ ΤΙΜΕΣ ΕΜΠΟΡΕΥΜΑΤΩΝ ΚΑΙ ΤΟΥΣ ΔΕΙΚΤΕΣ ΜΕΤΑΒΛΗΤΟΤΗΤΑΣ ΤΟΥΣ. ΕΜΠΕΙΡΙΚΗ ΔΙΕΡΕΥΝΗΣΗ ΓΙΑ ΤΟΝ ΧΡΥΣΟ ΚΑΙ ΤΟ ΠΕΤΡΕΛΑΙΟ

Π. Καρανικόλας[1], Π. Φουσέκης[2], Γ. Τσακλίδη[3]
[1]Τμήμα Μαθηματικών, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
panteliskaras@gmail.com
[2]Τμήμα Οικονομικών, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
fousekis@econ.auth.gr
[3]Τμήμα Μαθηματικών, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
tsaklidi@math.auth.gr

ΠΕΡΙΛΗΨΗ

Η εργασία εξετάζει τη συνδεσιμότητα ανάμεσα στις τιμές του χρυσού και του πετρελαίου και τους αντίστοιχους δείκτες μεταβλητότητας GVZ και OVX με τη χρήση του υποδείγματος Quantile Frequency Connectedness (QFC). Σύμφωνα με τα εμπειρικά αποτελέσματα η συνδεσιμότητα εξαρτάται τόσο από την κατάσταση της αγοράς (market state) όσο και από τον χρονικό ορίζοντα (frequency/time-scale). Ειδικότερα, διερευνάται και δείχνεται ότι: (α) Για το ζεύγος των μεταβλητών (τιμή χρυσού, GVZ) υπάρχει ισχυρή συνδεσιμότητα μόνο στα άκρα της από κοινού κατανομής ενώ για το ζεύγος (τιμή πετρελαίου, OVX) σε όλες τις καταστάσεις της αγοράς. (β) Και για τα δύο ζεύγη μεταβλητών η συνδεσιμότητα είναι πολύ ισχυρή στην υψηλή συχνότητα (<5 μέρες) και αδύναμη σε όλες τις χαμηλότερες συχνότητες που σημαίνει ότι οι έμποροι αντιδρούν γρήγορα στην εισροή νέας πληροφόρησης. (γ) Υπάρχει ασυμμετρία αναφορικά με την πηγή των διαταράξεων.

Λέξεις Κλειδιά: Εμπόρευμα, Τιμή, Δείκτης μεταβλητότητας, Συνδεσιμότητα.

1. ΕΙΣΑΓΩΓΗ

1.1 Τι είναι οι δείκτες μεταβλητότητας

Οι δείκτες μεταβλητότητας (volatility indexes) είναι χρηματοοικονομικά εργαλεία που μετρούν την αναμενόμενη μεταβλητότητα των αγορών για ένα συγκεκριμένο χρονικό διάστημα στο μέλλον. Η μεταβλητότητα αναφέρεται στο πόσο έντονες είναι οι διακυμάνσεις στην τιμή του περιουσιακού στοιχείου ή του δείκτη και μπορεί να είναι ένδειξη της αβεβαιότητας ή του ρίσκου που σχετίζεται με το εν λόγω περιουσιακό στοιχείο.

1.2 Υπολογισμός δεικτών μεταβλητότητας

Συνήθως ο υπολογισμός των δεικτών μεταβλητότητας προκύπτει χρησιμοποιώντας τις τιμές των δικαιωμάτων προαίρεσης (options) που σχετίζονται με το συγκεκριμένο περιουσιακό στοιχείο καθώς και την τρέχουσα τιμή αυτού. Τα δικαιώματα προαίρεσης είναι παράγωγα που δίνουν το δικαίωμα (αλλά όχι την υποχρέωση) στους επενδυτές να αγοράσουν ή να πουλήσουν το περιουσιακό στοιχείο σε μια συγκεκριμένη τιμή στο μέλλον. Οι δείκτες μεταβλητότητας υπολογίζονται μέσω μαθηματικών μοντέλων που βασίζονται στις τιμές δικαιωμάτων προαίρεσης αγοράς (call options) και πώλησης (put options) για ποικίλες ημερομηνίες λήξης και τιμές εξάσκησης. Με αυτή τη μεθοδολογία υπολογίζεται και δημοσιεύεται καθημερινά το μέτρο της αναμενόμενης μεταβλητότητας ενός περιουσιακού στοιχείου ή δείκτη.

1.3 Κατηγορίες δεικτών μεταβλητότητας

Ο πρώτος δείκτης που δημιουργήθηκε ήταν ο VIX, δείκτης μεταβλητότητας του CBOE (Chicago Board Options Exchange) και στηρίχθηκε σε εργασία του Whaley (1993). Έως τότε οι δείκτες μεταβλητότητας αποτελούσαν αποκλειστικά αντικείμενο ακαδημαϊκής έρευνας. Οι δείκτες μεταβλητότητας χωρίζονται σε κατηγορίες ανάλογα με το περιουσιακό στοιχείο στο οποίο αναφέρονται. Οι κυριότερες κατηγορίες είναι οι δείκτες μεταβλητότητας μετοχικών δεικτών, οι δείκτες μεταβλητότητας ETF μετοχών, δείκτες μεταβλητότητας νομισμάτων, δείκτες μεταβλητότητας επιτοκίων, δείκτες μεταβλητότητας δεικτών και τέλος οι δείκτες μεταβλητότητας εμπορευμάτων. Η τελευταία κατηγορία θα είναι αντικείμενο της παρούσας εργασίας, καθώς θα μελετηθεί εκτενώς η συνδεσιμότητα μεταξύ της τιμής των συμβολαίων μελλοντικής εκπλήρωσης του χρυσού και των συμβολαίων μελλοντικής εκπλήρωσης του αργού πετρελαίου σε σχέση με τους αντίστοιχους δείκτες μεταβλητότητας GVZ (Gold Volatility Index) και OVX (Crude Oil Volatility Index).

1.4 Συμβόλαια μελλοντικής εκπλήρωσης

Τα Συμβόλαια Μελλοντικής Εκπλήρωσης (ΣΜΕ) ανήκουν στην οικογένεια των παράγωγων χρηματοοικονομικών και είναι προθεσμιακά συμβόλαια δηλαδή μία δεσμευτική συμφωνία και υποχρέωση για αγοραπωλησία ενός περιουσιακού στοιχείου, μεταξύ ενός αγοραστή και ενός πωλητή, σε προκαθορισμένη μελλοντική χρονική στιγμή (maturity) και σε προκαθορισμένη τιμή (delivery price). Ειδικότερα, ο επενδυτής που αγοράζει ένα ΣΜΕ αναλαμβάνει την υποχρέωση να αγοράσει το υποκείμενο περιουσιακό στοιχείο την ημέρα λήξης του συμβολαίου στην προκαθορισμένη τιμή (futures price). Αντιθέτως, ο πωλητής ενός ΣΜΕ αναλαμβάνει την υποχρέωση να πουλήσει το υποκείμενο περιουσιακό στοιχείο την ημέρα λήξης στη συμφωνημένη τιμή, κάτι που συνεπάγεται ότι ο πωλητής προσδοκά πτώση της τιμής του περιουσιακού στοιχείου. Οι επενδυτές αυτοί δεν καταβάλουν ή εισπράττουν την αξία του ΣΜΕ αλλά υποχρεούνται να παρέχουν το περιθώριο ασφάλισης (margin). Τα ΣΜΕ αποσκοπούν στην προστασία του επενδυτή από τη μεταβλητότητα των τιμών καθώς αυτός λαμβάνει ένα λογικό κέρδος με την αγορά ενός ΣΜΕ όταν προβλέπεται αύξηση της τιμής του στο μέλλον.

Στην παρούσα εργασία χρησιμοποιούνται οι τιμές των ΣΜΕ για τον χρυσό και το αργό πετρέλαιο, αναφερόμενες στα συμβόλαια με τον πλησιέστερο μήνα λήξης. Οι επενδυτές που διαπραγματεύονται αυτό το συμβόλαιο συμφωνούν νωρίτερα από αυτή την ημερομηνία σε μια τιμή στην οποία θα γίνει η παράδοση του χρυσού ή η τακτοποίηση του συμβολαίου.

1.5 Συνδεσιμότητα

Η συνδεσιμότητα (connectedness) μεταξύ των τιμών ενός εμπορεύματος και του αντίστοιχου δείκτη μεταβλητότητας αναφέρεται στη σχέση ή τον συσχετισμό που υπάρχει μεταξύ των τιμών τους. Ορίζεται ως το ποσοστό της διακύμανσης μιας μεταβλητής του συστήματος, που εξηγείται από μεταβολές που έχουν προηγηθεί σε άλλες μεταβλητές του συστήματος.

1.6 Σχέση απόδοσης – μεταβλητότητας

Η σχέση μεταξύ απόδοσης και μεταβλητότητας αποτελεί έναν από τους θεμελιώδεις άξονες αξιολόγησης της απόδοσης επενδυτικών προϊόντων, ιδιαίτερα στην περίπτωση των ΣΜΕ σε εμπορεύματα. Η απόδοση αναφέρεται στο κέρδος ή τη ζημία που μπορεί να αποφέρει μια επένδυση, ενώ η μεταβλητότητα εκφράζει τη διακύμανση των τιμών ενός χρηματοοικονομικού μέσου μέσα στον χρόνο. Στην πράξη, υψηλότερη μεταβλητότητα συνεπάγεται μεγαλύτερο επενδυτικό κίνδυνο, αλλά ταυτόχρονα προσφέρει τη δυνατότητα για μεγαλύτερες αποδόσεις. Για τον λόγο αυτό, οι δείκτες μεταβλητότητας χρησιμοποιούνται ευρέως τόσο για την αποτίμηση των χρηματοοικονομικών μέσων όσο και για τη διαμόρφωση στρατηγικών διαχείρισης κινδύνου (risk management), όπως η αντιστάθμιση (hedging). Στην αγορά των εμπορευμάτων, για παράδειγμα, η τιμή του πετρελαίου ή του αργού πετρελαίου μπορεί να παρουσιάζει έντονες διακυμάνσεις λόγω γεωπολιτικών γεγονότων ή μεταβολών στην παγκόσμια ζήτηση, επηρεάζοντας έτσι σημαντικά τη μεταβλητότητα των σχετικών ΣΜΕ.

Η συστηματική παρακολούθηση της μεταβλητότητας μέσω δεικτών, όπως η "υπονοούμενη μεταβλητότητα" (implied volatility), παρέχει πολύτιμα εργαλεία σε επενδυτές και διαχειριστές χαρτοφυλακίων, διευκολύνοντάς τους να αξιολογούν αποτελεσματικότερα τη σχέση κινδύνου–απόδοσης πριν από τη λήψη επενδυτικών αποφάσεων.

2. Μεθοδολογία

Η μεθοδολογία που ακολουθήθηκε στην ανάλυση των ζευγών απόδοσης-μεταβλητότητας παρουσιάζεται παρακάτω.

Αρχικά, θεωρείται μία *N*- διάστατη στάσιμη στοχαστική διαδικασία της μορφής:

$$X_t = \left(X_{1,t}, \dots, X_{N,t}\right), όπου\ t = 1, .., T$$

Για να επιτευχθεί ακρίβεια στον υπολογισμό των μέτρων συνδεσιμότητας θα οριστεί ένα διανυσματικό αυτοπαλίνδρομο μοντέλο ποσοστημορίων $QVAR\ (\tau, p)$, ως εξής:

$$X_t = \Phi(\tau) + \sum_{k=1}^{p} \Phi_k(\tau) X_{t-k} + \varepsilon_t(\tau) \quad (1)$$

Στο παραπάνω μοντέλο, το $\tau \in (0,1)$ προσδιορίζει το ποσοστημόριο, το p είναι η τάξη υστέρησης, το $\Phi(\tau)$ είναι ένα διάνυσμα μέσων τιμών διάστασης $N \times 1$, $\Phi_k(\tau)$ είναι πίνακες συντελεστών που εξαρτώνται από το ποσοστημόριο $\varepsilon_t(\tau) \sim N\left(0, \Sigma_\varepsilon(\tau)\right)$ είναι ένα διάνυσμα λευκού θορύβου.

Για να υπολογιστούν οι πίνακες συνδεσιμότητας στα διάφορα ποσοστημόρια, θα μετατραπεί το μοντέλο (1) σε ένα αντίστοιχο μοντέλο κινούμενου μέσου με άπειρη τάξη $QVMA(\tau, \infty)$ με χρήση του θεωρήματος του Wold. Το QVMA βοηθά στην απομόνωση και ανάλυση της πραγματικής σχέσης μεταξύ μεταβλητών αφαιρώντας την επίδραση των αυτοσυσχετίσεων ή άλλων παρόμοιων παραγόντων. Έτσι, η αρχική χρονοσειρά λαμβάνει την μορφή:

$$X_t = \Theta(\tau) + \sum_{h=0}^{\infty} \Theta_h(\tau) \varepsilon_{t-h} \quad (2)$$

Πλέον μπορεί να εφαρμοστεί η Γενικευμένη Ανάλυση Διακύμανσης Σφάλματος Πρόβλεψης (Generalized Forecast Error Variance Decomposition 'GFEVD') των Pesaran και Shin, (1998), στην σχέση (2) για ορίζοντα πρόβλεψης H. Από αυτή τη διαδικασία προκύπτει λοιπόν, ο γενικευμένος πίνακας διακύμανσης σφάλματος πρόβλεψης (GFEV) διάστασης $N \times N$, του οποίου τα στοιχεία $i, j = 1,2, \dots, N$ είναι της μορφής

$$\theta_{i,j}^{H}(\tau) = \frac{\sigma_{jj}^{-1}(\tau) \sum_{h=0}^{H} (\Psi_h(\tau) \sum_\varepsilon(\tau))^2}{\sum_{h=0}^{H} (\Psi_h(\tau) \sum_\varepsilon(\tau) \Psi_h'(\tau))_{ii}} \quad (3)$$

Στην σχέση (3), $\sum_\varepsilon(\tau)$ είναι ο πίνακας συνδιακύμανσης σφάλματος στο ποσοστημόριο τ και $\sigma_{jj}(\tau)$ είναι το $j - στο$ διαγώνιο στοιχείο. Εφόσον τα στοιχεία $\theta_{i,j}^{H}(\tau)$ δεν αθροίζουν στην μονάδα ως προς τις γραμμές, πρέπει να κανονικοποιηθούν μέσω του αθροίσματος ως προς τις γραμμές. Όπως φαίνεται παρακάτω προκύπτουν τα στοιχεία

$$\widetilde{\theta_{i,j}^{H}(\tau)} = \frac{\theta_{i,j}^{H}(\tau)}{\sum_{j=1}^{N} \theta_{i,j}^{H}(\tau)}$$

τα οποία ικανοποιούν τις ιδιότητες (5) και (6)

$$\sum_{j=1}^{N} \widetilde{\theta_{i,j}^{H}(\tau)} = 1 \quad (5)$$

$$\sum_{i=1}^{N}\sum_{j=1}^{N}\widetilde{\theta_{i,j}^{H}(\tau)} = N \qquad (6)$$

Έτσι, $\widetilde{\theta_{i,j}^{H}(\tau)}$ είναι το μέρος του γενικευμένου πίνακα διακύμανσης σφάλματος πρόβλεψης (GFEVD) για τον ορίζοντα πρόβλεψης H βημάτων της $i-\sigma\tau\eta\varsigma$ στοχαστικής διαδικασίας και του $\tau-o\sigma\tau o\acute{\upsilon}$ ποσοστημορίου που μπορεί να αποδοθεί σε σοκ στην $j-\sigma\tau\eta$ διαδικασία, ή ισοδύναμα, η μετάδοση από την διαδικασία j στην διαδικασία i.

Με βάση το άθροισμα των στοιχείων του πίνακα GFEV, τα κανονικοποιημένα στοιχεία αποτελούν τα βασικά στοιχεία για αρκετά εναλλακτικά μέτρα συνδεσιμότητας μεταξύ στοχαστικών διαδικασιών στον τομέα του χρόνου Ando και Tiwari (2022).

Σύμφωνα με την μέθοδο φασματικής ανάλυσης (spectral decomposition) του Stiassny (1996), η ανάλυση της συνδεσιμότητας μπορεί να γίνει και στο πεδίο των συχνοτήτων με την χρήση του μετασχηματισμού Fourier. Ο μετασχηματισμός Fourier του $QVMA(\tau,\infty)$ για δοσμένη συχνότητα $\omega \in \mathbb{R}$ είναι

$$S_X^{\omega}(\tau) = \Psi(e^{-i\omega h})\sum_t \Psi(e^{+i\omega h})$$

όπου, $i=\sqrt{-1}$, $\Psi\left(e^{-i\omega h}\right)=\sum_{h=0}^{\infty} e^{-i\omega h}\,\Psi_h$, η συνάρτηση απόκρισης συχνότητας και Σ_t ο πίνακας συνδιακύμανσης του X_t.

Η συνάρτηση απόκρισης συχνότητας δηλώνει πόσο ενισχύεται ή μειώνεται το πλάτος ενός σήματος και πόσο μετατοπίζεται η φάση του καθώς διέρχεται από το σύστημα.

Έτσι προκύπτουν στοιχεία αντίστοιχα αυτών του (3) στον τομέα των συχνοτήτων, τα οποία είναι

$$\theta_{i,j}^{\omega}(\tau) = \frac{\sigma_{jj}^{-1}(\tau)\sum_{h=0}^{H}(\Psi(\tau)(e^{-i\omega h}\sum_{\varepsilon}(\tau))^2}{\sum_{h=0}^{H}(\Psi(e^{-i\omega h})\sum_{\varepsilon}(\tau)\,\Psi'(e^{+i\omega h})_{ii}}$$

Ο αντίστοιχος κανονικοποιημένος συντελεστής διάχυσης μεταξύ i και j είναι ,

$$\widetilde{\theta_{i,j}^{\omega}(\tau)} = \frac{\theta_{i,j}^{\omega}(\tau)}{\sum_{j=1}^{N}\theta_{i,j}^{\omega}(\tau)}$$

και αντιπροσωπεύει το τμήμα του φάσματος της i διαδικασίας στη συχνότητα και στο τ ποσοστημόριο που μπορεί να αποδοθεί σε κραδασμούς στην j διαδικασία. Τυπικά, η αξιολόγηση της συνδεσιμότητας στον τομέα της συχνότητας δεν πραγματοποιείται για μεμονωμένες συχνότητες αλλά σε εύρη συχνοτήτων (π.χ. βραχυπρόθεσμη, μεσαία και μακροχρόνια). Το μέτρο διάχυσης σε ένα δεδομένο εύρος συχνοτήτων είναι:

$$d = (a,b): a,b \in (-\pi,\pi) \text{ με } a<b$$

και

$$\widetilde{\theta_{i,j}^{d}(\tau)} = \int_a^b \widetilde{\theta_{i,j}^{\omega}}(\omega,\tau)d\omega.$$

Σύμφωνα με τους Diebold και Yilmaz (2014) μπορεί κανείς να ορίσει μια χρονοσειρά από εναλλακτικά μέτρα συνδεσιμότητας με βάση τα ποσοστημόρια και το εύρος των συχνοτήτων.

Το άθροισμα

$$TCI^{d}(\tau) = N^{-1}\sum_{i=1}^{N} TO(\tau)_i^{d} = N^{-1}\sum_{i=1}^{N} FROM(\tau)_i^{d}$$

απεικονίζει τη μέση επίδραση που έχει μια διατάραξη σε μια χρονοσειρά σε όλες τις άλλες. Όσο υψηλότερη είναι η τιμή TCI τόσο μεγαλύτερες είναι οι μεταβολές και αντίστοιχα όσο χαμηλότερη είναι η τιμή TCI τόσο μικρότερες είναι οι μεταβολές, πάνω στο εύρος συχνοτήτων d.

Το άθροισμα

$$TO(\tau)_i^d = \sum_{\substack{i=1 \\ i \neq j}}^{N} \widetilde{\theta_{j,i}^d(\tau)}$$

μετράει πώς μία διαταραχή στην χρονοσειρά i μεταδίδεται στις υπόλοιπες χρονοσειρές j, πάνω στο εύρος συχνοτήτων d.

Το άθροισμα

$$FROM(\tau)_i^d = \sum_{\substack{i=1 \\ i \neq j}}^{N} \widetilde{\theta_{i,j}^d(\tau)}$$

μετράει πώς οι διαταραχές στις χρονοσειρές j μεταδίδονται στην χρονοσειρά i, πάνω στο εύρος συχνοτήτων d.

Η διαφορά

$$NPC(\tau)_{ij}^d = \widetilde{\theta_{i,j}^d(\tau)} - \widetilde{\theta_{j,i}^d(\tau)}$$

είναι η καθαρή διάχυση ζεύγους (net pair spillover). Εάν $NPC(\tau)_{ij}^d > 0$, $(NPC(\tau)_{ij}^d < 0)$ σημαίνει ότι η χρονοσειρά j επηρεάζει την χρονοσειρά i περισσότερο (λιγότερο) από ότι το αντίστροφο, πάνω στο εύρος συχνοτήτων d.

Η διαφορά

$$NET_i^d(\tau) = TO(\tau)_i^d - FROM(\tau)_i^d$$

ερμηνεύεται ως η καθαρή επιρροή που έχει η χρονοσειρά i στο σύνολο των χρονοσειρών του δικτύου. Εάν $NET_i^d(\tau) > 0$ $\left(NET_i^d(\tau) < 0\right)$, η χρονοσειρά i επηρεάζει τις άλλες j σειρές περισσότερο (λιγότερο) από ότι επηρεάζεται αυτή από μεταβολές σε αυτές τις j σειρές.

Τα μέτρα συνδεσιμότητας στον τομέα χρόνου είναι τα αθροίσματα των αντίστοιχων στον τομέα συχνοτήτων σε όλα τα εξεταζόμενα εύρη.

$$NPC(\tau)_{ij}^H = \sum_d NPC(\tau)_{ij}^d$$

$$NET(\tau)_i^H = \sum_d NET(\tau)_i^d$$

$$TO(\tau)_i^H = \sum_d TO(\tau)_i^d$$

$$FROM(\tau)_i^H = \sum_d FROM(\tau)_i^d$$

$$TCI^H(\tau) = \sum_d TCI(\tau)^d$$

Για δίκτυα που αποτελούνται από 2 χρονοσειρές άμεσα προκύπτει ότι

$$TO(\tau)_i^d = FROM(\tau)_j^d$$

$$NPC(\tau)_{ij}^{d} = NET_{i}^{d}(\tau)$$

3. ΔΕΔΟΜΕΝΑ ΚΑΙ ΑΠΟΤΕΛΕΣΜΑΤΑ

3.1 Στασιμότητα για το ζεύγος χρυσού – δείκτη GVZ

Αρχικά, διενεργείται έλεγχος στασιμότητας στο πρώτο ζεύγος χρονοσειρών που θα μελετηθεί, αυτό των τιμών των συμβολαίων μελλοντικής εκπλήρωσης του Χρυσού και των τιμών του δείκτη GVZ. Τα αποτελέσματα των ελέγχων συνοψίζονται στον πίνακα 1.

Πίνακας 1. *Στασιμότητα Χρυσού – δείκτη GVZ*

	KPSS test	Dickey-Fuller test
Δεδομένα χρυσού	Μη στάσιμη	Μη στάσιμη
Λογ. δεδ. Χρυσού	Μη στάσιμη	Μη στάσιμη
Πρώτες διαφορές σε Λογ. Δεδ. χρυσού	**Στάσιμη**	**Στάσιμη**
Δεδομένα δείκτη GVZ	Μη στάσιμη	**Στάσιμη**
Λογ. δεδομένα δείκτη GVZ	Μη στάσιμη	**Στάσιμη**
Πρώτες διαφορές σε λογ. δεδομένα δείκτη GVZ	**Στάσιμη**	**Στάσιμη**

Από τον Πίνακα 1, προκύπτει ότι η ανάλυση της συνδεσιμότητας θα διενεργηθεί στις πρώτες διαφορές των λογαριθμημένων δεδομένων και των δύο χρονοσειρών ώστε να επιτευχθεί κοινή στασιμότητα των δύο χρονοσειρών και στους δύο ελέγχους.

3.2 Ανάλυση συνδεσιμότητας για το ζεύγος χρυσού – δείκτη GVZ

Για να εφαρμοστεί ένα QVAR μοντέλο είναι απαραίτητο να γίνει επιλογή μερικών βασικών παραμέτρων. Η επιλογή γίνεται σύμφωνα με την φύση του προβλήματος και τον στόχο της ανάλυσης. Αρχικά πρέπει να καθοριστεί ο χρονικός ορίζοντας πρόβλεψης του μοντέλου, όπου σε αυτή την περίπτωση είναι $H = 24$. Καθώς το H αυξάνεται οι πληροφορίες που παράγονται από το μοντέλο γίνονται λιγότερο πολύτιμες. Η επιλογή του Η βασίστηκε στον αριθμό χρηματιστηριακών ημερών που αντιστοιχούν περίπου σε έναν μήνα. Έπειτα, σημαντικά είναι τα ποσοστημόρια καθώς και οι χρονικές κλίμακες (συχνότητες) στις οποίες θα πραγματοποιηθεί η εκτίμηση του μοντέλου. Εδώ επιλέχθηκαν τρία ποσοστημόρια (0.05, 0.5 και 0.95) και τρείς συχνότητες (0-5 , 5-10 και >10 μέρες). Η επιλογή των ποσοστών είναι σύμφωνη με τους Qiu και Goodwin (2013) και των Fousekis και Tzaferi (2022). Η επιλογή των συχνοτήτων έγινε με στόχο να χωριστούν οι (κατά προσέγγιση) 22 με 24 χρηματιστηριακές ημέρες που έχεις ένας μήνας σε μία υψηλή συχνότητα (0-5 μέρες), μια χαμηλή συχνότητα (>10 μέρες) και μία ενδιάμεση (5-10 μέρες). Επίσης, σημαντική είναι η επιλογή κατάλληλων τιμών υστέρησης ανά συνδυασμό ποσοστημορίου και συχνότητας. Η παρούσα εργασία ακολουθεί τους Ando κ.ά. (2022) και Χατζηαντωνίου κ.ά. (2022). Υποθέτει ότι το κατάλληλο μήκος υστέρησης είναι ίσο με το βέλτιστο μήκος (εδώ 2), το οποίο επιλέγεται για το υπό όρους μοντέλο μέσου όρου στον τομέα του χρόνου με ελαχιστοποίηση του κριτηρίου πληροφοριών Schwarz.

Πίνακας 2. *Συνδεσιμότητα χρυσού - δείκτη GVZ*

	TCI	TO GVZ	FROM GVZ	NPC
Quantile 0.05				
Time-domain	28.22 (<**0.01**)	28.46 (<**0.01**)	27.99 (<**0.01**)	0.47 (0.25)
High-frequency	23.84 (<**0.01**)	24.02 (<**0.01**)	23.65 (<**0.01**)	0.37 (0.81)
Med-frequency	1.78 (<**0.01**)	1.80 (<**0.01**)	1.76 (<**0.01**)	0.04 (0.94)
Low-frequency	2.61 (<**0.01**)	2.64 (<**0.01**)	2.58 (<**0.01**)	0.06 (0.96)
Quantile 0.5				
	TCI	TO GVZ	FROM GVZ	NPC
Time-domain	1.34 (0.20)	1.36 (0.19)	1.31 (0.21)	0.05 (0.61)
High-frequency	1.01 (0.22)	1.01 (0.22)	1.01 (0.22)	0.00 (0.99)
Med-frequency	0.13 (0.15)	0.14 (0.12)	0.12 (0.22)	0.02 (0.68)
Low-frequency	0.20 (0.15)	0.21 (0.12)	0.18 (0.22)	0.03 (0.68)
Quantile 0.95				
	TCI	TO GVZ	FROM GVZ	NPC
Time-domain	30.31 (<**0.01**)	30.25 (<**0.01**)	30.36 (<**0.01**)	-0.11 (0.82)
High-frequency	23.65 (<**0.01**)	23.94 (<**0.01**)	23.37 (<**0.01**)	0.57 (0.70)
Med-frequency	2.64 (<**0.01**)	2.52 (<**0.01**)	2.77 (<**0.01**)	-0.25 (0.69)
Low-frequency	4.01 (<**0.01**)	3.80 (<**0.01**)	4.22 (<**0.01**)	-0.42 (0.69)

Στον Πίνακα 2 παρουσιάζονται τα αποτελέσματα συνδεσιμότητας για το ζεύγος των τιμών συμβολαίων μελλοντικής εκπλήρωσης του χρυσού και του δείκτη Οι τιμές p-value υπολογίστηκαν με Block Bootstrap (Politis και Romano, 1994) με 1000 επαναλήψεις. Ακόμα, προκύπτει ότι οι τιμές των μέτρων TCI, TO GVZ και FROM GVZ είναι στατιστικά σημαντικές για μόνο για τις ακραίες τιμές των κατανομών και όχι για το μεσαίο ποσοστημόριο. Για τα άκρα των κατανομών, υψηλές τιμές συνδεσιμότητας παρατηρούνται για τον τομέα του χρόνο και την υψηλή συχνότητα ενώ στην μεσαία συχνότητα και την χαμηλή οι τιμές των μέτρων είναι πολύ μικρότερες. Στο μέσο των κατανομών οι τιμές των μέτρων είναι χαμηλές ανεξάρτητα από την συχνότητα. Τέλος ο δείκτης NPC είναι σε κάθε συχνότητα μικρός και στατιστικά μη σημαντικός από όπου συμπεραίνεται ότι οι δύο χρονοσειρές επιδρούν συμμετρικά στο σύστημα. Συνολικά, οι πληροφορίες μεταδίδονται μεταξύ των δύο αγορών

ανάλογα με το μέρος της κατανομής όπου συμβαίνει η διαταραχή και ανεξάρτητα της συχνότητας.

Πίνακας 3. *Χρυσός - δείκτης GVZ. Έλεγχοι συμμετρίας στο φάσμα συχνοτήτων σε κάθε ποσοστημόριο (Ho: Η συνδεσιμότητα είναι ίδια στο φάσμα συχνοτήτων)*

	TCI	TO GVZ	FROM GVZ	NPC
Quantile 0.05				
Test statistic	22.06και -0.83 **(<0.01)**	22.23και -0.84 **(<0.01)**	21.88 και 0.83 **(<0.01)**	0.34 και -0.02 (0.25)
Quantile 0.5				
Test statistic	0.88 και -0.07 (0.14)	0.87 και -0.07 (0.12)	0.89και -0.06 (0.21)	-0.02και -0.01 (0.62)
Quantile 0.95				
Test statistic	21.00και -1.37 **(<0.01)**	21.42και -1.27 **(<0.01)**	20.59και -1.45 **(<0.01)**	0.83 και 0.17 (0.69)

Στον Πίνακα 3 παρουσιάζονται τα αποτελέσματα των ελέγχων συμμετρίας στο φάσμα των συχνοτήτων σε κάθε ποσοστημόριο. Η μηδενική υπόθεση απορρίπτεται στα ακραία ποσοστημόρια. Άμεσα προκύπτει ότι η συνδεσιμότητα είναι ίδια μεταξύ των μέτρων στο πεδίο των συχνοτήτων μόνο στο μεσαίο ποσοστημόριο καθώς στατιστικά σημαντικές τιμές έχουμε στα ποσοστημόρια 0.05 και 0.95.

Πίνακας 4. *Χρυσός - δείκτης GVZ. Έλεγχοι συμμετρίας στις άκρες των κατανομών σε κάθε συχνότητα (Ho: η συνδεσιμότητα είναι ίδια στο άνω και στο κάτω άκρο των κατανομών)*

	TCI	TO GVZ	FROM GVZ	NPC
Time-domain				
Test statistic	-2.09 (0.15)	-1.79 (0.22)	-2.37 (0.09)	0.57 (0.32)
High-frequency				
Test statistic	0.18 (0.92)	0.08 (0.96)	-2.39 (0.30)	0.20 (0.92)
Med-frequency				
Test statistic	-0.86 (0.07)	-0.72 (0.19)	-1.01 (0.14)	0.28 (0.70)
Low-frequency				
Test statistic	-1.40 (0.09)	-1.15 (0.20)	-1.64 (0.16)	0.48 (0.70)

Από τον Πίνακα 4 προκύπτει ότι η συνδεσιμότητα είναι ίδια στο άνω και στο κάτω άκρο της κατανομής στον τομέα του χρόνου και σε κάθε συχνότητα καθώς οι τιμές των p-value είναι παντού στατιστικά μη σημαντικές και κατά συνέπεια η μηδενική υπόθεση δεν απορρίπτεται σε κανέναν έλεγχο.

3.3 Στασιμότητα για το ζεύγος αργού πετρελαίου – δείκτη OVX

Πίνακας 5. *Στασιμότητα Αργού πετρελαίου – δείκτη OVX*

	KPSS test	Dickey-Fuller test
Δεδομένα αργ. πετρελαίου	Μη στάσιμη	Μη στάσιμη
Λογ. δεδ. αργ. πετρελαίου	Μη στάσιμη	Μη στάσιμη

Πρώτες διαφορές σε λογ. δεδ. αργ. πετρελαίου	**Στάσιμη**	**Στάσιμη**
Δεδομένα δείκτη OVX	Μη στάσιμη	**Στάσιμη**
Λογ. δεδομένα δείκτη OVX	Μη στάσιμη	**Στάσιμη**
Πρώτες διαφορές σε λογ. δεδομένα δείκτη OVX	**Στάσιμη**	**Στάσιμη**

Ομοίως με το ζεύγος Χρυσού – δείκτη GVZ, ο Πίνακας 5 παρουσιάζει τα αποτελέσματα των ελέγχων στασιμότητας για το ζεύγος αργού πετρελαίου – δείκτη OVX. Αντίστοιχα με το προηγούμενο ζεύγος η ανάλυση της συνδεσιμότητας θα γίνει στις πρώτες διαφορές των λογαριθμημένων δεδομένων και των δύο χρονοσειρών ώστε να επιτευχθεί κοινή στασιμότητα των δύο χρονοσειρών και στους δύο ελέγχους.

3.4 Ανάλυση συνδεσιμότητας για το ζεύγος αργού πετρελαίου – δείκτη OVX

Όμοια με το ζεύγος χρυσού – δείκτη GVZ οι παράμετροι που θα χρησιμοποιηθούν είναι, ορίζοντας πρόβλεψης $H = 24$, τρία ποσοστημόρια (0.05, 0.5 και 0.95), τρείς συχνότητες (0-5 , 5-10 και >10 μέρες) και μήκος υστέρησης 4 και πάλι σύμφωνα με το κριτήριο του Schwarz.

Πίνακας 6. *Συνδεσιμότητα Αργού πετρελαίου - δείκτη OVX (Ηο: Δεν υπάρχει συνδεσιμότητα μεταξύ του αργού πετρελαίου και του δείκτη OVX)*

	TCI	TO OVX	FROM OVX	NPC
Quantile 0.05				
Time-domain	15.01 **(<0.01)**	15.51 **(<0.01)**	14.51 **(<0.01)**	1.00 (0.52)
High-frequency	13.14 **(<0.01)**	14.27 **(0.01)**	12.01 **(<0.01)**	2.26 (0.34)
Med-frequency	0.76 (0.06)	0.52 (0.25)	1.01 **(0.04)**	-0.49 (0.34)
Low-frequency	1.11 (0.07)	0.72 (0.30)	1.50 **(0.05)**	-0.78 (0.36)
Quantile 0.05				
Time-domain	15.01 **(<0.01)**	15.51 **(<0.01)**	14.51 **(<0.01)**	1.00 (0.52)
High-frequency	15.01 **(<0.01)**	15.51 **(<0.01)**	14.51 **(<0.01)**	1.00 (0.52)
Med-frequency	0.76 (0.06)	0.52 (0.25)	1.01 **(0.04)**	-0.49 (0.34)
Low-frequency	1.11 (0.07)	0.72 (0.30)	1.50 **(0.05)**	-0.78 (0.36)
Quantile 0.95				
Time-domain	14.39 **(<0.01)**	14.66 **(<0.01)**	14.12 **(<0.01)**	0.54 (0.74)
High-frequency	11.79 **(<0.01)**	13.47 **(<0.01)**	10.12 **(0.02)**	3.35 (0.15)
Med-frequency	1.05 (0.06)	0.52 (0.45)	1.58 **(0.01)**	-1.06 (0.15)
Low-frequency	1.15 (0.08)	0.67 (0.55)	2.42 **(0.02)**	-1.75 (0.16)

Στον Πίνακα 6 φαίνονται τα αποτελέσματα για τα μέτρα συνδεσιμότητας του ζεύγους Αργό πετρέλαιο – δείκτης OVX. Εύκολα, συμπεραίνεται ότι υπάρχει συνδεσιμότητα σε όλο το

εύρος των τιμών των δύο χρονοσειρών. Στα άκρα των κατανομών τα μέτρα συνδεσιμότητας συμπεριφέρονται ακριβώς με τον ίδιο τρόπο. Συγκεκριμένα, τα μέτρα TCI, TO OVX είναι στατιστικά σημαντικά με υψηλές τιμές στο πεδίο των συχνοτήτων και την υψηλή συχνότητα ενώ στην μεσαία συχνότητα και την χαμηλή οι τιμές των μέτρων αυτών είναι πολύ χαμηλότερες και στατιστικά μη σημαντικές. Το μέτρο FROM OVX εμφανίζει στατιστική σημαντικότητα σε ολόκληρο το φάσμα συχνοτήτων. Ωστόσο, αυτό δεν επηρεάζει το μέτρο NPC, οι τιμές του οποίου υποδηλώνουν συμμετρική αλληλεπίδραση μεταξύ των δύο αγορών. Στο μέσο των κατανομών, δηλαδή για το ποσοστημόριο 0.5, τα μέτρα TCI, TO OVX και FROM OVX είναι σε κάθε συχνότητα στατιστικά σημαντικά υποδεικνύοντας υψηλή συνδεσιμότητας σε αυτό το σημείο των κατανομών. Υψηλές τιμές και πάλι παρατηρούνται στον τομέα του χρόνου και την υψηλή συχνότητα. Το μέτρο NPC παραμένει στατιστικά μη σημαντικό και σε αυτό το σημείο των κατανομών. Συνοψίζοντας, για τα άκρα των κατανομών οι επενδυτές αφομοιώνουν την πληροφορία και δρουν σε διάστημα το πολύ 5 ημερών καθώς στον τομέα του χρόνο και την υψηλή συχνότητα οι τιμές των δεικτών είναι στατιστικά σημαντικές, ενώ για το μέσο των κατανομών οι πληροφορίες μεταδίδονται ανεξάρτητα της συχνότητας.

Πίνακας 7. *Αργού πετρελαίου - δείκτη OVX. Έλεγχοι συμμετρίας στο φάσμα συχνοτήτων σε κάθε ποσοστημόριο (Ho: Η συνδεσιμότητα είναι ίδια στο φάσμα συχνοτήτων)*

	TCI	TO OVX	FROM OVX	NPC
Quantile 0.05				
Test statistic	12.37και -0.34 **(<0.01)**	13.75και -0.20 **(<0.01)**	10.99και -0.48 **(<0.01)**	2.75 και 0.29 (0.33)
Quantile 0.5				
Test statistic	15.64και -0.79 **(<0.01)**	15.93και -0.72 **(<0.01)**	15.35και -0.83 **(<0.01)**	0.58 και 0.11 (0.15)
Quantile 0.95				
Test statistic	10.74και -0.49 **(<0.01)**	12.95και -0.16 **(<0.01)**	8.54 και -0.83 **(<0.01)**	4.41 και 0.67 (0.12)

Από τον Πίνακα 7 προκύπτει ότι η συνδεσιμότητα είναι διαφορετική για κάθε μέτρο καθώς οι διαφορές των τιμών τους είναι μεγάλες κυρίως μεταξύ μεσαίας συχνότητας και χαμηλής σε σχέση με την υψηλή συχνότητα και τον τομέα του χρόνου. Εξαίρεση αποτελεί το μέτρο NPC του οποίου οι τιμές είναι παρόμοιες για κάθε ποσοστημόριο και σε κάθε συχνότητα.

Πίνακας 8. *Αργού πετρελαίου - δείκτη OVX. Έλεγχοι συμμετρίας στις άκρες των κατανομών σε κάθε συχνότητα (Ho: η συνδεσιμότητα είναι ίδια στο άνω και στο κάτω άκρο των κατανομών)*

	TCI	TO OVX	FROM OVX	NPC
Time-domain				
Test statistic	0.62 (0.75)	0.85 (0.72)	0.39 (0.85)	0.36 (0.84)
High-frequency				
Test statistic	1.34 (0.50)	0.80 (0.80)	0.39 (0.80)	-1.08 (0.72)
Med-frequency				
Test statistic	-0.28 (0.55)	0.02 (0.99)	-0.57 (0.36)	0.58 (0.44)
Low-frequency				
Test statistic	-0.43 (0.59)	0.05 (0.96)	-0.92 (0.38)	0.96 (0.44)

Από τον Πίνακα 8 προκύπτει ότι στα ακραία ποσοστημόρια η συνδεσιμότητα είναι ίδια για όλα τα μέτρα συνδεσιμότητας σε κάθε συχνότητα. Το αποτέλεσμα αυτό προκύπτει άμεσα από τον πίνακα 6 καθώς οι τιμές κάθε μέτρου ξεχωριστά είναι παρόμοιες σε κάθε

ποσοστημόριο. Επομένως, τα μέτρα συμπεριφέρονται με τον ίδιο τρόπο στο εύρος των τιμών των κατανομών ανεξάρτητα από το ποσοστημόριο.

4. ΣΥΜΠΕΡΑΣΜΑΤΑ

Άπό την ανάλυση της τιμής του χρυσού και του δείκτη GVZ προέκυψε ότι η συνδεσιμότητα είναι στατιστικά σημαντική μόνο στις μεταβολές που συμβαίνουν στις ακραίες τιμές των κατανομών, ενώ μεταβολές κοντά στη μέση τιμή φαίνεται να μην μεταφέρονται στο σύστημα από τη μία χρονοσειρά στην άλλη. Ακόμα, στον έλεγχο της συνδεσιμότητας μεταξύ των συχνοτήτων για τα τρία ποσοστημόρια, οι έλεγχοι έδειξαν ότι μόνο στο ποσοστημόριο 0.5 η συνδεσιμότητα είναι ίδια. Αναφορικά με τη συνδεσιμότητα μεταξύ των ακραίων ποσοστημορίων για κάθε συχνότητα, μόνο στην υψηλή συχνότητα προέκυψε ίδια συνδεσιμότητα Στο ζεύγος της τιμής του αργού πετρελαίου και του δείκτη μεταβλητότητας OVX, ισχυρή συνδεσιμότητα για κάθε συχνότητα στο ποσοστημόριο 0.5, ενώ στις ακραίες τιμές ισχυρή συνδεσιμότητα παρατηρήθηκε στον τομέα του χρόνου και την υψηλή συχνότητα. Στους ελέγχους μεταξύ των συχνοτήτων για κάθε ποσοστημόριο σε καμία περίπτωση η συνδεσιμότητα δεν προέκυψε ίδια μεταξύ του άνω άκρου της κατανομής και του μεσαίου η συνδεσιμότητά προέκυψε παντού ίδια.

ABSTRACT

This study investigates the connectedness between gold and oil prices and their respective volatility indices, GVZ and OVX, using the Quantile Frequency Connectedness (QFC) framework. The empirical findings suggest that connectedness depends on both the market state and the time horizon (frequency/time-scale). Specifically, the analysis reveals that: (a) For the pair (gold price, GVZ), strong connectedness is observed only at the tails of the joint distribution, whereas for the pair (oil price, OVX), it persists across all market conditions. (b) For both pairs of variables, connectedness is particularly strong at high frequencies (less than 5 days) and weak at lower frequencies, indicating that market participants react swiftly to new information. (c) There is evidence of asymmetry with respect to the source of shocks

ΑΝΑΦΟΡΕΣ

Fousekis, P. (2024). Connectedness between conventional and organic milk prices in the USA. The Journal of Economic Asymmetries, 30, e00367.

Fousekis, P. (2023). Contemporaneous dependence between euro, crude oil, and gold returns and their respective implied volatility changes: Evidence from the local Gaussian correlation approach. *Studies in Economics and Finance, 40*(5), 795–813.

Fousekis, P. (2019). Crude oil price and implied volatility: Insights from non-parametric quantile regressions. Studies in Economics and Finance, 36(2), 168-182.

Bekaert, G., & Wu, G. (2000). Asymmetric volatility and risk in equity markets. The Review of Financial Studies, 13(1), 1-42.

Bekiros, S., Jlassi, M., Naoui, K., & Uddin, G. S. (2017). The asymmetric relationshipbetween returns and implied volatility: Evidence from global stock markets. Journal of Financial Stability, 30, 156-174

Black, F. (1976). Studies of stock market volatility changes. Proceedings of the American Statistical Association, Business & Economic Statistics Section, 1976.

Chatziantoniou, I., Abakah, E. J. A., Gabauer, D., & Tiwari, A. K. (2022). Quantile time–frequency price connectedness between green bond, green equity, sustainable investments and clean energy markets. *Journal of Cleaner Production, 361*, 132088.

Chatziantoniou, I., Gabauer, D., & Gupta, R. (2023). Integration and risk transmission in the market for crude oil: New evidence from a time-varying parameter frequency connectedness approach. *Resources Policy, 84*, 103729

Cboe Global Markets. (2020). Volatility index methodology: Cboe volatility index (VIX).

Diebold, F. X., & Yılmaz, K. (2015). Financial and macroeconomic connectedness: A network approach to measurement and monitoring. Oxford University Press, USA.
Fassas, A. P., & Siriopoulos, C. (2021). Implied volatility indices–A review. The Quarterly Review of Economics and Finance, 79, 303-329.
Stiassny, A. (1996). A spectral decomposition for structural VAR models. Empirical Economics, 21, 535-555.

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 37ου Πανελληνίου & 3ου Διεθνούς Συνεδρίου Στατιστικής (2025) , σελ. 65–83

ΜΙΑ ΕΠΙΣΚΟΠΗΣΗ ΤΩΝ ΔΕΙΚΤΩΝ ΤΙΜΩΝ ΑΚΙΝΗΤΩΝ ΣΤΙΣ ΕΠΙΣΗΜΕΣ ΣΤΑΤΙΣΤΙΚΕΣ [1]

Δ. Λυμπερόπουλος[1]
[1] Τμήμα Μαθηματικών, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
dily@math.auth.gr

ΠΕΡΙΛΗΨΗ

Στην απλούστερη εκδοχή του, ένας δείκτης τιμών είναι ένας λόγος τιμών (ή συνολικών δαπανών) μεταξύ δύο χρονικών περιόδων και χρησιμεύει στην αποτύπωση της εξέλιξης ενός μεγέθους/αξίας/τιμής σε σύγκριση με ένα χρονικό ορόσημο. Με τον όρο «επίσημες στατιστικές» αναφερόμαστε γενικά σε στατιστικά εξαγόμενα (προϊόντα) που παράγονται και διαχέονται στο πλαίσιο ενός συστήματος εθνικών ή υπερεθνικών οργανισμών (π.χ. κυβερνητικών υπηρεσιών, ανεξάρτητων αρχών) σύμφωνα με βασικές αρχές που διασφαλίζουν ελάχιστα επαγγελματικά πρότυπα και υπό συγκεκριμένο νομικό πλαίσιο (βλ. π.χ. European Commission, Eurostat, Statistics Explained). Στον απόηχο της χρηματοοικονομικής κρίσης του 2008, το ενδιαφέρον για την κατάρτιση νέων επισήμων στατιστικών και ιδιαίτερα δεικτών τιμών ακινήτων ενισχύθηκε σημαντικά. Σε αυτή την εργασία, θα εστιάσουμε σε μεθοδολογικά θέματα των δεικτών τιμών οικιστικών ακινήτων που καταρτίζονται από ευρωπαϊκές Εθνικές Στατιστικές Υπηρεσίες (ιστορικό, βασικές έννοιες και ιδιαιτερότητες), θα αναφερθούμε στις προτεινόμενες μεθόδους κατάρτισής τους, όπως αυτή των Κυλιόμενων Χρονικών Ψευδομεταβλητών (RTD: Rolling Time Dummies), και θα γίνει μια σύντομη αναφορά στις επικρατούσες σήμερα τάσεις στον ευρωπαϊκό χώρο, βλ. π.χ. European Commission, Eurostat (2017) και Hill et al. (2018).

Λέξεις Κλειδιά: Επίσημη Στατιστική, δείκτης τιμών, σταθμίσεις, ποιοτική προσαρμογή, ηδονικός δείκτης, Μέθοδος Αναπροσαρμογής Τιμών (Repricing Method), Μέθοδος Κυλιόμενων Χρονικών Ψευδομεταβλητών (RTD - Rolling Time Dummy method).

1. ΕΙΣΑΓΩΓΗ

Η εκτίμηση αξιών ακινήτων και η μελέτη της διαχρονικής εξέλιξης τους αποτελούν δύο θέματα που έχουν αποσχολήσει εδώ και χρόνια τη διεθνή βιβλιογραφία, αρχικά

[1]Η παρούσα εργασία πραγματοποιήθηκε σχεδόν αποκλειστικά κατά την περίοδο απασχόλησης του συγγραφέα στην Ελληνική Στατιστική Αρχή (ΕΛΣΤΑΤ), προς την οποία θέλει να εκφράσει τις θερμές του ευχαριστίες.

την οικονομική ενώ με το πέρασμα των χρόνων το ενδιαφέρον υπήρξε αυξανόμενο για τις ποσοτικές πτυχές αυτών των δύο προβλημάτων, βλ. π.χ. Brooks and Tsolacos (2010) και Tsolacos and Andrew (2020), αντίστοιχα. Εκτός, όμως, από το καθαρά επιστημονικό ενδιαφέρον, είναι προφανές πως η διαμόρφωση των τιμών ακινήτων έχει και πρακτικές πλευρές και συνακόλουθα ενδιαφέρει πολλούς εμπλεκόμενους άμεσα ή έμμεσα με την αγορά ακινήτων, όπως πολιτικούς φορείς, διεθνείς οργανισμούς και κοινωνικές ομάδες για τη χάραξη (στεγαστικών και όχι μόνο) πολιτικών, αλλά και τραπεζικά ιδρύματα και επενδυτικές εταιρείες για τη διαχείριση των χαρτοφυλακίων τους.

Η χρηματοοικονομική κρίση του 2008 που πυροδοτήθηκε από τη «φούσκα» στην αγορά ακινήτων των Η.Π.Α. ήρθε να επιτείνει το ενδιαφέρον αυτό, αλλά και να καταδείξει την ανάγκη για την παραγωγή κατάλληλων, *συγκρίσιμων και αξιόπιστων* δεικτών τιμών ακινήτων και γενικότερα μακροοικονομικής σταθερότητας. Ειδικά στο χώρο της αγοράς ακινήτων, και παρά το γεγονός ότι υπήρχε διαθεσιμότητα διαφόρων δεικτών παρακολούθησης της (π.χ. από επενδυτικές εταιρείες και τράπεζες), η διαφορετική στόχευση και μεθοδολογία τους, η μη πλήρης κάλυψη του συνόλου των συναλλαγών ακινήτων, αλλά και η διαφιλονικούμενη μετά το 2008 αξιοπιστία και διαφάνεια στην κατάρτισή τους, κατέστησαν την προαναφερθείσα ανάγκη επιτακτική. Στην πράξη, το γεγονός αυτό εκδηλώθηκε με την άσκηση προτροπών και πιέσεων για την εμπλοκή των Εθνικών Στατιστικών Υπηρεσιών (ΣΥ) στην παραγωγή επισήμων στατιστικών ακινήτων.

Με τον όρο **επίσημες στατιστικές** αναφερόμαστε γενικά σε στατιστικά εξαγόμενα (προϊόντα) που παράγονται και διαχέονται στο πλαίσιο ενός συστήματος εθνικών ή υπερεθνικών οργανισμών (π.χ. κυβερνητικών υπηρεσιών, ανεξάρτητων αρχών) σύμφωνα με βασικές αρχές που διασφαλίζουν ελάχιστα επαγγελματικά πρότυπα (π.χ. ο Κώδικας Ορθής Πρακτικής στον οποίο έχουν οικοιοθελώς δεσμευτεί οι ευρωπαϊκές ΣΥ) και υπό συγκεκριμένο νομικό πλαίσιο (π.χ. εθνικούς νόμους, ευρωπαϊκούς κανονισμούς και οδηγίες), βλ. π.χ. European Commission, Eurostat, Statistics Explained. Χαρακτηριστικά παραδείγματα επισήμων στατιστικών αποτελούν ο *Εθνικός και ο Εναρμονισμένος Δείκτης Τιμών Καταναλωτή (ΔΤΚ και ΕνΔΤΚ, αντίστοιχα)*, καθώς και το *Ακαθάριστο Εγχώριο Προϊόν* που δημοσιεύει η Ελληνική Στατιστική Αρχή (ΕΛΣΤΑΤ), οι *δείκτες αγοραστικών ισοτιμιών (PPPs)* που δημοσιεύει η Eurostat, όπως και το *μήκος νοσηλείας* που δημοσιεύει ο ΟΟΣΑ (Οργανισμός Οικονομικής Συνεργασίας και Ανάπτυξης).

Στην παρούσα εργασία το ενδιαφέρον επικεντρώνεται σε μεθοδολογικά θέματα των επισήμων δεικτών τιμών οικιστικών ακινήτων. Πιο συγκεκριμένα, στην Ενότητα 2. παρουσιάζονται βασικές έννοιες των δεικτών τιμών, ενώ στην Ενότητα 3. γίνεται μια αναφορά στις ιδιαιτερότητες των δεικτών τιμών κατοικιών, μια σκιαγράφηση και συγκριτική μελέτη (βλ. Παρατήρηση 6) των βασικών μεθόδων υπολογισμού τους, καθώς και μια σύνοψη άλλων σχετικών θεμάτων (βλ. Υποενότητα 3.4). Στην Ενότητα 4. παρατίθενται ανοιχτά θέματα και μελλοντικές προκλήσεις, ενώ στην καταληκτική ενότητα της εργασίας δίνονται κάποια συνοπτικά συμπεράσματα.

2. ΓΕΝΙΚΑ ΠΕΡΙ ΔΕΙΚΤΩΝ ΤΙΜΩΝ

Στην παρούσα ενότητα συνοψίζονται βασικές έννοιες των δεικτών τιμών υπό μια πιθανοθεωρητική οπτική που γενικά δεν συναντάται στη σχετική βιβλιογραφία. Για μια πιο αναλυτική αναφορά σε αυτά τα θέματα παραπέμπουμε στο πόνημα του von der Lippe (2001).

Με $\mathbb{R}$ και $\mathbb{N}$ συμβολίζεται το σύνολο των πραγματικών και θετικών ακεραίων αριθμών, αντίστοιχα. Στο εξής, θεωρούμε έναν χώρο πιθανότητας (Ω, Σ, P). Για δύο οποιεσδήποτε τυχαίες μεταβλητές (τ.μ.) $X, Y : \Omega \longrightarrow \mathbb{R}$ γράφουμε

$$X = Y \;\; P - \text{σχεδόν βέβαια } (P\text{-σ.β.})$$

αν $P(X \neq Y) = 0$. Με $A \times B$ συμβολίζεται το καρτεσιανό γινόμενο των συνόλων A και B.

Επίσης θεωρούμε ένα σύνολο δεικτών $\mathbb{T} \subseteq \mathbb{N} \times \mathbb{N}$ των χρονικών στιγμών που μας ενδιαφέρουν, ενώ με $r \in \mathbb{T}$ θα συμβολίζεται μια **περίοδος αναφοράς**. Ας θεωρήσουμε επίσης τις στοχαστικές διαδικασίες (σ.δ.) $\{Y_t\}_{t\in\mathbb{T}}$ και $\{Z_t\}_{t\in\mathbb{T}}$ **τιμών** και **ποσοτήτων**, αντίστοιχα, όπου Y_t και Z_t γνήσια θετικές τ.μ. επάνω στο Ω. Με $\{V_t\}_{t\in\mathbb{T}}$ συμβολίζεται η στοχαστική διαδικασία (σ.δ.) **αξιών ή δαπανών**, όπου κάθε τ.μ. V_t αποτελεί συνάρτηση των $\{Y_t\}_{t\in\mathbb{T}}$, $\{Z_t\}_{t\in\mathbb{T}}$ με τιμές στο $(0, \infty)$.

Ορισμός 1. Ως ένας **δείκτης τιμών μιας περιόδου** t **ως προς μια περίοδο αναφοράς** r ορίζεται ο λόγος τιμών (*price relative*)

$$I_{t:r} := I_{t:r}(Y_r, Y_t) := 100\frac{Y_t}{Y_r}$$

ή ο λόγος δαπανών

$$I_{t:r} := I_{t:r}(V_r, V_t) := 100\frac{V_t}{V_r}.$$

Παρατηρήσεις 1. **(α)** Ένας δείκτης τιμών είναι ένας καθαρός αριθμός (χωρίς μονάδες μέτρησης) που παίρνει πάντα την τιμή 100 στην περίοδο αναφοράς (δηλ. ισχύει ότι $I_{r:r} = 100$). Συνεπώς, η διαφορά $I_{t:r} - I_{r:r}$ δηλώνει την ποσοστιαία (%) μεταβολή τιμών σε σχέση με την περίοδο αναφοράς r. Για παράδειγμα αν $\tau = (2024, 11)$, $r = (2023, 12)$ και $I_{t:r} = 103,5$, τότε το επίπεδο τιμών τον Νοέμβριο του 2024 παρουσίασε αύξηση κατά 3,5% σε σχέση με τον Δεκέμβριο του 2023.
(β) Στην περίπτωση του ΕνΔΤΚ — το γενικό πλαίσιο του οποίου διέπει και τους δείκτες τιμών ακινήτων — ουσιαστικά αναφερόμαστε σε λόγους δαπανών, για αυτό και η ορολογία «σ.δ. δαπανών» έχει επικρατήσει έναντι της «σ.δ. αξιών» για τη $\{V_t\}_{t\in\mathbb{T}}$.
(γ) Γενικά, ένας τέτοιος δείκτης τιμών $I_{t:r}$, $t \in \mathbb{T}$, καλύπτει n-κατηγορίες δαπανών ή n-προϊόντα ($n \in \mathbb{N}$ με $n \geq 2$) με τιμές και ποσότητες $Y_t(j)$ και $Z_t(j)$, $t \in \mathbb{T}$, αντίστοιχα, όπου $j \in \{1, \ldots, n\}$. Οι κατηγορίες αυτές καθορίζουν το πως συντίθεται ο δείκτης τιμών και συνιστούν το «καλάθι» του δείκτη. Για λόγους απλοποίησης,

θεωρούμε ότι *οι κατηγορίες αυτές είναι δεδομένες για το χρονικό διάστημα στο οποίο εστιάζουμε το ενδιαφέρον μας* χωρίς αυτό να σημαίνει ότι και η συνεισφορά κάθε κατηγορίας στο «καλάθι» παραμένει διαχρονικά σταθερή.

Ορισμός 2. Ένας δείκτης τιμών όπως αυτός της Παρατήρησης 1,(γ) ονομάζεται
(*i*) **τύπου Laspereyes** αν

$$\left[V_t = \sum_{j=1}^{n} Y_t(j) Z_r(j) \ \text{ και } \ V_r = \sum_{j=1}^{n} Y_r(j) Z_r(j) \right],$$

(*ii*) **τύπου Paasche** αν

$$\left[V_t = \sum_{j=1}^{n} Y_t(j) Z_t(j) \ \text{ και } \ V_r = \sum_{j=1}^{n} Y_r(j) Z_t(j) \right].$$

Η συνθήκη Laspereyes ουσιαστικά δηλώνει ότι η σύνθεση του δείκτη τιμών που αποτελεί τη σύνθεση αναφοράς/μέτρο σύγκρισης είναι αυτή της περιόδου αναφοράς r. Αντίστοιχα, σε έναν δείκτη τύπου Paasche το «καλάθι-μέτρο σύγκρισης» είναι αυτό της τρέχουσας περιόδου t.

Καθότι τόσο ο ΔΤΚ και ο ΕνΔΤΚ όσο και οι δείκτες τιμών ακινήτων αποτελούν δείκτες τύπου Laspereyes, στο εξής θα μας απασχολήσουν μόνο τέτοιου τύπου δείκτες τιμών.

Στο επόμενο αποτέλεσμα δίνεται η ανάλυση ενός δείκτη τιμών στα δομικά του στοιχεία: *τους υποδείκτες του και τις αντίστοιχες σταθμίσεις.*

Λήμμα 1. Για έναν δείκτη τιμών $I_{t:r}$, $t \in \mathbb{T}$, τύπου Laspereyes ισχύει ότι

$$I_{t:r} := I_{t:r}(V_t, V_r) = \sum_{j=1}^{n} W_r(j) I_{t:r}(Y_t(j), Y_r(j)) \quad \text{για κάθε } t \in \mathbb{T},$$

όπου

$$W_r(j) := \frac{Y_r(j) Z_r(j)}{\sum_{j=1}^{n} Y_r(j) Z_r(j)}.$$

Το Λήμμα 1 αποτελεί συνέπεια των Ορισμών 1 και 2.

Παρατηρήσεις & Ορισμοί 3. (α) Οι δείκτες $I_{t:r}(Y_t(j), Y_r(j))$ του Λήμματος 1 καλούνται **υποδείκτες του (γενικού) δείκτη** $I_{t:r}$, ενώ οι $W_j := W_r(j)$ αποτελούν τις **αντίστοιχες σταθμίσεις** τους.
(β) *Γενικά*, οι σταθμίσεις ενός δείκτη τιμών $I_{t:r}$, $t \in \mathbb{T}$, δύνανται να βασίζονται σε **(ονομαστικές) αξίες** μιας περιόδου $b \in \mathbb{T}$, δηλαδή έχουμε ότι

$$W_b(j) := \frac{Y_b(j) Z_b(j)}{\sum_{j=1}^{n} Y_b(j) Z_b(j)} \quad \text{για κάθε } \ j \in \{1, \ldots, n\}.$$

Η b **καλείται περίοδος βάσης** ενός δείκτη τιμών.
(γ) Οι (υπο)δείκτες τιμών τύπου Laspereyes είναι **σταθερής βάσης** αν $b = r$. Σε αντίθετη περίπτωση είναι μη σταθερής βάσης.
(δ) Σε περιπτώσεις, όμως, που επιθυμούμε την απαλλαγή των σταθμίσεων από τη μεταβολή των ονομαστικών τιμών, δηλαδή από την επίδραση του πληθωρισμού, οι σταθμίσεις ενός δείκτη τιμών $I_{t:r}$, $t \in \mathbb{T}$, δύνανται να βασίζονται σε **επικαιροποιημένες σε όρους/τιμές μιας περιόδου** $u \in \mathbb{T}$ **αξίες** μιας περιόδου $b \in \mathbb{T}$. Επισημαίνεται ότι οι *σταθμίσεις ενός δείκτη τιμών αναθεωρούνται ετησίως ή ισοδύναμα παραμένουν σταθερές εντός ενός έτους.*

Στη συνέχεια παρατίθεται η διαδικασία **επικαιροποίησης των σταθμίσεων** ενός δείκτη τιμών:

Βήμα 0: Έστω $b = y_0 \in \mathbb{N}$ και $r = (y_\rho, q_\rho)$, $u = (y_1, q_1)$, $t = (y, q) := (y_\tau, q_\tau)$ στο $\mathbb{T}$ (όπου $\rho, \tau \in \mathbb{N}$ με $\rho \leq \tau$), οι περίοδοι βάσης, αναφοράς, επικαιροποίησης και η περίοδος ενδιαφέροντος (π.χ. τρέχουσα περίοδος) του δείκτη $I_{t:r}$, $t \in \mathbb{T}$.
Στην πράξη και για τους σκοπούς των τριμηνιαίων δεικτών τιμών κατοικιών (αντ. μηνιαίων ΕνΔΤΚ), οι παραπάνω περίοδοι είναι τρίμηνα (αντ. μήνες) ενός έτους από το 2010 (αντ. 1997) και εντεύθεν, όπου υπήρξε νομική υποχρέωση κατάρτισης τους βάσει ευρωπαϊκών κανονισμών, δηλαδή έχουμε ότι $T \subseteq \{2010, \ldots, 2024, \ldots\} \times \{1, 2, 3, 4\}$ (αντ. $\mathbb{T} \subseteq \{1997, \ldots, 2024, \ldots\} \times \{1, \ldots, 12\}$).

Βήμα 1: Υπολογίζεται ο **μέσος ετήσιος δείκτης τιμών**

$$I_{y_0:(y_\rho,q_\rho)} = \frac{1}{4} \sum_{q=1}^{4} I_{(y_0,q):(y_\rho,q_\rho)}.$$

Βήμα 2: Υπολογίζονται οι **επικαιροποιημένες σε τιμές μιας περιόδου** u, αξίες περιόδου b:

$$S_{y_0:u}(j) := S_{y_0}(j) \frac{I_{u:(y_\rho,q_\rho)}}{I_{y_0:(y_\rho,q_\rho)}} \quad \text{για κάθε } j \in \{1, \ldots, n\}.$$

Βήμα 3: Υπολογίζονται οι **επικαιροποιημένες σε τιμές μιας περιόδου** u**, σταθμίσεις**:

$$W_{b:u}(j) := W_{y_0:u}(j) := \frac{S_{y_0:u}(j)}{\sum_{j=1}^{n} S_{y_0:u}(j)} \quad \text{για κάθε } j \in \{1, \ldots, n\}.$$

Βήμα 4: Υπολογίζονται ο **δείκτης τιμών της περιόδου** t **με περίοδο αναφοράς** r**, περίοδο βάσης** y_0 **και επικαιροποιημένες σταθμίσεις σε τιμές μιας περιόδου** u:

$$I_{t:r;y_0:u} := I_{t:r;y_0:u}(V_t, V_r) = \sum_{j=1}^{n} W_{y_0:u}(j) I_{t:r}(Y_t(j), Y_r(j)) \quad \text{για κάθε } t \in \mathbb{T}.$$

Με αφορμή το Βήμα 4, αξίζει να επισημάνουμε τα εξής: Αν και στο ευρύ κοινό των χρηστών των επισήμων στατιστικών, για τους δείκτες τιμών έχει επικρατήσει ακόμα και σήμερα η εντύπωση ότι υπάρχουν δύο μόνο περίοδοι που εμπλέκονται στον υπολογισμό τους (η τρέχουσα t και η «περίοδος βάσης» όπως αποκαλείται η r), το Βήμα 4 αποκαλύπτει ότι ένας δείκτης τιμών δύναται να εξαρτάται μέχρι και από τέσσερις διαφορετικές περιόδους! Την περίοδο που μας ενδιαφέρει, την περίοδο που αποτελεί μέτρο σύγκρισης και στην οποία ο δείκτης είναι εξ′ορισμού ίσος με 100 (περίοδος αναφοράς), την περίοδο από όπου ο δείκτης αντλεί τις σταθμίσεις του και άρα το υποκείμενο πρότυπο δαπανών (περίοδο/έτος βάσης) και μια περίοδο στην οποία χρειάζεται να εκφραστούν οι σταθμίσεις απαλλαγμένες από πληθωριστικές επιδράσεις (περίοδος επικαιροποίησης σταθμίσεων).

Η σύγχυση αυτή οφείλεται στο ότι οι περισσότεροι δείκτες (όχι κατ′ανάγκη τιμών) των επισήμων στατιστικών είναι ακόμα και σήμερα σταθερής βάσης χωρίς αυτό να υποδηλώνει κάτι προβληματικό. Ακόμη, όμως, και οι δείκτες τιμών καταναλωτή (ΔΤΚ και ΕνΔΤΚ), οι οποίοι αποτέλεσαν έναν από τους κύριους λόγους εισαγωγής των δεικτών τιμών μη σταθερής βάσης και ειδικότερα των λεγόμενων *αλυσωτών* (*chain-linked*) δεικτών τιμών (όπου η περίοδος βάσης είναι πλέον κυλιόμενη και αναθεωρείται διαρκώς) παρέμεναν στον ευρωπαϊκό χώρο μέχρι τα μέσα της δεκαετίας του 2010 δείκτες σταθερής βάσης. Το βασικό επιχείρημα για την μετατροπή τους σε αλυσωτούς είχε υπόβαθρο από την Οικονομική Θεωρία και αναφέρονταν στην ανάγκη συνεχούς επικαιροποίησης του «καλαθιού» των δεικτών τιμών καταναλωτή — δηλαδή το πρότυπο καταναλωτικών δαπανών που αντικατοπτρίζεται στις σταθμίσεις των ΔΤΚ και ΕνΔΤΚ θα έπρεπε να επικαιροποιείται ετησίως και όχι ανά πενταετία, όπως συνέβαινε όταν ήταν δείκτες σταθερής βάσης. Ενδιαφέρουσες στατιστικές πτυχές της μετάβασης αυτής παρατίθενται στον von der Lippe (2001, Chapters 0,1 και 2).

Επομένως, θα ήταν παράλειψη κλείνοντας την αναφορά σε βασικές έννοιες, χρήσιμες για την κατάρτιση δεικτών τιμών στα πλαίσια των επισήμων στατιστικών, να μην επανέλθουμε στην Παρατήρηση 3, (γ) και να παρουσιάσουμε την κατ′εξοχήν εναλλακτική των δεικτών τιμών σταθερής βάσης.

Ορισμός 4. Ένας **(τριμηνιαίος) αλυσωτός δείκτης τιμών (τύπου Laspereyes)** δίνεται από τύπο

$$\widetilde{I}_{t:r,u} = \prod_{k\in\{\rho+1,\ldots,\tau\}} I_{(y_k,q_k):(y_{k-1},q_{k-1});y_{k-1}:u} \quad \text{για κάθε } t \in \mathbb{T},$$

όπου $y_{k-1} := y_k - 1$, ενώ $q_{k-1} := q_k - 1$ για $k = 2, 3, 4$ και $q_{k-1} := 4$ για $k = 1$. Ιδιαιτέρως, αν δεν υφίσταται επικαιροποίηση σταθμίσεων, η παράμετρος u παραλείπεται. Αναλόγως ορίζεται και ο μηνιαίος αλυσωτός δείκτης.

3. ΔΕΙΚΤΕΣ ΤΙΜΩΝ ΚΑΤΟΙΚΙΩΝ

Όπως αναφέρθηκε ήδη στην Εισαγωγή, η ζήτηση για την κατάρτιση επισήμων στατιτιστικών ακινήτων υπήρξε ιδιαίτερα έντονη μετά το 2008. Στον ευρωπαϊκό χώρο,

η Eurostat είχε ανταποκριθεί στο αίτημα αυτό πολύ νωρίτερα, ξεκινώντας όμως από άλλη αφετηρία: *την κατάρτιση του ΕνΔΤΚ*.

Από τη δεκαετία του '90 παραμένει ανοιχτό το θέμα της ενσωμάτωσης στον ΕνΔΤΚ των δαπανών αγοράς και κυριότητας κατοικιών, η περαιτέρω αναφορά στο οποίο ξεφεύγει από τους σκοπούς της παρούσας εργασίας. Για να ανταποκριθεί στην πρόκληση αυτή η Ευρωπαϊκή Στατιστική Υπηρεσία, ξεκίνησε σε συνεργασία με τις ΣΥ των Κρατών-Μελών (Κ-Μ) ήδη από τα τέλη της δεκαετίας του 2000 την κατάρτιση δεικτών τιμών κατοικιών σε πιλοτική βάση και πιο συγκεκριμένα του Δείκτη Τιμών Κατοικιών (ΔΤΚατ γνωστού και ως HPI: House Price Index) και του Δείκτη Τιμών Ιδιοκατοίκησης (ΔΤΙΚ γνωστού και ως OOHPI: Own-Occupied Housing Price Index). Ο τελευταίος προϋποθέτει την κατάρτιση του πρώτου *στον οποίο και θα εστιάσουμε στο εξής*.

Η κατάρτιση των παραπάνω δύο δεικτών έγινε νομικά δεσμευτική για πρώτη φορά με τον Κανονισμό ΕΕ 93/2013, χωρίς αυτό να σημαίνει την άμεση ανταπόκριση όλων των Κ-Μ. Καθότι οι δείκτες τιμών κατοικιών εντάσσονται στο ευρύτερο μεθοδολογικό πλαίσιο των ΔΤΚ, ο Κανονισμός ΕΕ 2016/792 (του Ευρωπαϊκού Κοινοβουλίου και του Συμβουλίου, της 11ης Μαΐου 2016, για τους εναρμονισμένους δείκτες τιμών καταναλωτή και τον δείκτη τιμών κατοικιών και την κατάργηση του κανονισμού (ΕΚ) αριθ. 2494/95 του Συμβουλίου) αποτέλεσε ένα ακόμη ορόσημο για τους δείκτες τιμών κατοικιών. Το ισχύον σήμερα νομικό πλαίσιο είναι αυτό που καθορίζει ο Εκτελεστικός Κανονισμός ΕΕ 2023/1470, ο οποίος επικαιροποιεί και συμπληρώνει τους προκατόχους του.

Σύμφωνα με αυτόν, ο HPI πρέπει να τηρεί κάποιες ελάχιστες προδιαγραφές: **(π1)** Να βασίζεται σε δεδομένα πραγματικών αξιών συναλλαγής (όχι π.χ. σε εκτιμήσεις αξιών ακινήτων). **(π2)** Τριμηνιαία Συχνότητα Κατάρτισης. **(π3)** Ο Γενικός HPI να αποτελείται τουλάχιστον από τους εξής δύο υποδείκτες: (H11) Νέων Κατοικιών και (H12) Υφιστάμενων Κατοικιών. **(π4)** Να είναι ένας αλυσωτός δείκτης τιμών τύπου Laspereyes με ***επικαιροποιημένες*** σταθμίσεις σε τιμές 4ου τριμήνου προηγούμενου έτους. **(π5)** Να είναι **ποιοτικά προσαρμοσμένος.**

Παρατηρήσεις & Ορισμοί 5. Αξίζει να επισημάνουμε τα εξής:
(α) Στην (π3) αποτυπώνεται η ελάχιστη απαιτούμενη ανάλυση του Γενικού HPI. Στην πράξη ο γενικός δείκτης μπορεί να αναλύεται σε πολλές περισσότερες συνιστώσες και να αποτελείται από διάφορα επίπεδα/στρώματα υποδεικτών. Οι υποδείκτες που ανήκουν στο πιο αναλυτικό (κατώτερο) από αυτά ονομάζονται **στοιχειώδεις υποδείκτες**. Οι υποδείκτες ανωτέρων (λιγότερο αναλυτικών) επιπέδων προκύπτουν ως *σταθμισμένοι μέσοι των υποδεικτών του αμέσως κατωτέρου επιπέδου*, βλ. Λήμμα 1. Όπως προκύπτει και από τη διεθνή βιβλιογραφία, βλ. π.χ. Hill et al. (2018), ένα πρόσθετο επίπεδο ανάλυσης για παράδειγμα του υποδείκτη H12 θα μπορούσε να είναι εκείνο σε Υφιστάμενα Διαμερίσματα και Υφιστάμενες Μονοκατοικίες.
(β) Από την (π4) έπεται ότι για τις σταθμίσεις των (υπο)δεικτών του HPI της τρέχουσας περιόδου $t = (y, q) \in \mathbb{T}$ ισχύει ότι $u = (y - 1, 4)$. Επί πλέον στις περισ-

σότερες των περιπτώσεων, οι ονομαστικές σταθμίσεις (πριν την επικαιτοποίησή τους σε τιμές περιόδου u) υπολογίζονται από τα πιο πρόσφατα διαθέσιμα δεδομένα είτε των *συνολικών ετησίων αξιών συναλλαγής* ή άλλων πηγών δεδομένων (π.χ. στοιχεία εθνικών λογαριασμών), δηλ. είτε από δεδομένα του προηγούμενου ($y_0 = y - 1$) ή του προπροηγούμενου ($y_0 = y - 2$) έτους. Για περισσότερες λεπτομέρειες, παραπέμπουμε στην ιστοσελίδα των εθνικών μεταδεδομένων της Eurostat (βλ. European Commission, Eurostat, Reference Metadata).

(γ) Η (π5) αποτελεί την προδιαγραφή εκείνη που καθιστά τον HPI έναν ιδιαίτερο δείκτη. Με τον όρο **ποιοτική προσαρμογή** ενός δείκτη τιμών εννοούμε την απαλλαγή του δείκτη από την επίδραση της αλλαγής ποιότητας (μεταξύ δύο ή περισσοτέρων περιόδων) στο μείγμα των προϊόντων (στο πλαίσιο αυτής της εργασίας ακινήτων) που αποτελούν αντικείμενο συναλλαγής *ή ισοδύναμα* στον υπολογισμό μεταβολών τιμών σε ένα μείγμα προϊόντων/ακινήτων όπου επιτυγχάνεται με τεχνητό τρόπο η αναφορά σε προϊόντα/ακίνητα αντίστοιχης ποιότητας/τυπολογικών χαρακτηριστικών.

Το πρόβλημα της ποιοτικής προσαρμογής θα παρουσιαστεί αναλυτικότερα αμέσως παρακάτω. Στο σημείο αυτό, απλώς αναφέρουμε ότι για να είναι ένας (υπο)δείκτης τιμών *ποιοτικά προσαρμοσμένος, αρκεί να είναι τέτοιοι οι στοιχειώδεις δείκτες που τον αποτελούν. Η ποιοτική προσαρμογή δηλαδή εφαρμόζεται στο στοιχειώδες επίπεδο ανάλυσης ενός δείκτη.*

3.1 Το πρόβλημα της ποιοτικής προσαρμογής

Στην Παρατήρηση 5, (γ) περιγράφεται στο που αποσκοπεί η εφαρμογή της ποιοτικής προσαρμογής. Το κεντρικό πρόβλημα που καλείται να αντιμετωπίσει έγκειται στο ότι μια διαχρονική σύγκριση τιμών πρέπει να αφορά *σε όσο το δυνατόν πιο όμοια αγαθά*, αφού η τιμή ενός αγαθού καθορίζεται από τα επιμέρους ποιοτικά του χαρακτηριστικά. Εγείρεται, λοιπόν, το ερώτημα για το *πώς τιμολογείται ένα «περίεργο» (ανομοιογενές ως προς τα επιμέρους χαρακτηριστικά του) αγαθό;* Τα ακίνητα και οι Η/Υ αποτελούν χαρακτηριστικά παραδείγματα τέτοιων αγαθών.

Πιο συγκεκριμένα, ας υποθέσουμε ότι το πλήθος των ποιοτικών χαρακτηριστικών ενός αγαθού (π.χ. κατοικίας) είναι ίσο με $c \in \mathbb{N}$ και ότι το τυχαίο διάνυσμα $\mathbf{X}_t = (X_{1,t}, \ldots, X_{c,t})$, $t \in \mathbb{T}$, αποτελεί το διάνυσμα των ποιοτικών χαρακτηριστικών του τη χρονική περίοδο τ. Τότε, σύμφωνα με τα όσα αναφέρθηκαν παραπάνω, μπορούμε να ισχυριστούμε ότι $Y_t = Y_t(\mathbf{X}_t)$ για κάθε $t \in \mathbb{T}$, συνεπώς ο λόγος (ονομαστικών) τιμών $100Y_t/Y_r$ $r, t \in \mathbb{T}$ **δεν** λαμβάνει υπόψη τη διαφοροποίηση στην ποιότητα του αγαθού. Ιδανικά, ένας τέτοιος λόγος θα ήταν ο $100Y_t(\mathbf{X}_s)/Y_r(\mathbf{X}_s)$ για κάποιο $s \in \mathbb{T}$. Πρακτικά, όμως, ένας *εφικτός ποιοτικά προσαρμοσμένος λόγος τιμών* θα λαμβάνει υπόψη κάποιες τεχνητές τιμές $\widetilde{Y}_t$ που θα εξαρτώνται από τα τυχαία διανύσματα $\mathbf{X}_s$, $s, t \in \mathbb{T}$ ώστε s προγενέστερο t, καθώς και τις ονομαστικές τιμές Y_t, $t \in \mathbb{T}$.

Αυτά τα δύο είδη τιμών είναι που εμπλέκονται και στις 4+1 βασικές μεθόδους ποιοτικής προσαρμογής, οι οποίες τόσο από τη διεθνή βιβλιογραφία όσο και στη διεθνή πρακτική (βλ. π.χ. Hill et al. (2018) και European Commission, Eurostat

(2017)) έχουν επικρατήσει για την κατάρτιση ποιοτικά προσαρμοσμένων δεικτών τιμών ακινήτων:

(M0) Προσαρμογή Μείγματος/Στρωματοποίηση (*Mix Adjustment*).
(M1) Υποκατάσταση Τιμών (*Imputation*).
(M2) Μέσων Χαρακτηριστικών (*Average Characteristics*).
(M3) Μέθοδος Αναπροσαρμογής Τιμών (*RM: Repricing Method*).
(M4) Μέθοδος Κυλιόμενων Χρονικών Ψευδομεταβλητών (*RTD: Rolling Time Dummies*).

Η (M0) αποτελεί την πιο απλή μέθοδο ποιοτικής προσαρμογής, αφού η τελευταία επιτυγχάνεται μέσω του υπολογισμού των δεικτών τιμών βάσει μέσων/διαμέσων (ονομαστικών) τιμών ομογενών ως προς ποιοτικά χαρακτηριστικά (π.χ. επιφάνεια, ηλικία, τοποθεσία) στρωμάτων. Σήμερα θεωρείται πλέον ως μη βέλτιστη μέθοδος από τη Eurostat και ως λύση ανάγκης.

Αντίθετα, στις υπόλοιπες τέσσερις μεθόδους, η ποιοτική προσαρμογή επιτυγχάνεται με τη χρησιμοποίηση εκτιμώμενων μέσω υποδειγμάτων παλινδρόμησης τιμών. Οι μέθοδοι (M1), (M2) αποτελούν ειδικές περιπτώσεις ή παραλλαγές της (M3), την οποία και θα παρουσιάσουμε αναλυτικά αμέσως παρακάτω, ενώ το ίδιο θα γίνει και για τον κύριο ανταγωνιστή της τη (M4) σε μια άλλη υποενότητα.

3.2 Η Μέθοδος Αναπροσαρμογής Τιμών (RM: Repricing Method)

Η κεντρική ιδέα αυτής της μεθόδου έγκειται στον διαχωρισμό δύο βασικών συνιστωσών των (υπο)δεικτών HPI: Μίας που να αντικατοπτρίζει την εξέλιξη των ονομαστικών αξιών και μίας που να αποτυπώνει τις ποιοτικές μεταβολές στο μείγμα των συναλλασόμενων αγαθών. Από τον λόγο της πρώτης προς τη δεύτερη συνιστώσα, προκύπτει ο αντίστοιχος στοιχειώδης (και πλέον ποιοτικά προσαρμοσμένος) (υπο)δείκτης HPI.

Πιο συγκεκριμένα, *στο εξής υποθέτουμε* $h : (0,\infty) \longrightarrow (0,\infty)$ αντιστρέψιμη συνάρτηση, $H_t := h(Y_t)$ για κάθε $t \in \mathbb{T}$, όπου $\mathbb{T} \subseteq \{2010,\ldots,2024,\ldots\}\times\{1,2,3,4\}$, και N_t απαριθμήτρια συναλλαγών τ.μ. σε κάθε $t \in \mathbb{T}$.

Τότε ο υπολογισμός ενός ποιοτικά προσαρμοσμένου δείκτη τιμών σύμφωνα με την RM έχει ως εξής:

Βήμα 0: Καθορίζεται το σύνολο $\mathbb{T}_e \subseteq \mathbb{T}$ των χρονικών περιόδων βάσει των δεδομένων των οποίων θα εκτιμηθεί για κάθε $t \in \mathbb{T}_e$ και $n = 1,\ldots,N_t$ το υπόδειγμα πολλαπλής γραμμικής παλινδρόμησης

$$H_{t,n} = \beta_0 + \sum_{j=1}^{c} \beta_j X_{t,n} + \varepsilon_{t,n}, \qquad (RePM)$$

όπου $\beta_j \in \mathbb{R}$ οι παράμετροι του υποδείγματος και $\varepsilon_{t,n}$ τα τυχαία σφάλματα. Το σύνολο $\mathbb{T}_e$ ονομάζεται **περίοδος εκτίμησης** του υποδείγματος.

Βήμα 1: Υπολογίζεται το εκτιμώμενο (RePM): Για κάθε $t \in \mathbb{T}_e$ και $n = 1, \ldots, N_t$

$$\widehat{H}_{t,n} = \widehat{\beta}_0 + \sum_{j=1}^{c} \widehat{\beta}_j X_{t,n} \;\; P - \text{σ.β.}. \qquad (RePM*)$$

Βήμα 2: Σταθεροποιούμε τις εκτιμώμενες παραμέτρους $\widehat{\beta}_j$, $j = 0, 1, \ldots, c$, σύμφωνα με τη συχνότητα αναθεώρησης υποδείγματος (π.χ. ετήσια).

Βήμα 3: Υπολογίζονται οι τιμές $\widetilde{H}_t = \widehat{\beta}_0 + \sum_{j=1}^{c} \widehat{\beta}_j \overline{X}_t$ για κάθε $t \in \mathbb{T}$, όπου $\overline{X}_t := (1/N_t) \sum_{n=1}^{N_t} X_{t,n}$ για κάθε $\tau \in \mathbb{T}$.

Βήμα 4: Υπολογίζεται ο **Δείκτης Ονομαστικών Τιμών** (QUI: Quality Unadjusted Index)

$$100 \cdot \frac{1}{N_t} \sum_{j=1}^{N_t} H_t / \frac{1}{N_r} \sum_{j=1}^{N_r} H_r$$

του q-τριμήνου έτους y ως προς την περίοδο αναφοράς $r = (y_\rho, q_\rho) \in \mathbb{T}$.

Βήμα 5: Υπολογίζεται ο **Δείκτης Τεχνητών Τιμών** (QAF: Quality Adjustment Factor):

$$100 \cdot \frac{1}{N_t} \sum_{j=1}^{N_t} \widetilde{H}_t / \frac{1}{N_r} \sum_{j=1}^{N_r} \widetilde{H}_r$$

του q-τριμήνου έτους y ως προς την περίοδο αναφοράς r.

Βήμα 6: Υπολογίζεται ο αντίστοιχος *ποιοτικά προσαρμοσμένος δείκτης HPI*:

$$I_{t:r} = 100 \cdot QUI(t:r)/QAF(t:r) \;\; \text{για κάθε } t \in \mathbb{T}.$$

Οι εκτιμήσεις $\widehat{\beta}_j$ των παραμέτρων β_j του υποδείγματος (RePM) είναι γνωστές και ως «σκιώδεις τιμές», αφού δηλώνουν την «ηδονή» που αντλείται από την κατανάλωση μιας μονάδας του j-ποιοτικού χαρακτηριστικού, που συμβάλλει στον καθορισμό της αξίας του αγαθού. Για τον λόγο αυτό, τα σχετικά υποδείγματα παλινδρόμησης και οι επαγόμενοι από αυτά δείκτες τιμών είναι γνωστά στη βιβλιογραφία ως «ηδονικά υποδείγματα» και «ηδονικοί δείκτες» (*hedonic models, hedonic indices*), αντίστοιχα.

Κλείνοντας την παρούσα ενότητα παραθέτουμε ένα απλουστευμένο αλλά απτό παράδειγμα υπολογισμού ενός ποιοτικά προσαρμοσμένου HPI για να γίνει πιο κατανοητός ο τρόπος με τον οποίο επιτυγχάνεται η ποιοτική προσαρμογή μέσω της RM.

Παράδειγμα 1. Ας θεωρήσουμε δύο τρίμηνα, το $t = (2024, 2)$ και το τρίμηνο αναφοράς $r = (2023, 4)$ και ας υποθέσουμε ότι το μοναδικό ποιοτικό χαρακτηριστικό από το οποίο διαμορφώνεται η τιμή ενός ακινήτου στην αγορά κατά τα δύο αυτά τρίμηνα είναι η κύρια επιφάνειά του με «σκιώδη τιμή» $\widehat{\beta}_1 = 1000$ (ευρώ/τ.μ.). Έστω επίσης ότι η κοινή ελάχιστη τιμής κατοικίας (γνωστή και ως «αέρας») είναι η $\widehat{\beta}_0 = 20.000$ ευρώ. Αν, τώρα, οι μέσες ονομαστικές αξίες αγοραπωλησίας για το 2ο τρίμηνο του 2024 (24Q2) και το 4ο του 2023 (23Q4) είναι ίσες με 180.000 και 200.000 ευρώ, αντίστοιχα, τότε έχουμε τα εξής:

- Ο δείκτης ονομαστικών τιμών θα ισούται με

$$QUI(t:r) = 100 \cdot (180.000/200.000) = 90,$$

υποδηλώνοντας μια μείωση των ονομαστικών τιμών μεταξύ 23Q4 και 24Q2 της τάξης του 10%.
- Οι μέσες τεχνητές τιμές θα ισούνται με $20.000 + 1.000 \cdot 70 = 90.000$ και $20.000 + 1.000 \cdot 100 = 120.000$ ευρώ για το 23Q4 και το 24Q2, αντίστοιχα. Επομένως, ο δείκτης τεχνητών τιμών θα ισούται με

$$QAF(t:r) = 100 \cdot (90.000/120.000) = 75,$$

υποδηλώνοντας μια μείωση της τάξης του 25% στην ποιότητα του μείγματος των κατοικιών που αποτέλεσαν αντικείμενο αγοραπωλησίας μεταξύ των δύο τριμήνων.
- Ο αντίστοιχος ποιοτικά προσαρμοσμένος δείκτης HPI, τότε θα ισούται με

$$I_{t:r} = 100 \cdot QUI(t:r)/QAF(t:r) = 100 \cdot (90/75) = 120,$$

υποδηλώνοτας μια αύξηση της τάξης του 20% στις πραγματικές (ποιοτικά προσαρμοσμένες) τιμές των κατοικιών.

Το προηγούμενο παράδειγμα καταδεικνύει το ότι σε αγαθά όπως οι κατοικίες μια μεταβολή των ονομαστικών τιμών δύναται να είναι από μόνη της παραπλανητική, αφού μπορεί να είχαμε μια μείωση τους κατά 10%, αλλά για αρκετά μικρότερες (δηλ. χειρότερες ποιοτικά) κατοικίες. Συνεπώς, πωλήθηκε ένα αρκετά υποδεέστερο αγαθό φθηνότερα μεν όχι επαρκώς φθηνότερα δε, άρα στην πραγματικότητα είχαμε αύξηση τιμών κάτι που αποτυπώνεται στον ποιοτικά προσαρμοσμένο δείκτη HPI.

3.3 Η Μέθοδος Κυλιόμενων Χρονικών Ψευδομεταβλητών (RTD: Rolling Time Dummies)

Η μέθοδος RTD αποσκοπεί στην αξιοποίηση δεδομένων περισσοτέρων της μίας περιόδους για την κατάρτιση ποιοτικά προσαρμοσμένων δεικτών. Η μεταβολή της ποιότητας του μείγματος των συναλλασομένων ακινήτων δεν αποτυπώνεται άμεσα όπως στην RM μέσω του Δείκτη Τεχνητών Τιμών QAF, αλλά ενσωματώνεται έμμεσα στις εκτιμώμενες παραμέτρους του ηδονικού υποδείγματος που εμπλέκονται στον υπολογισμό του (ποιοτικά προσαρμοσμένου) δείκτη. Θεωρώντας, μάλιστα, τις «σκιώδεις» τιμές των επιμέρους χαρακτηριστικών ως παγιομένες σε καθένα από τα κυλιόμενα παράθυρα περιόδων που εμπλέκονται στον υπολογισμό του δείκτη, η ποιοτική μεταβολή αποτυπώνεται στην περίοδο που έλαβε χώρα η συναλλαγή του ακινήτου.

Η μέθοδος RTD προϋποθέτει την επιλογή ενός μήκους $\ell \in \mathbb{N}$ των κυλιόμενων παραθύρων. Αν τα τρίμηνα με διαθέσιμα δεδομένα για την κατάρτιση του ποιοτικά προσαρμοσμένου δείκτη τιμών είναι $d \in \mathbb{N}$ το πλήθος, τότε το πλήθος των κυλιόμενων

παραθύρων που θα συμμετάσχουν στον υπολογισμό του δείκτη θα ισούται με $d-\ell+1$. Επίσης το πλήθος των αγοραπωλησιών του παραθύρου

$$\langle t\rangle := \{t, \ldots, t+\ell-1\}$$

θα δίνεται από το άθροισμα

$$M_t := \sum_{s\in\{t,\ldots,t+\ell-1\}} N_s \ \text{ για κάθε } \ t \in \mathbb{T}.$$

Στη μέθοδο RM είδαμε ότι το ρόλο της εξαρτημένης μεταβλητή στο αντίστοιχο ηδονικό υπόδειγμα (RePM) μπορεί να έχει οποιοσδήποτε αντιστέψιμος μετασχηματισμός των ονομαστικών τιμών/αξιών των αγαθών (ακινήτων). Μια προφανής επιλογή για δεδομένα τιμών, οι κατανομές των οποίων συνηθίζεται να παρουσιάζουν έντονη δεξιά ασυμμετρία, είναι η λογαριθμική συνάρτηση. Ειδικά για την περίπτωση της μεθόδου RTD, θα θεωρήσουμε ότι $h(y) := \ln y$ για κάθε $y > 0$.

Τότε, ο υπολογισμός ενός ποιοτικά προσαρμοσμένου δείκτη τιμών σύμφωνα με τη μέθοδο RTD έχει ως εξής:

Βήμα 0: Καθορίζεται το μήκος παραθύρου ℓ. Τότε για κάθε $t \in \mathbb{T}$ το παρακάτω υπόδειγμα πολλαπλής γραμμικής παλινδρόμησης αναφέρεται στο παράθυρο $\langle t\rangle$:

$$\ln Y_{t,m} = \beta_{0,\langle t\rangle} + \sum_{j\in\{1,\ldots,c\}} \beta_{j,\langle t\rangle} X_{t,m} + \sum_{v\in\{t-\ell+1,\ldots,t\}} \delta_v^{\langle t\rangle} D_{v,m} + \varepsilon_{t,m}, \ (RTDM)$$

όπου

- $D_{v,m}$ η δείκτρια τ.μ. της m-αγοραπωλησίας του v-τριμήνου ($v \in \mathbb{T}$, $m = 1, \ldots, M_t$),
- $\beta_{j,\langle t\rangle} \in \mathbb{R}$ οι παράμετροι του υποδείγματος που αντιστοιχούν στα ποιοτικά χαρακτηριστικά του ακινήτου,
- $\delta_v^{\langle t\rangle} := \delta_v^{[t-\ell+1,t]} \in \mathbb{R}$ οι παράμετροι του υποδείγματος που αντιστοιχούν στις δείκτριες τ.μ. των τριμήνων που αποτελούν το κυλιόμενο παράθυρο $\langle t\rangle$, και
- $\varepsilon_{t,m}$ τα τυχαία σφάλματα.

Βήμα 1: Υπολογίζεται το εκτιμώμενο (RTDM): Για κάθε $t \in \mathbb{T}$ και $m = 1, \ldots, M_t$

$$\widehat{\ln Y_{t,m}} = \widehat{\beta}_{0,\langle t\rangle} + \sum_{j\in\{1,\ldots,c\}} \widehat{\beta}_{j,\langle t\rangle} X_{t,m} + \sum_{v\in\{t-\ell+1,\ldots,t\}} \widehat{\delta}_v^{\langle t\rangle} D_{v,m} \ \ P-\sigma.\beta.. \quad (RTDM*)$$

Βήμα 2: Υπό τις συνθήκες Laspereyes και των παγιομένων «σκιωδών» τιμών, υπολογίζονται οι λόγοι

$$\frac{\widehat{Y}_{v+1,m}}{\widehat{Y}_{v,m}} = e^{\widehat{\delta}_{v+1}^{\langle v+1\rangle} - \widehat{\delta}_v^{\langle v\rangle}} = e^{\widehat{\delta}_{v+1}^{[v+1,v+\ell]} - \widehat{\delta}_v^{[v,v+\ell-1]}}, \ \ \nu \in \{r, \ldots, t-1\}.$$

Βήμα 3: Τότε, ο ποιοτικά προσαρμοσμένος δείκτης τιμών (τύπου Laspereyes) του τριμήνου t ως προς το τρίμηνο αναφοράς r θα δίνεται από τον τύπο

$$I_{t:r} = \prod_{\nu\in\{r,\ldots,t-1\}} e^{\widehat{\delta}_{v+1}^{[v+1,v+\ell]} - \widehat{\delta}_{v}^{[v,v+\ell-1]}} \quad \text{για κάθε } t \in \mathbb{T}.$$

Ακολουθεί μια σύγκριση των δύο βασικών μεθόδων ποιοτικής προσαρμογής που αναφέρθηκαν στην Υποενότητα 3.1.

Παρατηρήσεις 6. (α) Η μέθοδος RTD υπερτερεί της RM στα εξής:

(α1) **Διαθεσιμότητα Δεδομένων για τον υπολογισμό των δεικτών,** αφού για τον υπολογισμό του δείκτη κάθε περιόδου (συνήθως τριμήνου) αξιοποιούνται δεδομένα τόσων περιόδων όσο το μήκος του κυλιόμενου παραθύρου που έχει επιλεγεί. Αντίθετα, στην RM αν και ο υπολογισμός των «σκιωδών» τιμών βασίζεται σε δεδομένα που μπορεί να εκτείνονται σε πολλές περιόδους (π.χ. σε δεδομένα έτους ή διετίας), ο υπολογισμός του δείκτη καθεαυτού βασίζεται σε δεδομένα μιας μόνο περιόδου. Το τελευταίο γεγονός μπορεί να αποδειχθεί προβληματικό για συγκεκριμένες συνιστώσες του HPI όπου το πλήθος των αντίστοιχων συναλλαγών κατοικιών είναι μικρό (π.χ. Νέες Μονοκατοικίες).

(α2) **Εφαρμοσιμότητα:** Το παραπάνω πλεονέκτημα καθιστά τη μέθοδο RTD καταλληλότερη για περισσότερες αγορές (π.χ. μικρότερου μεγέθους ή/και με περιορισμούς στην τυπολογία των συναλλασόμενων ακινήτων) σε σχέση με την RM. Επίσης οι εξαχθέντες δείκτες με την RTD αναμένεται να είναι λιγότερο ευμετάβλητοι σε σχέση με τους αντίστοιχους δείκτες που έχουν υπολογιστεί με την RM αν και κάτι τέτοιο δεν είναι πάντα εγγυημένο και ούτε αποτελεί μοναδικό κριτήριο αξιολόγησης ενός ποιοτικά προσαρμοσμένου δείκτη τιμών.

(α3) **Συχνότητα Κατάρτισης:** Οι RTD δείκτες είναι πιο εύκολο να καταρτιστούν και με μικρότερη της τριμηνιαίας συχνότητας (βλ. και Ενότητα 4.), σε αντίθεση με τους δείκτες που καταρτίζονται με την RM και για τους οποίους συνήθως μια πιο αραιή συχνότητα κατάρτισης μπορεί να γίνει μόνο εφικτή.

(β) Η μέθοδος RTD υστερεί της RM (και των συναφών μεθόδων (M1) και (M2), βλ. Υποενότητα 3.1) στα εξής:

(β1) **Ευαισθησία σε επιλογές:** Η μέθοδος RTD προφανώς εξαρτάται σε μεγάλο βαθμό από την επιλογή του μήκους ℓ του κυλιόμενου παραθύρου. Αν και έχει επικρατήσει η τιμή $\ell = 4$ για τριμηνιαίους δείκτες (δηλαδή κυλιόμενο παράθυρο ενός έτους, το οποίο ενδέχεται να εξομαλύνει και θέματα εποχικών κυμάνσεων) δεν απουσιάζουν και διαφορετικές επιλογές ή/και κατά καιρούς αλλαγές μήκους παραθύρων. Επομένως, θεωρητικά θα ήταν επιθυμητή μια εξαντλητική μελέτη σχετικά με την ευστάθεια/μεταβλητότητα των υπολογισθέντων δεικτών μέσω της RTD για διαφορετικά ℓ. Παρόμοια ζητήματα ανακύπτουν ακόμα πιο έντονα αναφορικά με την κατάρτιση δεικτών RTD υψηλότερης συχνότητας, όπως μηνιαίων ή εβδομαδιαίων, που τα τελευταία έτη

κερδίζουν συνεχώς έδαφος στη σχετική έρευνα (βλ. π.χ. Hill et al. (2018) και Hill et al. (2021)).

Από την άλλη, η μέθοδος RM δεν παρουσιάζει ανάλογα ζητήματα πέραν της επιλογής της περιόδου εκτίμησης $\mathbb{T}_e$ των ηδονικών υποδειγμάτων — κι άρα και των «σκιώδων» τιμών των ποιοτικών χαρακτηριστικών των αγαθών (κατοικιών) — η οποία συνδέεται και με την απόφαση για τη συχνότητα αναθεώρησής τους. Μάλιστα, επισημαίνεται πως εσχάτως φαίνεται να υιοθετείται η ετήσια ή διετής αναθεώρηση των υποδειγμάτων που εμπλέκονται στον υπολογισμό RM (βλ. European Commission, Eurostat, Reference Metadata) εγκαταλείποντας πρακτικές των πρώτων ετών εφαρμογής της μεθόδου που προέκριναν (τουλάχιστον) πενταετή αναθεώρηση.

(β2) **Ερμηνευσιμότητα:** Πρόκειται για το προφανέστερο και ισχυρότερο πλεονέκτημα της RM έναντι της RTD. Εκ κατασκευής, η πρώτη μέθοδος επιτρέπει την ξεχωριστή μελέτη της συνιστώσας των ονομαστικών τιμών ενός αγαθού και της μεταβολής στην ποιότητά του μέσω αντίστοιχων δεικτών (των QUI και QAF, βλ. Υποενότητα 3.2). Επί πλέον, η διαμόρφωση των σχετικών υποδειγμάτων επιτρέπει την καθαρή αποτύπωση της επίδρασης καθενός από τα επιμέρους ποιοτικά χαρακτηριστικά στη μεταβολή της ποιότητας του μείγματος — η μείωση κατά 25% των τεχνητών τιμών του Παραδείγματος 1 αποδίδεται στη μείωση των κύριων επιφανειών. Κάτι τέτοιο ισχύει και σε μη απλουστευμένα υποδείγματα όπως αυτό του Παραδείγματος 1, όπου η μεταβολή ποιότητας θα μπορούσε να αποδοθεί λ.χ. σε μεταβολές στην παλαιότητα και την τοποθεσία των κατοικιών.

Αντίθετα, και πάλι εκ κατασκευής, η μέθοδος RTD δεν προσφέρει τη δυνατότητα ανάλυσης του δείκτη τιμών σε ονομαστική και ποιοτική συνιστώσα και καθιστά εως πολύ δύσκολη τη μελέτη της επίδρασης των επιμέρους ποιοτικών χαρακτηριστικών λόγω της εμπλοκής των κυλιόμενων παραθύρων και των σχετικών δείκτριων των περιόδων τ.μ. στα ηδονικά υποδείγματα (βλ. και Υποενότητα 3.3).

Στο σημείο αυτό αξίζει να επισημανθεί το εξής: Το πλεονέκτημα (β2) της Παρατήρησης 6 έχει διαφορετική βαρύτητα ανάλογα με το πεδίο εφαρμογής των ηδονικών υποδειγμάτων. Αν για παράδειγμα, ένα υπόδειγμα χρησιμοποιείται για την εκτίμηση των τιμών πώλησης Η/Υ από μια τεχνολογική εταιρεία, η ερμηνευσιμότητά του ίσως δεν είχε τόσο μεγάλη σημασία όσο η αξιολόγηση του (π.χ. μέσω ενός υψηλού συντελεστή R^2). Ανάλογη θα μπορούσε να είναι η στάση μιας τράπεζας για το πρόβλημα της αποτίμησης του χαρτοφυλακίου ακινήτων της. Όταν, όμως, η στόχευση εστιάζεται στην αποτύπωση μακροοικονομικών τάσεων και τη χάραξη οικονομικών/φορολογικών πολιτικών — κάτι που ισχύει κατ′ εξοχήν στην περίπτωση των Επισήμων Στατιστικών Ακινήτων — το ισοζύγιο μεταξύ αποτελεσματικότητας και ερμηνευσιμότητας ενός υποδείγματος/εξαγόμενου κλίνει υπέρ της δεύτερης, κάτι που υπογραμμίζει ακόμα περισσότερο το συγκεκριμένο συγκριτικό πλεονέκτημα της

μεθόδου RM.

3.4 Άλλα σχετικά θέματα

Ο υπολογισμός του δείκτη HPI μέσω των δύο βασικών μεθόδων ποιοτικής προσαρμογής ήταν αυτός που απασχόλησε μέχρι τώρα την παρούσα ενότητα. Παρόλα αυτά, υπάρχουν ακόμα αρκετά σχετικά θέματα τα οποία θα μπορούσαν να αποτελέσουν αντικείμενο ξεχωριστών εργασιών, όπως τα εξής:

(I) **Η επιλογή ερμηνευτικών μεταβλητών στα ηδονικά υποδείγματα:** Αν εξαιρέσουμε την κύρια επιφάνεια και την παλαιότητα του ακινήτου, οι οποίες είναι σχεδόν θέσφατο ότι θα συμπεριληφθούν σε ένα ηδονικό υπόδειγμα με εξαρτημένη μεταβλητή την αξία ακινήτου, οι υπόλοιπες ερμηνευτικές μεταβλητές μπορεί να ποκίλουν ανάλογα με τη διαθεσιμότητα των δεδομένων και το είδος του υποδείγματος (με όρους αλληλεπίδρασης ή χωρίς, γεωχωρικό ή μη). Μεταβλητές σχετικά με την τοποθεσία του ακινήτου είναι αναμενόμενο να παίζουν καθοριστικό ρόλο, αλλά να μην είναι πάντα διαθέσιμες. Άλλα ποιοτικά χαρακτηριστικά όπως ο τύπος ακινήτου (διαμέρισμα ή όχι), ύπαρξη χώρου στάθμευσης, ο αριθμός των υπνοδωματίων έχουν αποδειχθεί επίσης σημαντικά στη διεθνή πρακτική και βιβλιογραφία (βλ. π.χ. Hill et al. (2018)).

(II) **Πηγές δεδομένων:** Η καταλληλότητα και η ποιότητα τους αποτελεί αναγκαία συνθήκη για την παραγωγή αξιόπιστων επισήμων στατιστικών. Ειδικά για τους δείκτες τιμών ακινήτων (HPI και OOHPI) υπάρχουν περιπτώσεις που απαιτούνται δύο κατηγορίες πήγων, μία για καθένα από τα δομικά στοιχεία τους, δηλαδή τους υποδείκτες και τις σταθμίσεις τους (βλ. και Λήμμα 1). Γενικά πάντως, και σύμφωνα με τα πιο πρόσφατα αναθεωρημένα μεταδεδομένα των K-M (βλ. European Commission, Eurostat, Reference Metadata), στην περίπτωση του HPI σταθμίσεις και υποδείκτες βασίζονται στα ίδια δεδομένα. Στην πλειοψηφία των περιπτώσεων πρόκειται για φορολογικά δεδομένα που προέρχονται από τις αρμόδιες εθνικές αρχές ή/και κτηματολογικά δεδομένα που προέρχονται από κτηματολογικές υπηρεσίες, υποθηκοφυλάκεια ή άλλους κρατικούς φορείς (π.χ. υπουργεία). Υπάρχουν περιπτώσεις όπου αντλούνται στοιχεία Εθνικών Λογαριασμών μέσα από τις ίδιες τις ΣΥ κυρίως για τον υπολογισμό σταθμίσεων (η σουηδική περίπτωση, βλ. European Commission, Eurostat, National Reference Metadata, Sweden). Ειδικές δειγματοληπτικές έρευνες και τραπεζικά ιδρύματα μπορεί επίσης να αποτελέσουν πηγές δεδομένων αν και η διεθνής πρακτική έχει αποδείξει ότι αυτό γίνεται πλέον με μικρότερη συχνότητα καθώς οι προαναφερθείσες πηγές διοικητικών δεδομένων κερδίζουν όλο και περισσότερο έδαφος.

(III) **Προεπεξεργασία δεδομένων:** Ακόμα, όμως, κι αν οι πηγές δεδομένων είναι οι καταλληλότερες δυνατές, στα αρχικά σύνολα δεδομένων που διαβιβάζονται προς τις ΣΥ δεν απουσιάζουν τα ελαττώματα, όπως τυπογραφικά λάθη, ελλιπούσες ή ακραίες τιμές. Επομένως, τα δεδομένα που εμπλέκονται τελικά στον υπολογισμό του HPI είναι υποσύνολα του αρχικού συνόλου δεδομένων

και δύνανται να διαφοροποιούνται ανάλογα με το στάδιο υπολογισμού του δείκτη. Για παράδειγμα, στην RM τα δεδομένα βάσει των οποίων υπολογίζονται τα μέσα διανύσματα (βλ. Υποενότητα 3.2, Βήμα 3) δεν ταυτίζονται κατ'ανάγκη με αυτά βάσει των οποίων εκτιμώνται οι «σκιώδεις» τιμές $\widehat{\beta}_j$.

Κλείνοντας την παρούσα ενότητα, κρίνεται σκόπιμο να γίνει μια καταγραφή της κύριας μεθόδου κατάρτισης του δείκτη HPI που εφαρμόζεται σε διάφορες ευρωπαϊκές χώρες. Ο όρος «κύρια» έχει την έννοια ότι αναφέρεται στη μέθοδο κατάρτισης της βασικής συνιστώσας του HPI, αφού υπάρχουν περιπτώσεις χωρών (Γαλλία, Σλοβενία και Σουηδία) οι οποίες χρησιμοποιούν διαφορετικές μεθόδους κατάρτισης για διαφορετικές συνιστώσες του HPI.

Πιο συγκεκριμένα, και σύμφωνα με τα πιο πρόσφατα μεταδεδομένα (βλ. European Commission, Eurostat, Reference Metadata), προκύπτει ότι 8 ευρωπαϊκές χώρες ακολουθούν την RTD, ενώ 6, 3 και 2 την RM και τις μεθόδους Μέσων Χαρακτηριστικών και Υποκατάστασης Τιμών, αντίστοιχα. Τέσσερις χώρες έχουν καταφύγει στη Προσαρμογή Μείγματος/Στρωματοποίηση, ενώ ιδιαίτερης μνείας πρέπει να γίνει στην ολλανδική, δανική και σουηδική περίπτωση όπου η κύρια συνιστώσα του HPI που αναφέρεται στις Υφιστάμενες Κατοικίες υπολογίζεται με τη μέθοδο Λόγου Εκτιμήσεων Τιμών Πώλησης (SPAR: Sales Price Appraisal Ratio).

Η τελευταία δεν ανήκει στην οικογένεια των μεθόδων ποιοτικής προσαρμογής (βλ. Υποενότητα 3.1), αλλά σε αυτή των μεθόδων που βασίζονται σε επαναλαμβανόμενες πωλήσεις της ίδιας κατοικίας. Οι εξαγόμενοι με αυτή τη μέθοδο δείκτες είναι κλασσικοί δείκτες τιμών — δεν έχουν δηλαδή ανάγκη ποιοτικής προσαρμογής — και βρίσκουν εφαρμογές σε συγκεκριμένου τύπου αγορές (με κατ'εξοχήν παράδειγμα αυτή των Η.Π.Α.) που χαρακτηρίζονται από υψηλή συχνότητα πωλήσεων της ίδιας κατοικίας.

Επομένως, θα μπορούσαμε να ισχυριστούμε ότι ένα παρεπόμενο τέτοιου τύπων αγορών είναι ότι κάνουν τη δουλειά των ΣΥ πιο εύκολη κατά την κατάρτιση των δεικτών τιμών κατοικιών, αφού για τους τελευταίους μπορεί να παρακαμφθεί η απαίτηση της ποιοτικής προσαρμογής. Ταυτόχρονα, όμως, συνήθως υποκρύπτουν ένα άλλο παρεπόμενο σε κοινωνικό επίπεδο, αφού σε τέτοιου τύπου αγορές οι κατοικίες δεν αντιμετωπίζονται ως *οίκοι* αλλά ως *στοιχεία ενεργητικού (assets)* με όποιες συνέπειες έχει κάτι τέτοιο σε κοινωνικό επίπεδο...

4. ΑΝΟΙΧΤΑ ΘΕΜΑΤΑ

Οι Στατιστικές (Τιμών) Ακινήτων αποτελούν ένα από τα πιο δυναμικά και καινοτόμα αντικείμενα στο χώρο των Επισήμων Στατιστικών, τόσο λόγω του μεγάλου ενδιαφέροντος για σχετικά στατιστικά προϊόντα από διάφορους φορείς (βλ. Ενότητα 3.) όσο και του γεγονότος ότι οι εθνικές ή υπερεθνικές στατιστικές υπηρεσίες καλούνται πλέον να δραστηριοποιηθούν σε έναν μη παραδοσιακό τομέα για αυτές. Είναι, λοιπόν, αναμενόμενο τα ανοιχτά θέματα και οι προκλήσεις στον χώρο να είναι αρκετές. Εδώ θα περιοριστούμε στο να ονομάσουμε μόνο μερικές από αυτές:

- Η παραγωγή μηνιαίων δεικτών τιμών κατοικιών: Πρόκειται για ένα ζητούμενο που αφορά αρχικά τον δείκτη HPI, αλλά αποσκοπεί τελικά στην παραγωγή ενός ΕνΔΤΚ με ενσωματομένες τις δαπάνες αγοράς και κυριότητας κατοικιών. Το πρώτο μέρος ικανοποιεί από μόνο του ανάγκες πληροφόρησης των χρηστών και φαντάζει πιο εφικτό. Και για τα δύο μέρη, πάντως, η μεγαλύτερη συχνότητα παραγωγής πρέπει να συνδυαστεί με διατήρηση της ποιότητας και της συγκρισιμότητας των παραγόμενων δεικτών.
- Εκτός από την εξέλιξη των τιμών των κατοικιών ενδιαφέρον παρουσιάζει και ο όγκος πωλήσεων κατοικιών. Ο Δείκτης Όγκου Πωλήσεων (HSI: House Sales Indicator) παράγεται ήδη πιλοτικά από κάποια Κ-Μ.
- Σε πιο τεχνικά θέματα, η παρουσία ή μη εποχικών κυμάνσεων/προτύπων στη χρονοσειρά του δείκτη HPI είναι κάτι που αξίζει περαιτέρω διερεύνησης. Κάτι τέτοιο βέβαια σχετίζεται με τη συμπεριφορά τόσο της αγοράς ακινήτων όσο και της ευρύτερης οικονομικής δραστηριότητας μιας χώρας. Σήμερα, οι μοναδικές ευρωπαϊκές χώρες που δημοσιεύουν εποχικά εξομαλυμένο δείκτη HPI είναι η Νορβηγία και η Ιρλανδία.
- Σύγκριση διαφόρων τύπων δεικτών τιμών κατοικιών: Εκτός από τους επίσημους δείκτες τιμών κατοικιών που παράγονται από τις ΣΥ δημοσιεύονται και άλλοι δείκτες τιμών κατοικιών (βλ. και Ενότητες 1. και 3.), οι οποίοι συνήθως βασίζονται σε δεδομένα εκτιμήσεων αξιών ακινήτων (*appraisals*) και ζητούμενων τιμών πώλησης από σχετικούς ιστότοπους (*asking prices*). Συγκριτικές μελέτες των διαφορετικών αυτών δεικτών τιμών κατοικιών προκαλούν το ενδιαφέρον.
- Αν και στην παρούσα εργασία επικεντρωθήκαμε στους δείκτες τιμών οικιστικών ακινήτων, δεν πρέπει να παραβλέψουμε την αλματώδη πρόοδο που παρατηρείται ειδικά κατά την τελευταία πενταετία σε αντίστοιχα στατιστικά προϊόντα εμπορικών ακινήτων, γνωστά και ως CREIs (Commercial Real Estate Indicators). Πρόκειται σαφώς για ένα νέο πεδίο έρευνας με ξεχωριστές προκλήσεις, όπως η διακριτή αντιμετώπιση διαφορετικών ειδών εμπορικών ακινήτων (π.χ. γραφεία, καταστήματα, βιομηχανικά κτίρια) και η εμπλοκή των λεγόμενων «τριγωνικών» συναλλαγών.

5. ΣΥΝΟΠΤΙΚΑ ΣΥΜΠΕΡΑΣΜΑΤΑ

Συνοψίζοντας, αναφέρουμε ότι οι δείκτες τιμών κατοικιών ακολουθούν τα πρότυπα του ΔΤΚ και του ΕνΔΤΚ και αποτελούν λόγους δαπανών. Είναι αλυσωτοί δείκτες τύπου Laspereyes με επικαιροποιημένες σταθμίσεις για καθένα από τους υποδείκτες που τους αποτελούν, οι οποίοι πρέπει να είναι τουλάχιστον δύο: αυτοί των Νέων και των Υφιστάμενων Κατοικιών. Η βασική ιδιαιτερότητα των δεικτών τιμών κατοικιών — η οποία απορρέει από το γεγονός ότι τα ακίνητα αποτελούν εξαιρετικά ανομοιογενή ως προς τα επιμέρους ποιοτικά τους χαρακτηριστικά αγαθά — είναι ότι πρέπει να είναι *ποιοτικά προσαρμοσμένοι*, απαλλαγμένοι δηλαδή από την επίδραση που ασκούν στη διαμόρφωση των τιμών τους οι μεταβολές (από περίοδο σε περίοδο) στην ποιότητα

των ακινήτων που αποτέλεσαν αντικείμενο συναλλαγής. Οι RM και RTD αποτελούν τις δύο βασικές, όχι όμως τις μοναδικές, μεθόδους μέσω της εφαρμογής των οποίων επιτυγχάνεται η ποιοτική προσαρμογή των δεικτών τιμών κατοικιών και εμπλέκουν την εφαρμογή υποδειγμάτων πολλαπλής γραμμικής παλινδρόμησης γνωστών και ως «ηδονικά υποδείγματα». Πέραν των μεθόδων αυτών, οι πηγές δεδομένων συναλλαγών ακινήτων και ο χειρισμός αυτών παίζουν καθοριστικό ρόλο για την παραγωγή κατάλληλων και αξιόπιστων δεικτών τιμών κατοικιών. Οι δείκτες όγκου πωλήσεων κατοικιών και εμπορικών ακινήτων αποτελούν δύο από τα πολλά νέα πεδία έρευνας στο χώρο των επισήμων στατιστικών ακινήτων.

ABSTRACT

In its simplest version, a price index is a price (or total expenses) relative between two time periods and is used to capture the evolution of a value/price compared to a time milestone. By the term "official statistics" we generally refer to statistical outputs that are produced and disseminated within a system of national or international organizations (e.g. government agencies, independent authorities) both according to basic principles that ensure minimum professional standards and under a specific legal framework (see e.g. European Commission, Eurostat, Statistics Explained). In the aftermath of the 2008 financial crisis, the interest in compiling new official statistics and in particular real estate price indices has increased significantly. In this paper, we will focus on methodological issues of residential property price indices compiled by European National Statistical Institutes (background, basic concepts and particularities), we will refer to the proposed compilation methods, such as that of Rolling Time Dummies (RTD), and we will make a brief reference to the current trends in Europe, see e.g. European Commission, Eurostat (2017) and Hill et al. (2018).

Δήλωση Ευθύνης: Οι αναλύσεις και οι απόψεις που περιέχονται σε αυτή την εργασία ανήκουν αποκλειστικά στον συγγραφέα και όχι στην Ελληνική Στατιστική Αρχή (ΕΛΣΤΑΤ) ή στο Ελληνικό Στατιστικό Ινστιτούτο (ΕΣΙ).

ΑΝΑΦΟΡΕΣ

Εκτελεστικός κανονισμός (ΕΕ) 2023/1470 της Επιτροπής της 17ης Ιουλίου 2023 για τον καθορισμό των μεθοδολογικών και τεχνικών προδιαγραφών σύμφωνα με τον κανονισμό (ΕΕ) 2016/792 του Ευρωπαϊκού Κοινοβουλίου και του Συμβουλίου όσον αφορά τον δείκτη τιμών κατοικιών και τον δείκτη τιμών για τις ιδιοκατοικούμενες κατοικίες, και για την τροποποίηση του κανονισμού (ΕΕ) 2020/1148 της Επιτροπής (Κείμενο που παρουσιάζει ενδιαφέρον για τον ΕΟΧ)
ΕΕ L 181 της 18.7.2023, p. 1–15 (BG, ES, CS, DA, DE, ET, EL, EN, FR, GA, HR, IT, LV, LT, HU, MT, NL, PL, PT, RO, SK, SL, FI, SV)

Κανονισμός (ΕΕ) αριθ. 93/2013 της Επιτροπής, της 1ης Φεβρουαρίου 2013 , σχετικά με τον καθορισμό λεπτομερών κανόνων εφαρμογής του κανονισμού (ΕΚ) αριθ. 2494/95

για τη θέσπιση εναρμονισμένων δεικτών τιμών καταναλωτή, όσον αφορά τον καθορισμό δεικτών τιμών για τις ιδιοκατοικούμενες κατοικίες Κείμενο που παρουσιάζει ενδιαφέρον για τον EOX
EE L 33 της 2.2.2013, p. 14–16 (BG, ES, CS, DA, DE, ET, EL, EN, FR, IT, LV, LT, HU, MT, NL, PL, PT, RO, SK, SL, FI, SV)

Κανονισμός (ΕΕ) 2016/792 του Ευρωπαϊκού Κοινοβουλίου και του Συμβουλίου, της 11ης Μαΐου 2016, για τους εναρμονισμένους δείκτες τιμών καταναλωτή και τον δείκτη τιμών κατοικιών και την κατάργηση του κανονισμού (ΕΚ) αριθ. 2494/95 του Συμβουλίου (Κείμενο που παρουσιάζει ενδιαφέρον για τον EOX)
EE L 135 της 24.5.2016, p. 11–38 (BG, ES, CS, DA, DE, ET, EL, EN, FR, GA, HR, IT, LV, LT, HU, MT, NL, PL, PT, RO, SK, SL, FI, SV)

Brooks, C. and Tsolacos, S. (2010). *Real Estate Modelling and Forecasting*, Cambridge University Press, Cambridge.

European Commission, Eurostat (2017). *Technical manual on Owner-Occupied Housing and House Price Indices.*

European Commission, Eurostat. House price and sales index (prc_hpi_inx), National Reference Metadata, Sweden.
https://ec.europa.eu/eurostat/cache/metadata/EN/prc_hpi_inx_esmshpi_se.htm.

European Commission, Eurostat. House price and sales index (prc_hpi_inx), Reference Metadata.
https://ec.europa.eu/eurostat/cache/metadata/en/prc_hpi_inx_esms.htm.

European Commission, Eurostat. Statistics Explained,
https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Beginners:Statistical_concept_-_What_are_official_statistics?.

Hill, R.J., Rambaldi, A.N. and Scholz, M. (2021). Higher frequency hedonic property price indices: a state-space approach. *Empirical Economics* **61**, 417-441.

Hill, R.J., Scholz, M., Shimizu, C. and Steurer, M. (2018). An evaluation of the methods used by European countries to compute their official house price indices. *Economie et Statistique / Economics and Statistics* **500-502**, 221-238.

Tsolacos, S. and Andrew, M. (2020). *Applied Quantitative Analysis for Real Estate*, 1st edition, Taylor & Francis, London.

von der Lippe, P. (2001). *Chain Indices. A Study in Price Index Theory*, Statistisches Bundesamt, Wiesbaden.

Greek Statistical Institute
Proceedings of the 37th Panhellenic Statistics Conference (2025), pp. 85–101

ΕΛΕΓΧΟΣ ΚΑΛΗΣ ΠΡΟΣΑΡΜΟΓΗΣ ΓΙΑ ΤΗ ΣΥΝΑΡΤΗΣΗ ΚΙΝΔΥΝΟΥ

Πασχάλης Μαυρίδης[1], *Δημήτριος Μπάγκαβος*[1]
[1] Τμήμα Μαθηματικών, Πανεπιστήμιο Ιωαννίνων, Ιωάννινα, Ελλάδα
p.mavridis@uoi.gr

ΠΕΡΙΛΗΨΗ

Η παρούσα μελέτη προτείνει έναν νέο έλεγχο καλής προσαρμογής για τη συνάρτηση κινδύνου, βασισμένο στη μέθοδο των πυρήνων, όταν τα δεδομένα υπόκεινται σε τυχαία δεξιά λογοκρισία. Η στατιστική συνάρτηση ελέγχου αναπτύσσεται με βάση την L_2 μετρική και η ασυμπτωτική της κατανομή προσδιορίζεται αναλυτικά υπό τη μηδενική υπόθεση. Η ισχύς του τεστ μελετάται για εναλλακτικές υποθέσεις τύπου Pitman μέσω προσομοιώσεων με δεδομένα από γνωστές κατανομές, για διάφορα μεγέθη δείγματος και ποσοστά λογοκρισίας.

Λέξεις Κλειδιά: Συνάρτηση κινδύνου, Έλεγχοι καλής προσαρμογής, Μέθοδος των πυρήνων, Τυχαία δεξιά λογοκρισία, Εναλλακτικές Pitman

1. ΕΙΣΑΓΩΓΗ

Η συνάρτηση κινδύνου (hazard function) κατέχει κεντρικό ρόλο σε διάφορα επιστημονικά πεδία όπως η ανάλυση επιβίωσης, η βιοστατιστική, η δημογραφία και η μηχανική καθώς ποσοτικοποιεί τον στιγμιαίο κίνδυνο αποτυχίας και αποτελεί βασικό εργαλείο τόσο για την εξαγωγή στατιστικών συμπερασμάτων όσο και για τη λήψη αποφάσεων. Συχνά, οι αναλυτές σε αυτούς τους χώρους ενδιαφέρονται να ελέγξουν την ορθότητα ενός παραμετρικού μοντέλου απευθείας εκφρασμένο με τη μορφή της συνάρτησης κινδύνου (βλ. Huh and Hutmacher (2016)). Αυτό συμβαίνει διότι η εκτίμηση της συνάρτησης κινδύνου είναι πρωταρχικής σημασίας, αφού εκείνο που ενδιαφέρει κατά κύριο λόγο τους ερευνητές σε αυτούς τους χώρους είναι ο ρυθμός με τον οποίο συμβαίνει το υπό μελέτη γεγονός ή, ακόμη περισσότερο, η αλλαγή στον ρυθμό με τον οποίο συμβαίνουν τα γεγονότα, η οποία εκφράζεται μέσω της συνάρτησης κινδύνου. Αυτός ήταν ουσιαστικά και ο λόγος για τον οποίο ο Hjort (1990) (σελ. 1222) πρότεινε μια γενική κλάση ελέγχων καλής προσαρμογής για την αξιολόγηση της επάρκειας παραμετρικών μοντέλων επιβίωσης, βασισμένων στη αθροιστική συνάρτηση κινδύνου. Στη συνέχεια, ο Peña (1998) ανέπτυξε μια

γενίκευση του ελέγχου του Neyman (1937) για σύνθετες υποθέσεις σε μοντέλα βασισμένα στη συνάρτηση κινδύνου υπό τυχαία λογοκρισία, ενώ οι Bagdonavičius and Nikulin (2011) επέκτειναν την προσέγγιση του Hjort (1990) σε ευρύτερες κατηγορίες μοντέλων, όπως τα μοντέλα επιταχυνόμενου χρόνου αποτυχίας, τα γενικευμένα μοντέλα αναλογικών κινδύνων και τα μοντέλα ευπάθειας. Πιο πρόσφατα, οι Vital και Patil (2020, 2021) πρότειναν δύο μη παραμετρικούς ελέγχους καλής προσαρμογής για τη συνάρτηση κινδύνου. Έναν έλεγχο που βασίζεται στο ιστόγραμμα ικανό να ανιχνεύει τόσο σταθερές όσο και τοπικές εναλλακτικές υποθέσεις, καθώς και έναν έλεγχο τύπου χ^2, του οποίου οι ασυμπτωτικές ιδιότητες είναι ανάλογες με αυτές του κλασικού ελέγχου του Pearson.

Η κατασκευή του ελέγχου ξεκινά (βλ. Ενότητα 2.) με τον ορισμό της στατιστικής συνάρτησης ελέγχου ως την απόσταση μεταξύ της πραγματικής συνάρτησης κινδύνου και της υποτιθέμενης, υπό τη μηδενική υπόθεση, καμπύλης. Για τη μέτρηση της απόστασης εδώ χρησιμοποιούμε την L_2 μετρική. Επειδή η συνάρτηση κινδύνου είναι άγνωστη στην πράξη, στη θέση της χρησιμοποιούμε τον κλασικό εκτιμητή πυρήνα που προτάθηκε στην εργασία των Tanner and Wong (1983). Εστιάζουμε σε σύνθετες εναλλακτικές υποθέσεις, δηλαδή σε υποθέσεις που η δεδομένη (υπό τη μηδενική υπόθεση) συνάρτηση κινδύνου ανήκει σε κάποια παραμετρική οικογένεια συναρτήσεων κίνδυνου, με τις άγνωστες παραμέτρους να έχουν προκύψει μέσω εκτίμησης με τη μέθοδο της μέγιστης πιθανοφάνειας (Maximum Likelihood Estimation). Όπως αναμένεται, υπό τη μηδενική υπόθεση η παραγόμενη ελεγχοσυνάρτηση ισούται με μηδέν. Αντίθετα, υπό την εναλλακτική υπόθεση η τιμή της ελεγχοσυνάρτησης αποκλίνει προς το άπειρο όσο η απόσταση μεταξύ των δυο συναρτήσεων κινδύνου αυξάνει.

Η δομή του άρθρου είναι η εξής. Στην Ενότητα 3. αποδεικνύουμε ότι η ελεγχοσυνάρτηση ασυμπτωτικά συγκλίνει προς την κανονική κατανομή με παραμέτρους οι οποίες διαφέρουν ανάλογα με το ποσό της εξομάλυνσης που εφαρμόζεται στα δεδομένα. Στην Ενότητα 4. μελετούμε, μέσω προσομοιώσεων, την ισχύ του προτεινόμενου ελέγχου για εναλλακτικές υποθέσεις τύπου Pitman, εκφραζόμενες ως προς τη συνάρτηση κινδύνου, ενώ η εργασία ολοκληρώνεται με την Ενότητα 5., η οποία περιλαμβάνει τα συμπεράσματα της παρούσας έρευνας. Στο Παράρτημα A παρατίθενται τα βοηθητικά λήμματα που απαιτούνται για την απόδειξη του κυρίου θεωρήματος.

Τέλος, σημειώνεται ότι η παρούσα εργασία αποτελεί επέκταση της μεταπτυχιακής διατριβής του πρώτου συγγραφέα, η οποία εκπονήθηκε υπό την επίβλεψη του δεύτερου συγγραφέα (βλ. Μαυρίδης (2023)). Σε αντίθεση με τη μεταπτυχιακή διατριβή, όπου ο έλεγχος αναπτύχθηκε για την πλήρως προσδιορισμένη (fully specified) περίπτωση, στην παρούσα μελέτη εξετάζεται η σύνθετη (composite) περίπτωση, με τις παραμέτρους να εκτιμώνται από τα δεδομένα.

2. ΣΥΜΒΟΛΙΣΜΟΙ ΚΑΙ ΟΡΙΣΜΟΣ ΕΛΕΓΧΟ-ΣΥΝΑΡΤΗΣΗΣ

Έστω $T_i, i = 1, \ldots, n$ ένα δείγμα μεγέθους n από ανεξάρτητους και ισόνομους χρόνους επιβίωσης, οι οποίοι λογοκρίνονται από τα δεξιά από n ανεξάρτητους και ισόνομους χρόνους λογοκρισίας $U_i, i = 1, \ldots, n$, με τα U_i να είναι ανεξάρτητα από τα T_i. Έστω f και F αντίστοιχα η πυκνότητα και η συνάρτηση κατανομής των T_i και H η συνάρτηση κατανομής των U_i. Τα δεδομένα που παρατηρούνται είναι τα ζεύγη $(X_i, \delta_i), i = 1, \ldots, n$ με $X_i = \min\{T_i, U_i\}$ και $\delta_i = \mathbf{1}_{\{Ti \leq U_i\}}$, όπου $\mathbf{1}_{\{\cdot\}}$ είναι μία δείκτρια τυχαία μεταβλητή του γεγονότος $\{\cdot\}$. Τα παρατηρούμενα δεδομένα αποτελούν δείγμα ανεξάρτητων και ισόνομων τυχαίων μεταβλητών με πυκνότητα g και συνάρτηση κατανομής G, η οποία ικανοποιεί τη σχέση $1 - G = (1 - F)(1 - H)$. Υποθέτουμε ότι η εκτίμηση λαμβάνει χώρα στο διάστημα $[0, m]$ με $m = \sup\{x : 1 - G(x) > \varepsilon\}$, για κάποιο μικρό $\varepsilon > 0$.

Ο κλασικός εκτιμητής πυρήνα της συνάρτησης κινδύνου $\lambda(x) = f(x)/\{1-F(x)\}$, ο οποίος συμβολίζεται με $\hat{\lambda}(x; h)$, εισήχθη στη βιβλιογραφία από τους Tanner and Wong (1983) και δίνεται από τον τύπο

$$\hat{\lambda}(x; h) = \frac{1}{nh} \sum_{i=1}^{n} \frac{\delta_i}{1 - G_n(X_i)} K\left(\frac{x - X_i}{h}\right), \tag{1}$$

όπου K είναι μια πραγματική συνάρτηση που ολοκληρώνει στο 1 και ονομάζεται πυρήνας (*kernel*), h είναι το εύρος ζώνης (*bandwidth*) το οποίο ελέγχει το ποσό της εξομάλυνσης που εφαρμόζεται στα δεδομένα και $G_n(X_i)$ είναι η εμπειρική συνάρτηση κατανομής των X_i.

Η παρούσα εργασία αναπτύσσει έναν έλεγχο καλής προσαρμογής για την υπόθεση

$$H_0 : \lambda = \lambda_0 \in P_\Theta \quad \text{έναντι της} \quad H_1 : \lambda = \lambda_1 \notin P_\Theta, \tag{2}$$

όπου $P_\Theta = \{\lambda(\cdot; \theta) : \theta \in \Theta \subseteq \mathbb{R}^p\}$ συμβολίζει μία παραμετρική οικογένεια συναρτήσεων κινδύνου της μορφής $\lambda(x) = \lambda(x; \theta)$, με παράμετρο $\theta \in \Theta \subseteq \mathbb{R}^p$, όπου p θετικός ακέραιος και Θ ο παραμετρικός χώρος. Στο πλαίσιο της παρούσας έρευνας περιοριζόμαστε σε τοπικές εναλλακτικές υποθέσεις τύπου Pitman. Ακολουθώντας την εργασία του Hjort (1990), υποθέτουμε μία ευρύτερη παραμετρική οικογένεια συναρτήσεων κινδύνου διάστασης $(p + 1)$, της μορφής $\lambda(\cdot; \theta, \eta)$, όπου η τιμή $\eta = \eta_0$ αντιστοιχεί στην απλούστερη οικογένεια $\lambda(\cdot; \theta)$. Οι εναλλακτικές θα είναι της μορφής

$$\lambda_n(x) = \lambda(x; \theta_0, \eta_0 + \gamma/\sqrt{n}) = \lambda_0(x; \theta_0)[1 + \phi_0(x; \theta_0)\gamma/\sqrt{n}], \quad x \in \mathbb{R},\ \theta_0 \in \Theta, \tag{3}$$

όπου η $\lambda_0(x; \theta_0)$ συμβολίζει τη συνάρτηση κινδύνου υπό τη μηδενική υπόθεση, η $\gamma \geq 0$ είναι μια σταθερά και η $\phi_0(x; \theta_0)$ εκφράζει την παράγωγο της $\log \lambda_0(x; \theta_0, \eta)$ ως προς την παράμετρο η, υπολογισμένη στο σημείο $\eta = \eta_0$.

Στην παρούσα μελέτη υποθέτουμε ότι η άγνωστη παράμετρος θ εκτιμάται με τη μέθοδο της μέγιστης πιθανοφάνειας. Έστω $\hat{\theta}$ ο εκτιμητής μέγιστης πιθανοφάνειας

της θ και $\lambda(x;\hat{\theta})$ ο αντίστοιχος εκτιμητής της $\lambda(x)$. Τότε, ένας έλεγχος καλής προσαρμογής για το πρόβλημα του ελέγχου υποθέσεων της (2) μπορεί να βασιστεί στην L_2 απόσταση μεταξύ της $\lambda(x)$ και της $\lambda(x;\hat{\theta})$, δηλαδή στην

$$T_n(\hat{\theta}) = \int \left\{\lambda(x) - \lambda(x;\hat{\theta})\right\}^2 w(x)\, dx, \tag{4}$$

όπου $w(x)$ είναι μια συνάρτηση βάρους. Παρατηρούμε ότι $T_n(\hat{\theta}) \geq 0$, με την ισότητα να ισχύει αν και μόνο αν η μηδενική υπόθεση H_0 είναι αληθής. Υπό την εναλλακτική υπόθεση H_1, η τιμή του $T_n(\hat{\theta})$ αναμένεται να αποκλίνει από το μηδέν και να αυξάνεται καθώς η απόκλιση από τη μηδενική αυξάνεται. Ωστόσο, η ποσότητα (4) περιλαμβάνει την πραγματική συνάρτηση κινδύνου $\lambda(x)$ και επομένως δεν είναι εφαρμόσιμη στην πράξη. Για τον λόγο αυτόν, η $\lambda(x)$ αντικαθίσταται από την $\hat{\lambda}(x;h)$. Έτσι, προκύπτει η πρακτικά χρήσιμη έκδοση της $T_n(\hat{\theta})$, η οποία συμβολίζεται με $\hat{T}_n(h;\hat{\theta})$ και δίνεται από τη σχέση

$$\hat{T}_n(h;\hat{\theta}) = \int \left\{\hat{\lambda}(x;h) - \lambda(x;\hat{\theta})\right\}^2 w(x)\, dx. \tag{5}$$

Πριν προχωρήσουμε στη μελέτη των ασυμπτωτικών ιδιοτήτων της $\hat{T}_n(h;\hat{\theta})$, δίνουμε τις παρακάτω υποθέσεις οι οποίες θα χρησιμοποιηθούν σε όλη την εργασία.

Έστω $\Lambda(x) = \int_0^x \lambda(t)\, dt$ η αθροιστική συνάρτηση κινδύνου της $\lambda(x)$ και έστω $\mu_i(K)$ η i-οστή ροπή του πυρήνα K, για $i = 0, 1, 2$. Επιπλέον, έστω $R(K)$ το ολοκλήρωμα της πραγματικής συνάρτησης K^2 επί του πεδίου ορισμού της.

- A.1. (a) Η $\lambda(x)$ είναι δύο φορές παραγωγίσιμη ως προς x και η $\lambda''(x)$ είναι φραγμένη και ομοιόμορφα συνεχής.
  (b) Οι $\lambda(x;\theta)$ και $\Lambda(x;\theta)$ είναι δύο φορές παραγωγίσιμες ως προς θ με τις $\lambda''(x;\theta)$ και $\Lambda''(x;\theta)$ να είναι φραγμένες, ομοιόμορφα συνεχείς και τετραγωνικά ολοκληρώσιμες.
- A.2. Για $l = 0, 1, 2$, η i-οστή παράγωγος του K, $K^{(l)}$, είναι φραγμένη και απόλυτα ολοκληρώσιμη με πεπερασμένες δεύτερες ροπές.
- A.3. $R(K) < +\infty$ και $\mu_0(K) = 1, \mu_1(K) = 0, \mu_2(K) < +\infty$, δηλαδή ο πυρήνας K είναι τάξης 2.
- A.4. Υποθέτουμε ότι $h \to 0$ και $nh \to \infty$ καθώς $n \to \infty$.
- A.5. Υπό τυχαία δεξιά λογοκρισία, ο λογάριθμος της συνάρτησης πιθανοφάνειας $l_n(x_i;\theta,\delta_i)$ διαθέτει δεύτερη παράγωγο, φραγμένη από μία πεπερασμένη συνάρτηση σε μία μικρή περιοχή γύρω από το θ. Δηλαδή για οποιοδήποτε $\theta \subset \Theta$ υπάρχει μία περιοχή $U(\theta,\nu) \subset \Theta$ και μία μετρήσιμη συνάρτηση $M(x_i;\theta,\delta_i)$ με $\mathbb{E}M(x_i;\theta,\delta_i) < \infty$ τέτοια ώστε

$$\left|\frac{\partial^2}{\partial\theta_i\partial\theta_j} l_n(x_i;\theta',\delta_i)\right| \leq M(x_i;\theta,\delta_i),$$

για κάθε $\theta' \in U(\theta,\nu)$.

A.6. Υπό τυχαία δεξιά λογοκρισία, η ορίζουσα του πίνακα πληροφορίας του Fischer $\mathcal{I}(\theta) = [\mathcal{I}(\theta)]_{i,j}$, $i, j = 1, \ldots, k$, (βλέπε Efron and Johnstone (1990)) είναι διάφορη του μηδενός για όλα τα $\theta \in \Theta$.

A.7. Υπό την H_0, $\left\|\hat{\theta} - \theta_0\right\| = O_p\left(n^{-1/2}\right)$, βλέπε ενότητα 3.4 Kalbfleisch and Prentice (2002).

3. ΚΥΡΙΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

Στην παρούσα ενότητα μελετάμε την κατανομή της στατιστικής συνάρτησης υπό την μηδενική υπόθεση. Αρχικά θέτουμε

$$k = \frac{\mu_2(K)}{2},\ B(u) = \int K(t)K(u+t)\,dt,\ \sigma_1^2 = R\left\{\frac{\lambda_T(x)w(x)}{1-G(x)}\right\} R\left\{B(u)\right\},$$

$$\sigma_4^2 = \left\{\int \frac{\lambda_T(x)}{1-G(x)}\left[\lambda''(x;\theta_0)\,w(x)\right]^2 dx\right\} - \left\{\int \lambda''(x;\theta_0)\lambda_T(x)w(x)\,dx\right\}^2,$$

όπου θ_0 είναι η πραγματική τιμή της θ υπό τη μηδενική υπόθεση. Επιπλέον, ακολουθώντας την τεχνική του Hall (1984), θέτουμε

$$d(n) = \begin{cases} nh^{1/2}, & \text{αν } nh^5 \to 0, \\ n^{1/2}h^{-2}, & \text{αν } nh^5 \to \infty, \\ n^{9/10}, & \text{αν } nh^5 \to \xi > 0, \end{cases}$$

μία ακολουθία θετικών σταθερών που αποκλίνουν στο άπειρο. Η περίπτωση $nh^5 \to 0$ αντιστοιχεί σε υποεξομάλυνση των δεδομένων, η περίπτωση $nh^5 \to \infty$ αντιστοιχεί σε υπερεξομάλυνση των δεδομένων και η περίπτωση $nh^5 \to \xi > 0$ αντιστοιχεί στο βέλτιστο ποσό εξομάλυνσης που εφαρμόζεται στα δεδομένα με την έννοια ότι το συγκεκριμένο h ελαχιστοποιεί το μέσο ολοκληρωμένο τετραγωνικό σφάλμα του εκτιμητή $\hat{\lambda}(x;h)$. Έστω Z μία ασυμπτωτικά τυπική κανονική τυχαία μεταβλητή. Η ασυμπτωτική κατανομή της $\hat{T}_n(h;\hat{\theta})$ δίνεται στο παρακάτω θεώρημα.

Θεώρημα 1. *Υπό τις υποθέσεις A.1-A.7 και την H_0, ισχύει ότι*

$$d(n)\left[\hat{T}_n(h;\hat{\theta}) - \mathbb{E}\left\{\hat{T}_n(h;\hat{\theta})\right\}\right] = \begin{cases} 2^{1/2}\sigma_1 Z, & \text{αν } nh^5 \to 0, \\ 2k\sigma_4 Z, & \text{αν } nh^5 \to \infty, \\ \left(2\sigma_1^2\xi^{-1/5} + 4k^2\sigma_4^2\xi^{4/5}\right)^{1/2} Z, & \text{αν } nh^5 \to \xi, \end{cases}$$

για $\xi \in (0, \infty)$.

Απόδειξη. Αρχικά,

$$\begin{aligned}\hat{T}_n(h;\hat{\theta}) &= \int \left\{\hat{\lambda}(x;h) - \lambda(x;\hat{\theta})\right\}^2 w(x)\,dx \\ &= \int \left\{\hat{\lambda}(x;h) - \lambda(x;\theta_0) + \lambda(x;\theta_0) - \lambda(x;\hat{\theta})\right\}^2 w(x)\,dx \\ &= \int \left\{\hat{\lambda}(x;h) - \lambda(x;\theta_0)\right\}^2 w(x)\,dx + \int \left\{\lambda(x;\hat{\theta}) - \lambda(x;\theta_0)\right\}^2 w(x)\,dx \\ &\quad - 2\int \left\{\hat{\lambda}(x;h) - \lambda(x;\theta_0)\right\}\left\{\lambda(x;\hat{\theta}) - \lambda(x;\theta_0)\right\} w(x)\,dx \\ &\equiv \hat{T}_1(h,\theta_0) + \hat{T}_2(\hat{\theta},\theta_0) - 2\hat{T}_3(h;\hat{\theta},\theta_0).\end{aligned}$$

Υπό τις υποθέσεις A.1-A.7 είναι εύκολο να αποδειχθεί ότι

$$\hat{T}_2(\hat{\theta},\theta_0) = O_p(n^{-1}) \quad \text{και} \quad \hat{T}_3(h;\hat{\theta},\theta_0) = O_p(h^2 n^{-1/2} + n^{-1}h^{-1/2}).$$

Κατά συνέπεια προκύπτει το εξής

$$\hat{T}_n(h;\hat{\theta}) = \hat{T}_1(h,\theta_0) + O_p(n^{-1}) + O_p(h^2 n^{-1/2} + n^{-1}h^{-1/2}). \tag{6}$$

Επομένως η ασυμπτωτική κατανομή του $\hat{T}_n(h;\hat{\theta})$ είναι ισοδύναμη με την ασυμπτωτική κατανομή του $\hat{T}_1(h,\theta_0)$. Η μελέτη της κατανομής του $\hat{T}_1(h,\theta_0)$, ξεκινάει με τη διάσπαση του όρου

$$\begin{aligned}\hat{T}_1(h,\theta_0) &= \int \left\{\hat{\lambda}(x;h) - \lambda(x;\theta_0)\right\}^2 w(x)\,dx \\ &= \int \left\{\hat{\lambda}(x;h) - \mathbb{E}\{\hat{\lambda}(x;h)\} + \mathbb{E}\{\hat{\lambda}(x;h)\} - \lambda(x;\theta_0)\right\}^2 w(x)\,dx \\ &= \int \left[\hat{\lambda}(x;h) - \mathbb{E}\{\hat{\lambda}(x;h)\}\right]^2 w(x)\,dx + \int \left[\mathbb{E}\{\hat{\lambda}(x;h)\} - \lambda(x;\theta_0)\right]^2 w(x)\,dx \\ &\quad + 2\int \left[\hat{\lambda}(x;h) - \mathbb{E}\{\hat{\lambda}(x;h)\}\right]\left[\mathbb{E}\{\hat{\lambda}(x;h)\} - \lambda(x;\theta_0)\right] w(x)\,dx.\end{aligned}$$

Έστω τώρα ο εκτιμητής $\tilde{\lambda}(x;h)$, ο οποίος ορίζεται ως

$$\tilde{\lambda}(x;h) = \frac{1}{nh}\sum_{i=1}^{n} \frac{\delta_i}{1 - G(X_i)} K\left(\frac{x - X_i}{h}\right).$$

Δεδομένου ότι προκύπτει ότι $\hat{\lambda}(x;h) = \tilde{\lambda}(x;h)\{1 + O_p(n^{-1/2})\}$, έπεται ότι είναι ασυμπτωτικά ισοδύναμο να αποδείξουμε το θεώρημα αντικαθιστώντας το $\hat{T}_1(h,\theta_0)$

με το

$$\tilde{T_1}(h;\theta_0) = \int \left\{\tilde{\lambda}(x;h) - \mathbb{E}\{\tilde{\lambda}(x;h)\}\right\}^2 w(x)\,dx + \int \left\{\mathbb{E}\{\tilde{\lambda}(x;h)\} - \lambda(x;\theta_0)\right\}^2 w(x)\,dx + 2\int \left\{\tilde{\lambda}(x;h) - \mathbb{E}\{\tilde{\lambda}(x;h)\}\right\}\left\{\mathbb{E}\{\tilde{\lambda}(x;h)\} - \lambda(x;\theta_0)\right\} w(x)\,dx.$$

Σύμφωνα με το Θεώρημα 2 στην εργασία του Bagkavos (2011) και δεδομένου ότι ο εκτιμητής $\hat{\lambda}_L(x)$ που μελετάται εκεί είναι πανομοιότυπος με αυτόν που χρησιμοποιείται στην παρούσα έρευνα έχουμε ότι

$$\mathbb{E}\{\hat{\lambda}(x;h)\} - \lambda(x;\theta_0) = \frac{h^2}{2}\lambda''(x;\theta_0)\mu_2(K) + o(h^2), \tag{7}$$

και κατά συνέπεια για βέλτιστο εύρος ζώνης $h \sim n^{-1/5}$ (βλέπε Watson and Leadbetter (1964)) ότι

$$\mathbb{E}\{\tilde{\lambda}(x;h)\} - \lambda(x;\theta_0) = \frac{h^2}{2}\lambda''(x;\theta_0)\mu_2(K) + o(h^2). \tag{8}$$

Από την παραπάνω εξίσωση, το $\tilde{T_1}(h;\theta_0)$ γίνεται

$$\tilde{T_1}(h;\theta_0) = \int \left\{\tilde{\lambda}(x;h) - \mathbb{E}\{\tilde{\lambda}(x;h)\}\right\}^2 w(x)\,dx + 2\int \left\{\tilde{\lambda}(x;h) - \mathbb{E}\{\tilde{\lambda}(x;h)\}\right\}\left\{\mathbb{E}\{\tilde{\lambda}(x;h)\} - \lambda(x;\theta_0)\right\} w(x)\,dx + O(h^4).$$

Έστω $Y_i = (X_i, \delta_i)$, $i = 1, \ldots, n$ το δείγμα που παρατηρείται και επιπλέον έστω

$$z_i(x, Y_i) = \frac{\delta_i}{1 - G(X_i)} K\left(\frac{x - X_i}{h}\right),\ R_i(x, Y_i) = z_i(x, Y_i) - \mathbb{E}\{z_i(x, Y_i)\},$$
$$H_n(Y_i, Y_j) = \int R_i(x, Y_i) R_j(x, Y_j) w(x)\,dx.$$

Τότε,

$$\int \left[\tilde{\lambda}(x;h) - \mathbb{E}\left\{\tilde{\lambda}(x;h)\right\}\right]^2 = (nh)^{-2}\sum_{i=1}^{n}\sum_{i\neq j}^{n} H_n(Y_i, Y_j) + (nh)^{-2}\sum_{i=1}^{n} H_n(Y_i, Y_i).$$

Επομένως το $\tilde{T}_1(h,\theta_0)$ γίνεται

$$\begin{aligned}\tilde{T}_1(h,\theta_0) &= (nh)^{-2}\sum_{i=1}^{n}\sum_{i\neq j}^{n} H_n(Y_i,Y_j) + (nh)^{-2}\sum_{i=1}^{n} H_n(Y_i,Y_i) + O(h^4)\\ &\quad + 2\int \left\{\tilde{\lambda}(x;h) - \mathbb{E}\{\tilde{\lambda}(x;h)\}\right\}\left\{\mathbb{E}\{\tilde{\lambda}(x;h)\} - \lambda(x;\theta_0)\right\} w(x)\,dx\\ &\equiv \tilde{T}_{1.1}(h) + \tilde{T}_{1.2}(h) + 2\tilde{T}_{1.3}(h,\theta_0) + O(h^4). \qquad (9)\end{aligned}$$

Εφαρμόζοντας τα Λήμματα 2, 3 και 5 στην (9) και κάνοντας χρήση του Λήμματος 4 λαμβάνουμε ότι

$$\hat{T}_1(h,\theta_0) - \mathbb{E}\left\{\hat{T}_1(h,\theta_0)\right\} = 2^{1/2}n^{-1}h^{-1/2}\sigma_1 Z + 2n^{-1/2}h^2 k\sigma_4 Z + O_p(h^4 + n^{-3/2}h^{-1}).$$

Αντικαθιστώντας το παραπάνω στην

$$\begin{aligned}\hat{T}_n(h;\hat{\theta}) - \mathbb{E}\left\{\hat{T}_n(h;\hat{\theta})\right\} &= \hat{T}_1(h,\theta_0) - \mathbb{E}\left\{\hat{T}_1(h,\theta_0)\right\}\\ &\quad + O_p(n^{-1}) + O_p\left(h^2 n^{-1/2} + n^{-1}h^{-1/2}\right),\end{aligned}$$

και για βέλτιστο εύρος ζώνης $h \sim n^{-1/5}$ ολοκληρώνεται η απόδειξη. □

Από το Θεώρημα 1 έχουμε ότι ο ρυθμός σύγκλισης της $\hat{T}_n(h;\hat{\theta})$ στην κανονική κατανομή εξαρτάται από το ποσό εξομάλυνσης που εφαρμόζεται στα δεδομένα.

Στη συνέχεια, το Θεώρημα 1 εφαρμόζεται για τη δημιουργία ενός ελέγχου καλής προσαρμογής. Σύμφωνα λοιπόν με το Θεώρημα 1 και θέτοντας

$$\hat{T}_n^*(h;\hat{\theta}) = d(n)\left[\hat{T}_n(h;\hat{\theta}) - \mathbb{E}\left\{\hat{T}_n(h;\hat{\theta})\right\}\right], \qquad (10)$$

για τον έλεγχο της H_0 έναντι της H_1 με επίπεδο σημαντικότητας α, ισχύει ότι

$$\hat{T}_n^*(h;\hat{\theta}) \Big/ \sqrt{\text{Var}\left\{\hat{T}_n^*(h;\hat{\theta})\right\}} \longrightarrow N(0,1),$$

όπου

$$\text{Var}\left\{\hat{T}_n^*(h;\hat{\theta})\right\} = \begin{cases} 2\sigma_1^2, & \text{αν } nh^5 \to 0,\\ 4k^2\sigma_4^2, & \text{αν } nh^5 \to \infty,\\ 2\sigma_1^2\xi^{-1/5} + 4k^2\sigma_4^2\xi^{4/5} & \text{αν } nh^5 \to \xi. \end{cases}$$

Κατά συνέπεια ο έλεγχος υποδηλώνει απόρριψη της H_0 όταν

$$\hat{T}_n^*(h;\hat{\theta}) \Big/ \sqrt{\text{Var}\left\{\hat{T}_n^*(h;\hat{\theta})\right\}} > z_\alpha,$$

όπου z_α είναι το (άνω) α-ποσοστιαίο σημείο της τυπικής κανονικής κατανομής και α είναι το επίπεδο σημαντικότητας. Η αριθμητική υλοποίηση του παραπάνω ελέγχου περιλαμβάνεται στην ενότητα που ακολουθεί.

4. ΑΡΙΘΜΗΤΙΚΑ ΠΑΡΑΔΕΙΓΜΑΤΑ

Η παρούσα ενότητα επικεντρώνεται στη μελέτη της συνάρτησης ισχύος του προτεινόμενου ελέγχου καλής προσαρμογής.

Αρχικά, χρησιμοποιώντας την συνάρτηση πυκνότητας πιθανότητας του δείγματος ως συνάρτηση βάρους στην $\hat{T}_n(h;\hat{\theta})$, προκύπτει η δειγματική της εκδοχή η οποία συμβολίζεται με $\tilde{T}_n(h;\hat{\theta})$ και δίνεται από τη σχέση

$$\tilde{T}_n(h;\hat{\theta}) = \int \left\{\hat{\lambda}\left(x;h\right) - \lambda(x,\hat{\theta})\right\}^2 g(x)\,dx = \mathbb{E}\left\{\left[\hat{\lambda}\left(X;h\right) - \lambda(X,\hat{\theta})\right]^2\right\}$$
$$\approx \frac{1}{n}\sum_{i=1}^{n}\left\{\hat{\lambda}\left(X_i;h\right) - \lambda(X_i,\hat{\theta})\right\}^2. \tag{11}$$

Στη συνέχεια, αντικαθιστώντας την (11) στην (10), προκύπτει επίσης η δειγματική της εκδοχή που συμβολίζεται με $\tilde{T}_n^*(h;\hat{\theta})$ και δίνεται από τη σχέση

$$\tilde{T}_n^*(h;\hat{\theta}) = d(n)\left[\tilde{T}_n(h;\hat{\theta}) - \mathbb{E}\left\{\tilde{T}_n(h;\hat{\theta})\right\}\right].$$

Για τον εκτιμητή $\hat{\lambda}(x;h)$ που δίνεται από την (1) χρησιμοποιείται ο πυρήνας της τυπικής κανονικής κατανομής (Gaussian kernel). Το εύρος ζώνης λαμβάνεται ως $h = b^* \cdot \hat{h}_{\text{opt}}$, όπου $\hat{h}_{\text{opt}}$ αποτελεί τη δειγματική εκδοχή του βέλτιστου εύρους ζώνης ως προς το μέσο ολοκληρωμένο τετραγωνικό σφάλμα (MISE), βασισμένο στη μέθοδο plug-in, όπως περιγράφεται στην εξίσωση (18), σελ. 51 της εργασίας των Hua et al. (2018). Ο συντελεστής b^*, αποτελεί μία σταθερά και επιλέγεται από τον χρήστη. Η λογική πίσω από αυτή την επιλογή έγκειται στο γεγονός ότι το $\hat{h}_{\text{opt}}$ είναι βέλτιστο για τον εκτιμητή $\hat{\lambda}(x;h)$, όχι όμως απαραίτητα για το ολοκληρωμένο τετραγωνικό σφάλμα (ISE) της στατιστικής συνάρτησης. Ο συντελεστής b^* ρυθμίζει το επίπεδο εξομάλυνσης ώστε η στατιστική συνάρτηση να συγκλίνει ταχύτερα στην κανονική κατανομή υπό την H_0. Οι συναρτήσεις $\hat{\lambda}(x;h)$ και $\hat{h}_{\text{opt}}$ είναι διαθέσιμες στην R μέσω του πακέτου `NPHazardRate`.

Στη συνέχεια, για τον υπολογισμό των κρίσιμων σημείων (critical points), η κατανομή της στατιστικής συνάρτησης $\tilde{T}_n^*(h;\hat{\theta})$ υπολογίζεται υπό τη μηδενική υπόθεση μέσω 1000 επαναλήψεων προσομοίωσης, για διάφορα μεγέθη δείγματος και ποσοστά λογοκρισίας. Για τον προσδιορισμό του σημείου αποκοπής, εφαρμόζεται ο ορισμός 7 της εργασίας των Hyndman and Fan (1996), όπως αυτός έχει υλοποιηθεί στην R μέσω της συνάρτησης `quantile()` του πακέτου `stats`, επί των παραγόμενων τιμών της $\tilde{T}_n^*(h;\hat{\theta})$. Η παραπάνω διαδικασία επαναλαμβάνεται 500 φορές, με αποτέλεσμα την παραγωγή 500 σημείων αποκοπής. Η τελική κρίσιμη τιμή προκύπτει ως ο μέσος όρος των 500 αυτών σημείων.

Επιπλέον, όσων αφορά την αριθμητική προσέγγιση της ισχύος του ελέγχου, για κάθε τιμή της παραμέτρου i, παράγονται 100 τιμές της $\tilde{T}_n^*(h;\hat{\theta})$ υπό την εναλλακτική υπόθεση τύπου Pitman που δίνεται από την (3). Οι προσομοιώσεις πραγματοποιούνται

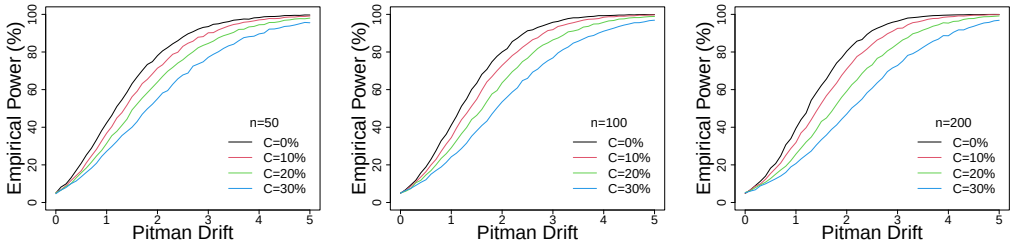
για διάφορα μεγέθη δείγματος και ποσοστά λογοκρισίας. Η ισχύς, στη συνέχεια, εκτιμάται ως το ποσοστό απορρίψεων της μηδενικής υπόθεσης επί των 100 επαναλήψεων και προσεγγίζεται από τον τύπο

$$P\left(\tilde{T}_n^*(h;\hat{\theta}) > \hat{c}_\alpha | H_1\right) \approx \frac{\#\left\{\tilde{T}_n^*(h;\hat{\theta}) > \hat{c}_\alpha\right\}}{100},$$

όπου $\hat{c}_\alpha$ είναι η κρίσιμη τιμή που προέκυψε από το προηγούμενο στάδιο. Η διαδικασία αυτή επαναλαμβάνεται 100 φορές και η τελική καμπύλη ισχύος υπολογίζεται ως ο μέσος όρος των 100 επιμέρους εκτιμήσεων για κάθε τιμή του γ.

Για την αριθμητική αξιολόγηση της προτεινόμενης μεθόδου, εξετάζονται δύο χαρακτηριστικά παραδείγματα όπου η συνάρτηση κινδύνου υπό τη μηδενική υπόθεση προέρχεται από την εκθετική κατανομή με $\lambda_0(x,\theta) = \theta$, όπου $\theta > 0$ και από την Weibull κατανομή με $\lambda_0(x;\theta_1,\theta_2) = \theta_1\theta_2^{-\theta_1}x^{\theta_1-1}$, όπου $\theta_1,\theta_2 > 0$. Για τον συντελεστή b^* επιλέγονται οι τιμές 4 και 8 για την εκθετική και για την Weibull περίπτωση αντίστοιχα. Ως εναλλακτική υπόθεση θεωρούμε την τοπική εναλλακτική τύπου Pitman που δίνεται από την (3) για $\phi_0(x;\theta) = \lambda_0(x;\cdot)$ με την επιλογή αυτήν να διασφαλίζει ότι η εναλλακτική που παράγεται αποτελεί επίσης συνάρτηση κινδύνου. Η τιμή γ η οποία καθορίζει την απόσταση από την συνάρτηση κινδύνου υπό την μηδενική υπόθεση λαμβάνεται ως $\gamma \in \{0, 0.1, 0.2, \ldots, 4\}$ για την περίπτωση της εκθετικής κατανομής και ως $\gamma \in \{0, 0.1, 0.2, \ldots, 1.5\}$ για την περίπτωση της Weibull κατανομής. Προφανώς για $\gamma = 0$ έχουμε τη μηδενική υπόθεση. Οι προσομοιώσεις πραγματοποιούνται για μεγέθη δείγματος $n = 50, 100, 200$ και ποσοστά τυχαίας δεξιάς λογοκρισίας $C = 0\%, 10\%, 20\%, 30\%$. Το επίπεδο σημαντικότητας έχει οριστεί στο 5%. Τα παραγόμενα αποτελέσματα παρουσιάζονται στις εικόνες 1 και 2 παρακάτω.

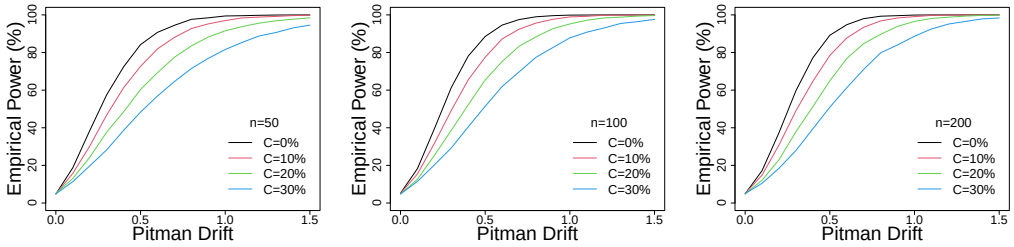
Εικόνα 1: *Εμπειρική ισχύς για την περίπτωση της Εκθετικής κατανομής,* $n = 50, 100, 200$*, λογοκρισία (%)* $= 0, 10, 20, 30$



(a) $n = 50$ ***(b)*** $n = 100$ ***(c)*** $n = 200$

Από τις καμπύλες ισχύος στις εικόνες 1 και 2 παρατηρούμε ότι η ισχύς του ελέγχου αυξάνεται όσο μεγαλώνει η απόκλιση από τη μηδενική υπόθεση, όπως αναμένεται. Επιπλέον, για κάθε επίπεδο λογοκρισίας, ο έλεγχος είναι συνεπής ως προς το ονομαστικό επίπεδο σημαντικότητας. Εξετάζοντας στη συνέχεια την επίδραση του μεγέθους

Εικόνα 2: *Εμπειρική ισχύς για την περίπτωση της κατανομής Weibull,* $n = 50, 100, 200$*, λογοκρισία (%)* $= 0, 10, 20, 30$



(a) $n = 50$ ***(b)*** $n = 100$ ***(c)*** $n = 200$

δείγματος, διαπιστώνεται ότι η εμπειρική ισχύς αυξάνεται σημαντικά με την αύξηση του n. Τέλος, όσον αφορά τα επίπεδα λογοκρισίας, προκύπτει ότι η αύξηση του ποσοστού λογοκρισίας επηρεάζει αρνητικά την ισχύ του ελέγχου. Συγκεκριμένα, για κάθε n, όσο αυξάνεται η λογοκρισία, απαιτείται μεγαλύτερη απόκλιση από την H_0 προκειμένου να ανιχνευθεί H_1.

5. ΣΥΜΠΕΡΑΣΜΑΤΑ

Η παρούσα μελέτη εισάγει έναν νέο μη παραμετρικό έλεγχο καλής προσαρμογής για τη συνάρτηση κινδύνου, βασισμένο στη μέθοδο των πυρήνων, για συνεχή δεδομένα υπό τυχαία δεξιά λογοκρισία. Αποδείχθηκαν οι βασικές ασυμπτωτικές ιδιότητες της στατιστικής συνάρτησης υπό τη μηδενική υπόθεση και μελετήθηκε η επίδοση του ελέγχου μέσω προσομοιώσεων για εναλλακτικές υποθέσεις τύπου Pitman. Τα αποτελέσματα των προσομοιώσεων επιβεβαίωσαν ότι ο έλεγχος είναι συνεπής και παρουσιάζει ικανοποιητική ισχύ ακόμη και σε περιπτώσεις με υψηλά επίπεδα λογοκρισίας.

ABSTRACT

This study introduces a novel kernel-based goodness-of-fit test for the hazard rate function under random right censoring. The test statistic is defined via the L_2 metric, and its asymptotic distribution under the null hypothesis is derived analytically. The power performance of the test is examined against Pitman-type local alternatives via a simulation study based on well-known lifetime distributions, across various sample sizes and amounts of censoring.

ΑΝΑΦΟΡΕΣ

Bagdonavičius, V., and Nikulin, M. (2011). Chi-squared tests for general composite hypotheses from censored samples, *Comptes Rendus Mathematique*, **349**(3–4), 219–223.

Bagkavos, D. (2011). Local linear hazard rate estimation and bandwidth selection, *Annals of the Institute of Statistical Mathematics*, **63**(5), 1019–1046.

Efron, B., and Johnstone, I. M. (1990). Fisher's information in terms of the hazard rate, *The Annals of Statistics*, **18**(1), 38–62.

Hall, P. (1984). Central limit theorem for integrated square error of multivariate nonparametric density estimators, *Journal of Multivariate Analysis*, **14**(1), 1–16.

Hall, P., and Heyde, C. C. (1980). *Martingale Limit Theory and Its Application*, Academic Press.

Hjort, N. L. (1990). Goodness of fit tests in models for life history data based on cumulative hazard rates, *The Annals of Statistics*, **18**(3), 1221–1258.

Hua, H., Bagkavos, D., and Patil, P. N. (2018). An L^1 analysis of a kernel-based hazard rate estimator, *Australian & New Zealand Journal of Statistics*, **60**(1), 43–64.

Huh, Y., and Hutmacher, M. M. (2016). Application of a Hazard-based Visual Predictive Check to Evaluate Parametric Hazard Models, *Journal of Pharmacokinetics and Pharmacodynamics*, **43**, 57–71.

Hyndman, R.J., and Fan, Y. (1996). Sample quantiles in statistical packages, *The American Statistician*, **50**(4), 361–365.

Kalbfleisch, J.D., and Prentice, R.L. (2002). *The Statistical Analysis of Failure Time Data*, John Wiley & Sons.

Μαυρίδης, Π. (2023). *Έλεγχος καλής προσαρμογής για τη συνάρτηση ρυθμού επικινδυνότητας. Μεταπτυχιακή Διατριβή*, Πανεπιστήμιο Ιωαννίνων, Σχολή Θετικών Επιστημών, Τμήμα Μαθηματικών.

Neyman, J. (1937). "Smooth test" for goodness of fit, *Skand. Aktuarietidskr.*, **20**, 149–199.

Peña, E. A. (1998). Smooth goodness-of-fit tests for composite hypothesis in hazard-based models, *The Annals of Statistics*, **26**(5), 1935–1971.

Tanner, M. A., and Wong, W. H. (1983). The estimation of the hazard function from randomly censored data by the kernel method, *The Annals of Statistics*, **11**(3), 989–993.

Vital, R.-A., and Patil, P. (2020). Goodness-of-fit test for hazard rate, *Journal of Nonparametric Statistics*, **32**(2), 403–427.

Vital, R.-A., and Patil, P. N. (2021). Chi-square type goodness-of-fit test for hazard rate, *Statistics in Biosciences*, **13**(1), 77–89.

Watson, G. S., and Leadbetter, M. (1964). Hazard analysis II, *Sankhyā: The Indian Journal of Statistics, Series A*, **26**(1), 101–116.

Α ΒΟΗΘΗΤΙΚΑ ΛΗΜΜΑΤΑ

Στο παρόν Παράρτημα περιλαμβάνονται τα βοηθητικά λήμματα που απαιτούνται για την απόδειξη του Θεωρήματος 1.

Λήμμα 1. *Για i, j σταθερά με $i \neq j$ ισχύει ότι*

$$\mathbb{E}\left\{H_n^2\left(Y_i, Y_j\right)\right\} = h^3 R\left\{\lambda_T(x) w(x)\left(1 - G(x)\right)^{-1}\right\} R\left\{B(u)\right\} + O\left(h^4\right), \quad (12)$$

$$\mathbb{E}\left\{H_n^4\left(Y_i, Y_j\right)\right\} = O\left(h^5\right). \quad (13)$$

Απόδειξη. Η απόδειξη βασίζεται σε αναγωγή υπό συνθήκη ως προς τον αριθμό των μη λογοκριμένων παρατηρήσεων στο δείγμα $Y_i = (X_i, \delta_i)$, $i = 1, \ldots, n$. Εάν N είναι ο αριθμός των μη λογοκριμένων παρατηρήσεων τότε $N \sim \text{Binomial}(n, p)$, όπου $p = \int f(x)\left[1 - H(x)\right] dx$. Δεδομένου ότι $N = \nu$ το $(X_i : \delta_i = 1)$ αποτελεί ένα σύνολο ανεξάρτητων και ισόνομων τυχαίων μεταβλητών με πυκνότητα $f(x)\left[1 - H(x)\right]/p$, για $\nu = 1, 2, \ldots, n$. Αρχικά,

$$\mathbb{E}\left\{H_n^2\left(Y_i, Y_j\right)\right\} = \iint \left[\mathbb{E}\left\{R_i\left(x\right) R_j\left(y\right)\right\}\right]^2 w(x) w(y)\, dxdy. \quad (14)$$

Ακολουθώντας την απόδειξη του Λήμματος 4.2.5 (σελ. 94) στην διατριβή του Μαυρίδη (2023) προκύπτει ότι

$$\begin{aligned}\mathbb{E}\left\{R_i\left(x, Y_i\right) R_j\left(y, Y_i\right)\right\} = h \int \frac{\lambda_T\left(x - ht\right)}{1 - G\left(x - ht\right)} K(t) K\left(\frac{y - x + ht}{h}\right) dt \\ - h^2 \left(\int \lambda_T\left(x - ht\right) K(t)\, dt\right) \\ \times \left(\int \lambda_T\left(y - ht\right) K(t)\, dt\right). \quad (15)\end{aligned}$$

Αντικαθιστώντας τη (15) στην (14), λαμβάνουμε

$$\begin{aligned}\mathbb{E}\{H_n^2\left(Y_i, Y_j\right)\} = \iint \Bigg\{ h \int \frac{\lambda_T\left(x - ht\right)}{1 - G\left(x - ht\right)} K(t) K\left(\frac{y - x + ht}{h}\right) dt \\ - h^2 \left(\int \lambda_T\left(x - ht\right) K(t)\, dt\right) \\ \times \left(\int \lambda_T\left(y - ht\right) K(t)\, dt\right) \Bigg\}^2 w(x) w(y)\, dxdy.\end{aligned}$$

Με αλλαγή μεταβλητής $y - x = uh$ και ακολουθώντας την ίδια διαδικασία όπως στον Μαυρίδη (2023) προκύπτει ότι

$$\begin{aligned}&\mathbb{E}\left\{H_n^2\left(Y_i, Y_j\right)\right\} \\ &= h^3 \left\{\int \left(\frac{\lambda_T(x) w(x)}{1 - G(x)}\right)^2 dx\right\} \left\{\int \left(\int K(t) K(u + t)\, dt\right)^2 du\right\} + O(h^4),\end{aligned}$$

και επομένως η απόδειξη της (12) ολοκληρώθηκε. Για την απόδειξη της (13), αρχικά έχουμε

$$\mathbb{E}\left\{H_n^4\left(Y_i,Y_j\right)\right\} = \iiiint \left[\mathbb{E}\left\{R_i\left(x\right)R_i\left(y\right)R_i\left(z\right)R_i\left(u\right)\right\}\right]^2 \times w(x)w(y)w(z)w(u)\,dxdydzdu.$$

Το αποτέλεσμα της παραπάνω ποσότητας αποτελείται από πολλαπλούς όρους, καθένας από τους οποίους είναι τάξης h^5. Ακολουθώντας την ίδια διαδικασία όπως στο Λήμμα 4.2.5 στη διατριβή του Μαυρίδη (2023), αποδεικνύουμε ενδεικτικά ότι ο κυρίαρχος όρος της παραπάνω ποσότητας είναι επίσης τάξης h^5, το οποίο οδηγεί και στην απόδειξη της (13). □

Λήμμα 2. *Καθώς $n \to +\infty$, έχουμε ότι*

$$\tilde{T}_{1.1}(h) \sim N(0, 2n^{-2}h^{-1}\sigma_1^2).$$

Απόδειξη. Η απόδειξη ακολουθεί παρόμοια προσέγγιση με αυτή του Θεωρήματος 1 στην εργασία του Hall (1984). Αρχικά έχουμε ότι

$$H_n\left(Y_i,Y_j\right) = \int R_i\left(x\right)R_j\left(x\right)w(x)\,dx = \int R_j\left(x\right)R_i\left(x\right)w(x)\,dx = H_n\left(Y_j,Y_i\right),$$

και κατά συνέπεια το $\tilde{T}_{1.1}(h)$ είναι ένα συμμετρικό U-στατιστικό. Επομένως

$$\tilde{T}_{1.1}(h) = \frac{1}{n^2h^2}\sum_{i=1}^{n}\sum_{i\neq j}^{n} H_n\left(Y_i,Y_j\right) = \frac{2}{n^2h^2}\sum_{i=2}^{n}\sum_{j=1}^{i-1} H_n\left(Y_i,Y_j\right) = \frac{2}{n^2h^2}\sum_{i=2}^{n} P_i,$$

όπου

$$P_i = \sum_{j=1}^{i-1} H_n\left(Y_i,Y_j\right).$$

Για να προσδιοριστεί η ασυμπτωτική κατανομή του $\tilde{T}_{1.1}(h)$, θα εφαρμοστεί το Πόρισμα 3.1 της εργασίας των Hall and Hayde (1980). Από το γεγονός ότι $\mathbb{E}\left\{H_n\left(Y_i,Y_j\right)H_n\left(Y_i,Y_k\right)\right\} = 0$ και τις υποθέσεις του Θεωρήματος 1 στην εργασία του Hall (1984), είναι εύκολο να δειχθεί ότι η ποσότητα

$$\sum_{i=2}^{k} P_i,\ k = 2,\ldots,n$$

είναι ένα τετραγωνικώς ολοκληρώσημο martingale με διαφορές P_i και μηδενική μέση τιμή. Επιπλέον, στην περίπτωσή μας, η συνθήκη Lindeberg (βλέπε εξίσωση (3.7), Hall and Hayde (1980), σελ. 54) που χρησιμοποιείται στο Πόρισμα 3.1 της εργασίας των Hall and Hayde (1980) αντικαθίσταται από τη συνθήκη Lindeberg (βλέπε εξίσωση (3.6), Hall and Hayde (1980), σελ. 53), η οποία είναι ισοδύναμή της. Επομένως η απόδειξη θα ολοκληρωθεί δείχνοντας ότι

(a) Για κάθε $\varepsilon > 0$, όταν $n \to +\infty$,

$$s_n^{-2} \sum_{i=1}^{n} \mathbb{E}\left\{P_i^2 I\left(|P_i| > \varepsilon s_n\right)\right\} \xrightarrow{p} 0,$$

όπου

$$s_n^2 = \mathbb{E}\left\{U_n^2\right\},$$

(b)

$$s_n^{-2} V_n^2 = \sum_{i=1}^{n} \mathbb{E}\left\{P_i^2 \mid Y_1, Y_2, \ldots, Y_{i-1}\right\} \xrightarrow{p} 1.$$

Είναι προφανές ότι οι (a) και (b) συνεπάγονται ότι

$$\tilde{T}_{1.1}(h)/s_n \sim N(0,1),$$

το οποίο, συνεπώς, υποστηρίζει τη διατύπωση του λήμματος. Ακολουθώντας την απόδειξη του Λήμματος 4.2.6 (σελ. 101) της διατριβής του Μαυρίδη (2023) προκύπτει το αποτέλεσμα. □

Λήμμα 3. *Καθώς $n \to +\infty$, έχουμε ότι*

$$\tilde{T}_{1.2}(h) = \sigma_2^2 + O_p(n^{-3/2}h^{-1})$$

Απόδειξη. Αρχικά παρατηρούμε ότι

$$\tilde{T}_{1.2}(h) = \frac{1}{n^2h^2} \sum_{i=1}^{n} \int \left[z_i(x) - \mathbb{E}\{z_i(x)\}\right]^2 w(x)\, dx \equiv \frac{1}{n^2h^2} \sum_{i=1}^{n} L_i,$$

όπου

$$L_i = \int \left[z_i(x) - \mathbb{E}\{z_i(x)\}\right]^2 w(x)\, dx.$$

Ακολουθώντας την ίδια διαδικασία με την απόδειξη του Λήμματος 4.2.7 (σελ. 109) της διατριβής του Μαυρίδη (2023), προκύπτουν τα εξής

$$\mathbb{E}\{L_i\} = hR(K) \int \frac{\lambda_T(x)}{1 - G(x)} w(x)\, dx - h^2 \left(\iint K(t)K(t+u)\, dtdu\right) \times \int \lambda_T(x)\lambda_T\left(x + hu\right) w(x)\, dx + O\left(h^2\right), \quad (16)$$

και

$$\mathbb{E}\left\{L_i^2\right\} = O\left(h^2\right).$$

Από τις παραπάνω σχέσεις λαμβάνουμε

$$\text{Var}\left\{\tilde{T}_{1.2}(h)\right\} = n^{-3}h^{-4}\left[\mathbb{E}\left\{L_i^2\right\} - \left(\mathbb{E}\{L_i\}\right)^2\right] = O\left(n^{-3}h^{-2}\right). \quad (17)$$

Τέλος, από τη (16), θέτοντας $\sigma_2^2 = \left(nh^2\right)^{-1} \mathbb{E}\{L_i\}$ και χρησιμοποιώντας την (17), ολοκληρώνεται η απόδειξη. □

Λήμμα 4.

$$\mathbb{E}\left\{\tilde{T}_1(h,\theta_0)\right\} = \sigma_2^2 + O_p(n^{-3/2}h^{-1} + h^4).$$

Απόδειξη. Μέσω άμεσου υπολογισμού και χρησιμοποιώντας την (8) έχουμε ότι

$$\begin{aligned}
\mathbb{E}\left\{\tilde{T}_1(h,\theta_0)\right\} &= \mathbb{E}\left\{\int \left[\tilde{\lambda}(x;h) - \lambda(x;\theta_0)\right]^2 w(x)\,dx\right\} \\
&= \mathbb{E}\left\{\int \left[\tilde{\lambda}(x;h) - \mathbb{E}\{\tilde{\lambda}(x;h)\} + \mathbb{E}\{\tilde{\lambda}(x;h)\} - \lambda(x;\theta_0)\right]^2 w(x)\,dx\right\} \\
&= \mathbb{E}\left\{\int \left[\tilde{\lambda}(x;h) - \mathbb{E}\{\tilde{\lambda}(x;h)\}\right]^2 w(x)\,dx\right\} \\
&\quad + h^2\mu_2(K)\mathbb{E}\left\{\int \left[\tilde{\lambda}(x;h) - \mathbb{E}\left\{\tilde{\lambda}(x;h)\right\}\right] \lambda''(x;\theta_0)w(x)\,dx\right\} + O(h^4).
\end{aligned}$$

Παρατηρούμε ότι

$$\int \left[\tilde{\lambda}(x;h) - \mathbb{E}\{\tilde{\lambda}(x;h)\}\right]^2 w(x)\,dx = \tilde{T}_{1.2}(h),$$

και ότι

$$\mathbb{E}\left\{\int \left[\tilde{\lambda}(x;h) - \mathbb{E}\{\tilde{\lambda}(x;h)\}\right] \lambda''(x,\theta_0)w(x)\,dx\right\} = 0.$$

Κατά συνέπεια

$$\mathbb{E}\left\{\tilde{T}_1(h,\theta_0)\right\} = \tilde{T}_{1.2}(h) + O(h^4),$$

και εφαρμόζοντας το Λήμμα 3 το αποτέλεσμα ακολουθεί. □

Λήμμα 5. *Υπό τις υποθέσεις A.1, A.2, A.3 και την $H_0 : \lambda(x;\theta) = \lambda(x;\theta_0)$, ισχύει ότι*

$$\tilde{T}_{1.3}(h,\theta_0) \sim N(0, n^{-1}h^4k^2\sigma_4^2).$$

Απόδειξη. Αρχικά, έστω

$$\begin{aligned}
\tilde{T}_{1.3}(h,\theta_0) &= \int \left[\tilde{\lambda}(x;h) - \mathbb{E}\{\tilde{\lambda}(x;h)\}\right] \left[\mathbb{E}\{\tilde{\lambda}(x;h)\} - \lambda(x;\theta_0)\right] w(x)\,dx \\
&= (nh)^{-1}\sum_{i=1}^{n} Q_i,
\end{aligned}$$

όπου

$$Q_i = \int [z_i(x) - \mathbb{E}\{z_i(x)\}] \left[\mathbb{E}\{\tilde{\lambda}(x;h)\} - \lambda(x;\theta_0)\right] w(x)\,dx.$$

Έστω τώρα $Q_i = D_i - \mathbb{E}\{D_i\}$, με

$$D_i = \int z_i(x) \left[\mathbb{E}\left\{\tilde{\lambda}(x;h)\right\} - \lambda(x;\theta_0)\right] w(x)\,dx.$$

Είναι προφανές ότι $\mathbb{E}\{Q_i\} = 0$. Τώρα, θέτουμε $D_{i,1} = \mathbb{E}\{D_i\}$, $D_{i,2} = \mathbb{E}\left\{D_i^2\right\}$ και παρατηρούμε το εξής

$$\mathbb{E}\left\{Z_i^2\right\} = \mathbb{E}\left\{\left[D_i - \mathbb{E}\{D_i\}\right]^2\right\} = \mathbb{E}\left\{D_i^2\right\} - \left[\mathbb{E}\{D_i\}\right]^2 = D_{i,2} - D_{i,1}^2. \tag{18}$$

Ακολουθώντας την ίδια διαδικασία με την απόδειξη του Λήμματος 4.2.9 (σελ. 116) της διατριβής του Μαυρίδη (2023), λαμβάνουμε ότι

$$D_{i,1} = \frac{h^3}{2}\mu_2(K)\int \lambda''(x;\theta_0)\lambda_T(x)w(x)\,dx + o\left(h^3\right), \tag{19}$$

$$D_{i,2} = \frac{h^6}{4}\mu_2^2(K)\left\{\int \frac{\lambda_T(x)}{1-G(x)}\left[\lambda''(x;\theta_0)\right]^2 w^2(x)\,dx\right\} + o\left(h^6\right). \tag{20}$$

Αντικαθιστώντας τις (19) και (20) στην (18) παίρνουμε ότι

$$\mathbb{E}\left\{Q_i^2\right\} = \frac{h^6}{4}\mu_2^2(K)\sigma_4^2 + o\left(h^6\right).$$

Δουλεύοντας ακριβώς με τον ίδιο τρόπο δείχνουμε ότι $\mathbb{E}\left\{Q_i^4\right\} = O\left(h^{12}\right)$. Θέτοντας

$$s_n^2 = \sum_{i=1}^{n}\mathbb{E}\left\{Q_i^2\right\},$$

και χρησιμοποιώντας την ανισότητα του Chebyshev έχουμε ότι

$$s_n^{-2}\sum_{i=1}^{n}\mathbb{E}\left\{Q_i^2 I\left(|Q_i| > \varepsilon s_n\right)\right\} \leq \varepsilon^{-2}s_n^{-4}\sum_{i=1}^{n}\mathbb{E}\left\{Q_i^4\right\} \to 0,$$

καθώς το $n \to +\infty$. Συνεπώς, το $\tilde{T}_{1.3}(h,\theta_0)$ ακολουθεί την κανονική κατανομή με μέση τιμή μηδέν και διακύμανση

$$\text{Var}\{\tilde{T}_{1.3}(h,\theta_0)\} = n^{-1}h^{-2}\mathbb{E}\left\{Q_i^2\right\} = n^{-1}h^4\left[\frac{\mu_2(K)}{2}\right]^2\sigma_4^2 = n^{-1}h^4k^2\sigma_4^2.$$

□

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 37ου Πανελληνίου & 3ου Διεθνούς Συνεδρίου Στατιστικής (2025), σελ.103-122

ΠΡΟΓΝΩΣΤΙΚΗ ΙΚΑΝΟΤΗΤΑ ΤΕΧΝΙΚΩΝ ΣΥΝΑΛΛΑΚΤΙΚΩΝ ΚΑΝΟΝΩΝ ΣΤΑ ΧΡΗΜΑΤΙΣΤΗΡΙΑ ΑΘΗΝΩΝ ΚΑΙ ΝΕΑΣ ΥΟΡΚΗΣ ΚΑΙ ΑΠΟΤΕΛΕΣΜΑΤΙΚΟΤΗΤΑ ΚΕΦΑΛΑΙΑΓΟΡΩΝ.

Αλέξανδρος Ε. Μηλιώνης [1] , Βασίλειος Θ. Βαρλάγκας [1, 2]

[1] Πανεπιστήμιο Αιγαίου, Τμήμα Στατιστικής και Αναλογιστικών-Χρηματοοικονομικών Μαθηματικών, Σάμος, Ελλάδα
[2] Υπουργείο Εθνικής Άμυνας, Αθήνα, Ελλάδα
{amilionis, v.varlagkas}@aegean.gr

ΠΕΡΙΛΗΨΗ

Η υπόθεση των αποτελεσματικών κεφαλαιαγορών συνιστά θεμελιώδη αρχή της σύγχρονης χρηματοοικονομικής, διαμορφώνοντας τη θεωρητική έρευνα και τις επενδύσεις. Συνήθης τρόπος ελέγχου της είναι η σύγκριση της απόδοσης τεχνικών συναλλακτικών κανόνων με την απόδοση της αγοράς και διακράτησης περιουσιακών στοιχείων. Ωστόσο, σπάνια εξετάζεται αν η απόδοση των τεχνικών κανόνων έχει πραγματική αξία ή οφείλεται στην τύχη, φαινόμενο γνωστό ως data-snooping. Στην παρούσα εργασία εφαρμόζουμε τον απλό κινητό μέσο στα Χρηματιστήρια Αθηνών και Νέας Υόρκης λαμβάνοντας υπόψη όλες τις παραμέτρους του τεχνικού κανόνα και ελέγχουμε τυχόν ύπαρξη φαινομένου data-snooping με μια εναλλακτική μέθοδο τυχαίων συναλλαγών με χρήση προσομοίωσης Monte Carlo. Τα αποτελέσματα τεκμηριώνουν την προβλεπτική ικανότητα του απλού κινητού μέσου για το Χρηματιστήριο Αθηνών, αλλά όχι για το Χρηματιστήριο της Νέας Υόρκης. Ταυτόχρονα, αποκαλύπτονται μεροληψίες στην σύγκριση της απόδοσης ενός τεχνικού κανόνα με την απόδοση της αγοράς και διακράτησης, καθώς το αποτέλεσμα επηρεάζεται από την πορεία του χρηματιστηριακού δείκτη εφαρμογής. Το γεγονός αυτό έχει άμεσες επιπτώσεις τόσο στην αξιολόγηση της προβλεπτικής ικανότητας των τεχνικών κανόνων όσο και στον έλεγχο της υπόθεσης των αποτελεσματικών κεφαλαιαγορών.

Λέξεις Κλειδιά: Τεχνικοί Συναλλακτικοί Κανόνες, Τεχνική Ανάλυση, Data-Snooping, Χρηματιστήριο Αξιών Αθηνών, Αποτελεσματικότητα Κεφαλαιαγορών.

1. ΕΙΣΑΓΩΓΗ

Η υπόθεση των αποτελεσματικών κεφαλαιαγορών είναι ένας εκ των ακρογωνιαίων λίθων της σύγχρονης χρηματοοικονομικής, τόσο για την θεωρητική της αξία όσο και για τις συνέπειες της στις επενδύσεις. Αν και η αποτελεσματικότητα της αγοράς ορίζεται διαφορετικά από διάφορους συγγραφείς ((π.χ. (Black, 1986), (Malkiel, 1992), (Milionis, 2007)), ο ορισμός που οφείλεται στον (Fama, 1970, p. 388) είναι αυτός που έχει καθιερωθεί. Σύμφωνα με αυτόν τον ορισμό, μια αγορά είναι αποτελεσματική εάν «...οι τιμές των τίτλων σε κάθε χρονική στιγμή "αντανακλούν πλήρως" όλες τις διαθέσιμες πληροφορίες». Η κλασσική κατηγοριοποίηση των διαθέσιμων πληροφοριών (Fama, 1970), κατατάσσει την αποτελεσματικότητα σε ασθενούς μορφής, όταν το σύνολο των πληροφοριών περιλαμβάνει τις τιμές του παρελθόντος, σε μέσης ισχύος, όταν το σύνολο των πληροφοριών περιλαμβάνει όλες τις δημόσια διαθέσιμες πληροφορίες, και σε ισχυρής μορφής, όταν το σύνολο των πληροφοριών περιλαμβάνει όλες τις δημόσια ή ιδιωτικά διαθέσιμες πληροφορίες. Σε μία πιο διευρυμένη εκδοχή της ασθενούς ισχύος, για την προβλεψιμότητα των αποδόσεων (Fama, 1991) το διαθέσιμο σύνολο πληροφοριών, εκτός από τις παρελθούσες τιμές, μπορεί επίσης να περιλαμβάνει ειδικά χαρακτηριστικά της επιχείρησης, μακροοικονομικές μεταβλητές ή ακόμη και ημερολογιακές επιδράσεις (Fama, 1991). Σε μια αποτελεσματική αγορά τα αποτελέσματα των ελέγχων για την προβλεψιμότητα των αποδόσεων δεν θα πρέπει να απορρίπτουν τη μηδενική υπόθεση της μη προβλεψιμότητας.

Μέχρι τις αρχές της δεκαετίας του 1990, το γενικό συμπέρασμα που ανέκυπτε από τα αποτελέσματα των περισσότερων εμπειρικών ελέγχων για την αποτελεσματικότητα της αγοράς ήταν ότι, πλην λίγων εξαιρέσεων, η υπόθεση ότι οι κεφαλαιαγορές είναι αποτελεσματικές δεν απορρίπτεται, τουλάχιστον στην ασθενή και μέσης ισχύος μορφή της [(Elton & Gruber, 1995), (Fama, 1970), (Fama, 1991)]. Ωστόσο, σε νεότερες ερευνητικές εργασίες (από τις αρχές έως τα μέσα της δεκαετίας του 1990) η υπόθεση των αποτελεσματικών αγορών, ακόμη και στην ασθενή της μορφή, απορρίφθηκε αρκετά συχνά. Για τον εμπειρικό έλεγχο της αποτελεσματικότητας της αγοράς στην ασθενή μορφή της (WFME) έχει χρησιμοποιηθεί μια σειρά μεθοδολογικών προσεγγίσεων [(Fama, 1970), (Fama, 1991)], οι οποίες μπορούν να ταξινομηθούν σε δύο κύριες κατηγορίες: α) αμιγώς στατιστικο-οικονομετρικοί έλεγχοι, β) έλεγχοι που βασίζονται σε συναλλακτικούς κανόνες της τεχνικής ανάλυσης.

Στην πρώτη κατηγορία, η αποτελεσματικότητα της αγοράς δεν μπορεί να ελεγχθεί αυτοτελώς, αλλά από κοινού με ένα υπόδειγμα αποτίμησης που παράγει τις υπό όρους προσδοκίες των αποδόσεων των περιουσιακών στοιχείων. Ωστόσο, η αξιολόγηση αυτού του ελέγχου προσκρούει σε ένα πρόβλημα κοινής υπόθεσης. Μια ενδεχόμενη διαπίστωση ανωμαλιών στις τιμές ενός περιουσιακού στοιχείου δεν μπορεί να αποδοθεί αποκλειστικά στην αναποτελεσματικότητα της αγοράς, αλλά πάντα θα εγείρονται αμφιβολίες για το αν οι ανωμαλίες οφείλονται, εν όλω ή εν μέρει, σε ένα ακατάλληλο υπόδειγμα αποτίμησης περιουσιακών στοιχείων.

Στη δεύτερη κατηγορία, προκειμένου να ξεπεραστεί το αδιέξοδο του παραπάνω προβλήματος, η απόδοση που προκύπτει με τη χρήση συναλλακτικών κανόνων συγκρίνεται άμεσα με την αντίστοιχη απόδοση της αγοράς και διακράτησης του περιουσιακού στοιχείου για μια προκαθορισμένη χρονική περίοδο. Οι τεχνικοί αναλυτές πιστεύουν ότι οι τεχνικοί κανόνες έχουν προβλεπτική ικανότητα, και αυτό απορρέει από το γεγονός ότι οι κανόνες αυτοί πιθανόν να εμπεριέχουν με εμπειρικό τρόπο μη γραμμικές εξαρτήσεις στις χρονοσειρές των αποδόσεων των αξιογράφων, η ακριβής φύση των οποίων είναι άγνωστη. Ακολουθώντας αυτή την προσέγγιση, η υπόθεση της αποτελεσματικής αγοράς ασθενούς ισχύος εξαρτάται λιγότερο από ένα υπόδειγμα αποτίμησης, καθώς η μόνη υπόθεση που γίνεται είναι ότι οι τιμές ακολουθούν μια διαδικασία submartingale ((δηλαδή, $E(R_{t+1} \mid \Phi_t) \geq 0$ όπου Ε είναι ο τελεστής αναμενόμενης τιμής και $E(R_{t+1} \mid \Phi_t)$ είναι η αναμενόμενη απόδοση τη χρονική στιγμή t+1 με βάση τις διαθέσιμες πληροφορίες μέχρι τη χρονική στιγμή t (Φ_t)). Όπως σημειώνει ο (Fama, 1970), είναι επιθυμητό για πολλούς λόγους να ελέγχεται άμεσα η αποδοτικότητα διαφόρων συναλλακτικών κανόνων.

Λεκτέο ακόμη, ότι πιο πρόσφατα έχουν εμφανιστεί και πολλές μελέτες που χρησιμοποιούν μεθόδους μηχανικής μάθησης για την πρόβλεψη των αποδόσεων (π.χ. Oikonomou et al., 2025).

Αν και οι πρώτες έρευνες σχετικά με τον έλεγχο της αποτελεσματικότητας με τη χρήση τεχνικών συναλλακτικών κανόνων έδειξαν μη απόρριψη της WFME [(Cowles, 1933), (Fama & Blume, 1966)] και οι ακαδημαϊκές μελέτες ήταν λιγοστές, υπήρξε μια αναζωπύρωση της έρευνας μετά την επιδραστική εργασία των (Brock et al., 1992) στην οποία τεκμηριώθηκε η προβλεπτική ικανότητα των τεχνικών συναλλακτικών κανόνων. Μάλιστα, υπάρχουν εμπειρικές ενδείξεις ότι τροποποιημένες εκδόσεις των τυπικών συναλλακτικών κανόνων μπορούν να βελτιώσουν ακόμη περισσότερο την προβλεπτική τους ισχύ, ιδίως σε λιγότερο ανεπτυγμένες αγορές και σε αγορές με χαμηλότερη κεφαλαιοποίηση [(Ma et al., 2021), (Marshall et al., 2017), (Papailias & Thomakos, 2015)].

Μεταξύ των κανόνων της τεχνικής ανάλυσης, οι οποίοι είναι μαθηματικά καλά ορισμένοι κατά την έννοια του (Neftci, 1991), αυτός που χρησιμοποιείται συχνότερα από τους ερευνητές για τον έλεγχο της αποτελεσματικότητας της αγοράς είναι ο κινητός μέσος (ΜΑ). Πράγματι, ο κανόνας ΜΑ έχει χρησιμοποιηθεί εκτενώς από πολλούς ερευνητές και για πολλές κεφαλαιαγορές, αγορές συναλλάγματος και κρυπτονομισμάτων (π.χ. [(Brock et al., 1992), (Corbet et al., 2019), (Hudson et al., 1996), (Kwon & Kish, 2002), (Luukka et al., 2016), (Papailias & Thomakos, 2015)]).

Η πιο κοινή εκδοχή του χρησιμοποιεί δύο κινητούς μέσους με διαφορετικό εύρος έκαστος, οι οποίοι υπολογίζονται από την χρονοσειρά των τιμών ενός αξιογράφου ή ενός δείκτη:

$$MAS_t = \left(\frac{1}{M} \sum_{i=0}^{M-1} \theta_i B^i P_t \right) \quad (1)$$

$$MAL_t = \left(\frac{1}{N}\sum_{i=0}^{N-1}\theta_i B^i P_t\right) with\ N > M, \quad (2)$$

όπου ο MAS_t αντιπροσωπεύει τον βραχύ MA με μήκος M υπολογιζόμενος στο χρόνο t και ο MAL_t αντιπροσωπεύει τον εκτενέστερο MA με μήκος N. P_t είναι η τιμή του αξιογράφου ή του δείκτη στο χρόνο t, θ_i είναι μη χρονικοί παράμετροι και B είναι ο τελεστής χρονικής υστέρησης. Τα σήματα αγοράς δημιουργούνται στους χρόνους τ_j^B, όπου:

$$\tau_j^B \equiv inf\{t: t > \tau_j^B, MAS_t - MAL_t > DP_{t-1}\} \quad (3)$$

Τα σήματα πώλησης δημιουργούνται στους χρόνους τ_j^S, όπου:

$$\tau_j^S \equiv inf\{t: t > \tau_j^S, MAS_t - MAL_t < -DP_{t-1}\} \quad (4)$$

Οι αρχικοί χρόνοι τ_0^B και τ_0^S ορίζονται ίσοι με μηδέν και το D είναι το γνωστό εύρος ζώνης (μια προκαθορισμένη μη μηδενική σταθερά).

Στην παρούσα εργασία χρησιμοποιούμε τον κλασικό απλό κινητό μέσο MA (SMA εφεξής) και εξετάζουμε τις διαδοχικές αποδόσεις χρησιμοποιώντας όλα τα μήκη των κινητών μέσων (εξισώσεις (1) και (2)). Ουσιαστικά υπολογίζουμε την απόδοση του κανόνα για κάθε συνδυασμό ή παράμετρο προκειμένου να ανακύψει η αναμενόμενη απόδοσή του απαλλαγμένη από την μεροληπτική επιλογή ενός συγκεκριμένου συνδυασμού βραχύτερου και εκτενέστερου κινητού μέσου (MAS-MAL). Συνεπώς, λαμβάνοντας υπόψη όλους τους συνδυασμούς του SMA, η μέση απόδοση που ανακύπτει αντιπροσωπεύει την αναμενόμενη απόδοσή του και όχι την βέλτιστη παραμετροποιημένη εκδοχή του με την οποία επιτυγχάνεται η υψηλότερη απόδοση πολύ πιθανόν μόνο στα συγκεκριμένα δεδομένα, όπως συνηθίζεται σε παρόμοιες μελέτες [π.χ., (Brock et al., 1992), (Hudson et al., 1996), (Mills, 1997), (Papailias & Thomakos, 2015), (Luukka et al., 2016), (Miralles-Quirós et al., 2018)]. Με αυτόν τον τρόπο δυνάμεθα να συγκρίνουμε αμερόληπτα τον SMA με την απόδοση της αγοράς και διακράτησης και εντέλει να αξιολογήσουμε την προβλεπτική του ικανότητα. Με τον όρο απόδοση της αγοράς και διακράτησης νοούμε την απόδοση που αποκομίζουμε από την αγορά και διακράτηση ενός περιουσιακού στοιχείου μέχρι τη λήξη μιας προκαθορισμένης χρονικής περιόδου, στην λήξη της οποίας υλοποιείται η πώλησή του. Ο εν λόγω όρος έχει την ίδια έννοια με την εξίσου γνωστή έννοια της παθητικής επενδυτικής στρατηγικής.

Ωστόσο, κατά την αξιολόγηση των μοντέλων πρόβλεψης που προκύπτουν από την μελέτη των δεδομένων πάντα εγείρεται ένα σημαντικό ζήτημα, το οποίο σπάνια διευθετείται, και είναι γνωστό ως 'data-snooping'. Με τον όρο αυτό νοούμε την περίπτωση που η προβλεπτική ικανότητα ενός μοντέλου οφείλεται μόνο στη χρήση των συγκεκριμένων δεδομένων και δεν έχει πραγματική αξία. Δεν είναι λίγοι οι ερευνητές που επισημαίνουν τον κίνδυνο που εμφωλεύει στις αξιοσημείωτες αποδόσεις ενός μοντέλου πρόβλεψης ότι δηλαδή οι αποδόσεις αυτές απλά προέκυψαν κατά τύχη και όχι λόγω ικανότητας του μοντέλου.

Όσον αφορά την ανάδειξη πιθανών μεροληψιών και κίβδηλων αποδόσεων στην τεχνική ανάλυση και ειδικότερα στην απόδοση των τεχνικών συναλλακτικών κανόνων ο (White, 2000) είναι ο πρώτος που κατασκεύασε μια τεκμηριωμένη και εφαρμόσιμη μέθοδο, την οποία ονομάζει Reality Check (RC), για να διακρίνει αν η απόδοση ενός τεχνικού συναλλακτικού κανόνα είναι συγκριτικά ανώτερη από την απόδοση αναφοράς (benchmark). Μάλιστα με τη χρήση του Reality Check τα συμπεράσματα της εργασίας των (Brock et al., 1992) τίθενται υπό αμφισβήτηση. Ωστόσο, και το Reality Check του (White, 2000) έχει δεχθεί κριτική, καθώς απαιτεί την χρήση μιας πληθώρας («σύμπαν» όπως αναφέρει ο ίδιος) τεχνικών συναλλακτικών κανόνων για να υπάρχει επαρκής ποσοτικοποίηση των δεδομένων και να αναδειχθεί μια πιθανή μεροληψία (Sullivan et al., 1999, p. 1649). Περαιτέρω, ο (Hansen, 2005) παρείχε μια βελτίωση του ελέγχου RC, ο οποίος είναι πιο ισχυρός και λιγότερο ευαίσθητος στην ένταξη ακατάλληλων και ασύνδετων εναλλακτικών συναλλακτικών κανόνων βασιζόμενος όμως στο ίδιο πλαίσιο με αυτό του (White, 2000). Συνεπώς, η χρήση χιλιάδων συναλλακτικών κανόνων για την ανάδειξη μιας πιθανής μεροληψίας στην απόδοση του υπό εξέταση συναλλακτικού κανόνα είναι αναπόφευκτη. Ενδεικτικά αναφέρουμε ότι οι (Sullivan et al., 1999, p. 1654) χρησιμοποιούν 7.846 τεχνικούς συναλλακτικούς κανόνες για την επιτυχία του ελέγχου, η επιλογή των οποίων καθίσταται κρίσιμη.

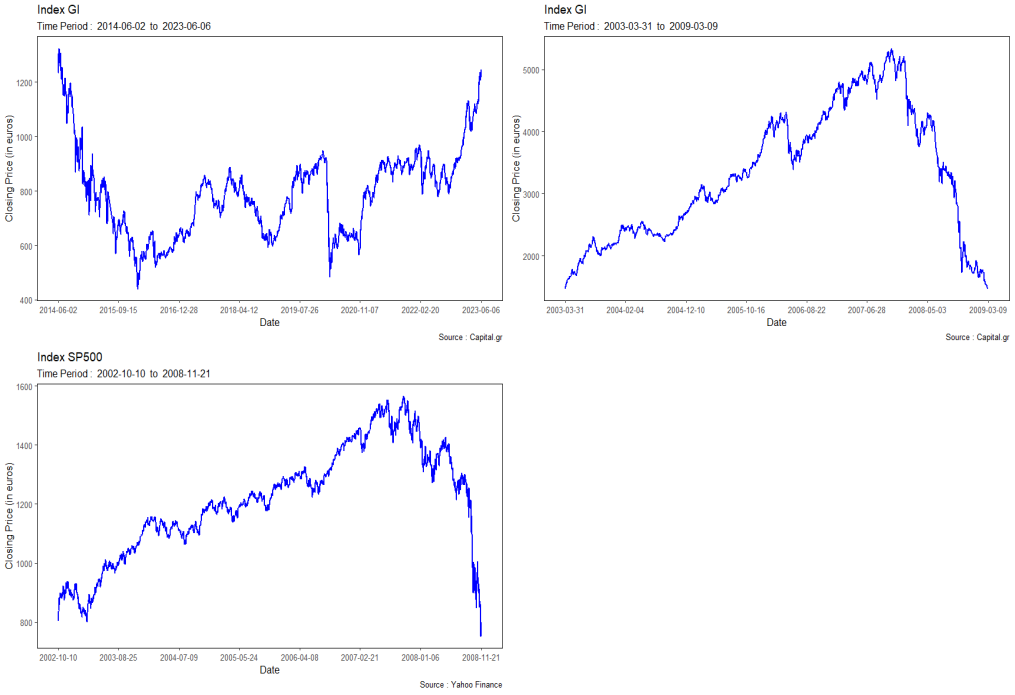
Συμβάλλοντας σε κάποιο βαθμό στην αντιμετώπιση του προβλήματος του data snooping εισάγουμε ένα νέο τρόπο ελέγχου για να διαπιστώσουμε αν η προγνωστική ικανότητα του υπό εξέταση τεχνικού συναλλακτικού κανόνα δύναται να αποδοθεί στην επιλογή της συγκεκριμένης χρονικής περιόδου, δηλαδή στην τύχη. Αρχικά θα αντλήσουμε τα δεδομένα μας από την κεφαλαιαγορά του Χρηματιστηρίου Αξιών Αθηνών (ΧΑΑ εφεξής). Περαιτέρω, θα χρησιμοποιήσουμε και δεδομένα από το Χρηματιστήριο της Νέας Υόρκης. Συγκεκριμένα, αναπαράγουμε σε τυχαίες ημερομηνίες σήματα αγοράς-πώλησης ισάριθμα με αυτά του SMA για κάθε εύρος των MAS-MAL (εξισώσεις (1) και (2)) και δημιουργούμε εμπειρικές κατανομές των αποδόσεών τους με προσομοίωση Monte Carlo (MC εφεξής). Στη συνέχεια συγκρίνουμε την απόδοση του κανόνα με την αντίστοιχη απόδοση από τη μέθοδο MC σε επίπεδο σημαντικότητας 5%. Με αυτόν τον τρόπο ουσιαστικά ελέγχουμε αν η προγνωστική ικανότητα του SMA διαφέρει στατιστικά από ισάριθμες τυχαίες προβλέψεις των τιμών στις υπό εξέταση κεφαλαιαγορές. Τέλος, αξιολογούμε τόσο την απόδοση του SMA με τις αντίστοιχες αποδόσεις τυχαίων συναλλαγών όσο και με την απόδοση της αγοράς και διακράτησης, ώστε να αποκαλυφθούν τυχόν μεροληψίες στην μεταξύ τους σύγκριση.

Η δομή της εργασίας έχει ως εξής: στην ενότητα 2 παραθέτουμε ορισμένες πληροφορίες για τις εξεταζόμενες κεφαλαιαγορές και περιγράφουμε τα δεδομένα· στην ενότητα 3 εξηγούμε τη μεθοδολογική μας προσέγγιση· στην ενότητα 4 παρουσιάζουμε και αναλύουμε τα αποτελέσματά μας· στην ενότητα 5 συνοψίζουμε και παραθέτουμε τα συμπεράσματά μας.

2. ΚΕΦΑΛΑΙΑΓΟΡΑ ΚΑΙ ΔΕΔΟΜΕΝΑ

Τα δεδομένα που χρησιμοποιήθηκαν στην παρούσα μελέτη αποτελούνται από τις ημερήσιες τιμές στο κλείσιμο του Γενικού Δείκτη (εφεξής ΓΔ) του Χρηματιστηρίου Αξιών Αθηνών και του Δείκτη S&P 500 του Χρηματιστηρίου της Νέας Υόρκης (NYSE) (εφεξής SP500). Οι χρονικοί περίοδοι που μελετήθηκαν σε κάθε δείκτη επιλέχθηκαν με βάση τα παρακάτω κριτήρια: (α) Η τιμή στην εκκίνηση και στη λήξη της εκάστοτε περιόδου να βρίσκεται στο ίδιο επίπεδο. (β) Το επίπεδο των τιμών να παρουσιάζει είτε ένα ελάχιστο είτε ένα μέγιστο σημείο, ώστε να δύναται να χωριστεί σε δύο υποπεριόδους, ήτοι μια ανοδική και μια καθοδική. Ειδικότερα, για τον ΓΔ μελετήθηκαν οι χρονικοί περίοδοι 02 Ιουνίου 2014 – 06 Ιουνίου 2023 και 31 Μαρτίου 2003 – 09 Μαρτίου 2009, ενώ για τον SP 500 η περίοδος 10 Οκτωβρίου 2002 – 21 Νοεμβρίου 2008. Σε αυτές τις περιόδους περιλαμβάνονται σημαντικές οικονομικές κρίσεις που αποτυπώθηκαν στο επίπεδο των τιμών τους, όπως η παγκόσμια χρηματοπιστωτική κρίση που πυροδοτήθηκε από την κατάρρευση της Lehman Brothers το 2008, οι περιορισμοί κεφαλαίων του 2015 με την παρατεταμένη αστάθεια της ελληνικής οικονομίας και η παγκόσμια κρίση από την πανδημία του κορονοϊού το 2020. Στην Εικόνα 1 παρουσιάζεται η πορεία του ΓΔ και του SP500 στις υπό εξέταση χρονικές περιόδους.

Εικόνα 1. *Διάγραμμα του Γενικού Δείκτη του Χρηματιστηρίου Αξιών Αθηνών και του S&P 500.*



3. ΜΕΘΟΔΟΛΟΓΙΑ

Όπως αναφέρεται στην εισαγωγή, στις περισσότερες έρευνες, ο στατιστικός έλεγχος της απόδοσης, άρα και της προβλεπτικής ικανότητας του συναλλακτικού κανόνα του κινητού μέσου υλοποιείται με την επιλογή συγκεκριμένων συνδυασμών του μήκους τόσο του βραχύτερου όσο και του εκτενέστερου κινητού μέσου. Δεδομένου ότι η επιλογή των παραπάνω συνδυασμών θεωρείται, τουλάχιστον σε κάποιο βαθμό, αυθαίρετη, οι (Milionis & Papanagiotou, 2008) καταλήγουν στο συμπέρασμα ότι τα ευρήματα για την υπερνίκηση ή μη της αγοράς, χρησιμοποιώντας συγκεκριμένους συνδυασμούς του κανόνα συναλλαγών SMA, δεν παρέχουν ασφαλή συμπεράσματα ούτε για την προβλεπτική ικανότητα του συναλλακτικού κανόνα ούτε για την εγκυρότητα της υπόθεσης του WFME. Επιπροσθέτως, επιλέγοντας συγκεκριμένους συνδυασμούς του MAS-MAL ελλοχεύει πάντοτε ο κίνδυνος οι καλές αποδόσεις του υπό εξέταση κανόνα να είναι προϊόν των επιλεγμένων δεδομένων (data snooping) και όχι της ικανότητας του κανόνα. Προκειμένου να αντιμετωπιστεί ως ένα βαθμό αυτό το πρόβλημα, εξετάζονται όλοι οι συνδυασμοί του βραχέως (MAS) με τον εκτενέστερο (MAL) κινητό μέσο, ακολουθώντας ουσιαστικά την ίδια μεθοδολογία των (Milionis Alexandros E & Varlagkas Vasileios, 2021).

Συγκεκριμένα, ο βραχύς κινητός μέσος (MAS) ορίζεται σταθερός με μήκος ίσο με ένα, ενώ το μήκος του εκτενέστερου κινητού μέσου (MAL) μεταβάλλεται από 5 έως 100 με βήμα τη μονάδα. Στις εξισώσεις υπολογισμού του κινητού μέσου ((βλέπε εξισώσεις (1) έως (4)), θέτουμε όλες τις παραμέτρους θ_i ίσες με τη μονάδα και το εύρος ζώνης D ίσο με το μηδέν. Οι διαδοχικές σωρευτικές ποσοστιαίες αποδόσεις του SMA για όλα τα μήκη του εκτενέστερου κινητού μέσου δημιουργούν μια σειρά αποδόσεων. Σημειώνεται ότι η χρήση λιγότερου ιστορικού τιμών επιτυγχάνεται με βραχέα μήκη MAL.

Στην συνέχεια υπολογίζουμε ομοίως την απόδοση που ανακύπτει από την εκτέλεση τυχαίων ισάριθμων συναλλαγών με τρόπο που να είναι άμεσα συγκρίσιμη με την απόδοση του SMA. Λεπτομερέστερα, ακολουθούμε τα παρακάτω βήματα:

(α) Αρχικά χρησιμοποιούμε ως δεδομένο τον αριθμό των συναλλαγών που σηματοδότησε ο κανόνας SMA για κάθε εύρος του MA.

(β) Επιλέγουμε από την χρονοσειρά των τιμών κλεισίματος του ΓΔ τυχαίες ημερομηνίες, χωρίς αντικατάσταση, ισάριθμες με τον αριθμό των συναλλαγών που σηματοδότησε ο κανόνας, για κάθε εύρος του MAS-MAL.

(γ) Κατατάσσουμε τις τυχαία επιλεγμένες ημερομηνίες σε χρονικά αύξουσα σειρά και αντιστοιχίζουμε σε αυτές διαδοχικά σήματα αγοράς-πώλησης.

(δ) Εκτελούμε τις αντίστοιχες συναλλαγές για να παραχθεί η εκάστοτε σωρευτική απόδοση για κάθε τιμή εύρους του MA. Με αυτόν τον τρόπο δημιουργείται μια σειρά αποδόσεων ίδιου μήκους με αυτή του εξεταζόμενου τεχνικού συναλλακτικού κανόνα, του οποίου ‘αντιγράψαμε’ τον αριθμό των συναλλαγών. Για λόγους ευκολίας ονομάζουμε «SMA-Randomly Signaled» (SMA_{RS} εφεξής) τον συναλλακτικό κανόνα

του οποίου ο αριθμός των τυχαίων συναλλαγών εκπηγάζει από τον τεχνικό συναλλακτικό κανόνα SMA.

(ε) Διενεργούμε προσομοίωση Monte Carlo προκειμένου να προβλέψουμε την αναμενόμενη απόδοσή του SMA_{RS} εντός Διαστήματος Εμπιστοσύνης 95%.

Όσον αφορά τον απαιτούμενο αριθμό προσομοιώσεων για τη διαδικασία Monte Carlo θεωρείται επαρκής ένας αριθμός προσομοιώσεων ώστε να επιτυγχάνεται αφενός σταθεροποίηση, αφετέρου συρρίκνωση του διαστήματος εμπιστοσύνης της μέσης απόδοσης του εκάστοτε κανόνα. Η ικανοποίηση των ανωτέρω συνθηκών κατά την προσομοίωση Monte Carlo απολήγει στην αποδοχή της ευρεθείσας μέσης απόδοσης και την θεώρησή της ως την αναμενόμενη απόδοση από τη χρήση του κανόνα τυχαίων συναλλαγών στην συγκεκριμένη χρονική περίοδο. Έτσι, προκειμένου να βεβαιωθούμε ότι ο αριθμός των εκτελούμενων προσομοιώσεων MC είναι επαρκής, καταγράφουμε την μεταβολή της αναμενόμενης απόδοσης, ως συνάρτηση του αριθμού των προσομοιώσεων, μέχρι να εξασφαλίσουμε ότι η σχετική μεταβολή είναι μικρότερη του 1‰.

Το πρώτο βήμα στην ανάλυση μας είναι να εξετάσουμε το ενδεχόμενο ύπαρξης μοναδιαίας ρίζας στις παραγόμενες «σειρές» των διαδοχικών αποδόσεων, ώστε να εξακριβωθεί αν οι σειρές αυτές είναι στάσιμες. Σε περίπτωση ύπαρξης στασιμότητας η σειρά θα κυμαίνεται έχοντας ως σημείο αναφοράς ένα μέσο επίπεδο, βάσει μιας σταθερής διακύμανσης. Επομένως, μπορεί να χρησιμοποιηθεί το ύψος του μέσου από τον SMA, ως το ύψος της αναμενόμενης απόδοσής του και στη συνέχεια να συγκριθεί με το ύψος της απόδοσης από την στρατηγική της αγοράς και διακράτησης. Προκειμένου αυτή η σύγκριση να καταστεί αξιόπιστη πραγματοποιείται ένας έλεγχος σημαντικότητας σε διάστημα εμπιστοσύνης 95%. (λεπτομέρειες δίνονται στους (Milionis Alexandros E & Varlagkas Vasileios, 2021)). Αναφορικά με τον κανόνα SMA_{RS}, ο έλεγχος στασιμότητας διενεργείται αρχικά στις σειρές των σωρευτικών διαδοχικών αποδόσεων και εφόσον απορριφθεί η μηδενική υπόθεση ύπαρξης μοναδιαίας ρίζας, τότε και μόνο τότε προχωρούμε στην μέθοδο MC. Τέλος, διενεργούμε έλεγχο στασιμότητας στις σειρές των μέσων αποδόσεων που παράχθηκαν από τις επαναλήψεις της προσομοίωσης MC.

Η διενέργεια του ελέγχου για τυχόν ύπαρξη μοναδιαίας ρίζας στις σειρές των αποδόσεων εκτελέστηκε ακολουθώντας ως επί των πλείστων την μεθοδολογία των (Milionis Alexandros E & Varlagkas Vasileios, 2021). Αρχικά, χρησιμοποιήθηκε ο πιο κοινός και δημοφιλής έλεγχος, αυτός του Augmented Dickey-Fuller (ADF), ενώ για να ενδυναμωθεί η αξιοπιστία των αποτελεσμάτων διενεργήθηκε επιπρόσθετα η τροποποίηση του ανωτέρω ελέγχου από τους (Elliott et al., 1996), ο γνωστός έλεγχος ERS. Τέλος, λαμβάνοντας υπόψη ότι αφενός η ισχύς αμφότερων των ελέγχων είναι χαμηλή, αφετέρου βρέθηκε στατιστικά σημαντική αυτοσυσχέτηση στις σωρευτικές αποδόσεις των τεχνικών συναλλακτικών κανόνων διενεργήθηκε επιπρόσθετα ο γνωστός έλεγχος μοναδιαίας ρίζας των (Phillips & Perron, 1988). Ο εν λόγω έλεγχος μοναδιαίας ρίζας αντιμετωπίζει το πρόβλημα της αυτοσυσχέτησης χρησιμοποιώντας τους εκτιμητές των (Newey & West, 1987), οι οποίοι είναι ανθεκτικοί στην

αυτοσυσχέτηση και στην ετεροσκεδαστικότητα, αντί της προσθήκης υστερήσεων στο υπόδειγμα της παλινδρόμησης, με την οποία διευθετείται από την μέθοδο ADF.

Ωστόσο, για τον καθορισμό του διαστήματος εμπιστοσύνης γύρω από την εν λόγω αναμενόμενη απόδοση, στις περιπτώσεις όπου η σειρά των αποδόσεων είναι στάσιμη, απαιτείται ο υπολογισμός της διακύμανσης. Όμως, στην περίπτωση του SMA η διακύμανση δεν μπορεί να εκτιμηθεί με το γνωστό θεώρημα του δειγματικού μέσου για τυχαία δείγματα (Sample Mean Theorem), διότι οι αποδόσεις της σειράς εμφανίζουν μεταξύ τους ισχυρές συσχετίσεις και δεν αποτελούν τυχαίο δείγμα. Προκειμένου να ξεπεραστεί το ανωτέρω πρόβλημα υιοθετήθηκε το επαυξημένο θεώρημα του δειγματικού μέσου (Augmented Sample Mean Theorem, ASMT) για τον υπολογισμό της διακύμανσης, στο οποίο λαμβάνονται υπόψη και οι περιπτώσεις ύπαρξης γραμμικών αλληλεξαρτήσεων μεταξύ των δειγματικών παρατηρήσεων (λεπτομέρειες δίνονται στους (Milionis & Papanagiotou, 2013). Αν οι τυχαίες μεταβλητές X_1, X_2, …, X_N εμφανίζουν γραμμικές αλληλεξαρτήσεις, τότε αποτελούν ένα μη τυχαίο δείγμα μεγέθους N, από ένα πληθυσμό με μέσο μ και διακύμανση σ^2. Επιπρόσθετα, αν ρ_κ με κ = 1, 2, … συμβολίζει την συνάρτηση αυτοσυσχέτισης για τις X_1, X_2, …, X_N, τότε ο δειγματικός μέσος $\bar{X}$ είναι ένας αμερόληπτος εκτιμητής του μέσου μ και η διακύμανση του δειγματικού μέσου δίνεται από τον τύπο:

$$VAR(\bar{X}) = \frac{\sigma^2}{N}\left(1 + 2\frac{(N-1)}{N}\rho_1 + 2\frac{(N-2)}{N}\rho_2 + \cdots + \frac{2}{N}\rho_{N-1}\right) \qquad (5)$$

Σημειώνουμε ότι στις αποδόσεις του SMA_{RS} δεν παρατηρείται το ίδιο φαινόμενο, καθώς οι αποδόσεις βρέθηκαν ασυσχέτιστες αποτελώντας λευκό θόρυβο. Συνεπώς, η διακύμανση των αποδόσεων υπολογίζεται με το γνωστό θεώρημα του δειγματικού μέσου για τυχαία δείγματα.

Δεδομένου ότι οι αποδόσεις που δημιουργούνται από την προσομοίωση MC δεν ακολουθούν πάντοτε την κανονική κατανομή υπολογίζουμε το διάστημα εμπιστοσύνης της αναμενόμενης απόδοσης βρίσκοντας τις κρίσιμες τιμές από την κατανομή των δεδομένων, εν προκειμένω των αποδόσεων. Συγκεκριμένα, για πληθυσμό N και επίπεδο εμπιστοσύνης α η άνω (κάτω) κρίσιμη τιμή είναι η $N\alpha$ μεγαλύτερη (μικρότερη) απόδοση. Τέλος, συγκρίνουμε τις αποδόσεις που ανακύπτουν από τις προσομοιώσεις MC με τις αποδόσεις τόσο του τεχνικού κανόνα SMA, όσο και με την απόδοση της παθητικής επενδυτικής στρατηγικής διενεργώντας μη παραμετρικό στατιστικό έλεγχο σε επίπεδο σημαντικότητας 5%.

Η ανωτέρω μεθοδολογία εφαρμόστηκε στους υπό εξέταση χρηματιστηριακούς δείκτες σε χρονικές περιόδους, σταθερών, ανοδικών και καθοδικών αγορών, ώστε να συγκριθεί η απόδοση ενός εκ των δημοφιλέστερων τεχνικών κανόνων τόσο με την απόδοση τυχαίων ισάριθμων συναλλαγών χωρίς καμία προβλεπτική ικανότητα, όσο και με την απόδοση της αγοράς και διακράτησης. Ο αρχικός στόχος είναι να εξερευνηθεί με μια εναλλακτική μέθοδο εάν η προβλεπτική ικανότητα του κανόνα SMA στους συγκεκριμένους χρηματιστηριακούς Δείκτες και στις υπό εξέταση χρονικές περιόδους έχει πραγματική αξία ή δύναται να αποδοθεί στην τύχη (data

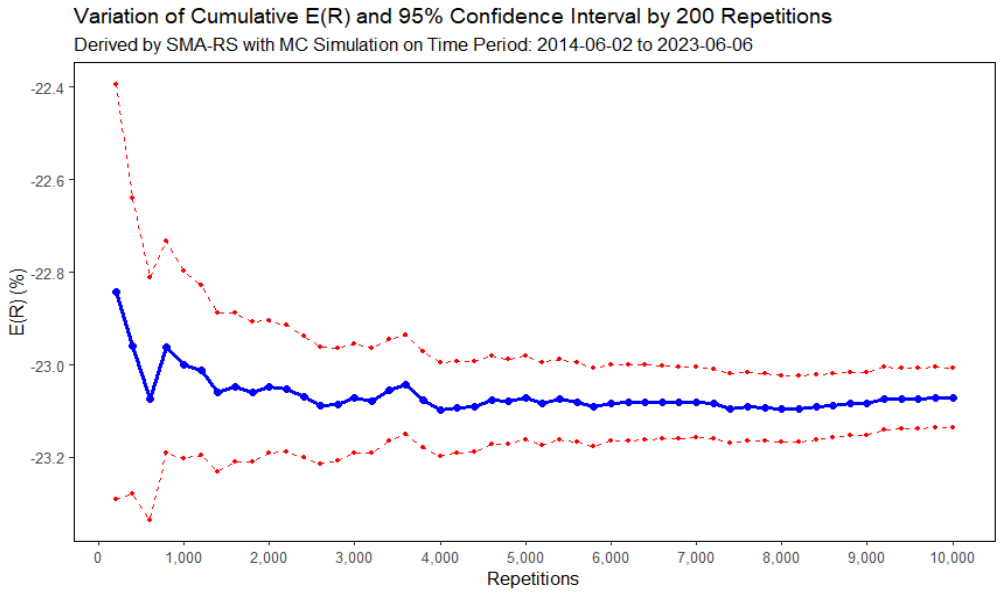
snooping). Ο απώτερος στόχος είναι να αποκαλυφθούν τυχόν μεροληψίες στην επικρατούσα μεθοδολογία αξιολόγησης των τεχνικών συναλλακτικών κανόνων, ήτοι τη σύγκριση της απόδοσής τους με την απόδοση της αγοράς και διακράτησης, καθώς η πορεία της αγοράς στην οποία αυτοί εφαρμόζονται διαδραματίζει σε πολλές περιπτώσεις καθοριστικό ρόλο στο αποτέλεσμα της μεταξύ τους σύγκρισης. Το γεγονός αυτό έχει άμεσες επιπτώσεις τόσο στο εύρημα υπερνίκησης ή μη της αγοράς από έναν τεχνικό συναλλακτικό κανόνα όσο και στην απόρριψη ή μη της υπόθεσης των αποτελεσματικών κεφαλαιαγορών ασθενούς ισχύος (WFME).

4. ΑΠΟΤΕΛΕΣΜΑΤΑ ΚΑΙ ΣΧΟΛΙΑΣΜΟΣ

4.1 Συναλλακτικός Κανόνας Τυχαίας Σηματοδότησης SMA_{RS}

Όπως επεξηγήθηκε στην προηγούμενη ενότητα, αρχικά εξετάζουμε την αναμενόμενη τιμή της απόδοσης του SMA_{RS} σε συνάρτηση με το πλήθος των προσομοιώσεων κατά τη διαδικασία MC. Τα αποτελέσματα φαίνονται στην Εικόνα 2.

Εικόνα 2. *Η σωρευτική μεταβολή της αναμενόμενης απόδοσης του SMA_{RS} και το αντίστοιχο ΔΕ 95% από την εφαρμογή του στον ΓΔ του Χρηματιστηρίου Αξιών Αθηνών.*



Έχοντας ταυτοποιήσει τον απαιτούμενο αριθμό επαναλήψεων (5000) υπολογίζουμε το 95% διάστημα εμπιστοσύνης των αποδόσεων από τυχαίες συναλλαγές. Τα αποτελέσματα των στατιστικών ελέγχων Jarque-Bera δείχνουν ότι οι αποδόσεις των προσομοιώσεων MC δεν ακολουθούν την κανονική κατανομή. Συνεπώς, υπολογίζουμε το 95% ΔΕ βάσει της εμπειρικής κατανομής των αποδόσεων τυχαίων συναλλαγών μη υποθέτοντας κανονικότητα. Επιπλέον, για λόγους συγκρισιμότητας,

υπολογίζουμε και τις κρίσιμες τιμές του ΔΕ 95% υποθέτοντας κανονικότητα. Από τα αποτελέσματα του Πίνακα 1 επιβεβαιώνεται η μικρή απόκλιση των εκάστοτε ορίων του ΔΕ μεταξύ των δύο μεθόδων υπολογισμού.

***Πίνακας 1**. Αποτελέσματα ελέγχου κανονικότητας των αποδόσεων και οι κρίσιμες τιμές του ΔΕ 95% υποθέτοντας και μη κανονικότητα στις 5000 επαναλήψεις προσομοίωσης MC.*

Δείκτης / Περίοδος	Κανονική Κατανομή LL	Κανονική Κατανομή UL	Εμπειρική Κατανομή LL	Εμπειρική Κατανομή UL	Αποτέλεσμα Ελέγχου Jarque-Bera
ΓΔ 2014-23	-2.68	18.00	-2.30	18.59	H0: Rejected $X^2 = 17,18$ (5,99)
ΓΔ 2003-09	-2.14	11.67	-1.86	11.72	H0: Rejected $X^2 = 17,82$ (5,99)
SP500 2002-08	-2.25	6.21	-2.04	6.21	H0: Rejected $X^2 = 10,69$ (5,99)

Σημείωση: Το LL συμβολίζει το κάτω όριο του Διαστήματος Εμπιστοσύνης 95%, το UL το άνω όριο και στο αποτέλεσμα του ελέγχου Jarque-Bera παρέχονται τα αποτελέσματα του ελέγχου και η κρίσιμη τιμή του X^2 με δύο βαθμούς ελευθερίας.

4.2 Έλεγχος Μοναδιαίας Ρίζας

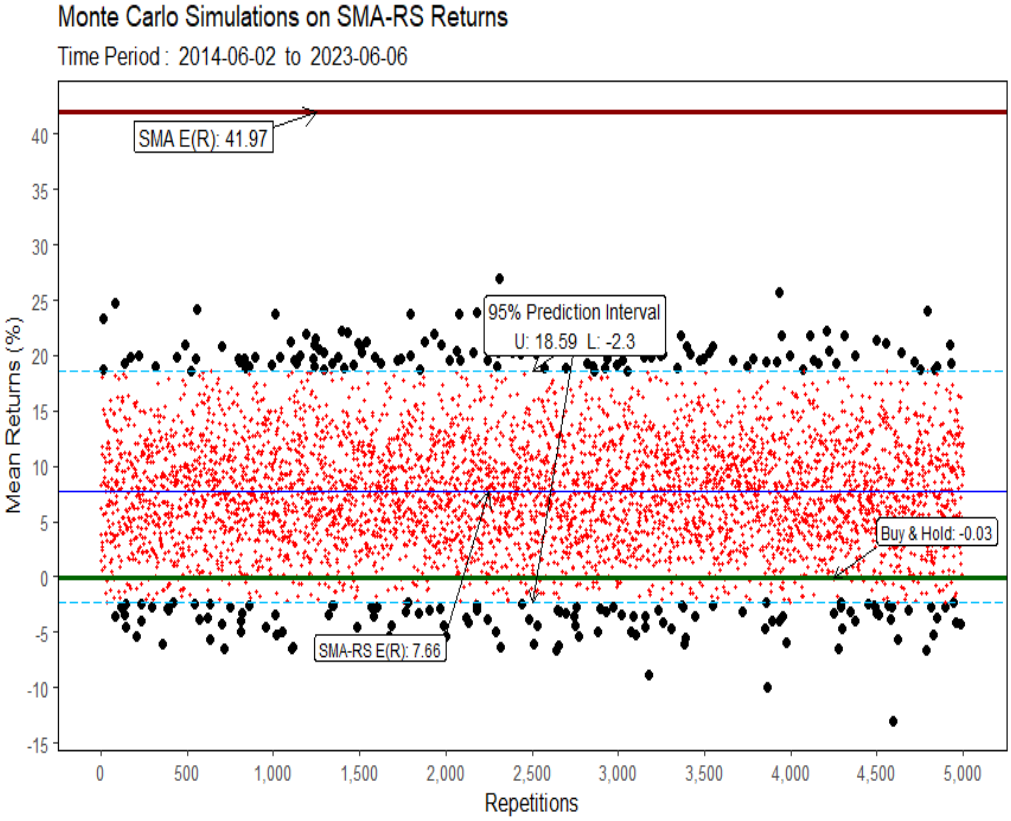
Στον πίνακα 2 παρουσιάζονται τα αποτελέσματα των ελέγχων για την ύπαρξη μοναδιαίας ρίζας στις σειρές τόσο των διαδοχικών σωρευτικών αποδόσεων του SMA όσο και των αποδόσεων του SMA_{RS} από τις προσομοιώσεις MC. Σύμφωνα με τα αποτελέσματα, όλες οι σειρές των αποδόσεων είναι στάσιμες σε διάστημα εμπιστοσύνης 95% πλην δύο περιπτώσεων του SMA, οι οποίες είναι στάσιμες σε διάστημα εμπιστοσύνης 90%.

***Πίνακας 2**. Αποτελέσματα ελέγχου μοναδιαίας ρίζας στις σειρές των αποδόσεων των συναλλακτικών κανόνων.*

Δείκτης	Χρονική Περίοδος	Συναλλακτικός Κανόνας	Αποτέλεσμα Ελέγχου Στασιμότητας
ΓΔ	2014-2023	SMA	Στάσιμη σειρά (PP 95%)
		SMA_{RS}	Στάσιμη σειρά (ADF 95%)
	2003-2009	SMA	Στάσιμη σειρά (ERS 90%)
		SMA_{RS}	Στάσιμη σειρά (ADF 95%)
SP 500	2002-2008	SMA	Στάσιμη σειρά (ADF 90%)
		SMA_{RS}	Στάσιμη σειρά (ADF 95%)

Συνεπώς, καθίσταται εφικτή η σύγκριση μεταξύ της αναμενόμενης απόδοσης του SMA τόσο με το διάστημα εμπιστοσύνης των αποδόσεων τυχαίων συναλλαγών, όσο με την απόδοση της αγοράς και διακράτησης.

Εικόνα 3. *Η αναμενόμενη απόδοση του SMA, η απόδοση της παθητικής επενδυτικής στρατηγικής και οι αποδόσεις τυχαίων συναλλαγών με το αντίστοιχο Διάστημα Εμπιστοσύνης 95% στον ΓΔ του ΧΑΑ την περίοδο 2014 - 2023.*



Σημείωση: Το SMA E(R) συμβολίζει την αναμενόμενη απόδοση του τεχνικού κανόνα SMA, το Buy & Hold την απόδοση της αγοράς και διακράτησης, οι διακεκομμένες μπλε γραμμές το Διάστημα Εμπιστοσύνης 95% των αποδόσεων του κανόνα τυχαίων συναλλαγών SMA_{RS} και η μπλε συνεχόμενη γραμμή την μέση απόδοση του εν λόγω κανόνα, ενώ οι κόκκινες κουκίδες είναι οι αποδόσεις του SMA_{RS} που βρίσκονται εντός του ΔΕ 95% και οι μαύρες κουκκίδες οι αποδόσεις που βρίσκονται εκτός του εν λόγω ΔΕ.

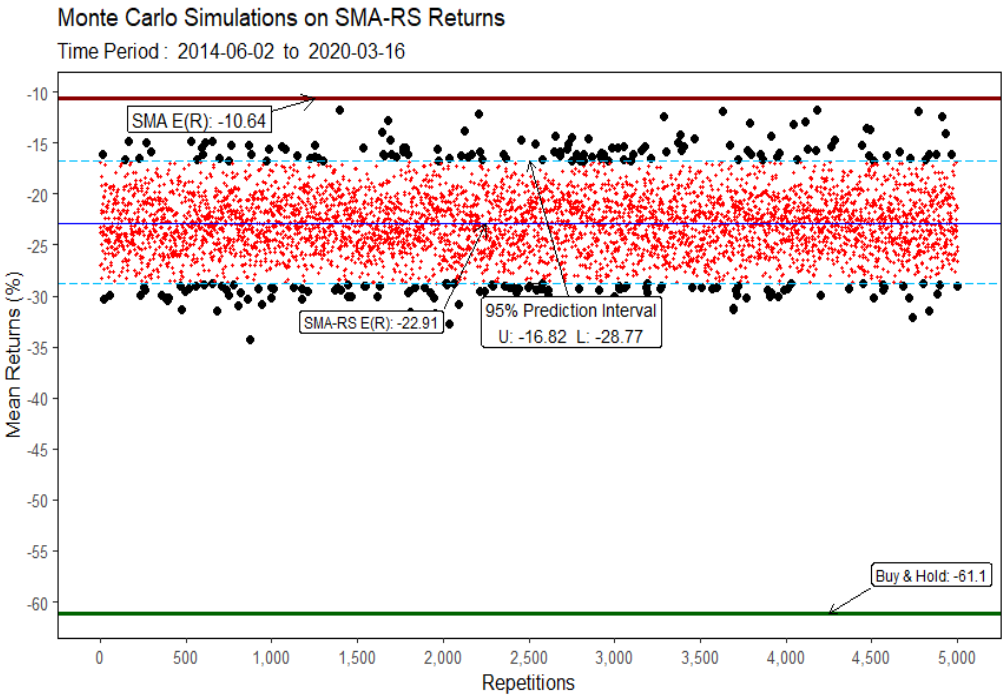
4.3 Έλεγχος Σημαντικότητας στο Χρηματιστήριο Αξιών Αθηνών

Στις Εικόνες 3 έως 5 συγκρίνουμε την αναμενόμενη απόδοση του SMA με το Διάστημα Εμπιστοσύνης (εφεξής ΔΕ) 95% της απόδοσης του κανόνα SMA_{RS} και την απόδοση της αγοράς και διακράτησης τόσο σε ολόκληρη την περίοδο 2014 – 2023, όσο και στις υποπεριόδους 2014-2020 και 2020-2023, όπου ο ΓΔ παρουσιάζει σταθερή, καθοδική και ανοδική κίνηση αντίστοιχα. Για πρακτικούς λόγους όταν θα αναφερόμαστε εφεξής στην απόδοση του SMA_{RS} νοούμε το Διάστημα Εμπιστοσύνης 95% της απόδοσης του SMA_{RS}. Σε ολόκληρη την περίοδο, όπου το επίπεδο του ΓΔ είναι το ίδιο στην έναρξη

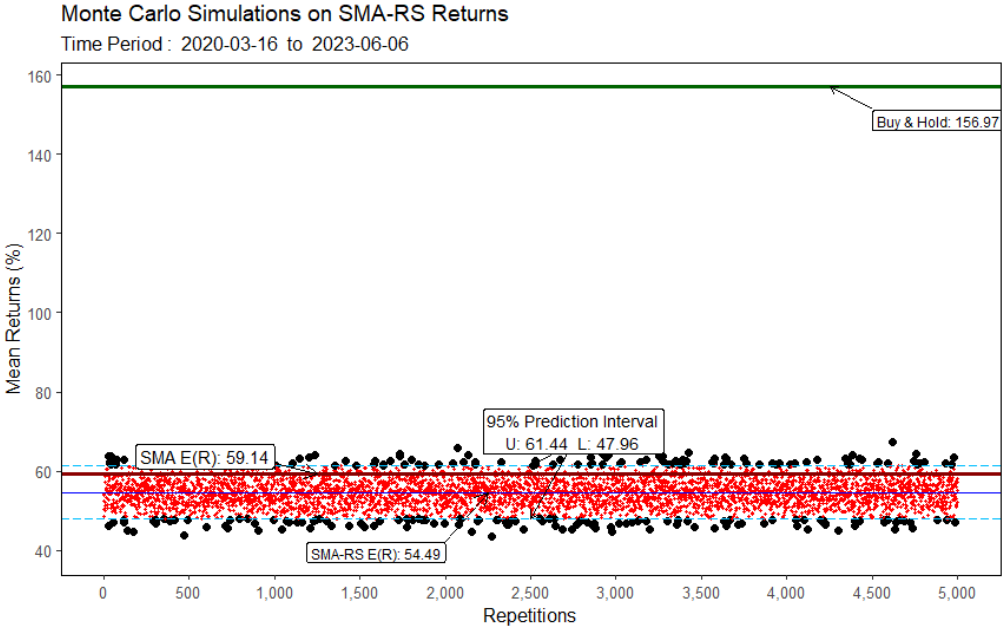
και στη λήξη της περιόδου (σταθερή περίοδος), η απόδοση του SMA είναι υψηλότερη από την απόδοση του SMA_{RS} και από την απόδοση της αγοράς και διακράτησης, η οποία βρίσκεται μέσα στο ΔΕ του SMA_{RS}. Συνεπώς, η απόδοση του τεχνικού κανόνα SMA έχει προγνωστική ικανότητα στη συγκεκριμένη χρονική περίοδο, ενώ η απόδοση της παθητικής στρατηγικής δε διαφέρει από την απόδοση των τυχαίων συναλλαγών.

Στην υποπερίοδο 2014 – 2020, όπου η αγορά ήταν εν γένει καθοδική, η απόδοση του SMA είναι υψηλότερη από την απόδοση του SMA_{RS} (τυχαίες συναλλαγές), το οποίο είναι αναμενόμενο καθώς η προγνωστική ικανότητα των τεχνικών κανόνων ανίχνευσης τάσης μεγιστοποιείται σε πτωτικές αγορές.

Εικόνα 4. *Η αναμενόμενη απόδοση του SMA, η απόδοση της παθητικής επενδυτικής στρατηγικής και οι αποδόσεις τυχαίων συναλλαγών με το αντίστοιχο Διάστημα Εμπιστοσύνης 95% στον ΓΔ του ΧΑΑ στην καθοδική υποπερίοδο 2014 - 2020.*



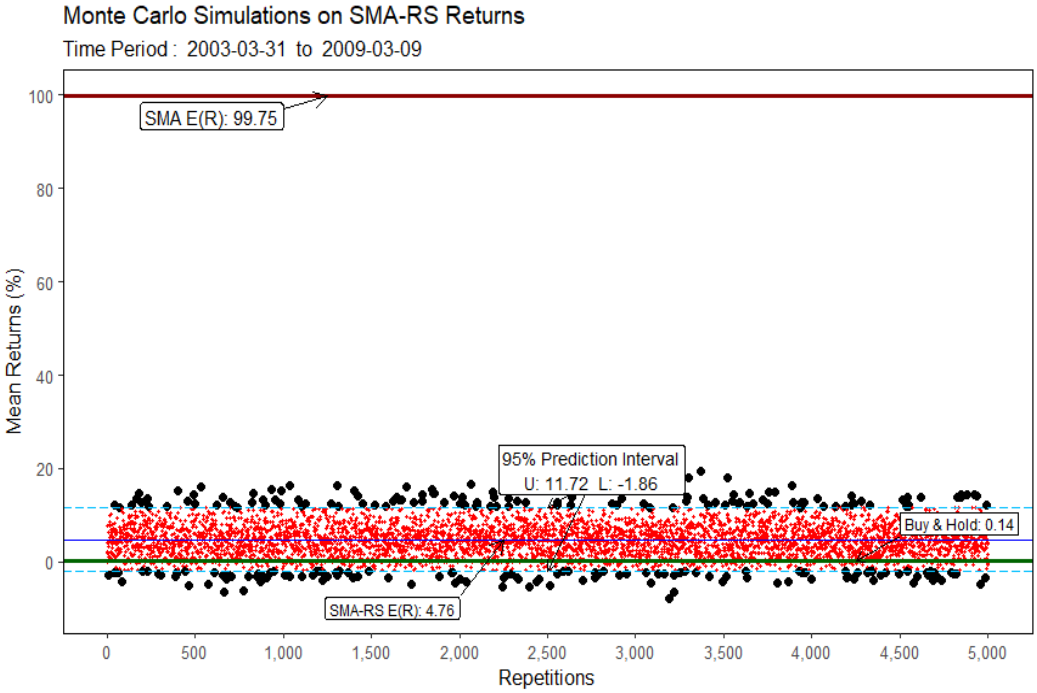
Εικόνα 5. *Η αναμενόμενη απόδοση του SMA, η απόδοση της παθητικής επενδυτικής στρατηγικής και οι αποδόσεις τυχαίων συναλλαγών με το αντίστοιχο Διάστημα Εμπιστοσύνης 95% στον ΓΔ του ΧΑΑ στην ανοδική υποπερίοδο 2020 - 2023.*



Τα αποτελέσματα διαφέρουν σημαντικά στην ανοδική υποπερίοδο 2020 – 2023, όπου η απόδοση αμφότερων των συναλλακτικών κανόνων SMA και SMA_{RS} είναι χαμηλότερη από την απόδοση της αγοράς και διακράτησης, ενώ η απόδοση του SMA είναι οριακά υψηλότερη από την απόδοση του SMA_{RS} σε ΔΕ 90%.

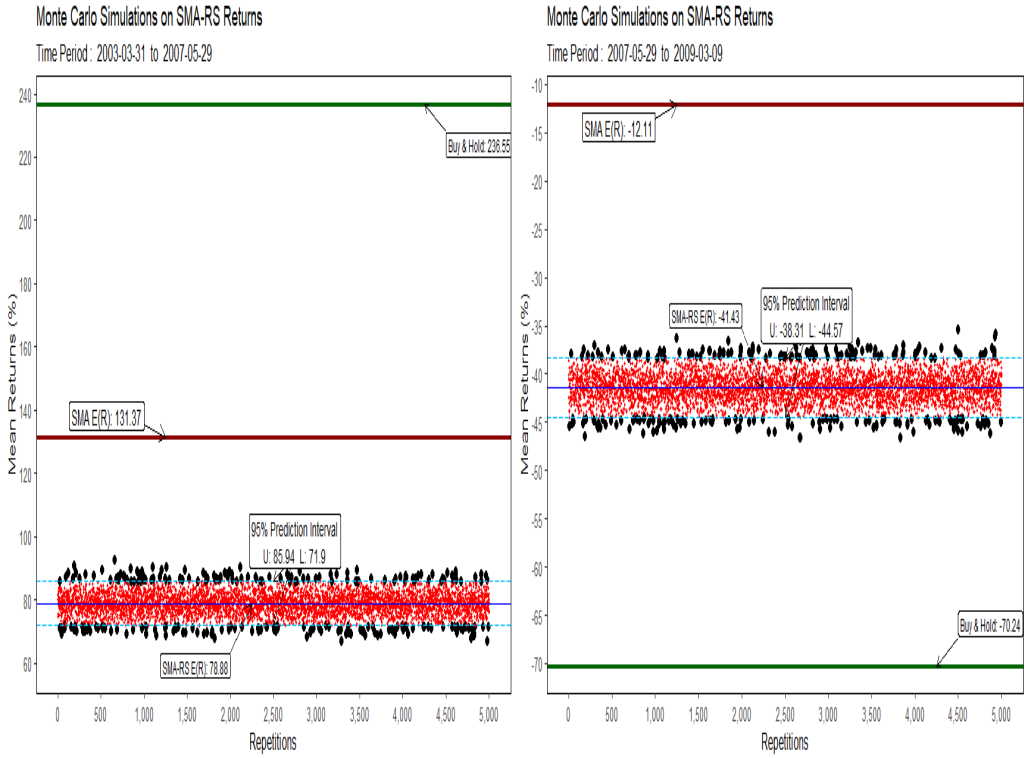
Στη συνέχεια διενεργούμε στατιστικό έλεγχο σημαντικότητας στην χρονική περίοδο 2003 – 2009 και στις υποπεριόδους 2003-2007 και 2007-2009, όπου ο ΓΔ παρουσιάζει ανοδική και καθοδική τάση αντίστοιχα. Στην Εικόνα 6 παρατηρούμε ότι σε ολόκληρη την περίοδο ο SMA επιτυγχάνει σημαντικά υψηλότερη απόδοση συγκριτικά τόσο με την απόδοση του SMA_{RS} όσο και με την απόδοση της αγοράς και διακράτησης, ενώ η τελευταία δε διαφέρει στατιστικά από την απόδοση του SMA_{RS}. Ωστόσο, στο αριστερό γράφημα της Εικόνας 7, όπου η πορεία του ΓΔ είναι ανοδική (2003-2007), τόσο ο SMA όσο και ο SMA_{RS} εμφανίζουν σημαντικά χαμηλότερη απόδοση από την απόδοση της παθητικής επενδυτικής στρατηγικής. Εντούτοις, η απόδοση του SMA είναι υψηλότερη από αυτή του SMA_{RS} υποδηλώνοντας ότι ο τεχνικός κανόνας έχει προβλεπτική ικανότητα παρά το γεγονός ότι δεν καταφέρνει να υπερνικήσει την αγορά. Αντίθετα, στο δεξί γράφημα της Εικόνας 7, όπου ο ΓΔ εμφανίζει καθοδική πορεία (2007-2009), ο SMA επιτυγχάνει σημαντικά υψηλότερη απόδοση από την απόδοση του SMA_{RS} και από την απόδοση της αγοράς και διακράτησης αλλά και ο SMA_{RS} επιτυγχάνει υψηλότερη απόδοση από αυτή της παθητικής στρατηγικής.

Εικόνα 6. *Η αναμενόμενη απόδοση του SMA, η απόδοση της παθητικής επενδυτικής στρατηγικής και οι αποδόσεις τυχαίων συναλλαγών SMA_{RS} με το αντίστοιχο Διάστημα Εμπιστοσύνης 95% στον ΓΔ του ΧΑΑ στην περίοδο 2003 - 2009.*



Τα ανωτέρω αποτελέσματα επιβεβαιώνουν ότι ο SMA έχει προγνωστική ικανότητα στην συγκεκριμένη αγορά και στις συγκεκριμένες χρονικές περιόδους, η οποία δεν δύναται να αποδοθεί στην τύχη αφού επιτυγχάνει υψηλότερη απόδοση από την απόδοση που παράγουν ισάριθμες τυχαίες συναλλαγές. Ταυτόχρονα όμως, αποκαλύπτεται μια μεροληψία στην σύγκριση του SMA με την απόδοση της αγοράς και διακράτησης τόσο στις ανοδικές όσο και στις καθοδικές αγορές που δύναται να οδηγήσει σε παραπλανητικά συμπεράσματα. Συγκεκριμένα, στις καθοδικές αγορές διενεργώντας τυχαίες συναλλαγές επιτυγχάνεται υψηλότερη απόδοση από την απόδοση της αγοράς και διακράτησης. Συνεπώς, για να πιστωθεί η προγνωστική ικανότητα του τεχνικού κανόνα δεν αρκεί αυτός να καταγράφει υψηλότερη απόδοση από την παθητική στρατηγική αλλά η απόδοση του πρέπει να είναι υψηλότερη από αυτή των τυχαίων συναλλαγών. Αντίθετα, στις ανοδικές αγορές η προγνωστική ικανότητα του τεχνικού κανόνα επικυρώνεται επιτυγχάνοντας υψηλότερη απόδοση από αυτήν των τυχαίων συναλλαγών και όχι κατ' ανάγκη από την απόδοση της αγοράς και διακράτησης. Αυτό είναι εύλογο καθώς η συνεχώς ανοδική κίνηση του Δείκτη προσφέρει την μέγιστη απόδοση, την οποία κανένας συναλλακτικός κανόνας δε δύναται να επιτύχει σηματοδοτώντας έστω και ελάχιστα κίβδηλα σήματα ημερών εκτός αγοράς.

Εικόνα 7. *Η αναμενόμενη απόδοση του SMA, η απόδοση της παθητικής επενδυτικής στρατηγικής και οι αποδόσεις τυχαίων συναλλαγών με το αντίστοιχο Διάστημα Εμπιστοσύνης 95% στον ΓΔ του ΧΑΑ στις υποπεριόδους 2003 – 2007 και 2007 - 2009.*

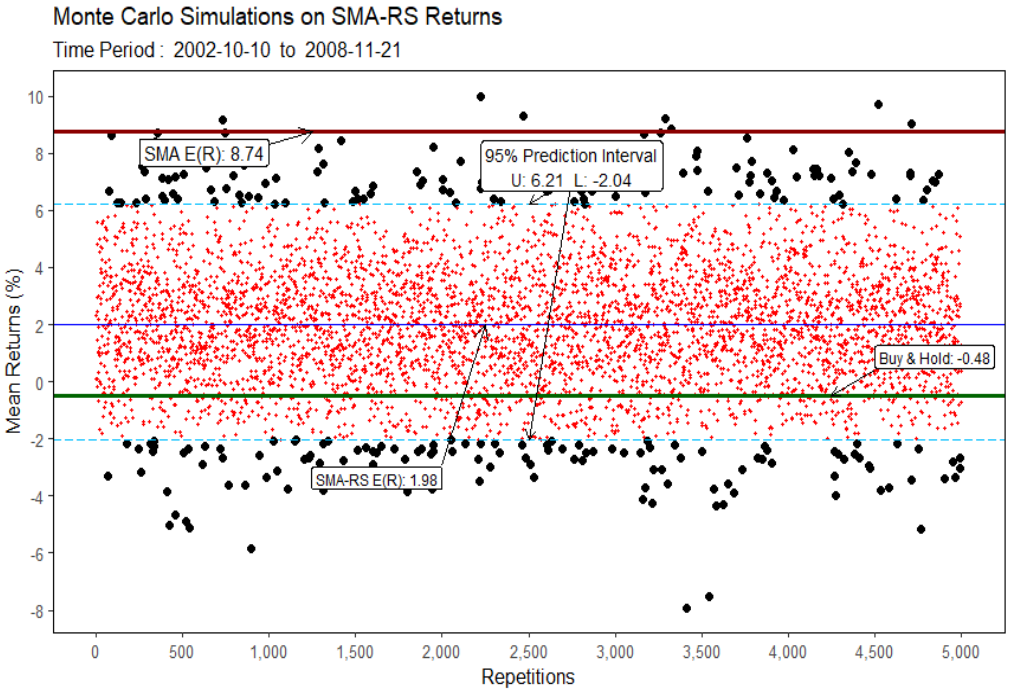


Τα ανωτέρω αποτελέσματα επιβεβαιώνουν ότι ο SMA έχει προγνωστική ικανότητα στην συγκεκριμένη αγορά και στις συγκεκριμένες χρονικές περιόδους, η οποία δε δύναται να αποδοθεί στην τύχη αφού επιτυγχάνει υψηλότερη απόδοση από την απόδοση που παράγουν ισάριθμες τυχαίες συναλλαγές. Ταυτόχρονα όμως, αποκαλύπτεται μια μεροληψία στην σύγκριση του SMA με την απόδοση της αγοράς και διακράτησης τόσο στις ανοδικές όσο και στις καθοδικές αγορές που δύναται να οδηγήσει σε παραπλανητικά συμπεράσματα. Συγκεκριμένα, στις καθοδικές αγορές διενεργώντας τυχαίες συναλλαγές επιτυγχάνεται υψηλότερη απόδοση από την απόδοση της αγοράς και διακράτησης. Συνεπώς, για να πιστωθεί η προγνωστική ικανότητα του τεχνικού κανόνα δεν αρκεί αυτός να καταγράφει υψηλότερη απόδοση από την παθητική στρατηγική αλλά η απόδοση του πρέπει να είναι υψηλότερη από αυτή των τυχαίων συναλλαγών. Αντίθετα, στις ανοδικές αγορές η προγνωστική ικανότητα του τεχνικού κανόνα επικυρώνεται επιτυγχάνοντας υψηλότερη απόδοση από αυτήν των τυχαίων συναλλαγών και όχι κατ' ανάγκη από την απόδοση της αγοράς και διακράτησης. Αυτό είναι εύλογο καθώς η συνεχώς ανοδική κίνηση του Δείκτη προσφέρει την μέγιστη απόδοση, την οποία κανένας συναλλακτικός κανόνας δε δύναται να επιτύχει σηματοδοτώντας έστω και ελάχιστα κίβδηλα σήματα ημερών εκτός αγοράς.

4.4 Διενέργεια Ελέγχου Σημαντικότητας στο Χρηματιστήριο της Νέας Υόρκης

Στην Εικόνα 8 διενεργούμε στατιστικό έλεγχο σημαντικότητας στον Δείκτη SP500 στην περίοδο 2002 – 2008. Παρατηρούμε ότι σε αυτή την περίοδο, όπου ο SP500 διατηρεί το επίπεδο της τιμής εκκίνησης στη λήξη της περιόδου, ο SMA εμφανίζει οριακά μόνο υψηλότερη απόδοση από την απόδοση του SMA_{RS} και της απόδοσης της αγοράς και διακράτησης, ενώ η τελευταία βρίσκεται μέσα στο ΔΕ 95% του SMA_{RS}. Το γεγονός αυτό δείχνει ότι ο SMA στην συγκεκριμένη αγορά έχει οριακή μόνο προγνωστική ικανότητα, το οποίο συμφωνεί με τα ευρήματα της υπάρχουσας βιβλιογραφίας για τον εν λόγω χρηματιστηριακό Δείκτη.

Εικόνα 8. *Η αναμενόμενη απόδοση του SMA, η απόδοση της παθητικής επενδυτικής στρατηγικής και οι αποδόσεις τυχαίων συναλλαγών SMA_{RS} με το αντίστοιχο Διάστημα Εμπιστοσύνης 95% στον SP500 στην περίοδο 2002 - 2008.*



5. ΣΥΝΟΨΗ ΚΑΙ ΣΥΜΠΕΡΑΣΜΑΤΑ

Σκοπός της παρούσας εργασίας είναι να αξιολογηθεί η προβλεπτική ικανότητα ενός εκ των δημοφιλέστερων τεχνικών συναλλακτικών κανόνων και να διερευνηθεί εάν η ευρεθείσα ικανότητα δύναται να αποδοθεί στην τύχη. Για την επίτευξη του σκοπού εισάγεται ένας εναλλακτικός τρόπος ελέγχου για την ανάδειξη τυχόν ύπαρξης του φαινομένου «υπερεκμετάλλευσης των δεδομένων», ευρέως γνωστό ως «data-snooping», το οποίο σπάνια διευθετείται σε παρόμοιες μελέτες. Η προτεινόμενη μεθοδολογία τυχαίας σηματοδότησης συναλλαγών είναι απλούστερη και ως ένα βαθμό παρέχει μια πιο άμεση σύγκριση στη διαδικασία ελέγχου για data-snooping, ούσα

απαλλαγμένη από την απαίτηση εφαρμογής χιλιάδων συναλλακτικών κανόνων από τους υπάρχοντες ελέγχους. Η κεφαλαιαγορά που μελετήθηκε είναι το Χρηματιστήριο Αξιών Αθηνών, και επικουρικά το Χρηματιστήριο της Νέας Υόρκης.

Τα αποτελέσματα δείχνουν ότι στο Χρηματιστήριο Αθηνών ο απλός κινητός μέσος αποδίδει υψηλότερα των τυχαία σηματοδοτούμενων συναλλαγών σε περιόδους σταθερών, καθοδικών και σε πολλές περιπτώσεις και ανοδικών αγορών, αποδεικνύοντας ότι η προβλεπτική του ικανότητα έχει πραγματική αξία και δεν δύναται να αποδοθεί στην τύχη. Όσον αφορά το Χρηματιστήριο της Νέας Υόρκης ο εν λόγω τεχνικός κανόνας έχει οριακή μόνο προγνωστική ικανότητα.

Τα ανωτέρω συμπεράσματα συμφωνούν με την κλασσική μέθοδο αξιολόγησης τεχνικών κανόνων, δηλαδή την σύγκρισή τους με την απόδοση της αγοράς και διακράτησης, μόνο όταν η απόδοση της παθητικής επενδυτικής στρατηγικής δε διαφέρει στατιστικά από την απόδοση των τυχαία σηματοδοτούμενων συναλλαγών.

Τα εμπειρικά ευρήματα της παρούσας προδήλως επιβεβαιώνουν ότι η σύγκριση ενός τεχνικού συναλλακτικού κανόνα αποκλειστικά με την απόδοση της αγοράς και διακράτησης είναι σε πολλές περιπτώσεις μεροληπτική και δύναται να οδηγήσει σε παραπλανητικά συμπεράσματα στον έλεγχο αφενός της αξιολόγησης ενός τεχνικού συναλλακτικού κανόνα ως προς την προγνωστική του ικανότητα, αφετέρου της υπόθεσης των αποτελεσματικών κεφαλαιαγορών ασθενούς ισχύος. Ωστόσο, η προτεινόμενη μέθοδος αξιολόγησης τυχαίων συναλλαγών, ούσα απαλλαγμένη από τις ευρεθείσες μεροληψίες, δύναται να αποτελέσει ένα αμερόληπτο σημείο αναφοράς.

ABSTRACT

The efficient market hypothesis constitutes a fundamental principle of modern finance, shaping theoretical research and investment practices. A common testing procedure compares the performance of technical trading rules with buy-and-hold return. However, it is rarely examined whether the performance of these trading rules holds genuine value or stems from luck, a phenomenon well-known as data-snooping. In this paper, we apply the simple moving average (SMA) trading rule to the Athens and New York Stock Exchanges, considering all trading rule parameters, and test for data-snooping biases using an alternative random trading method based on Monte Carlo simulation. The results show that the SMA's predictive ability clearly holds real value outperforming the Athens market but does so only marginally at the New York Stock Exchange. Furthermore, we expose significant biases in the assessment of a trading rule using the common testing procedure, as the outcome is significantly affected by the trend of the applied Index. This phenomenon has direct implications for both assessing trading rules' predictive performance and testing weak-form market efficiency. Nevertheless, the proposed methodology of random trading provides an unbiased benchmark for evaluating trading rules while testing, to an extent, for potential data-snooping.

ΑΝΑΦΟΡΕΣ

Black, F. (1986). Noise. *The Journal of Finance, 41*(3), 529–543.

Brock, W., Lakonishok, J., & LeBaron, B. (1992). Simple Technical Trading Rules and the Stochastic Properties of Stock Returns. *The Journal of Finance, 47*(5), 1731–1764.

Corbet, S., Eraslan, V., Lucey, B., & Sensoy, A. (2019). The effectiveness of technical trading rules in cryptocurrency markets. *Finance Research Letters, 31*, 32–37.

Cowles, A. 3. (1933). CAN STOCK MARKET FORECASTERS FORECAST? *Econometrica, 1*(3), 309–324.

Elliott, G., Rothenberg, T. J., & Stock, J. H. (1996). EFFICIENT TESTS FOR AN AUTOREGRESSIVE UNIT ROOT. *Econometrica, 64*(4), 813–836.

Elton, E. J., & Gruber, M. J. (1995). *Modern Portfolio Theory and Investment Analysis* (5th ed.). Wiley.

Fama, E. F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. *The Journal of Finance, 25*, 383–417.

Fama, E. F. (1991). Efficient Capital Markets: II. *The Journal of Finance, 46*(6), 1575–1617.

Fama, E. F., & Blume, M. E. (1966). Filter Rules and Stock-Market Trading. *The Journal of Business, 39*(1), 226–241.

Hansen, P. R. (2005). A Test for Superior Predictive Ability. *Journal of Business & Economic Statistics, 23*(4), 365–380.

Hudson, R., Dempsey, M., & Keasey, K. (1996). A note on the weak form efficiency of capital markets: The application of simple technical trading rules to UK stock prices -1935 to 1994. *Journal of Banking & Finance, 20*, 1121–1132.

Kwon, K., & Kish, R. J. (2002). Technical trading strategies and return predictability: NYSE. *Applied Financial Economics, 12*(9), 639–653.

Luukka, P., Pätäri, E., Fedorova, E., & Garanina, T. (2016). Performance of Moving Average Trading Rules in a Volatile Stock Market: The Russian Evidence. *Emerging Markets Finance and Trade, 52*(10), 2434–2450.

Ma, Y., Yang, B., & Su, Y. (2021). Stock return predictability: Evidence from moving averages of trading volume. *Pacific-Basin Finance Journal, 65*, 101494.

Malkiel, B. (1992). In Newman P., Milgate M. and Eatwell J.(Eds.), *Efficient Market Hypothesis* (New Palgrave Dictionary of Money and Finance ed.). McMillan.

Marshall, B. R., Nguyen, N. H., & Visaltanachoti, N. (2017). Time series momentum and moving average trading rules. *Quantitative Finance, 17*(3), 405–421.

Milionis Alexandros E, & Varlagkas Vasileios. (2021). Moving median against moving average, institutional investors against private investors and market efficiency. Paper presented at the *33rd Panhellenic Statistics Conference,* Larissa. 178–194.

Milionis, A. E. (2007). Efficient capital markets: A statistical definition and comments. *Statistics & Probability Letters, 77*(6), 607–613.

Milionis, A. E., & Papanagiotou, E. (2008). On the use of the moving average trading rule to test for Weak Form Efficiency in Capital Markets. *Economic Notes by Banca Monte Dei Paschi Di Siena SpA, 37*(2), 181–201.

Milionis, A. E., & Papanagiotou, E. (2013). Decomposing the predictive performance of the moving average trading rule of technical analysis: the contribution of linear and non-linear dependencies in stock returns. *Journal of Applied Statistics, 40*(11), 2480–2494.

Mills, T. C. (1997). Technical analysis and the London Stock Exchange: testing trading rules using the FT30. *International Journal of Finance & Economics, 2*, 319–331.

Miralles-Quirós, J. L., Miralles-Quirós, M. D. M., & Gonçalves, L. M. V. (2018). The Profitability of Moving Average Rules: Smaller Is Better in the Brazilian Stock Market. *Emerging Markets Finance and Trade, 55*(1), 150–167.

Neftci, S. N. (1991). Naive Trading Rules in Financial Markets and Wiener-Kolmogorov Prediction Theory: A Study of Technical Analysis. *Journal of Business, 64*(4), 549–571.

Newey, W. K., & West, K. D. (1987). A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica, 55*(3), 703–708.

Oikonomou, K., Damigos, D., & Dimitriou, D. (2025). Globality in the metal markets: Leveraging cross-learning to forecast aluminum and copper prices. *Resources Policy,* 103. 1-14.

Papailias, F., & Thomakos, D. D. (2015). An improved moving average technical trading rule. *Physica A: Statistical Mechanics and its Applications, 428*, 458–469.

Phillips, P. C. B., & Perron, P. (1988). Testing for a Unit Root in Time Series Regression. *Biometrica, 75*(2), 335–346.

Sullivan, R., Timmermann, A., & White, H. (1999). Data-Snooping, Technical Trading Rule Performance, and the Bootstrap. *The Journal of Finance, 54*(5), 1647–1691.

White, H. (2000). A REALITY CHECK FOR DATA SNOOPING. *Econometrica, 68*(5), 1097–1126.

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 37ου Πανελληνίου & 3ου Διεθνούς Συνεδρίου Στατιστικής (2025), σελ.123-142

ΕΓΚΥΡΟΤΗΤΑ ΚΑΙ ΕΠΙΠΤΩΣΕΙΣ ΤΩΝ ΟΙΚΟΝΟΜΕΤΡΙΚΩΝ ΕΛΕΓΧΩΝ ΓΙΑ ΤΗΝ ΥΠΟΘΕΣΗ ΤΩΝ ΑΠΟΤΕΛΕΣΜΑΤΙΚΩΝ ΑΓΟΡΩΝ

Α. Μηλιώνης[1], Ν. Γαλανόπουλος[1]
[1]Τμήμα Στατιστικής και Αναλογιστικών-Χρηματοοικονομικών Μαθηματικών
Πανεπιστήμιο Αιγαίου
{amilionis, ngalan}@aegean.gr

ΠΕΡΙΛΗΨΗ

Στην υπάρχουσα βιβλιογραφία για τον έλεγχο της αποτελεσματικότητας των αγορών χρησιμοποιούνται πολύ συχνά οι χρηματιστηριακές αποδόσεις, εκφρασμένες ως πρώτες διαφορές των φυσικών λογαρίθμων των τιμών, θεωρείται δε δεδομένη η στασιμότητα ως προς την διακύμανση των αντίστοιχων χρονοσειρών. Με την παρούσα μελέτη προκύπτει ότι η εφαρμογή του λογαριθμικού μετασχηματισμού στις χρονοσειρές των τιμών των αξιογράφων είναι σε πολλές περιπτώσεις ανεπαρκής για την σταθεροποίηση της διακύμανσης. Συνεπώς, για αυτές τις περιπτώσεις, οι έλεγχοι αυτοσυσχέτισης, που χρησιμοποιούνται ευρέως για τον έλεγχο της αποτελεσματικότητας των αγορών, αυστηρά δεν είναι έγκυροι. Στην παρούσα μελέτη αξιοποιώντας τη μεθοδολογία των Milionis & Galanopoulos (2017), τεκμηριώνεται ότι οι έλεγχοι αυτοσυσχέτισης δεν μπορούν να χρησιμοποιηθούν έγκυρα ως στατιστικό εργαλείο για τον έλεγχο της αποτελεσματικότητας των αγορών στην πλειοψηφία των εξεταζόμενων περιπτώσεων.

Λέξεις Κλειδιά: μετασχηματισμοί χρονοσειρών, μη στασιμότητα στη διακύμανση χρονοσειρών, έλεγχοι αυτοσυσχέτισης, αποτελεσματικότητα αγορών

1. ΕΙΣΑΓΩΓΗ

Ο πιο ευρέως αποδεκτός ορισμός των αποτελεσματικών αγορών δόθηκε από τον Fama (1970), δηλώνοντας ότι μία αγορά είναι αποτελεσματική όταν οι τιμές αντικατοπτρίζουν πλήρως όλες τις διαθέσιμες πληροφορίες. Με βάση τις διαθέσιμες πληροφορίες διακρίνονται τρεις μορφές αποτελεσματικότητας της αγοράς (Roberts, 1959). Πιο συγκεκριμένα, η αποτελεσματικότητα της αγοράς υπό την μορφή ασθενούς ισχύος (Weak-Form Market Efficiency – WFME) συνδέεται με καταστάσεις όπου οι διαθέσιμες πληροφορίες αφορούν αποκλειστικά τις ιστορικές τιμές των χρηματοοικονομικών περιουσιακών στοιχείων, η αποτελεσματικότητα μέσης ισχύος περιλαμβάνει όλες τις δημόσιες πληροφορίες, και η ισχυρή αποτελεσματικότητα

περιέχει τόσο τις δημόσιες όσο και τις ιδιωτικές πληροφορίες. Η παρούσα εργασία θα επικεντρωθεί στην αποτελεσματικότητα ασθενούς ισχύος (WFME).

Παρόλο που η θεωρητική θεμελίωση της αποτελεσματικότητας των αγορών τέθηκε στο άρθρο του Fama (1970), η στατιστική επεξήγηση που παρουσίασε έχει δεχθεί κριτική (Leroy, 1976). Για αυτόν το λόγο, διάφοροι ορισμοί έχουν προταθεί με το πέρασμα των χρόνων (Malkiel, 1992; Milionis, 2007).

Αναγνωρίζοντας πιθανές παρερμηνείες που θα μπορούσαν να προκύψουν από τον ορισμό που έδωσε ο Fama στο πρωτοποριακό του άρθρο το 1970, ο Fama (1976; 1991) διευθετεί ένα από τα πιο καίρια ζητήματα, το οποίο αφορά τον από κοινού έλεγχο της αποτελεσματικότητας της αγοράς και του υποδείγματος τιμολόγησης. Το 1991 αναθεώρησε την WFME εισάγοντας την έννοια της προβλεψιμότητας των αποδόσεων (Fama, 1991), αναγνωρίζοντας ότι σε μια αποτελεσματική αγορά, οι επενδυτές δεν μπορούν να επιτύχουν υπερβάλλουσες αποδόσεις ή κέρδη.

Περαιτέρω, έχει παρατηρηθεί το φαινόμενο ότι πολλοί ερευνητές χρησιμοποιούν συχνά και συνδέουν λανθασμένα τη στατιστική μεθοδολογία για τον έλεγχο της προβλεψιμότητας των αποδόσεων με την υπόθεση της αποτελεσματικότητας των αγορών. Οι συνθήκες που πρέπει να ισχύουν στους στατιστικούς ελέγχους, και πιο συγκεκριμένα στους ελέγχους αυτοσυσχέτισης, ώστε τα αποτελέσματα τους να συνδέονται ορθώς με την αποτελεσματικότητα των αγορών παρουσιάζονται στο άρθρο του Milionis (2007).

Πιο αναλυτικά, για να ελεγχθεί εμπειρικά η WFME, αρχικά θεωρείται ότι οι συνθήκες ισορροπίας μπορούν να περιγραφούν σε όρους αναμενόμενων αποδόσεων. Αυτές γράφονται ως $E\left(\tilde{P}_{j,t+1} \,/\, \Phi_t\right) = \left[1 + E\left(\tilde{R}_{j,t+1} \,/\, \Phi_t\right)\right] P_{j,t}$, όπου το $P_{j,t}$ αναπαριστά την τιμή του αξιογράφου j την χρονική στιγμή t, $R_{j,t+1}$ είναι η ποσοστιαία απόδοση του αξιογράφου j μεταξύ των χρονικών στιγμών t και $t+1$, και Φ_t συμβολίζει το πληροφοριακό σύνολο που αντικατοπτρίζεται πλήρως στην τιμή $P_{j,t}$. Η χρήση των περισπωμένων δηλώνει τυχαίες μεταβλητές μια οποιαδήποτε χρονική στιγμή t.

Οι προαναφερθείσες αναμενόμενες αποδόσεις καθορίζονται αξιοποιώντας ένα υπόδειγμα τιμολόγησης, κάνοντας τον έλεγχο της WFME ένα από κοινού έλεγχο της WFME και του υποδείγματος τιμολόγησης. Όταν το υπόδειγμα τιμολόγησης υποθέτει σταθερές αναμενόμενες αποδόσεις σε ένα πλαίσιο όπου δεν λαμβάνεται υπόψη ο κίνδυνος, είναι σύνηθες να εφαρμόζονται έλεγχοι αυτοσυσχέτισης στις αποδόσεις των αξιογράφων. Ωστόσο, είναι σημαντικό να τονιστεί ότι στατιστικά σημαντικές συσχετίσεις από μόνες τους δεν συνεπάγονται αναγκαία την απόρριψη της υπόθεσης της WFME, καθώς η από κοινού υπόθεση ελέγχεται, η οποία περιλαμβάνει και την WFME και το υπόδειγμα τιμολόγησης με τις σταθερές αναμενόμενες αποδόσεις.

Οι Elton et al. (2014) παρέχουν μια ανασκόπηση των αποτελεσμάτων από τους ελέγχους αυτοσυσχέτισης που έχουν διεξαχθεί από πολλούς ερευνητές. Οι έλεγχοι αυτοί αναλύουν τις πρώτες διαφορές των τιμών ή των λογαρίθμων των τιμών. Σε ένα πλαίσιο όπου δεν λαμβάνεται υπόψη ο κίνδυνος, είναι σημαντικό να σημειωθεί ότι αυτοί οι έλεγχοι είναι έγκυροι μόνο όταν οι σειρές των λογαρίθμων των τιμών

εμφανίζουν ασθενή στασιμότητα, δηλαδή στασιμότητα τόσο ως προς το επίπεδο όσο και ως προς διακύμανση-συνδιακυμάνσεις. Στο πλαίσιο της παρούσας μελέτης με τον όρο στασιμότητα εννοείται η ασθενής στασιμότητα. Εάν αυτή η συνθήκη δεν πληρείται, ο έλεγχος σημαντικότητας των συντελεστών αυτοσυσχέτισης στους ευρέως χρησιμοποιουμένους ελέγχους αυτοσυσχέτισης δεν είναι έγκυρος. Περαιτέρω επισημαίνεται ότι οι έλεγχοι αυτοσυσχέτισης έχουν νόημα μόνο στις περιπτώσεις που η υπόθεση των σταθερών αναμενόμενων αποδόσεων, και κατ' επέκταση, της στασιμότητας της διακύμανσης ισχύει (Milionis, 2007).

Οι χρονοσειρές των τιμών των χρηματοοικονομικών περιουσιακών στοιχείων (μετοχών, ομολόγων και χρηματιστηριακών δεικτών) είναι συνήθως μη-στάσιμες. Ωστόσο, στην πλειοψηφία των περιπτώσεων, η εμπειρική έρευνα εξετάζει μόνο την πρώτη ροπή (δηλαδή, τα επίπεδα των τιμών) και δεν ελέγχει τη δεύτερη ροπή. Αυτό ενδεχομένως συμβαίνει επειδή οι στατιστικοί έλεγχοι διεξάγονται χρησιμοποιώντας τις αποδόσεις των περιουσιακών στοιχειών και όχι τις τιμές τους. Καθώς όμως οι αποδόσεις των περιουσιακών στοιχείων εκφράζονται ως οι διαφορές στους λογαρίθμους των αντίστοιχων τιμών, σιωπηρά υιοθετείται ότι εξαιτίας του λογαριθμικού μετασχηματισμού οι αποδόσεις είναι στάσιμες και ως προς τη διακύμανση. Ωστόσο, αν δεν ισχύει η στασιμότητα και ως προς τη διακύμανση, οι συνθήκες για τον έλεγχο της υπόθεσης της WFME με τους ελέγχους αυτοσυσχέτισης δεν πληρούνται, όπως τονίστηκε προηγουμένως.

Ως εκ τούτου, πρωταρχικός στόχος αυτής της εργασίας είναι να ελεγχθεί αν αυτή η σιωπηρή υπόθεση ισχύει, διότι οι ερευνητές χρησιμοποιούν ευρέως τις αποδόσεις των αξιογράφων ως ένα τυπικό μέτρο, χωρίς να ελέγχουν αν αυτές οι χρονοσειρές είναι στάσιμες και ως προς τη διακύμανση. Ενώ ο λογαριθμικός μετασχηματισμός είναι ο κατάλληλος από πλευράς χρηματοοικονομικής, δεν υπάρχει καμιά εγγύηση ότι αυτό είναι επίσης ορθό από πλευράς στατιστικής. Επιπρόσθετα, είναι ενδιαφέρον να ελεγχθεί κατά πόσο ο βαθμός ωριμότητας μιας χρηματοοικονομικής αγοράς συνδέεται με μοτίβα μη-στασιμότητας στη διακύμανση, και πως αυτό επηρεάζει τους ελέγχους αποτελεσματικότητας της αγοράς.

Η δομή του άρθρου είναι η ακόλουθη: Η επόμενη ενότητα παρέχει μία σύντομη περιγραφή της μεθοδολογίας που θα αξιοποιηθεί για τον έλεγχο της μη-στασιμότητας της διακύμανσης, η οποία είναι απαραίτητη προϋπόθεση για τον έλεγχο της υπόθεσης της αποτελεσματικότητας της αγοράς. Στην ενότητα 3 περιγράφονται οι χρηματοοικονομικές αγορές και τα δεδομένα που χρησιμοποιούνται στην παρούσα εμπειρική έρευνα και ανάλυση. Η ενότητα 4 παρουσιάζει τα αποτελέσματα της εμπειρικής ερευνάς, και στην ενότητα 5 συνοψίζονται τα αποτελέσματα.

2. ΠΕΡΙΓΡΑΦΗ ΜΕΘΟΔΟΛΟΓΙΑΣ

Μία συνθήκη για την στασιμότητα μίας χρονοσειράς απαιτεί τουλάχιστον το επίπεδο και η διακύμανση της χρονοσειράς να είναι σταθερά με την πάροδο του χρόνου. Ωστόσο, αν η διακύμανση σχετίζεται συναρτησιακά με το επίπεδο, και το επίπεδο είναι μη-στάσιμο, τότε η διακύμανση είναι μη-σταθερή τόσο υπό συνθήκη, όσο και χωρίς συνθήκη (μη-δεσμευμένη). Ως εκ τούτου, η διαδικασία είναι μη-ομοιογενώς μη-στάσιμη στο πλαίσιο των Box και Jenkins (1976) και δεν μπορεί να γίνει στάσιμη απλώς διαφορίζοντας την. Για να αντιμετωπιστεί η μη-στασιμότητα της διακύμανσης, μια προσέγγιση είναι η εφαρμογή μετασχηματισμών, όπως είναι οι ευρέως γνωστοί μετασχηματισμοί των Box και Cox (1964). Για παράδειγμα, ο ακόλουθος μετασχηματισμός χρησιμοποιείται αρκετά συχνά:

$$f(z_t) = \begin{cases} \frac{z_t^{\lambda}-1}{\lambda} \; if \; \lambda \neq 0 \\ lnz_t \; if \; \lambda = 0 \end{cases} \quad (1)$$

όπου z_t είναι η αρχική χρονοσειρά και λ η παράμετρος του μετασχηματισμού Box-Cox. Στις πρακτικές εφαρμογές, οι υπάρχουσες οικονομετρικές προσεγγίσεις για την ανίχνευση και διόρθωση της μη-στασιμότητας της διακύμανσης εμφανίζουν αρκετά μειονεκτήματα, καθώς: (i) είναι ανεπαρκείς, διότι η επιλογή ενός συγκεκριμένου μετασχηματισμού συχνά είναι υποκειμενική και αυθαίρετη (Hay και McLeary, 1979; Milionis και Davies, 1994; Mills, 1990 κεφάλαιο 4), και (ii) είναι μεροληπτικές ως προς την υπερβολική απόρριψη της μηδενικής υπόθεσης της μη-δεσμευμένης σταθερής διακύμανσης (Milionis και Galanopoulos, 2019; 2020).

Μία επίσημη οικονομετρική προσέγγιση, που όχι μόνο ανιχνεύει τη μη-στασιμότητα της διακύμανσης όταν αυτή διασυνδέεται με το επίπεδο της χρονοσειράς (όπως αυτή περιγράφεται στη σχέση (2)), αλλά και προτείνει κατάλληλους μετασχηματισμούς για τη διόρθωση της, παρουσιάστηκε στην εργασία των Milionis και Galanopoulos (2017). Μία σύντομη περιγραφή της προαναφερθείσας οικονομετρικής προσέγγισης παρατίθεται σε αυτό το σημείο.

Αρχικά, οι χρονοσειρές διαμερίζονται σε τμήματα (υποσύνολα) ίσου μήκους, και για κάθε τμήμα ο (τοπικός) μέσος (Local Mean – LM) και η (τοπική) τυπική απόκλιση (Local Standard Deviation – LSD) υπολογίζονται. Η τοπική τυπική απόκλιση θεωρείται ότι εξαρτάται συναρτησιακά από τον τοπικό μέσο με την ακόλουθη μη γραμμική σχέση:

$$LSD = aLM^{\beta}e^{u} \quad (2)$$

όπου a, β είναι οι παράμετροι του υποδείγματος, e είναι η βάση του φυσικού λογάριθμου, και u η στοχαστική διαταραχή. Οι παράμετροι a, β του υποδείγματος εκτιμώνται με τη μέθοδο των ελαχίστων τετραγώνων (Ordinary Least Squares – OLS) χρησιμοποιώντας το αντίστοιχο log-log υπόδειγμα. Η εκτιμώμενη τιμή του β ($\hat{\beta}$) παρέχει την απαραίτητη πληροφόρηση για την ύπαρξη (ή μη-ύπαρξη) και τον τύπο του μετασχηματισμού δεδομένων που απαιτείται για να εξασφαλιστεί η στασιμότητα της

διακύμανσης (για παράδειγμα, ο λογαριθμικός μετασχηματισμός αντιστοιχεί στη τιμή $\beta = 1$).

Για να εξασφαλιστεί η ανθεκτικότητα ως προς τη διαμέριση και την πιθανή ύπαρξη ακραίων τιμών, καθώς αυτό προηγείται της ανίχνευσης ακραίων τιμών, η διαδικασία επαναλαμβάνεται για διαφορετικές διαμερίσεις. Ο αριθμός των διαφορετικών διαμερίσεων είναι τουλάχιστον ίσος με τον αριθμό των διαιρετών του μήκους της σειράς, δίνοντας ένα πηλίκο (μήκος σειράς προς διαιρέτη) ≥ 5 και περιορίζοντας το μέγεθος των υποσυνόλων να είναι ≥ 5.Διευκρινίζεται ότι αριθμός πέντε (5) για το πηλίκο επιλέχθηκε ως ένα λογικό κατώτερο όριο τόσο για το μέγεθος κάθε υποσυνόλου, όσο και για τον αριθμό των υποσυνόλων σε οποιαδήποτε διαμέριση της αρχικής σειράς.

Τέλος, τα προηγούμενα βήματα επαναλαμβάνονται με τα μετασχηματισμένα δεδομένα. Ο σκοπός αυτού του τελευταίου βήματος είναι να ελέγξει εάν ο προτεινόμενος μετασχηματισμός είναι ικανός να σταθεροποιήσει τη διακύμανση της χρονοσειράς. Για περισσότερες λεπτομέρειες βλέπε Milionis και Galanopoulos (2017).

2.1 Αλγόριθμος TRAMO-SEATS

Στην εργασία των Milionis και Galanopoulos (2017) αποδείχθηκε ότι ο ευρέως χρησιμοποιούμενος αλγόριθμος του λογισμικού TSW, ήταν μεροληπτικός ως προς την εφαρμογή λογαριθμικού μετασχηματισμού στα αρχικά δεδομένα. Το TSW αντιπροσωπεύει το TRAMO-SEATS για Windows, μια έκδοση για Windows των DOS προγραμμάτων TRAMO και SEATS των Gómez και Maravall (1996). Το TRAMO είναι αρκτικόλεξο για το "Time series Regression with ARIMA noise Missing observations and Outliers" και το SEATS αντίστοιχα για το "Signal Extraction in ARIMA Time Series" και ως αλγόριθμος είναι ενσωματωμένος και σε άλλα ευρέως χρησιμοποιούμενα στατιστικά λογισμικά, όπως το EVIEWS. Επίσης, το TSW προσφέρει μόνο δύο εναλλακτικές για τον μετασχηματισμό των δεδομένων: τον λογαριθμικό μετασχηματισμό ή κανένα μετασχηματισμό.

Αντίθετα, η προτεινόμενη μεθοδολογία των Milionis και Galanopoulos (2017), εφεξής M-G, επιτρέπει την εφαρμογή εναλλακτικών μετασχηματισμών (όπως η τετραγωνική ρίζα, ο λογάριθμος, ο αρνητικός αντίστροφος) για τη σταθεροποίηση της διακύμανσης όταν αυτή είναι μη-στάσιμη.

3. ΔΕΔΟΜΕΝΑ ΚΑΙ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ ΤΩΝ ΑΓΟΡΩΝ

Η κατηγοριοποίηση των χρηματοοικονομικών αγορών χρησιμοποιείται ευρέως από τους επενδυτές για να αξιολογήσουν και να λάβουν επενδυτικές αποφάσεις σε διάφορες αγορές. Ένα από τα πιο διαδεδομένα συστήματα κατηγοριοποίησης των αγορών είναι το πλαίσιο κατηγοριοποίησης της αγοράς από την Morgan Stanley Capital International (MSCI). Σύμφωνα με αυτό το πλαίσιο (MSCI, 2023), μία χρηματοοικονομική αγορά ταξινομείται ως ανεπτυγμένη (mature), αναδυόμενη (emerging), περιθωριακή (frontier), ή μεμονωμένη (standalone) με βάση τρία κριτήρια.

Το πρώτο κριτήριο είναι η οικονομική ανάπτυξη της χώρας, και πιο συγκεκριμένα, η διατήρηση της οικονομικής ανάπτυξης. Μία αγορά κατηγοριοποιείται ως ανεπτυγμένη, όταν το Κατά Κεφαλήν Ακαθάριστο Εθνικό Εισόδημα είναι 25% υψηλότερο από το όριο υψηλού εισοδήματος (high-income threshold) για 3 διαδοχικές χρονιές. Το όριο υψηλού εισοδήματος καθορίζεται από την Παγκόσμια Τράπεζα χρησιμοποιώντας την μέθοδο Atlas (https://datahelpdesk.worldbank.org/). Η διάκριση μεταξύ αναδυόμενων και περιθωριακών αγορών δεν είναι ουσιώδης, καθώς πολλά στάδια κατηγοριοποίησης σε μία από τις δύο προαναφερθείσες αγορές επικαλύπτονται.

Το δεύτερο κριτήριο σχετίζεται με τις εταιρείες, οι οποίες πρέπει να πληρούν κάποιες ελάχιστες προϋποθέσεις σχετικά με τη ρευστότητα και το μέγεθος τους. Το τρίτο κριτήριο έχει σχέση με την πρόσβαση της αγοράς από διεθνείς θεσμικούς επενδυτές και αποτελείται από πέντε υπό-κριτήρια, τα οποία είναι: i) ευνοϊκό καθεστώς για την ξένη ιδιοκτησία, ii) ευκολία εισροών/εκροών κεφαλαίου, iii) αποτελεσματικότητα του λειτουργικού πλαισίου, iv) διαθεσιμότητα επενδυτικών εργαλείων, v) σταθερότητα του θεσμικού πλαισίου.

Οι αγορές που είχαν κατηγοριοποιηθεί στο παρελθόν ως ανεπτυγμένες, αναδυόμενες, ή περιθωριακές, και τώρα κατηγοριοποιούνται ως μεμονωμένες, οφείλεται στο γεγονός ότι έχουν υποστεί μια μεγάλη υποβάθμιση είτε στις απαιτήσεις μεγέθους και ρευστότητας των εταιρειών είτε στη προσβασιμότητα της αγοράς. Επιπρόσθετα, ένας δεύτερος λόγος για να ταξινομηθεί μία αγορά ως μεμονωμένη είναι να πληρούνται και τα τρία κριτήρια που έχουν τεθεί από την MSCI, τα οποία όμως δεν πληρούνταν από τη συγκεκριμένη αγορά τα προηγούμενα χρόνια, ή επειδή η αγορά ήταν υποκείμενη σε μία συγκεκριμένη κατηγορία επενδυτών.

Οι Πίνακες 1 και 2 παρουσιάζουν την κατηγοριοποίηση των αγορών, με βάση το πόσο επαρκώς μία χώρα ικανοποιεί τα τρία προαναφερθέντα κριτήρια στο πλαίσιο μεθοδολογίας της MSCI (MSCI, 2023). Πιο αναλυτικά, ο Πίνακας 1 απεικονίζει τις αγορές που ικανοποιούσαν τα κριτήρια για τις ανεπτυγμένες αγορές, ενώ στον Πίνακα 2 παρατίθενται οι αγορές που είχαν κατηγοριοποιηθεί ως αναδυόμενες, περιθωριακές ή μεμονωμένες. Τα ακρώνυμα EMEA και APAC στην πρώτη γραμμή των Πινάκων 1 και 2, υποδηλώνουν την Ευρώπη (Europe) με την Μέση Ανατολή (Middle East) και την Αφρική (Africa), και την Ασία (Asia) με τον Ειρηνικό (Pacific), αντίστοιχα.

Οι αγορές που ανήκουν στον δείκτη MSCI των μεμονωμένων αγορών, παρόλο που δεν περιλαμβάνονται στις κατηγορίες του δείκτη MSCI των αναδυόμενων αγορών και στον δείκτη MSCI των περιθωριακών αγορών, χρησιμοποιούν τα ίδια μεθοδολογικά κριτήρια σχετικά με το μέγεθος και τη ρευστότητα των εταιρειών (κριτήριο 2), όπως οι αγορές που εντάσσονται είτε στον δείκτη MSCI των αναδυόμενων αγορών είτε στον δείκτη MSCI των περιθωριακών αγορών.

Με βάση την αλληλοσυσχέτιση μεταξύ των αναδυόμενων, περιθωριακών και μεμονωμένων αγορών που προαναφέρθηκε προηγουμένως για το κριτήριο 2, όπως και για το κριτήριο 3 που αφορά τη διαθεσιμότητα των επενδυτικών εργαλείων και τη σταθερότητα του θεσμικού πλαισίου (για περισσότερες λεπτομέρειες, MSCI (2023)), σε συνδυασμό με τον διακριτό διαχωρισμό που προτείνεται από το κριτήριο 1 μεταξύ

ανεπτυγμένων αγορών και όλων των υπόλοιπων κατηγοριών, για τους σκοπούς αυτής της εργασίας εφεξής, οι αναδυόμενες αγορές, οι περιθωριακές αγορές, και οι μεμονωμένες αγορές θα αναφέρονται συλλογικά ως αναδυόμενες αγορές.

Πίνακας 1. *Κατηγοριοποίηση Αγορών MSCI – Ανεπτυγμένες Αγορές*

Αμερική	**EMEA**		**APAC**
ΗΠΑ	Αυστρία	Βέλγιο	Αυστραλία
Καναδάς	Γαλλία	Γερμανία	Ιαπωνία
	Δανία	Ελβετία	Νέα Ζηλανδία
	Ηνωμένο Βασίλειο	Ιρλανδία	Σιγκαπούρη
	Ισπανία	Ισραήλ	Χονγκ Κονγκ
	Ιταλία	Νορβηγία	
	Ολλανδία	Πορτογαλία	
	Σουηδία	Φινλανδία	

Στην παρούσα εργασία το σύνολο των δεδομένων αποτελείται από 25 χρηματιστηριακούς δείκτες, εκ των οποίων οι 15 μετοχικοί δείκτες ανήκουν στις ανεπτυγμένες αγορές, ενώ οι 10 ανήκουν στις αναδυόμενες αγορές (Πίνακας 3). Το σύνολο των 25 μετοχικών δεικτών καλύπτει την χρονική περίοδο από το 1987 έως το 2016, εκτός από τους PCOMP, JCI, SMI (1988-2016) και MERVAL (1989-2016), χρησιμοποιώντας ημερήσιες τιμές κλεισίματος των μετοχικών δεικτών (πηγή δεδομένων: βάση δεδομένων Bloomberg). Το έτος 1987 επιλέχθηκε ως το αρχικό έτος για την παρούσα έρευνα, διότι το 1987 ξεκίνησε η τακτική δημοσίευση δεδομένων για την τιμή του Γενικού Δείκτη του Χρηματιστήριου Αθηνών. Η συνολική χρονική περίοδος θα χωριστεί σε έξι πενταετή διαστήματα (1987-1991, 1992-1996, 1997-2001, 2002-2006, 2007-2011, 2012-2016), ώστε να εξεταστεί κατά πόσο μεταβάλλονται τα συμπεράσματα που αφορούν την αποτελεσματικότητα υπό την μορφή ασθενούς ισχύος στις ανεπτυγμένες και αναδυόμενες αγορές, κάτω από διαφορετικές συνθήκες που επικρατούσαν στο παγκόσμιο οικονομικό γίγνεσθαι. Ο έλεγχος διαφορετικών χρονικών περιόδων έχει το πλεονέκτημα ότι επιτρέπει την ανίχνευση διαρθρωτικών αλλαγών, έτσι ώστε η αγορά να είναι ενδεχομένως αποτελεσματική σε ορισμένες περιόδους, ενώ σε άλλες να μην είναι (Borges, 2010). Για αυτόν το λόγο, η επιλογή πενταετών χρονικών διαστημάτων είναι ευρέως χρησιμοποιούμενη στη χρηματοοικονομική βιβλιογραφία (Beck και Levine, 2002; Borges, 2010).

Για τους μετοχικούς δείκτες PCOMP, JCI, και SMI, η αρχική χρονική περίοδος εκτείνεται από το 1988 έως το 1991, ενώ για τον μετοχικό δείκτη MERVAL, η αρχική χρονική περίοδος είναι από το 1989 έως το 1991. Αυτό οφείλεται στο γεγονός ότι οι τέσσερις προαναφερθέντες δείκτες άρχισαν να διαπραγματεύονται είτε το 1988 είτε το 1989. Τα υπόλοιπα πέντε πενταετή διαστήματα παραμένουν ίδια για τους 25 μετοχικούς δείκτες.

Πίνακας 2. *Κατηγοριοποίηση Αγορών MSCI – Αναδυόμενες, Περιθωριακές και Μεμονωμένες Αγορές*

Αναδυόμενες Αγορές			**Περιθωριακές Αγορές**		
Αμερική	**ΕΜΕΑ**	**APAC**	**Αμερική**	**ΕΜΕΑ**	**APAC**
Βραζιλία	Αίγυπτος	Ινδία	-	Ακτή Ελεφαντοστού	Βιετνάμ
Κολομβία	Ελλάδα	Ινδονησία		Εσθονία	Μπαγκλαντές
Μεξικό	Ηνωμένα Αραβικά Εμιράτα	Κίνα		Ιορδανία	Πακιστάν
Περού	Κατάρ	Κορέα		Ισλανδία	Σρι Λάνκα
Χιλή	Κουβέιτ	Μαλαισία		Καζακστάν	
	Νότιος Αφρική	Ταιβάν		Κένυα	
	Ουγγαρία	Ταϊλάνδη		Κροατία	
	Πολωνία	Φιλιππίνες		Λιθουανία	
	Σαουδική Αραβία			Μαρόκο	
	Τουρκία			Μαυρίκιος	
	Τσεχία			Μπαχρέιν	
				Μπενίν	
				Μπουρκίνα Φάσο	
				Νιγηρία	
				Ομάν	
				Ρουμανία	
				Σενεγάλη	
				Σερβία	
				Σλοβενία	
				Τυνησία	

Μεμονωμένες Αγορές			
Αμερική	**ΕΜΕΑ**		**APAC**
Αργεντινή	Βοσνία και Ερζεγοβίνη	Βουλγα-ρία	-
Παναμάς	Ζιμπάμπουε	Λίβανος	
Τζαμάικα	Μποτσουά-να	Μάλτα	
Τρινιντάντ και Τομπάγκο	Ουκρανία	Παλαι-στίνη	

Πίνακας 3. *Δεδομένα*

Μετοχικοί Δείκτες	Γεωγραφική Περιοχή	Χώρα
MSCI Ανεπτυγμένες Αγορές		
AEX	Ευρώπη	Ολλανδία
ATX	Ευρώπη	Αυστρία
CAC	Ευρώπη	Γαλλία
CCMP – NASDAQ	Αμερική	ΗΠΑ
DAX	Ευρώπη	Γερμανία
INDU – Dow Jones Industrial Average	Αμερική	ΗΠΑ
UKX – FTSE 100	Ευρώπη	Ηνωμένο Βασίλειο (Λονδίνο)
HEX	Ευρώπη	Φινλανδία
HIS	Ειρηνικός	Κίνα (Χονγκ Κονγκ)
IBEX	Ευρώπη	Ισπανία
NKY – Nikkei	Ειρηνικός	Ιαπωνία
OMX – Stockholm 30	Ευρώπη	Σουηδία
SPX – S&P 500	Αμερική	ΗΠΑ
SMI	Ευρώπη	Ελβετία
SXXP – STOXX Europe 600	Ευρώπη	-
MSCI Αναδυόμενες Αγορές		
ASE	Ευρώπη	Ελλάδα
FBMKLCI	Ασία	Μαλαισία
JCI	Ασία	Ινδονησία
KOSPI	Ασία	Κορέα
PCOMP	Ασία	Φιλιππίνες
SET	Ασία	Ταϊλάνδη
TWSE	Ασία	Ταιβάν
MSCI Μεμονωμένες Αγορές		
JMSMX	Αμερική	Τζαμάικα
MERVAL	Αμερική	Αργεντινή
MSCI Περιθωριακές Αγορές		
CSEALL	Ασία	Σρι Λάνκα

Οι ανεπτυγμένες αγορές, που χαρακτηρίζονται από υψηλότερο βαθμό ωριμότητας σε σύγκριση με τις αναδυόμενες αγορές, έχει αποδειχθεί ότι είναι αποτελεσματικές πιο συχνά από τις αναδυόμενες αγορές, όπου κατά κύριο λόγο δεν είναι αποτελεσματικές. Το εύρημα αυτό μπορεί να οφείλεται στο γεγονός ότι οι αναδυόμενες αγορές έχουν χαμηλότερη αξία κεφαλαιοποίησης, χαμηλότερο όγκο καθημερινών συναλλαγών, λιγότερες εγγεγραμμένες εταιρίες στο χρηματιστήριο, και γενικά πιο αδιαφανείς λειτουργίες και θεσμικό πλαίσιο σε σχέση με τις ανεπτυγμένες αγορές.

Πιο συγκεκριμένα, οι ανεπτυγμένες αγορές εμφανίζουν μικτά αποτελέσματα ως προς την αποτελεσματικότητα τους, τα οποία εξαρτώνται από την χρονική περίοδο και την

συγκεκριμένη χρηματοοικονομική αγορά που εξετάζει ο εκάστοτε ερευνητής. Μέχρι τη δεκαετία του 1990, είχε εδραιωθεί ότι η υπόθεση της WFME δεν απορριπτόταν στις ανεπτυγμένες αγορές (Fama, 1970; Dryden, 1970; Fama, 1991). Ωστόσο, υπήρχαν και μελέτες που υποδείκνυαν ότι η υπόθεση της WFME απορρίπτεται (Conrad και Jüttner, 1973). Όσον αφορά τις αναδυόμενες αγορές, τα ευρήματα της βιβλιογραφίας καταλήγουν στο συμπέρασμα ότι η υπόθεση της WFME απορρίπτεται (Ayadi και Pyun, 1994; Malafeyev et al., 2019).

Στις μελέτες που εξετάζουν ταυτόχρονα τις αναδυόμενες και ανεπτυγμένες αγορές, οι Milionis et al. (1998) συμπέραναν για το Χρηματιστήριο Αθηνών (ASE), το οποίο ανήκει στις αναδυόμενες αγορές (βλέπε Πίνακα 3), ότι η υπόθεση της WFME απορρίπτεται. Σε αντίθεση, για τους μετοχικούς δείκτες Standard and Poor's 500 (SPX) και FTSE 100 (UKX), οι οποίοι χρησιμοποιούνται για την Νέα Υόρκη και το Λονδίνο αντίστοιχα, οι οποίες ανήκουν στις ανεπτυγμένες αγορές, αποδείχθηκε ότι η υπόθεση της WFME δεν απορρίπτεται.

Η Borges (2010) κατέληξε στο ίδιο συμπέρασμα για τον ASE. Παρόμοιο συμπέρασμα, δηλαδή την απόρριψη της υπόθεσης της WFME, προέκυψε για τους μετοχικούς δείκτες της Πορτογαλίας, της Γαλλίας και του Ηνωμένου Βασιλείου. Αντίθετα, βρέθηκε ότι οι αγορές μετοχών της Γερμανίας και της Ισπανίας είναι αποτελεσματικές αγορές. Από τον Πίνακα 1, είναι προφανές ότι οι μετοχικοί δείκτες της Πορτογαλίας, της Γαλλίας, του Ηνωμένου Βασιλείου, της Γερμανίας και της Ισπανίας κατηγοριοποιούνται ως ανεπτυγμένες αγορές. Οι Dias et al. (2020) κατέληξαν στο συμπέρασμα ότι η υπόθεση της WFME απορρίπτεται μεταξύ 16 μετοχικών δεικτών που εξέτασαν (7 Ευρωπαϊκοί, 6 Ασιατικοί, και 3 Αμερικάνικοι), εκ των οποίων κάποιοι ανήκουν στις ανεπτυγμένες αγορές και κάποιοι στις αναδυόμενες αγορές.

Επομένως, αντικρουόμενα εμπειρικά συμπεράσματα ενδέχεται να προκύψουν σχετικά με τον έλεγχο της WFME μεταξύ διαφορετικών χρονικών περιόδων για τις ίδιες χώρες. Συνεπώς, καθώς η αποτελεσματικότητα της αγοράς παρατηρείται να μεταβάλλεται με τον χρόνο, είναι λογικό το ενδιαφέρον των ερευνητών να παραμένει αμείωτο για τον έλεγχο της αποτελεσματικότητας της αγοράς για τις ίδιες μετοχικές αγορές. Αυτό το ενδιαφέρον εκφράζεται αξιοποιώντας διαφορετικά σύνολα δεδομένων και εφαρμόζοντας διάφορους στατιστικούς ελέγχους, επεκτείνοντας την ήδη εκτεταμένη υπάρχουσα βιβλιογραφία. Ο στόχος αυτής της εργασίας, όπως αναφέρθηκε και προηγουμένως, είναι να εξεταστεί κατά πόσο η στατιστική μεθοδολογία για τον έλεγχο της αποτελεσματικότητας της αγοράς έχει εφαρμοσθεί έγκυρα στη δημοσιευμένη βιβλιογραφία. Μία τέτοια ενδελεχής εξέταση πραγματοποιείται στην ενότητα 4.

4. ΑΝΑΛΥΣΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

Πριν αξιοποιηθούν οι χρονοσειρές για την αξιολόγηση της WFME, αυτές οι σειρές απαιτούν ορισμένη στατιστική προετοιμασία και προ-επεξεργασία, διότι συνήθως δεν είναι αμέσως κατάλληλες στην αρχική τους μορφή. Αυτό συμβαίνει επειδή οι χρονοσειρές των αρχικών δεδομένων συχνά εμφανίζουν μη-στασιμότητα στη διακύμανση. Επομένως, η παρουσία της μη-στάσιμης διακύμανσης μπορεί να επηρεάσει αρνητικά την απόφαση ως προς το αν μια αγορά είναι αποτελεσματική ή δεν είναι αποτελεσματική, το οποίο με τη σειρά του θα έχει δυσμενή αντίκτυπο στις αποφάσεις των επενδυτών σχετικά με την επενδυτική στρατηγική που θα επιδιώξουν να υλοποιήσουν.

4.1 Μετασχηματισμός δεδομένων

Είναι σημαντικό ότι κατά την τυπική διαδικασία ανάλυσης μίας χρονοσειράς τιμών μετοχικού δείκτη, το αρχικό βήμα είναι να αξιολογηθεί εάν τα αρχικά δεδομένα απαιτούν μετασχηματισμό για να γίνει η διακύμανση σταθερή. Αυτή η αξιολόγηση πραγματοποιείται πριν από οποιαδήποτε άλλη ενέργεια, συμπεριλαμβανομένης της κατασκευής του μονομεταβλητού ARIMA υποδείγματος. Ως αποτέλεσμα, είναι σαφές ότι τα αποτελέσματα αυτών των επόμενων ενεργειών επηρεάζονται από την επιλογή που γίνεται σχετικά με τον μετασχηματισμό των δεδομένων.

Τονίζεται το γεγονός ότι οι έλεγχοι αποτελεσματικότητας της αγοράς πραγματοποιούνται συνήθως στις αποδόσεις και όχι στις τιμές, και για αυτό χρησιμοποιείται ο λογαριθμικός μετασχηματισμός. Παρόλο που αυτός ο μετασχηματισμός είναι ο κατάλληλος σε όρους χρηματοοικονομικής, δεν υπάρχει καμία εγγύηση ότι ο λογαριθμικός μετασχηματισμός είναι η σωστή επιλογή σε όρους στατιστικής.

Για να καθοριστεί αν η χρονοσειρά είναι στάσιμη ως προς τη διακύμανση, δύο εξειδικευμένοι αλγόριθμοι θα αξιοποιηθούν για να αξιολογήσουν αν ο μετασχηματισμός των αρχικών δεδομένων είναι απαραίτητος. Ο ένας αλγόριθμος περιλαμβάνεται στο λογισμικό JDemetra+, ενώ ο άλλος αλγόριθμος είναι αυτός που παρουσιάστηκε στην ενότητα 2.

Το JDemetra+ αναπτύχθηκε από την Εθνική Τράπεζα του Βελγίου (National Bank of Belgium) σε συνεργασία με την Eurostat, τη Γερμανική Ομοσπονδιακή Τράπεζα (Deutsche Bundesbank) και το Εθνικό Ινστιτούτο Στατιστικής και Οικονομικών Μελετών της Γαλλίας (National Institute of Statistics and Economic Studies - INSEE), και διατίθεται δωρεάν. Πιο αναλυτικά, το JDemetra+ χρησιμοποιεί τον έλεγχο για τη στασιμότητα της διακύμανσης που περιέχεται στο TRAMO, το οποίο βασίζεται στην εκτίμηση της παραμέτρου λ στο πλαίσιο μετασχηματισμών των Box και Cox (ενότητα 2). Η εκτίμηση αυτή διεξάγεται με την μέθοδο μεγίστης πιθανοφάνειας. Αν $\lambda = 1$, προτείνεται ο λαγαριθμικός μετασχηματισμός, ενώ αν $\lambda = 0$, δεν προτείνεται κάποιος μετασχηματισμός.

Ως εκ τούτου, εφόσον το JDemetra+ επιτρέπει μόνο τον λογαριθμικό μετασχηματισμό, η στατιστική μεθοδολογία M-G που παρουσιάστηκε στην ενότητα 2 θα αξιοποιηθεί. Η προσέγγιση αυτή επιτρέπει τον μετασχηματισμό και με τετραγωνική ρίζα και με αρνητικό αντίστροφο. Επιπρόσθετα, χρησιμοποιείται ο M-G αλγόριθμος, καθώς έχει αποδειχθεί ότι το TRAMO είναι μεροληπτικό ως προς την εφαρμογή του λογαριθμικού μετασχηματισμού (Grudkowska, 2016; Milionis και Galanopoulos, 2018).

Η σύγκριση των αποφάσεων που προκύπτουν από την νέα προτεινόμενη μεθοδολογία και από το JDemetra+ παρουσιάζονται στον Πίνακα 4. Πιο συγκεκριμένα, ο Πίνακας 4 απεικονίζει τα αποτελέσματα σχετικά με την επιλογή εφαρμογής ή μη-εφαρμογής του λογαριθμικού μετασχηματισμού στα δεδομένα των αρχικών χρονοσειρών.

Από τα αποτελέσματα του Πίνακα 4, είναι προφανές ότι η εφαρμογή του λογαριθμικού μετασχηματισμού στις τιμές των χρηματιστηριακών αγορών, που χρησιμοποιείται στην υπάρχουσα χρηματοοικονομική βιβλιογραφία όπως προαναφέρθηκε παραπάνω, είναι λανθασμένη στη πλειοψηφία των περιπτώσεων. Αυτό προκύπτει από το γεγονός ότι με την μεθοδολογία M-G ο λογαριθμικός μετασχηματισμός προτείνεται μόνο στις 25 από τις 150 περιπτώσεις (ή 1/6 περιπτώσεις) που εξετάστηκαν. Σε όλες τις άλλες περιπτώσεις, συγκεκριμένα στις 125 από τις 150 περιπτώσεις (ή 5/6 περιπτώσεις), είτε κάποιος άλλος μετασχηματισμός έπρεπε να εφαρμοστεί (όπως η τετραγωνική ρίζα ή ο αρνητικός αντίστροφος), είτε δεν έπρεπε να εφαρμοστεί κανένας μετασχηματισμός. Πιο αναλυτικά, ο λογαριθμικός μετασχηματισμός στις ανεπτυγμένες αγορές έπρεπε, με βάση τη μεθοδολογία των Milionis και Galanopoulos (2017) να εφαρμοστεί μόνο στο 8.9% των περιπτώσεων (8/90 περιπτώσεις), και στις αναδυόμενες αγορές στο 26.7% των περιπτώσεων (17/60 περιπτώσεις).

4.2 Έλεγχος της WFME

Η ύπαρξη ενός σημείου αναφοράς διευκολύνει τη διαδικασία της σύγκρισης μεταξύ διαφορετικών προσεγγίσεων. Επομένως, προτείνοντας ένα σημείο αναφοράς εξυπηρετείται ο σκοπός για τη δημιουργία μίας ενιαίας βάσης για τη διεξαγωγή συγκρίσεων. Όπως αναφέρθηκε προηγουμένως, οι πρώτες διαφορές των φυσικών λογαρίθμων των τιμών είναι μια κοινή χρησιμοποιούμενη τεχνική στη χρηματοοικονομική και θα αποτελέσει το σημείο αναφοράς στην παρούσα έρευνα. Οι Πρώτες Διαφορές των φυσικών Λογαρίθμων (First Differences of the natural Logarithms) των τιμών θα συμβολίζονται με FDL.

Στις περιπτώσεις όπου προτείνεται ο λογαριθμικός μετασχηματισμός, επειδή όλες οι χρονοσειρές στο εξεταζόμενο σύνολο δεδομένων είναι $I(1)$ (τα πλήρη αποτελέσματα των ελέγχων είναι διαθέσιμα από τους συγγραφείς), το συμπέρασμα που αφορά την αποτελεσματικότητα της αγοράς θα είναι το ίδιο με το συμπέρασμα της μεθόδου FDL.

Πίνακας 4. *Απόφαση σχετικά με τον μετασχηματισμό των δεδομένων*

Μετασχηματισμός δεδομένων		
Μετοχικοί Δείκτες	**M-G**	**JDemetra+**
MSCI Ανεπτυγμένες Αγορές	Log	Log
AEX	1	2
ATX	0	3
CAC	0	3
CCMP	1	3
DAX	1	4
DOWJONES	0	3
FTSE 100	1	2
HEX	1	5
HSI	0	3
IBEX	0	2
Nikkei	0	3
OMX	2	3
S&P 500	1	3
SMI	0	3
SXXP	0	2
Άθροισμα 1	**8 / 90**	**44 / 90**
MSCI Αναδυόμενες Αγορές	**M-G**	**JDemetra+**
ASE	3	6
FBMKLCI	1	3
JCI	0	5
KOSPI	0	3
PCOMP	3	5
SET	4	5
TWSE	1	4
Άθροισμα 2	**12 / 42**	**31 / 42**
MSCI Μεμονωμένες Αγορές	**M-G**	**JDemetra+**
JMSMX	2	6
MERVAL	2	4
Άθροισμα 3	**4 / 12**	**10 / 12**
MSCI Περιθωριακές Αγορές	**M-G**	**JDemetra+**
CSEALL	1	5
Άθροισμα 4	**1 / 6**	**5 / 6**
Άθροισμα 2 + Άθροισμα 3 + Άθροισμα 4	**17 / 60**	**46 / 60**
Συνολικό Άθροισμα	**25 / 150**	**90 / 150**

Αν δεν προτείνεται ο λογαριθμικός μετασχηματισμός (δηλαδή κανένας μετασχηματισμός των δεδομένων στο πλαίσιο μεθοδολογίας του JDemetra+), τότε αν η μεθοδολογία M-G δεν προτείνει επίσης κανένα μετασχηματισμό, το συμπέρασμα που αφορά την αποτελεσματικότητα της αγοράς θα συμπίπτει με την απόφαση της μεθόδου M-G. Τέλος, αν το JDemetra+ δεν προτείνει κάποιον μετασχηματισμό και η μεθοδολογία M-G προτείνει κάποιον μετασχηματισμό (υποθετικά, για παράδειγμα, τον μετασχηματισμό με τετραγωνική ρίζα), τότε εξετάζονται ξεχωριστά και οι τρεις μεθοδολογίες.

Η αξιολόγηση και σύγκριση των αποτελεσμάτων ως προς την αποτελεσματικότητα των αγορών μεταξύ των τριών εναλλακτικών μεθόδων, βασίζεται στην πιθανή ύπαρξη σειριακής συσχέτισης των μετασχηματισμένων και στάσιμων σειρών που προέρχονται από τις αρχικές χρονοσειρές των τιμών των μετοχικών δεικτών, για όλες τις υπό εξέταση κεφαλαιακές αγορές (βλέπε Πίνακα 3). Ο έλεγχος LBQ (Ljung και Box, 1978) θα χρησιμοποιηθεί, διότι λαμβάνει υπόψη του από κοινού αυτοσυσχετίσεις για αρκετές χρονικές υστερήσεις. Εν προκειμένω, αν δεν ανιχνευθούν κάποιες στατιστικά σημαντικές αυτοσυσχετίσεις μέχρι την $30^{η}$ χρονική υστέρηση, τότε η υπόθεση της WFME δεν απορρίπτεται. Αντίθετα, αν ανιχνευθούν στατιστικά σημαντικές αυτοσυσχέτισεις, τότε η υπόθεση της WFME απορρίπτεται, Τα αποτελέσματα που αφορούν την απόφαση σχετικά με τη WFME παρουσιάζονται στον Πίνακα 5.

Υπενθυμίζεται ότι η αξιολόγηση της υπόθεσης της WFME εξετάζεται για κάθε μετοχικό δείκτη σε έξι διαφορετικές χρονικές περιόδους και χρησιμοποιώντας τρεις διαφορετικές μεθόδους. Επιπρόσθετα, το σύμβολο * που εμφανίζεται στα αποτελέσματα του Πίνακα 5 υποδεικνύει ότι: i) είτε δεν υπάρχουν στατιστικά σημαντικές αυτοσυσχετίσεις στις πρώτες 10 (και ενδεχομένως 20) χρονικές υστερήσεις, ενώ υπάρχουν στατιστικά σημαντικά αυτοσυσχετίσεις σε χρονικές υστερήσεις υψηλότερης τάξης. Σε αυτή την περίπτωση, με επιφύλαξη, το αναγραφόμενο αποτέλεσμα είναι ότι η υπόθεση της WFME δεν απορρίπτεται*, ii) είτε ότι υπάρχουν στατιστικά σημαντικά αυτοσυσχετίσεις στις πρώτες 10 (και ενδεχομένως 20) υστερήσεις, ενώ δεν υπάρχουν στατιστικά σημαντικά αυτοσυσχετίσεις σε υστερήσεις υψηλότερης τάξης. Σε αυτή την περίπτωση, με επιφύλαξη, το αναγραφόμενο αποτέλεσμα είναι ότι η υπόθεση της WFME απορρίπτεται*.

Παρουσιάζοντας ένα παράδειγμα, στην έκτη γραμμή του Πίνακα 5, όπου παρουσιάζονται τα αποτελέσματα για τον δείκτη CAC, χρησιμοποιώντας την μέθοδο M-G, στις 4 από τις 6 υπό εξέταση χρονικές περιόδους (δεύτερη στήλη του Πίνακα 5) η υπόθεση της WFME απορρίπτεται, σε αντίθεση με τις 2 από τις 6 χρονικές περιόδους (τρίτη στήλη του Πίνακα 5) όπου δεν απορρίπτεται. Με την μέθοδο FDL, ο συμβολισμός 6 (2*) (τέταρτη στήλη του Πίνακα 5) υποδεικνύει ότι η υπόθεση της WFME απορρίπτεται σε όλες τις υπό εξέταση 6 χρονικές περιόδους, εκ των οποίων στις 2 φορές με επιφύλαξη. Παρόμοια ερμηνεία των αποτελεσμάτων με την μέθοδο FDL προκύπτουν για την μέθοδο JDemetra+ (έκτη και έβδομη στήλη του Πίνακα 5).

Από τα αποτελέσματα παρακάτω, στις ανεπτυγμένες αγορές παρατηρείται ότι με την μέθοδο M-G, η υπόθεση της WFME δεν απορρίπτεται στο 51.1% των εξεταζόμενων περιπτώσεων, ενώ με την εδραιωμένη μέθοδο FDL, η υπόθεση δεν απορρίπτεται στο 26.7% των εξεταζόμενων περιπτώσεων. Χρησιμοποιώντας το λογισμικό JDemetra+, όπου μόνο ο λογαριθμικός μετασχηματισμός επιτρέπεται, η υπόθεση της WFME δεν απορρίπτεται στο 44.4% των εξεταζόμενων περιπτώσεων. Στην υπάρχουσα βιβλιογραφία, ο έλεγχος της υπόθεσης της WFME στις ανεπτυγμένες αγορές δίνει μικτά αποτελέσματα, παρόλο που στα αρχικά χρόνια μελέτης των ανεπτυγμένων αγορών, είχε εδραιωθεί η αποτελεσματικότητα της αγοράς (Brealey και Mayers, 1988). Το συμπέρασμα αυτό υποστηρίζεται κυρίως από την προτεινόμενη μεθοδολογία M-G και έως κάποιο βαθμό από το JDemetra+. Αντίθετα, το σημείο αναφοράς υποδεικνύει μία μεροληψία ως προς την απόρριψη της υπόθεσης της WFME.

Στις αναδυόμενες αγορές, η υπόθεση της WFME απορρίπτεται στο 73.3%, 78.3% και 75% των εξεταζόμενων περιπτώσεων χρησιμοποιώντας τη μεθοδολογία M-G, το σημείο αναφοράς και το JDemetra+, αντίστοιχα. Το εύρημα αυτό επιβεβαιώνει την υπάρχουσα βιβλιογραφία και ισχύει και για τις τρεις διαφορετικές μεθοδολογίες, υποδεικνύοντας ότι η υπόθεση της WFME απορρίπτεται κατά κύριο λόγο στις αναδυόμενες αγορές.

Ιδιαίτερη έμφαση πρέπει να δοθεί στις περιπτώσεις όπου με τον λογαριθμικό μετασχηματισμό, οι μετασχηματισμένες χρονοσειρές των τιμών δεν είναι στάσιμες ως προς τη διακύμανση, και για αυτό τον λόγο, η ανάλυση με τον έλεγχο LBQ που χρησιμοποιείται ευρέως από τους ερευνητές, δεν είναι έγκυρη. Μία τέτοια περίπτωση στην οποία γίνεται εμφανής η πλήρης δυναμική της προτεινόμενης μεθοδολογίας M-G παρουσιάζεται στον Πίνακα 6. Πιο συγκεκριμένα, σε αυτή την περίπτωση, i) με τον λογαριθμικό μετασχηματισμό που προτείνεται από το JDemetra+, η μη-στασιμότητα ως προς τη διακύμανση δεν εξαλείφεται, και ii) προκύπτει διαφορετικό συμπέρασμα ως προς την αποτελεσματικότητα της αγοράς. Δηλαδή, με την μεθοδολογία M-G η αγορά είναι αποτελεσματική, ενώ με τις μεθοδολογίες JDemetra+ και του σημείου αναφοράς, η αγορά δεν είναι αποτελεσματική. Παρόμοια συμπεράσματα προκύπτουν και για τον δείκτη S&P 500 κατά την περίοδο 1992-1996. Περαιτέρω, για τους μετοχικούς δείκτες ATX και CCMP τις χρονικές περιόδους 2002-2006 και 1997-2001 αντίστοιχα, η υπόθεση της WFME απορρίπτεται με τη μεθοδολογία M-G, σε αντίθεση με τις μεθοδολογίες JDemetra+ και του σημείου αναφοράς που υποδεικνύουν τη μη-απόρριψη της υπόθεσης της WFME (Galanopoulos, 2023).

Πίνακας 5. *Απόφαση σχετικά με την WFME*

WFME						
Μετοχικοί Δείκτες	**M-G**		**FDL**		**JDemetra+**	
MSCI Ανεπτυγμένες Αγορές	**Απόρριψη**	**Μη-απόρριψη**	**Απόρριψη**	**Μη-απόρριψη**	**Απόρριψη**	**Μη-απόρριψη**
AEX	3 (1*)	3 (1*)	6 (1*)	0	3 (1*)	3 (1*)
ATX	4	2	5	1 (1*)	5	1
CAC	4	2	6 (2*)	0	6 (2*)	0
CCMP	2	4 (2*)	3	3 (2*)	2	4 (2*)
DAX	3 (1*)	3 (1*)	3	3 (3*)	3 (1*)	3 (2*)
DOWJONES	1	5 (2*)	3	3 (2*)	1	5 (3*)
FTSE 100	5	1	6	0	5	1
HEX	4	2 (1*)	4 (1*)	2 (1*)	4 (1*)	2 (1*)
HSI	2 (1*)	4 (2*)	3 (1*)	3 (2*)	2 (1*)	4 (2*)
IBEX	4	2	6	0	4	2
Nikkei	1	5 (1*)	3 (1*)	3 (1*)	3 (1*)	3
OMX	3 (1*)	3 (1*)	4	2 (2*)	3	3 (1*)
S&P 500	1	5 (3*)	3	3 (2*)	2	4 (2*)
SMI	4	2 (1*)	5	1	4	2 (1*)
SXXP	3	3 (2*)	6	0	3	3 (2*)
Άθροισμα 1	**44 / 90 (4*)**	**46 / 90 (17*)**	**66 / 90 (6*)**	**24 / 90 (16*)**	**50 / 90 (7*)**	**40 / 90 (17*)**
MSCI Αναδυόμενες Αγορές	**Απόρριψη**	**Μη-απόρριψη**	**Απόρριψη**	**Μη-απόρριψη**	**Απόρριψη**	**Μη-απόρριψη**
ASE	4	2 (2*)	5	1	5	1
FBMKLCI	5	1 (1*)	5	1 (1*)	5	1 (1*)
JCI	6	0	6	0	6	0
KOSPI	3	3	3	3	3	3
PCOMP	6	0	6	0	6	0
SET	4	2	4	2	4	2
TWSE	2 (1*)	4 (2*)	3	3	2	4 (2*)
Άθροισμα 2	**30 / 42 (1*)**	**12 / 42 (5*)**	**32 / 42**	**10 / 42 (1*)**	**31 / 42**	**11 / 42 (3*)**
MSCI Μεμονωμένες Αγορές	**Απόρριψη**	**Μη-απόρριψη**	**Απόρριψη**	**Μη-απόρριψη**	**Απόρριψη**	**Μη-απόρριψη**
JMSMX	5	1	5	1	5	1
MERVAL	3	3 (1*)	4 (1*)	2 (1*)	3	3 (2*)
Άθροισμα 3	**8 / 12**	**4 / 12 (1*)**	**9 / 12 (1*)**	**3 / 12 (1*)**	**8 / 12**	**4 / 12 (2*)**

MSCI Περιθωριακές Αγορές						
CSEALL	6	0	6	0	6	0
Άθροισμα 4	**6 / 6**	**0 / 6**	**6 / 6**	**0 / 6**	**6 / 6**	**0 / 6**
Άθροισμα 2 + Άθροισμα 3 + Άθροισμα 4	**44 / 60 (1*)**	**16 / 60 (6*)**	**47 / 60 (1*)**	**13 / 60 (2*)**	**45 / 60**	**15 / 60 (5*)**
Συνολικό Άθροισμα	**88 / 150 (5*)**	**62 / 150 (23*)**	**113 / 150 (7*)**	**37 / 150 (18*)**	**95 / 150 (7*)**	**55 / 150 (22*)**

Πίνακας 6. *Απόφαση σχετικά με τον μετασχηματισμό δεδομένων και την WFME για την σειρά ATX 2002-2006*

Μετασχηματισμός Δεδομένων				
JDemetra+				
Χρονοσειρά με	Αρχικά Δεδομένα			
Απόφαση	Λογάριθμος			
M-G				
Χρονοσειρά με	Αρχικά Δεδομένα			
$\hat{\beta}$	1.21			
Απόφαση	Αρνητικός Αντίστροφος			
Χρονοσειρά με	Αρνητικό Αντίστροφο			
$\hat{\beta}$	Μη στατιστικά σημαντικός			
Στασιμότητα ως προς τη διακύμανση				
Χρονοσειρά με	Αρχικά Δεδομένα	Λογάριθμο	Αρνητικό Αντίστροφο	
Απόφαση	Όχι	Όχι	Ναι	
Χρονοσειρά με	Πρώτες Διαφορές του Αρνητικού Αντίστροφου (M-G)		Πρώτες Διαφορές του Λογαρίθμου (FDL και JDemetra+)	
Υστέρηση	LBQ	p-τιμή	LBQ	p-τιμή
10	6,586	0,764	24,523	0,006
20	23,091	0,284	38,170	0,008
30	35,716	0,218	47,462	0,022
WFME	Δεν απορρίπτεται		Απορρίπτεται	

5. ΣΥΜΠΕΡΑΣΜΑΤΑ, ΠΕΡΙΟΡΙΣΜΟΙ ΚΑΙ ΜΕΛΛΟΝΤΙΚΗ ΕΡΕΥΝΑ

Ενώ η έννοια της αποτελεσματικότητας των αγορών έχει αποτελέσει αντικείμενο μελέτης για αρκετές δεκαετίες, οι ερευνητές χρησιμοποιούν εκτενώς τις αποδόσεις των περιουσιακών στοιχείων, οι οποίες ουσιαστικά είναι οι πρώτες διαφορές των φυσικών λογαρίθμων των τιμών, ως ένα πρότυπο μέτρο, χωρίς να ελέγχουν εάν αυτές οι χρονοσειρές είναι στάσιμες ως προς τη διακύμανση.

Σύμφωνα με την παρούσα μελέτη, είναι σαφές ότι η χρήση του λογαριθμικού μετασχηματισμού για τιμές χρηματιστηριακών δεικτών είναι ανεπαρκής για τη σταθεροποίηση της διακύμανσης στις περισσότερες από τις περιπτώσεις που εξετάστηκαν. Ως αποτέλεσμα, η χρήση των ελέγχων αυτοσυσχέτισης, που συνήθως χρησιμοποιούνται για τον έλεγχο της αποτελεσματικότητας της αγοράς, μπορεί να μην είναι έγκυρη. Πράγματι, στην παρούσα εργασία καταγράφηκε ότι οι έλεγχοι αυτοσυσχέτισης δεν μπορούν να χρησιμοποιηθούν έγκυρα ως στατιστικό εργαλείο για τον έλεγχο της αποτελεσματικότητας της αγοράς στο 83.3% των εξεταζόμενων περιπτώσεων.

Επιπλέον, διαπιστώθηκε ότι η χρήση των πρώτων διαφορών των φυσικών λογαρίθμων των τιμών μπορεί να εισάγει μεροληψία ως προς την απόρριψη της αποτελεσματικότητας ασθενούς ισχύος στις ανεπτυγμένες αγορές. Αυτό το αποτέλεσμα υποστηρίχθηκε από την προτεινόμενη μεθοδολογία M-G και, έως ένα βαθμό, από το λογισμικό JDemetra+. Είναι σημαντικό να σημειωθεί ότι μετά τη διενέργεια των στατιστικών ελέγχων χρησιμοποιώντας τη μεθοδολογία M-G, τα συμπεράσματα σχετικά με την WFME υπόθεση ήρθαν σε αντίθεση με αυτά που προέκυψαν μέσω της εδραιωμένης μεθοδολογίας στο 27.7% των περιπτώσεων για τις ανεπτυγμένες αγορές κεφαλαίου.

Στην περίπτωση των αναδυόμενων αγορών, η χρήση και των τριών μεθόδων που εξετάστηκαν, επανεπιβεβαίωσε την πολυετή ομοφωνία στην υπάρχουσα βιβλιογραφία, δηλαδή, ότι οι αναδυόμενες αγορές δεν είναι συνήθως αποτελεσματικές. Ωστόσο, αξίζει να σημειώσουμε ότι σε αρκετές περιπτώσεις, η εδραιωμένη μεθοδολογία δεν μπορούσε να εφαρμοστεί έγκυρα. Σε αυτές τις αγορές, ο βαθμός της μη-αποτελεσματικότητας είναι τόσο έντονος ώστε να ανιχνεύεται ανεξαρτήτως της μεθόδου που χρησιμοποιείται. Αυτό επιβεβαιώνει τις υπάρχουσες ερευνητικές εργασίες που υποστηρίζουν ότι το επίπεδο ωριμότητας μιας χρηματοοικονομικής αγοράς επηρεάζει την αποτελεσματικότητα της. Επιπλέον, η προηγουμένως αναφερθείσα μεροληψία του JDemetra+ στη χρήση του λογαριθμικού μετασχηματισμού επαληθεύτηκε για μια ακόμη φορά.

Τέλος, σημειώνεται ότι τα συμπεράσματα που σχετίζονται με τη μη-στασιμότητα της διακύμανσης των τιμών των μετοχικών δεικτών και, κατά συνέπεια, με τους ελέγχους αυτοσυσχέτισης στις αποδόσεις των μετοχικών δεικτών, συνδέονται αναπόφευκτα όχι μόνο με το συγκεκριμένο είδος περιουσιακών στοιχείων (δείκτες μετοχών), αλλά και με το συγκεκριμένο χρονικό διάστημα δειγματοληψίας που χρησιμοποιήθηκε (πέντε έτη). Επομένως, δεν πρέπει απαραίτητα να θεωρείται δεδομένο ότι αυτά τα συμπεράσματα ισχύουν πανομοιότυπα για άλλα είδη περιουσιακών στοιχείων, και

ακόμη σημαντικότερα, για τα συγκεκριμένα μεν περιουσιακά στοιχεία, αλλά σε μεγαλύτερα χρονικά διαστήματα. Αυτό αποτελεί ένα πεδίο για μελλοντική έρευνα.

ABSTRACT

In the existing literature on market efficiency testing, stock returns (expressed as first differences of the natural logarithms of prices) are very commonly used, and the stationarity of the corresponding time series in terms of variance is generally taken for granted. The present study shows that the application of the logarithmic transformation to the time series of security prices is, in many cases, insufficient for stabilizing the variance. Consequently, in such cases, the autocorrelation tests widely used to assess market efficiency are not valid. In the present study, by applying the methodology proposed by Milionis – Galanopoulos (Proceedings of the 30th Panhellenic Statistics Conference), it is documented that autocorrelation tests cannot be validly used as a statistical tool for assessing market efficiency in the majority of cases examined.

ΑΝΑΦΟΡΕΣ

Ayadi O. F. and Pyun C. S. (1994). An application of variance ratio test to the Korean securities market. *Journal of Banking and Finance*, **18**, 643–658.

Beck T. and Levine R. (2004). Stock markets, banks, and growth: Panel evidence. *Journal of Banking and Finance*, **28**, 423-442.

Borges M. R. (2010). Efficient market hypothesis in European stock markets. *European Journal of Finance*, **16**, 711–726.

Box G. E. P. and Cox D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society*, **26**, 211-243.

Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control*, San Francisco: Holden-Day.

Brealey, R. A. and Myers, S. C. (1988). *Principles of Corporate Finance*, New York: McGraw Hill.

Conrad K. and Jüttner D. J. (1973). Recent behaviour of stock market prices in Germany and the random walk hypothesis. *Kyklos*, **26**, 576–599.

Dias R., Heliodoro P., Teixeira N. and Godinho T. (2020). Testing the Weak Form of Efficient Market Hypothesis: Empirical Evidence from Equity Markets. *International Journal of Accounting, Finance and Risk Management*, **5**, 40-51.

Dryden M. M. (1970). A statistical study of U.K. share prices. *Scottish Journal of Political Economy*, **17**, 369–389.

Elton, E.J., Gruber, M.J., Brown, S. J. and Goetzmann, W. N. (2014). *Modern Portfolio Theory and Investment Analysis*, New Jersey: John Wiley.

Fama E. F. (1970). Efficient Capital Markets: A review of theory and Empirical work. *The Journal of Finance*, **25**, 383-417.

Fama E. F. (1976). Efficient Capital Markets: Reply. *The Journal of Finance*, **31**, 143-145.

Fama E. F. (1991). Efficient Capital Markets: II. *The Journal of Finance*, **46**, 1575–1617.

Galanopoulos N. G. (2023). *A study of time series with time-varying variance: a practical approach with applications in actuarial science and finance*. Doctoral Thesis.

Gómez V. and Maravall A. (1996). Programmes SEATS and TRAMO: instructions for the User'. *Bank of Spain Working Paper* 9628.

Grudkowska S. (2016). JDemetra+ Reference Manual Version 2.1. *Narodowy Bank Polski, Department of Statistics*, Warsaw, Poland.

Hay A. R. and McCleary R. (1979). Box-Tiao Time Series models for impact assessment. *Evaluation Review*, **3**, 277–314.

LeRoy S. F. (1976). Efficient Capital Markets: Comment. *The Journal of Finance*, **31**, 139-141.

Ljung G. M. and Box G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika,* **65**, 297–303.

Malafeyev O., Awasthi A., Kambekar K. S. and Kupinskaya A. (2019). Random walks and market efficiency in Chinese and Indian equity markets. *Statistics, Optimization and Information Computing*, **7**, 1-25.

Malkiel, B. G. (1992). *Efficient market hypothesis*, In: Newman, P., Milgate, M. and Eatwell, J., Eds., New Palgrave Dictionary of Money and Finance, London: McMillan.

Milionis A. E. (2007). Efficient capital markets: A statistical definition and comments. *Statistics & Probability Letters*, **77**, 607–613.

Milionis A. E. and Davies T. D. (1994). Regression and Stochastic models for Air Pollution, Part II. *Atmospheric Environment*, **28**, 2801-2810.

Milionis A. E., Moschos D. and Xanthakis M. (1998). The influence of foreign markets on the Athens Stock Exchange. *SPOUDAI Journal of Economics and Business*, **48**, 140-156.

Milionis A. E. and Galanopoulos N. G. (2017). Time series with interdependent level and second moment: testing, consequences for modelling; preliminary results. *Proceedings of the 30th Panhellenic Statistics conference*, 109-123.

Milionis A. E. and Galanopoulos N. G. (2018). Time series with interdependent level and second moment: testing, consequences for modelling; applications (part II). *Proceedings of the 31th Panhellenic Statistics conference,* 302-313.

Milionis A. E. and Galanopoulos N. G. (2019). Forecasting economic time series in the presence of variance instability and outliers. *Theoretical Economic Letters*, **9**, 2940-2964.

Milionis A. E. and Galanopoulos N. G. (2020). A study of the effect of data transformation and "linearization" on time series forecasts. A practical approach. *Bank of Greece Working Paper* 280.

Mills, T. C. (1990). *Time Series Techniques for Economists*, Cambridge: Cambridge University Press.

MSCI (2023). MSCI market classification framework, June 2023.

Roberts, H. (1959). Stock-Market "Patterns" and Financial Analysis: Methodological suggestions. *The Journal of Finance*, **14**, 1-10.

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 37ου Πανελληνίου & 3ου Διεθνούς Συνεδρίου Στατιστικής (2025),
σελ.143-154

ΜΟΝΤΕΛΟΠΟΙΗΣΗ ΕΝΟΣ ΟΜΟΓΕΝΟΥΣ ΜΑΡΚΟΒΙΑΝΟΥ ΣΥΣΤΗΜΑΤΟΣ ΩΣ ΙΞΩΔΟΕΛΑΣΤΙΚΟΥ ΜΕΣΟΥ

Α. Μπουκουβάλα[1], Γ. Τσακλίδης[2]
[1]Τμήμα Μαθηματικών, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
annampkv@gmail.com
[2]Τμήμα Μαθηματικών, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
tsaklidi@math.auth.gr

ΠΕΡΙΛΗΨΗ

Στην παρούσα εργασία μελετάται ένα Ομογενές Μαρκοβιανό Σύστημα (ΟΜΣ) και εξετάζεται αν η εξέλιξη των προσιτών δομών (attainable structures) μπορεί να περιγραφεί μέσω της θεωρίας των ιξωδοελαστικών μέσων και συγκεκριμένα του μηχανικού προτύπου Maxwell. Το ΟΜΣ θεωρείται συνεχές μέσο, του οποίου η κίνηση αντιστοιχεί στην εξέλιξη της δομής του συστήματος. Υπό την παραδοχή ότι η μεταβολή κάθε προσιτής δομής του ΟΜΣ καθορίζεται από την αλληλεπίδρασή της με το περιβάλλον, η συνολική εξέλιξη του ΟΜΣ μπορεί να προσεγγιστεί ως κίνηση ενός συνεχούς μέσου. Σε αυτό το πλαίσιο, μελετώνται οι εξισώσεις του Kolmogorov και υπολογίζονται οι συνιστώσες του τανυστή ρυθμού παραμόρφωσης και του τανυστή τάσης. Τα αποτελέσματα δείχνουν ότι η ερμηνεία της εξέλιξης ενός ΟΜΣ δύο διαστάσεων ως ιξωδοελαστικού μέσου τύπου Maxwell δύο διαστάσεων απορρίπτεται, καθιστώντας αδύνατη την ανάλυση ορισμένων σημαντικών ιδιοτήτων, όπως η ενεργειακή κατάσταση και η εντροπία του συστήματος, υπό αυτό το πρίσμα.

Λέξεις Κλειδιά: Ιξωδοελαστικό Μέσο, Maxwell, Μαρκοβιανό Σύστημα, Μοντελοποίηση

1. ΕΙΣΑΓΩΓΗ

Η δυναμική ενός Ομογενούς Μαρκοβιανού Συστήματος (ΟΜΣ) περιγράφεται μέσω του συνόλου των προσιτών δομών του. Εφόσον θεωρηθεί ότι κάθε προσιτή δομή του ΟΜΣ αντιστοιχεί σε ένα σημείο - σωματίδιο του $\mathbb{R}^n$, καθίσταται δυνατή η μελέτη της εξέλιξης του συστήματος ως κίνηση των σημείων - σωματιδίων στο χώρο. Υπό την παραδοχή ότι η κίνηση κάθε σωματιδίου εξαρτάται από την αλληλεπίδρασή του με το περιβάλλον, το ΟΜΣ προσεγγίζεται ως συνεχές μέσο, η παραμόρφωση του οποίου απεικονίζει τη μεταβολή του συστήματος.

Η εν λόγω προσέγγιση εισάγει τη δυνατότητα εφαρμογής εννοιών της μηχανικής των συνεχών μέσων στη μελέτη Μαρκοβιανών Συστημάτων. Ειδικότερα, εξετάζεται η

σύνδεση της εξέλιξης του ΟΜΣ με τη θεωρία των ιξωδοελαστικών υλικών, μέσω του μηχανικού προτύπου Maxwell. Εύλογα τίθεται το ερώτημα αν το σύνολο των προσιτών δομών ενός n-διαστατού ΟΜΣ, θεωρούμενου ως συνεχές μέσο, παραμορφώνεται κατά την κίνησή του ως ιξωδοελαστικό μέσο.

Αντικείμενο της παρούσας εργασίας αποτελεί η διερεύνηση ενός διδιάστατου ΟΜΣ ως συνεχούς μέσου, ώστε να διαπιστωθεί αν η χρονική του εξέλιξη περιγράφεται με όρους της θεωρίας των συνεχών μέσων με ιξωδοελαστική συμπεριφορά. Προς επίτευξη του ανωτέρω σκοπού, μελετώνται οι διαφορικές εξισώσεις του Kolmogorov, υπολογίζονται η παραμόρφωση, η ταχύτητα, η επιτάχυνση του ΟΜΣ και εξετάζεται η συμβατότητα των ανωτέρω με την κίνηση ενός ιξωδοελαστικού μέσου κινούμενου βάσει του προτύπου Maxwell .

Η σύνδεση μεταξύ στοχαστικών διαδικασιών και θεωρίας συνεχών μέσων συμβάλλει ουσιαστικά στην κατανόηση πολύπλοκων συστημάτων, παρέχοντας ένα εργαλείο μοντελοποίησης, που συσχετίζει την εξέλιξη ενός στοχαστικού μοντέλου με τη φυσική περιγραφή της παραμόρφωσης και της τάσης ενός ρευστού.

Στην παρούσα εργασία αποδεικνύεται ότι η ερμηνεία της εξέλιξης ενός ΟΜΣ δύο διαστάσεων ως ιξωδοελαστικού μέσου τύπου Maxwell δύο διαστάσεων απορρίπτεται. Επομένως, δεν είναι δυνατή η εξέταση σημαντικών ιδιοτήτων της εξέλιξης υπό αυτό το πρίσμα, όπως της ενεργειακής κατάστασης του ΟΜΣ και της εντροπίας του. Το παρόν συμπέρασμα προκύπτει υπό ορισμένες γενικές παραδοχές και συγκεκριμένα υπό την υπόθεση χρονικής ομογένειας του ΟΜΣ, της αδιαχωριστότητας του πίνακα τάσεων Q, της απουσίας μαζικών δυνάμεων στο συνεχές ανάλογο, καθώς και της χωρικής ομογένειας της πυκνότητας, η οποία λαμβάνεται ως συνάρτηση μόνο του χρόνου.

Το μη αποδεκτό της συγκεκριμένης θεώρησης έρχεται σε αντίθεση με τις περιπτώσεις εξέτασης της εξέλιξης ενός ΟΜΣ ως ελαστικού μέσου (Maaita, Tsaklidis, & Meletidou, 2013) ή ιξωδοελαστικού μέσου τύπου Kelvin-Voigt (Loumponias & Tsaklidis, 2017).

2. ΒΑΣΙΚΕΣ ΑΡΧΕΣ ΚΙΝΗΣΗΣ ΤΟΥ ΟΜΣ ΥΠΟ ΤΗ ΘΕΩΡΗΣΗ ΤΟΥ ΣΥΝΕΧΟΥΣ ΜΕΣΟΥ

2.1 Η Έννοια του συνεχούς μέσου

Το συνεχές μέσο αποτελεί μοντέλο στο οποίο η ύλη θεωρείται κατανεμημένη συνεχώς στον όγκο της, ώστε να καλύπτει ολοκληρωτικά τον χώρο που καταλαμβάνει, χωρίς διακοπές ή κενά - ενώ, με πιο αυστηρή διατύπωση, η κίνηση ενός σημείου του συνεχούς μέσου καθορίζεται από την επίδραση της γειτονίας του πάνω σε αυτό. Η θεμελιώδης αυτή παραδοχή επιτρέπει την περιγραφή της μακροσκοπικής συμπεριφοράς των υλικών μέσω μαθηματικών συναρτήσεων. Στο πλαίσιο αυτό, φυσικά μεγέθη, όπως η τάση και η μετατόπιση εκφράζονται ως συνεχείς και διαφορίσιμες συναρτήσεις του χώρου και του χρόνου.

Η μελέτη του συνεχούς μέσου πραγματοποιείται με την εισαγωγή του όρου σωματίδιο, ως στοιχειώδη όγκο του μέσου, ο οποίος προσδιορίζεται από τις χωρικές συντεταγμένες του. Επομένως, η δομή του μέσου συνίσταται από το σύνολο των σωματιδίων του και η διαμόρφωσή του κάθε χρονική στιγμή ορίζεται μέσω της αντιστοίχισης των σωματιδίων σε σημεία του χώρου. Επιπλέον, η έννοια της παραμόρφωσης δηλώνει την αλλαγή σχήματος του συνεχούς μέσου, συγκρίνοντας την αρχική και τελική διαμόρφωσή του (Χατζηδημητρίου & Μπόζης, 1997).

2.2 Περιγραφή της κίνησης

Η περιγραφή της κίνησης ή παραμόρφωσης ενός συνεχούς μέσου πραγματοποιείται ως προς ένα σύστημα αναφοράς, το οποίο λαμβάνεται ως Καρτεσιανό. Η μελέτη υλοποιείται, είτε μέσω σωματιδιακής περιγραφής κατά Lagrange, παρακολουθώντας τη μεταβολή των μεγεθών κατά μήκος της τροχιάς ενός συγκεκριμένου σωματιδίου, είτε μέσω χωρικής περιγραφής κατά Euler, εξετάζοντας τις μεταβολές σε σταθερό σημείο του χώρου. Συνεπώς η κίνηση ενός συνεχούς μέσου στον $\mathbb{R}^3$ με χρήση σωματιδιακών συντεταγμένων X_i (περιγραφή κατά Lagrange) δίνεται από την σχέση

$$x_i(X_1, X_2, X_3, t), i = 1,2,3,$$

όπου x_i είναι η τρέχουσα θέση του σωματιδίου που βρισκόταν στο σημείο (X_1, X_2, X_3) τη χρονική στιγμή $t = 0$. Επίσης, η κίνηση ενός συνεχούς μέσου με χρήση χωρικών συντεταγμένων x_i (περιγραφή κατά Euler) δίνεται από την σχέση

$$\mathrm{X}_i(x_1, x_2, x_3, t), i = 1,2,3,$$

η οποία περιγράφει την τροχιά του σωματιδίου, που στην παρούσα φάση καταλαμβάνει το σημείο (x_1, x_2, x_3) (Χατζηδημητρίου & Μπόζης, 1997).

Κάθε χρονική στιγμή t, το συνεχές μέσο καταλαμβάνει έναν όγκο V, καλύπτει μία οριοθετημένη επιφάνεια S και εκτείνεται σε μια συγκεκριμένη περιοχή R του φυσικού χώρου. Η αναγνώριση των τμημάτων του συνεχούς μέσου και των σημείων του χώρου που καταλαμβάνουν τη χρονική στιγμή t σε σχέση με ένα κατάλληλο σύστημα συντεταγμένων, ορίζει την διαμόρφωση του συνεχούς μέσου στον χρόνο t. Με τον όρο παραμόρφωση εννοούμε την αλλαγή του σχήματος του συνεχούς μέσου μεταξύ μιας αρχικής και μιας επόμενης διαμόρφωσης. Αν η αρχική και η τελική διαμόρφωση του συνεχούς μέσου απεικονίζονται σε δύο Καρτεσιανά συστήματα συντεταγμένων, με άξονες $OX_1X_2..X_n$ και $Ox_1x_2..x_n$ αντίστοιχα, γειτονικά σωματίδια που βρίσκονται στα σημεία P_0, Q_0 στην αρχική κατάσταση, μετακινούνται στα σημεία P και Q κατά την τελική διαμόρφωση. Η ποσότητα $(dx)^2 - (dX)^2$, που εκφράζει τη διαφορά των τετραγώνων των αποστάσεων δύο γειτονικών σωματιδίων πριν και μετά την παραμόρφωση, χρησιμοποιείται ως μέτρο της παραμόρφωσης στην περιοχή μεταξύ τους και υποδεικνύει άκαμπτη μετατόπιση αν λάβει την τιμή μηδέν (Mase, 1970).

Ορίζεται ο τανυστής τροπών κατά Lagrange ως

$$l_{ij} = \frac{1}{2}\left(\frac{dx_k}{dX_i}\frac{dx_k}{dX_j} - \delta_{ij}\right) = \left(\frac{du_i}{dX_j} + \frac{du_j}{dX_i} - \frac{du_m}{dX_i}\frac{du_m}{dX_j}\right)$$

και ο τανυστής τροπών κατά Euler ως

$$\varepsilon_{ij} = \frac{1}{2}\left(\delta_{ij} - \frac{dX_k}{dx_i}\frac{dX_k}{dx_j}\right) = \left(\frac{du_i}{dx_j} + \frac{du_j}{dx_i} - \frac{du_m}{dx_i}\frac{du_m}{dx_j}\right),$$

όπου

$$u_i = x_i - X_i.$$

Τότε, η διαφορά των τετραγώνων των αποστάσεων δύο γειτονικών σωματιδίων πριν και μετά την παραμόρφωση δίνεται από

$$(dx)^2 - (dX)^2 = 2l_{ij}dX_i \cdot dX_j = 2\varepsilon_{ij}dx_i \cdot dx_j$$

όπου

$(dX)^2 = \delta_{ij}dX_i \cdot dX_{j,}$ $(dx)^2 = \delta_{ij}dx_i \cdot dx_j$ (Reddy, 2013).

2.3 Τάση

Έστω $\boldsymbol{n} = (n_i)\,, i = 1,2, \dots, n$ το μοναδιαίο εξωτερικό διάνυσμα στο σημείο P, το οποίο ανήκει σε ένα στοιχειώδες τμήμα ΔS της επιφάνειας του μέσου. Επιπλέον, ας θεωρήσουμε ότι Δf_i είναι η δύναμη που ασκείται επί της στοιχειώδους επιφάνειας ΔS από το εξωτερικό περιβάλλον κατά τη διεύθυνση του άξονα i. Ως διάνυσμα τάσης ορίζεται το

$$t_i^{(\boldsymbol{n})} = \lim_{\Delta S \to 0} \frac{\Delta f_i}{\Delta S} = \frac{df_i}{dS}.$$

Το σύνολο των πιθανών ζευγών διανύσματος τάσης $t_i^{(\boldsymbol{n})}$- μοναδιαίου διανύσματος $\boldsymbol{n}$, που μπορούν να οριστούν για συγκεκριμένο σημείο P, καλείται κατάσταση τάσης του σημείου και περιγράφεται πλήρως μέσω του τανυστή τάσης t_{ij} (Capaldi, 2012). Ο τανυστής τάσης συνδέεται με το διάνυσμα τάσης μέσω της σχέσης, $t_i^{(\boldsymbol{n})} = t_{ji}n_j$, η οποία σε αναλυτική μορφή γράφεται

$$\begin{bmatrix} t_1^{\boldsymbol{n}} \\ t_2^{\boldsymbol{n}} \\ t_3^{\boldsymbol{n}} \end{bmatrix} = \begin{bmatrix} t_{11} & t_{12} & t_{13} \\ t_{21} & t_{22} & t_{23} \\ t_{31} & t_{32} & t_{33} \end{bmatrix} \begin{bmatrix} n_1 \\ n_2 \\ n_3 \end{bmatrix}.$$

Τα διαγώνια στοιχεία του τανυστή τάσης ονομάζονται ορθές τάσεις και καθορίζουν την αλλαγή του όγκου, ενώ τα υπόλοιπα στοιχεία του τανυστή ονομάζονται διατμητικές τάσεις και σχετίζονται με τη μεταβολή της γωνίας τεμνόμενων ευθειών του συνεχούς μέσου κατά την παραμόρφωση. Μια ορθή τάση είναι θετική όταν προκαλεί διόγκωση του συνεχούς μέσου (εφελκυστική) και αρνητική όταν προκαλεί συμπίεση του συνεχούς μέσου (θλιπτική) (Χατζηδημητρίου & Μπόζης, 1997).

2.4 Εξίσωση ισορροπίας και κίνησης

Στο πλαίσιο της μηχανικής των συνεχών μέσων, οι δυνάμεις διακρίνονται σε μαζικές και επιφανειακές. Οι μαζικές δυνάμεις δρουν σε όλο τον όγκο του μέσου και εκφράζονται, είτε ανά μονάδα μάζας ($\boldsymbol{b}$), είτε ανά μονάδα όγκου ($\boldsymbol{p}$). Αντίθετα, οι επιφανειακές δυνάμεις ($\boldsymbol{f}$) επιδρούν σε συγκεκριμένα σημεία, είτε περιοχών της επιφάνειας, είτε εσωτερικών επιφανειών και εκφράζονται ως δυνάμεις ανά μονάδα επιφάνειας.

Η συνθήκη ισορροπίας για ένα συνεχές μέσο, το οποίο υπόκειται σε επιφανειακές και μαζικές δυνάμεις, διατυπώνεται με την απαίτηση το άθροισμα των ολοκληρωμάτων των επιφανειακών και μαζικών δυνάμεων να είναι μηδέν, δηλαδή

$$\int_S t_i^{(\boldsymbol{n})}\, dS + \int_V \rho b_i\, dV = 0.$$

Αντικαθιστώντας $t_i^{(n)} = t_{ji} n_j$, *προκύπτει*

$$\int_S t_{ji}\, n_j\, dS + \int_V \rho b_i\, dV = 0,$$

Ενώ, με χρήση του θεωρήματος του Gauss η σχέση γράφεται

$$\int_V t_{ji,j}\, dV + \int_V \rho b_i\, dV = 0.$$

Αφού η προηγούμενη σχέση ισχύει για κάθε όγκο V, είναι επιτρεπτό να γίνει απαλοιφή του ολοκληρώματος, οπότε τελικά η συνθήκη ισορροπίας είναι

$$t_{ji,j} + \rho b_i = 0.$$

Βάσει της προηγούμενης διαπίστωσης αποδεικνύεται η εξίσωση ισορροπίας ως

$$div\boldsymbol{T}(\boldsymbol{z},t) + \rho(\boldsymbol{z},t)\boldsymbol{b}(\boldsymbol{z},t) = \boldsymbol{0},$$

όπου ρ η πυκνότητα του συνεχούς μέσου, **z** το διάνυσμα θέσης του σημείου P, όπου **T**(**z**,t) είναι ο τανυστής τάσεων του Cauchy με συνιστώσες t_{ij} και t ο χρόνος (Mase, 1970). Εν συνεχεία, η προσθήκη του όρου της επιτάχυνσης ($\boldsymbol{\alpha}$) στο ισοζύγιο δυνάμεων οδηγεί τελικά στη διατύπωση της εξίσωσης κίνησης του Cauchy,

$$div\boldsymbol{T}(\boldsymbol{z},t) + \rho(\boldsymbol{z},t)\boldsymbol{b}(\boldsymbol{z},t) = \rho\boldsymbol{\alpha}.$$

Θεώρημα (Gauss)
Το ολοκλήρωμα του όγκου V σχετίζεται με το ολοκλήρωμα της επιφάνειας S και για ένα διάνυσμα του χώρου $v = v(x)$, εκφράζεται ως:

$$\int_V v_{i,i}\, dV = \int_S v_i\, n_i\, dS$$

όπου n είναι το εξωτερικό μοναδιαίο διάνυσμα κάθετο στο σύνορο της επιφάνειας S.

2.5 Ιξωδοελαστική συμπεριφορά υλικών

Τα ελαστικά και τα ιξώδη υλικά εμφανίζουν διαφορετικές ιδιότητες ως προς την παραμόρφωση και τη χρονική απόκριση υπό εφαρμοζόμενη τάση. Η ιξωδοελαστική συμπεριφορά περιγράφει υλικά που συνδυάζουν χαρακτηριστικά τόσο ελαστικών όσο και ιξωδών σωμάτων, με αποτέλεσμα η παραμόρφωση να εξαρτάται από το χρόνο και να μην υπακούει σε αποκλειστικά γραμμική σχέση. Η πολυπλοκότητα της ανάλυσης τέτοιων υλικών απαιτεί τη χρήση μηχανικών προτύπων, τα οποία αποτελούνται από ιδεατά ελατήρια και αποσβεστήρες, προκειμένου να προσομοιωθούν η ελαστική απόκριση και η ιξώδης ροή, αντίστοιχα *(Mase, 1970; Ferry, 1980)*.

Η ελαστική συμπεριφορά του ελατηρίου περιγράφεται από τον νόμο του Hooke,

$$\sigma = G\varepsilon,$$

όπου σ είναι η τάση, ε η παραμόρφωση και G η παράμετρος του ελατηρίου, ενώ η ιξώδης συμπεριφορά του αποσβεστήρα περιγράφεται από τον νόμο του Newton για τα ιξώδη ρευστά,

$$\sigma = \eta \frac{d\varepsilon}{dt},$$

όπου η είναι το ιξώδες. Εάν μία από τις δύο αυτές ιδιότητες παραλειφθεί, το αντίστοιχο μοντέλο εκφυλίζεται, είτε σε καθαρά ελαστικό, είτε σε καθαρά ιξώδες (νευτώνειο ή μη) πρότυπο (Tschoegl, 1989; Christensen, 2012).

Το απλό γραμμικό ελαστικό πρότυπο αποτελεί την απλούστερη προσέγγιση της μηχανικής συμπεριφοράς ενός υλικού και βασίζεται αποκλειστικά στον νόμο του Hooke. Στο πλαίσιο αυτό, η παραμόρφωση εξαρτάται άμεσα και γραμμικά από την εφαρμοζόμενη τάση, χωρίς να λαμβάνονται υπόψη χρονικά εξαρτώμενα φαινόμενα ή μηχανισμοί απόσβεσης. Παρά την περιορισμένη του ικανότητα στην περιγραφή ιξωδοελαστικών φαινομένων, το γραμμικό ελαστικό πρότυπο χρησιμοποιείται συχνά ως σημείο αναφοράς ή ως πρώτη προσέγγιση στη μελέτη της μηχανικής απόκρισης υλικών (Mase, 1970; Christensen, 2012).

Το μηχανικό πρότυπο Maxwell είναι ένα από τα θεμελιώδη πρότυπα ιξωδοελαστικής συμπεριφοράς και χρησιμοποιείται για την περιγραφή χρονικά εξαρτώμενων φαινομένων υπό εφαρμοζόμενη τάση. Το πρότυπο απαρτίζεται από ένα ελατήριο με παράμετρο G και έναν αποσβεστήρα με παράμετρο η, συνδεδεμένους σε διάταξη σειράς *(Ferry, 1980; Christensen, 2012)*. Στην περίπτωση αυτή, η τάση που αναπτύσσεται στα δύο στοιχεία είναι κοινή, επομένως

$$\sigma = \sigma_{αποσβεστήρα} = \sigma_{ελατηρίου},$$

ενώ η συνολική παραμόρφωση ισούται με το άθροισμα των επιμέρους παραμορφώσεων του ελατηρίου και του αποσβεστήρα, δηλαδή

$$\varepsilon = \varepsilon_{αποσβεστήρα} + \varepsilon_{ελατηρίου}.$$

Η καταστατική εξίσωση του προτύπου Maxwell, η οποία συσχετίζει την τάση και την παραμόρφωση, προκύπτει από τον συνδυασμό των επιμέρους νόμων (Ferry, 1980) και λαμβάνει τη μορφή

$$\frac{d\sigma}{dt} + \frac{G}{\eta}\sigma = G\frac{d\varepsilon}{dt}.$$

Το πρότυπο Maxwell περιγράφει με ακρίβεια τη χαλάρωση τάσης υπό σταθερή παραμόρφωση, ωστόσο υπό σταθερή εφαρμοζόμενη τάση προβλέπει απεριόριστη αύξηση της παραμόρφωσης, γεγονός που περιορίζει την ικανότητά του να αποδώσει ρεαλιστικά φαινόμενα ερπυσμού (Findley et al., 1976; Ferry, 1980).

Ένα εναλλακτικό ιξωδοελαστικό πρότυπο αποτελεί το πρότυπο Kelvin–Voigt, στο οποίο το ελατήριο και ο αποσβεστήρας συνδέονται σε διάταξη παραλληλίας (Ferry, 1980; Christensen, 2012). Στο συγκεκριμένο μοντέλο, η παραμόρφωση είναι κοινή και στα δύο στοιχεία, επομένως

$$\varepsilon = \varepsilon_{αποσβεστήρα} = \varepsilon_{ελατηρίου},$$

ενώ η συνολική τάση ισούται με το άθροισμα των τάσεων που αναπτύσσονται στο ελατήριο και στον αποσβεστήρα, δηλαδή

$$\sigma = \sigma_{αποσβεστήρα} + \sigma_{ελατηρίου}.$$

Η αντίστοιχη καταστατική εξίσωση (Ferry, 1980) δίνεται από τη σχέση

$$\sigma = G\varepsilon + \eta\frac{d\varepsilon}{dt}.$$

Το πρότυπο Kelvin–Voigt περιγράφει ικανοποιητικά φαινόμενα ερπυσμού, καθώς προβλέπει πεπερασμένη παραμόρφωση υπό σταθερή τάση, ωστόσο αδυνατεί να αποδώσει φαινόμενα χαλάρωσης τάσης, γεγονός που περιορίζει την εφαρμοσιμότητά του σε συγκεκριμένα προβλήματα (Findley et al., 1976; Ferry, 1980).

Για την πληρέστερη περιγραφή της ιξωδοελαστικής συμπεριφοράς πραγματικών υλικών έχουν προταθεί οι γενικευμένες μορφές των προτύπων Maxwell και Kelvin–Voigt. Το γενικευμένο πρότυπο Maxwell προκύπτει από τη σύνδεση πολλαπλών στοιχειωδών προτύπων Maxwell σε παραλληλία, ενώ το γενικευμένο πρότυπο Kelvin–Voigt από τη σύνδεση στοιχειωδών προτύπων Kelvin–Voigt σε σειρά (Serra-Aguila, Puigoriol-Forcada, Reyes & Menacho, 2019).

3. ΜΟΝΤΕΛΟΠΟΙΗΣΗ ΤΟΥ ΟΜΣ ΩΣ ΙΞΩΔΟΕΛΑΣΤΙΚΟ ΜΕΣΟ

3.1 Το ΟΜΣ σε συνεχή χρόνο ως συνεχές μέσο

Στην παρούσα ενότητα, εξετάζεται η συμπεριφορά ενός ΟΜΣ σε συνεχή χρόνο, με χώρο καταστάσεων S={1,2,...,n} και με δεδομένο ότι κάθε μέλος είναι εφικτό να βρίσκεται σε μία μόνο κατάσταση κάθε χρονική στιγμή t.

Για κάθε μέλος είναι δυνατή η μετάβαση από μία κατάσταση i σε μία άλλη κατάσταση j, εντός του χρονικού διαστήματος $[t, T]$, με πιθανότητα μετάβασης $p_{ij}[\Delta t]$ και το πλήθος μελών του συστήματος λαμβάνεται ως σταθερό.

Θεωρώντας ότι κάθε εφικτή δομή του συστήματος αντιστοιχεί σε ένα σημείο του $\mathbb{R}^n$, η δυναμική εξέλιξη του συστήματος απεικονίζεται ως η κίνηση των σημείων στον $\mathbb{R}^n$. Εφόσον η κίνηση κάθε σημείου επηρεάζεται διαρκώς από το περιβάλλον του, το ΟΜΣ δύναται να αντιμετωπιστεί ως συνεχές μέσο, με την εξέλιξή του να εκφράζεται ως παραμόρφωση του μέσου.

Συγκεκριμένα, η εξίσωση του Kolmogorov

$$\dot{\boldsymbol{x}}^T(t) = \boldsymbol{x}^T(t)\boldsymbol{Q}$$

(όπου $\boldsymbol{x}(t)\, nx1$ το διάνυσμα κατάστασης του ΟΜΣ, $\boldsymbol{Q}$ ο πίνακας τάσεων των πιθανοτήτων μετάβασης) περιγράφει την κίνηση μιας στοχαστικής διάταξης στον χώρο $\mathbb{R}^n$, ενώ κάθε διάταξη του ΟΜΣ που ικανοποιεί την εν λόγω εξίσωση αντιμετωπίζεται ως σημείο του χώρου $\mathbb{R}^n$. Άμεσα προκύπτει ότι η κίνηση (ταχύτητα) κάθε σημείου καθορίζεται αποκλειστικά από τη θέση του, $\boldsymbol{x}(t)$, ή - ισοδύναμα - από την επίδραση της γειτονίας του $\boldsymbol{x}(t)$ πάνω του. Η θεώρηση αυτή επιτρέπει την υιοθέτηση της μοντελοποίησης του ΟΜΣ ως συνεχές μέσο.

Η λύση της διαφορικής εξίσωσης του Kolmogorov $\boldsymbol{x^T}(t) = \boldsymbol{x^T}(0)e^{\boldsymbol{Q}t},\ t > 0$ προσδιορίζει την τροχιά κάθε διανύσματος κατάστασης $\boldsymbol{x}(t)$, για δεδομένο αρχικό διάνυσμα $\boldsymbol{x}(0)$. Λαμβάνοντας υπόψη τη στοχαστική ιδιότητα έπεται ότι η κίνηση εξελίσσεται επί της υπερεπιφάνειας (Π): $x_1 + \cdots + x_n = 1$.

Χωρίς βλάβη της γενικότητας, ο πίνακας $\boldsymbol{Q}$ θεωρείται αδιαχώριστος. Στο στοχαστικό σημείο ισορροπίας $\boldsymbol{\pi}$ ορίζεται Καρτεσιανό σύστημα συντεταγμένων $\{\boldsymbol{f}_1, \boldsymbol{f}_2, \dots, \boldsymbol{f}_n\}$, τέτοιο ώστε $\boldsymbol{f}_i \epsilon (\Pi), i = 1,2, \dots, n-1$ και $\boldsymbol{f}_{\mathrm{n}} \perp (\Pi)$. Έστω

$$\boldsymbol{F} = [\boldsymbol{f}_1 | \dots | \boldsymbol{f}_n] = [\boldsymbol{F}_1 | \dots | \boldsymbol{f}_n],$$

όπου οι πίνακες $\boldsymbol{F}$ και $\boldsymbol{F}_1$ έχουν ως στήλες τα διανύσματα $\boldsymbol{f}_1, \dots, \boldsymbol{f}_n$ και $\boldsymbol{f}_1, \dots, \boldsymbol{f}_{n-1}$, αντίστοιχα. Η εξίσωση του Kolmogorov στο νέο σύστημα συντεταγμένων λαμβάνει την μορφή

$$\dot{\boldsymbol{z}}^{\boldsymbol{T}}(t) = \boldsymbol{z}^{\boldsymbol{T}}(t)\boldsymbol{W}$$

όπου $\boldsymbol{z}(t)\ (n-1)x1$ διάνυσμα και $\mathbf{W}$=$\boldsymbol{F}_1{}^T\boldsymbol{Q}\boldsymbol{F}_1$.

Με τον προαναφερθέντα μετασχηματισμό, διαπιστώνεται ότι έχει επιτευχθεί μείωση κατά μία μονάδα στο πλήθος των διαφορικών εξισώσεων που περιγράφουν τη δυναμική του μέσου. Επιπλέον, δεδομένου ότι εκ κατασκευής του πίνακα τάσεων $\boldsymbol{Q}$, ισχύει $tr(\boldsymbol{W}) = tr(\boldsymbol{Q}) < 0$, το πεδίο είναι συμπιεστό. Τέλος, δίνεται η δυνατότητα υπολογισμού της ταχύτητας $\boldsymbol{v}(\boldsymbol{z}, t)$ και της επιτάχυνσης $\boldsymbol{\alpha}(\boldsymbol{z}, t)$ στο νέο σύστημα συντεταγμένων με βάση την $\{\boldsymbol{f}_1, \dots, \boldsymbol{f}_{n-1}\}$ (συνεπώς συναρτήσει του πίνακα $\boldsymbol{W}$) βάσει των τύπων,

$$\boldsymbol{\upsilon}(\boldsymbol{z},t) = \boldsymbol{W}\cdot\boldsymbol{z}(t)$$

$$\boldsymbol{\alpha}(\boldsymbol{z},t) = (\boldsymbol{W}^2)^T\cdot\boldsymbol{z}(t).$$

3.2 Διερεύνηση της δυνατότητας ερμηνείας της εξέλιξης ενός 2D ΟΜΣ ως ιξωδοελαστικού μέσου

Έστω ΟΜΣ σε συνεχή χρόνο, με δισδιάστατο χώρο καταστάσεων, S={1,2}. Τότε ο πίνακας τάσεων πιθανοτήτων μετάβασης έχει μορφή $\boldsymbol{Q} = \begin{pmatrix} -\alpha & \alpha \\ \beta & -\beta \end{pmatrix}, \alpha,\beta > 0$ και η στάσιμη κατανομή είναι $\boldsymbol{\pi} = \left(\frac{\beta}{\alpha+\beta} \quad \frac{\alpha}{\alpha+\beta}\right)^T$.

Εισάγουμε νέο ορθοκανονικό σύστημα συντεταγμένων με αρχή το σημείο $\boldsymbol{\pi}$ και βάση $\{\boldsymbol{f}_1,\boldsymbol{f}_2\}$, όπου έστω $\boldsymbol{f}_1 = \left(\frac{\sqrt{2}}{2} \quad -\frac{\sqrt{2}}{2}\right)^T$, $\boldsymbol{f}_2 = \left(\frac{\sqrt{2}}{2} \quad \frac{\sqrt{2}}{2}\right)^T$. Όπως προαναφέρθηκε, η εξίσωση του Kolmogorov στο νέο σύστημα συντεταγμένων λαμβάνει πλέον τη μορφή

$$\dot{z}(t) = z(t)W \Leftrightarrow z(t) = z(0)e^{Wt},$$

όπου $z(t)$ παριστά εδώ και στο εξής $1x1$ διάνυσμα (αριθμό) και όμοια

$$W = \boldsymbol{f}_1^T\cdot\boldsymbol{Q}\cdot\boldsymbol{f}_1 = -(\alpha+\beta) < 0.$$

Επιπλέον, υπολογίζονται: 1. η ταχύτητα $\upsilon(z,t) = -(\alpha+\beta)\cdot z(t)$,
2. η επιτάχυνση $\alpha(z,t) = (\alpha+\beta)^2 z(t)$,
3. η μετατόπιση $u(z;t,t+\Delta t) = z(t)(e^{W\Delta t}-1)$,
4. η αριθμητική έκφραση της τροπής $\varepsilon = \frac{1}{2}(e^{2Wt}-1)$.

Ακολούθως, χρησιμοποιείται η εξίσωση συνέχειας, η οποία περιγράφει τη μεταβολή της πυκνότητας ενός συνεχούς μέσου σε συνάρτηση με το πεδίο ταχύτητας. Η εξίσωση διατυπώνεται υπό την υπόθεση ότι δεν υπάρχουν μαζικές δυνάμεις,

$$\frac{d\rho}{dt} + \rho\cdot div\upsilon = 0.$$

Θεωρώντας την πυκνότητα ανεξάρτητη της θέσης και συνάρτηση του χρόνου προκύπτει ότι $\rho(t) = \rho(0)e^{-Wt}$. Επιπλέον, λαμβάνεται $\rho(0) = 1$, οπότε εν τέλει η πυκνότητα δίνεται ως $\rho(t) = e^{-Wt}$.

Βάσει των ανωτέρω διατυπώνεται και αποδεικνύεται η ακόλουθη πρόταση.

Πρόταση Δεν υφίστανται συναρτήσεις $G(z)$ και $\eta(z)$ (συναρτήσεις παραμέτρων ελατηρίου και αποσβεστήρα αντίστοιχα), συμβατές με τις θεμελιώδεις φυσικές και μαθηματικές απαιτήσεις, ώστε να ικανοποιούνται οι εξισώσεις που διέπουν το μοντέλο περιγραφής του Ομογενούς Μαρκοβιανού Συστήματος (ΟΜΣ) ως ιξωδοελαστικού μέσου τύπου Maxwell.

Απόδειξη
Ξεκινώντας από την εξίσωση κίνησης του Cauchy και εφαρμόζοντας κατάλληλες αντικαταστάσεις προκύπτει,

$$\frac{\partial \sigma}{\partial z} = \rho(t)a(z,t) \Leftrightarrow \frac{\partial \sigma}{\partial z} = e^{-wt}(\alpha+\beta)^2 z,$$

από όπου ολοκληρώνοντας ως προς z λαμβάνεται

$$\sigma(z,t) = \frac{1}{2}e^{-wt}(\alpha+\beta)^2 z^2 + c(t).$$

Επιπλέον, παραγωγίζοντας ως προς t, προκύπτει

$$\frac{\partial \sigma}{\partial t} = -\frac{1}{2}we^{-wt}(\alpha+\beta)^2 z^2 + c'(t).$$

Στη συνέχεια, η καταστατική εξίσωση του προτύπου Maxwell γράφεται ως

$$\left(-w + \frac{G}{\eta}\right)e^{-wt}z^2\frac{k}{2} + \frac{G}{\eta}c(t) + c'(t) = Gwe^{2wt},$$

όπου $w = -(\alpha+\beta) < 0$ και $k = (\alpha+\beta)^2$.

Η προηγούμενη σχέση αποτελεί θεμελιώδη εξίσωση της διερεύνησης. Συγκεκριμένα, διερευνώνται ασυμπτωτικά δύο μορφές της $c(t)$, οι $c(t) = c$, με $c\epsilon\mathbb{R}$, ως η απλούστερη δυνατή μορφή και $c(t) = \sum_{\kappa=0}^{\nu} \alpha_\kappa t^\kappa$, ως υπολογιστικά διαχειρίσιμη προσέγγιση συναρτήσεων $c(t)$, οι οποίες αναπτύσσονται ως σειρές Taylor γύρω από το $t = 0$, με άπειρη ακτίνα σύγκλισης.

Περίπτωση 1 $c(t) = c$, με $c\epsilon\mathbb{R}$ (σταθερή συνάρτηση)
Το πρώτο μέλος της εξίσωσης τείνει στο άπειρο εκθετικά καθώς $t \to +\infty$, λόγω της μορφής του W. Αντιθέτως, το δεύτερο μέλος τείνει στο μηδέν. Επομένως, η εξίσωση δεν μπορεί να ικανοποιηθεί ασυμπτωτικά και η υπόθεση $c(t) = c$ απορρίπτεται ως μη συμβατή με τις συνθήκες του μοντέλου.

Περίπτωση 2 $c(t) = \sum_{\kappa=0}^{\nu} \alpha_\kappa t^\kappa$ (πολυωνυμική προσέγγιση συναρτήσεων αναλυτικών στο $t = 0$ με άπειρη ακτίνα σύγκλισης)
Εξετάζονται τα ενδεχόμενα $\alpha_\nu > 0$ και $\alpha_\nu < 0$.
Και στις δύο περιπτώσεις, το πρώτο μέλος της εξίσωσης τείνει στο άπειρο εκθετικά, καθώς $t \to +\infty$, ενώ το δεύτερο μέλος εξακολουθεί να μηδενίζεται ασυμπτωτικά. Συνεπώς, ούτε αυτές οι μορφές της $c(t)$ είναι συμβατές με τη θεμελιώδη εξίσωση. Δηλαδή, υπό την υπόθεση ότι η $c(t)$ είναι αναλυτική στο $t = 0$ με άπειρη ακτίνα σύγκλισης και προσεγγίσιμη μέσω πεπερασμένων πολυωνύμων Taylor, δεν υπάρχει μορφή της $c(t)$ τέτοια ώστε να ικανοποιείται η υπό διερεύνηση εξίσωση.

Κατά συνέπεια, στο πλαίσιο των ανωτέρω γενικών υποθέσεων, η μοντελοποίηση του ΟΜΣ ως ιξωδοελαστικού μέσου βάσει του προτύπου Maxwell δεν επιτρέπει ύπαρξη λύσης, συμβατής με τις βασικές φυσικές και μαθηματικές απαιτήσεις του μοντέλου και απορρίπτεται. ■

Συμπερασματικά, διαπιστώθηκε ότι δεν είναι δυνατός ο προσδιορισμός των συναρτήσεων των παραμέτρων, ώστε να ικανοποιούν ταυτόχρονα τις εξισώσεις του μοντέλου και τις απαιτούμενες συνθήκες του συστήματος. Επομένως, η περιγραφή του ΟΜΣ ως ιξωδοελαστικού μέσου τύπου Maxwell δεν είναι μαθηματικά συνεπής, οπότε οποιαδήποτε περαιτέρω σχετική θεώρηση, π.χ. για την ενεργειακή εξέλιξη του συστήματος ή την εντροπία του στο πλαίσιο της συγκεκριμένης μοντελοποίησης δεν είναι επιτρεπτή. Το παρόν συμπέρασμα προκύπτει υπό ορισμένες γενικές παραδοχές και συγκεκριμένα υπό την υπόθεση χρονικής ομογένειας του ΟΜΣ, της αδιαχωριστότητας του πίνακα τάσεων Q, της απουσίας μαζικών δυνάμεων στο συνεχές ανάλογο, καθώς και της χωρικής ομογένειας της πυκνότητας, η οποία λαμβάνεται ως συνάρτηση μόνο του χρόνου.

Η απόρριψη της ερμηνείας του ΟΜΣ ως ιξωδοελαστικού μέσου τύπου Maxwell οφείλεται στις - γενικές μεν αλλά - καθοριστικές προαναφερθείσες προϋποθέσεις σε συνδυασμό με την επιπλέον χαρακτηριστική δομική συνθήκη ότι το πρότυπο Maxwell ενσωματώνει μηχανισμό μακροχρόνιας μνήμης μέσω της διαφορικής εξίσωσης εξέλιξης της τάσης, ενώ η μαρκοβιανή δυναμική χαρακτηρίζεται από εκθετική απώλεια πληροφορίας και απουσία ιστορικής εξάρτησης. Αντιθέτως, τα ελαστικά (Maaita, Tsaklidis, & Meletidou, 2013) και Kelvin–Voigt (Loumponias & Tsaklidis, 2017) πρότυπα, τα οποία δεν περιλαμβάνουν μακροχρόνια μνήμη, παραμένουν συμβατά.

Περαιτέρω, η χρήση πεπερασμένων πολυωνυμικών προσεγγίσεων για τη συνάρτηση $c(t)$ δεν αποτελεί σημαντική βλάβη της γενικότητας του αποτελέσματος, καθώς οι προσεγγίσεις αυτές προκύπτουν από την ανάπτυξη Taylor αναλυτικών συναρτήσεων και επιτρέπουν τη μελέτη της ασυμπτωτικής συμπεριφοράς του συστήματος. Το αρνητικό αποτέλεσμα δεν οφείλεται στην αποκοπή της σειράς, αλλά στη δομική ασυμμετρία μεταξύ των εκθετικών όρων που εμφανίζονται στη θεμελιώδη εξίσωση. Θεωρητικά, μόνο συναρτήσεις με εκθετική ή υπερεκθετική αύξηση θα μπορούσαν να εξισορροπήσουν τους σχετικούς όρους, ωστόσο τέτοιες μορφές παραβιάζουν βασικές φυσικές απαιτήσεις, όπως η ύπαρξη πεπερασμένης ενέργειας και η ασυμπτωτική σταθερότητα, ενώ οι ταλαντωτικές συναρτήσεις οδηγούν σε δυναμικές ασύμβατες με τη μαρκοβιανή σύγκλιση. Κατά συνέπεια, το συμπέρασμα της ασυμβατότητας με το πρότυπο Maxwell δεν εξαρτάται από την πολυωνυμική προσέγγιση, αλλά αποτελεί δομικό χαρακτηριστικό της υπό μελέτη δυναμικής ενός συνεχούς μέσου τύπου Maxwell κάτω από τις δεδομένες προαναφερθείσες συνθήκες ομογένειας και απουσίας μαζικών δυνάμεων. Υπό το πρίσμα αυτό, αναμένεται ότι παρόμοια εμπόδια θα προκύπτουν και σε υψηλότερες διαστάσεις.

ABSTRACT

In this paper we study a Homogeneous Markovian System (HMS) and examine whether the evolution of attainable structures can be described by the theory of viscoelastic media, and in particular by the Maxwell mechanical model. The HMS is considered as a continuous medium, whose motion corresponds to the evolution of the system structure. Under the assumption that the change of each attainable structure of the HMS is determined by its interaction with the environment, the overall evolution of the HMS can be approximated as the motion of a continuous medium. In this context, Kolmogorov's equations are investigated and the components of the strain rate tensor and the stress tensor are calculated. The results indicate that interpreting the evolution of a two-dimensional HMS as a two-dimensional Maxwell-type viscoelastic medium is rejected, making it impossible to analyze certain important properties, such as the system's energy state and entropy, from this perspective.

ΑΝΑΦΟΡΕΣ

Capaldi, F. M. (2012). *Continuum Mechanics: Constitutive Modeling of Structural and Biological Materials*. Cambridge University Press.

Christensen, R. M. (2012). *Theory of viscoelasticity: An introduction* (2nd ed.). Academic Press.

Ferry, J. D. (1980). *Viscoelastic properties of polymers* (3rd ed.). John Wiley & Sons.

Findley, W. N., Lai, J. S., & Onaran, K. (1976). *Creep and relaxation of nonlinear viscoelastic materials*. North-Holland.

Maaita, J., Tsaklidis, G., & Meletlidou, E. (2013). The homogeneous Markov system as an elastic medium: The 3D case. *Communications in Statistics – Theory and Methods, 42*, 2259–2270. https://doi.org/10.1080/03610926.2013.766337

Loumponias, K., & Tsaklidis, G. (2017). Interpretation of the evolution of the homogeneous Markov system (or, equivalently, of the embedded Markov chain) as the deformation of a viscoelastic medium: The 3-D case. *Methodology and Computing in Applied Probability, 19*, 137–156. https://doi.org/10.1007/s11009-017-9568-1

Mase, G. (1970). *Continuum Mechanics*. Schaums Outline Series. McGraw-Hill.

Reddy, J. N. (2013). *An Introduction to Continuum Mechanics* (2nd ed.). Cambridge University Press.

Serra Aguila, A., Puigoriol Forcada, J. M., Reyes, G., & Menacho, J. (2019). Viscoelastic models revisited: Characteristics and interconversion formulas for generalized Kelvin–Voigt and Maxwell models. *Acta Mechanica, 230*, 2435–2454. https://doi.org/10.1007/s10409-019-00895-6

Tschoegl, N. W. (1989). *The phenomenological theory of linear viscoelastic behavior*. Springer.

Χατζηδημητρίου, Ι., & Μπόζης, Γ. (1997). *Εισαγωγή στη Μηχανική συνεχών μέσων* (2η έκδ.). Α. Τζιόλα Ε.

Ελληνικό Στατιστικό Ινστιτούτο

Πρακτικά 37ου Πανελληνίου & 3ου Διεθνούς Συνεδρίου Στατιστικής (2025), σελ.155-176

ΕΜΠΕΙΡΙΚΗ ΔΙΕΡΕΥΝΗΣΗ ΜΕΤΡΩΝ ΣΥΣΧΕΤΙΣΗΣ ΜΕ ΤΕΧΝΙΚΕΣ ΚΑΙ ΕΡΓΑΛΕΙΑ ΑΝΑΛΥΤΙΚΗΣ ΔΕΔΟΜΕΝΩΝ

Θεόδωρος Ξανθόπουλος[1] και Κυριακή Τσιλίκα[1]

[1]ΠΜΣ Εφαρμοσμένη Οικονομική (κατεύθυνση Αναλυτικής Δεδομένων), Τμήμα Οικονομικών Επιστημών, Πανεπιστήμιο Θεσσαλίας

{txanthopoulos, ktsilika}@uth.gr

ΠΕΡΙΛΗΨΗ

Η ανάλυση συσχέτισης διαθέτει έναν αριθμό τεχνικών και μέτρων συσχέτισης, ανάλογα με το είδος των δεδομένων που αναφέρεται. Πέρα από τους συντελεστές συσχέτισης Pearson, Spearman, Kendall's Tau, Hoeffding's D, νέες μετρικές, όπως η συσχέτιση απόστασης (Distance Correlation) και ο μέγιστος συντελεστής πληροφορίας (Maximal Information Coefficient) έχουν αναπτυχθεί για την ανάλυση πολύπλοκων και μη γραμμικών σχέσεων. Με τη διάδοση των ελεύθερων λογισμικών προγραμματισμού, οι τεχνικές συσχέτισης έχουν εξελιχθεί υπολογιστικά. Η εύληπτη παρουσίαση των πινάκων συσχέτισης με εικόνες και η πολυμεταβλητή ανάλυση συσχέτισης σε ένα output, είναι από τα πλεονεκτήματα που μας παρέχει η χρήση εργαλείων αναλυτικής δεδομένων. Στην παρούσα εργασία τέτοια αποτελέσματα παρουσιάζονται στη γλώσσα R. Στην πλατφόρμα αναλυτικής δεδομένων KNIME, διερευνώνται οι υπολογιστικές προσεγγίσεις της συσχέτισης με οπτικό προγραμματισμό και παρουσιάζεται η παραγωγή αποτελεσμάτων με ενσωμάτωση της R στο KNIME.

Λέξεις Κλειδιά: ανάλυση συσχέτισης, γλώσσα υπολογιστικής στατιστικής R, οπτικός προγραμματισμός, KNIME analytics platform.

1. ΕΙΣΑΓΩΓΗ

Η ανάλυση συσχέτισης αποτελεί μια θεμελιώδη στατιστική τεχνική, με εκτεταμένη χρήση στη διερεύνηση και στην ανάλυση δεδομένων, επιτρέποντας την κατανόηση της σχέσης μεταξύ μεταβλητών και την εξαγωγή κρίσιμων συμπερασμάτων. Αν και η ποσοτικοποίηση της συσχέτισης γίνεται συνήθως με τα παραδοσιακά μέτρα, όπως ο συντελεστής Pearson ή οι μη παραμετρικοί Spearman's Rho και Kendall's Tau, οι αυξανόμενες ανάγκες ανάλυσης σύνθετων, μη γραμμικών σχέσεων οδήγησαν στην ανάπτυξη νέων μετρικών όπως ο συντελεστής Hoeffding's D (Hoeffding, 1948), η Συσχέτιση Απόστασης (Distance Correlation) (Clark, 2013) και ο Μέγιστος Συντελεστής Πληροφορίας (Maximal Information Coefficient) (Reshef et al., 2011). Τα τελευταία 5 χρόνια αναπτύσσονται όλο και ανθεκτικότερα μέτρα συσχέτισης που δεν επηρεάζονται από ακραίες τιμές. Συντελεστές όπως οι Biweight Midcorrelation, Percentage Bend Correlation, Shepherd's Pi και Gamma Coefficient, είναι σχεδιασμένοι ώστε να εμφανίζουν μεγαλύτερη ανθεκτικότητα σε ακραίες τιμές και παρατηρήσεις που προέρχονται από εξωτερικούς ή ασυνήθιστους παράγοντες (Clark, 2013). Αυτές οι μετρικές επιτρέπουν την αξιόπιστη εκτίμηση της συσχέτισης ακόμη και σε περιπτώσεις που τα παραδοσιακά μέτρα ενδέχεται να επηρεάζονται αρνητικά από ακραία δεδομένα, ενισχύοντας έτσι την ακρίβεια της ανάλυσης.

Παράλληλα, η πρόοδος των εργαλείων ανάλυσης δεδομένων, όπως η γλώσσα προγραμματισμού R και η πλατφόρμα αναλυτικής δεδομένων KNIME, έχει ενισχύσει σημαντικά τις δυνατότητες υπολογισμού των μέτρων συσχέτισης, προσφέρει δυνατότητες οπτικοποίησης των συσχετίσεων μεταξύ δύο ή περισσότερων μεταβλητών και παρέχεται η δυνατότητα συνδυαστικών αναλύσεων.

Η παρούσα εργασία εξετάζει τη χρήση αυτών των εργαλείων για τον υπολογισμό και τη γραφική απόδοση παραδοσιακών και σύγχρονων μέτρων συσχέτισης, αναδεικνύοντας τις δυνατότητες και τους περιορισμούς τους στην ανάλυση οικονομικών και επιχειρηματικών δεδομένων (Halkos et al., 2022; Halkos & Tsilika, 2018; Tinbergen, 1947).

2. ΜΕΘΟΔΟΛΟΓΙΑ

Για τη μελέτη χρησιμοποιήθηκαν δύο εργαλεία ανάλυσης δεδομένων: η γλώσσα προγραμματισμού R (R Core Team, 2025) και η πλατφόρμα αναλυτικής δεδομένων KNIME (Berthold et al., 2007). Στο περιβάλλον της R αξιοποιήθηκαν εξειδικευμένα πακέτα (libraries), μεταξύ των οποίων τα Hmisc (Frank E, 2025), psych (Revelle W, 2025), Minerva (Albanese et al., 2013), energy (Rizzo & Székely, 2019), cca (González & Déjean, 2025), car (Fox & Weisberg, 2019), και correlation (Makowski et al., 2020), PerformannceAnalytics (Peterson & Carl, 2022), tseries (Trapletti and Hornik, 2024), rcompanion (Mangiafico, 2025) και corrplot (Wei et al., 2017), τα οποία επιτρέπουν τον υπολογισμό τόσο κλασικών όσο και σύγχρονων συντελεστών συσχέτισης με χρήση κώδικα. Στο KNIME με την προσέγγιση του οπτικού προγραμματισμού (visual programming) ο χρήστης σχεδιάζει προγράμματα χρησιμοποιώντας γραφικά στοιχεία, όπως κουτιά, διαγράμματα και συνδέσεις (ροές), αντί να γράφει παραδοσιακό κώδικα. Χρησιμοποιούνται εικονικά blocks ή κόμβοι που αναπαριστούν λειτουργίες ή εντολές. Η σύνδεση των στοιχείων (π.χ. με βέλη ή γραμμές) καθορίζει τη ροή εκτέλεσης. Το περιβάλλον είναι συνήθως drag-and-drop και είναι πολύ φιλικό προς αρχάριους χρήστες.

Στις ενότητες που ακολουθούν, μέτρα συσχέτισης όπως των Pearson, Spearman's Rho και Kendall's Tau, Hoeffding's D, Biweight Midcorrelation, Percentage Bend Correlation, Shepherd's Pi, ο συντελεστής απόστασης, ο μέγιστος συντελεστής πληροφορίας, ο συντελεστής Φ (Phi Coefficient), διαγνωστικά μέτρα πολυσυγγραμμικότητας, η ανάλυση κανονικής συσχέτισης, υπολογίστηκαν στα περιβάλλοντα αναλυτικής δεδομένων R και KNIME, με και χωρίς κώδικα, ώστε να συγκριθούν τα αποτελέσματα και να παρουσιαστεί η συνεργασία των δύο εργαλείων. Η διερεύνηση αυτή μας επιτρέπει να δούμε πως τα μέτρα συσχέτισης εφαρμόζονται σε διαφορετικού τύπου δεδομένα και τι πληροφορίες μπορεί να προσφέρει η χρήση του κάθε εργαλείου.

3. ΥΠΟΛΟΓΙΣΤΙΚΗ ΠΡΟΣΕΓΓΙΣΗ ΤΩΝ ΜΕΤΡΩΝ ΣΥΣΧΕΤΙΣΗΣ ΜΕΣΩ ΤΗΣ ΓΛΩΣΣΑΣ ΥΠΟΛΟΓΙΣΤΙΚΗΣ ΣΤΑΤΙΣΤΙΚΗΣ R

Ο συντελεστής συσχέτισης Pearson γνωστός και ως Pearson's r μετρά τη γραμμική σχέση μεταξύ δύο συνεχών, ποσοτικών, γραμμικών μεταβλητών και λαμβάνει τιμές από -1 έως 1 (Lind et al., 2012). Τιμές κοντά στο +1 ή στο -1 υποδεικνύουν ισχυρή θετική ή αρνητική συσχέτιση αντίστοιχα, ενώ τιμές κοντά στο 0 δείχνουν έλλειψη γραμμικής σχέσης (Clark, 2013).

Στην εμπειρική διερεύνηση του Pearson Correlation Coefficient, θα εξετάσουμε τις σχέσεις μεταξύ των βασικών οικονομικών μεταβλητών

(Ακαθάριστου Εγχώριου Προϊόν, Ανεργία, Εισαγωγές, Εξαγωγές, Επενδύσεις και Όγκος λιανικών πωλήσεων) της Ευρωπαϊκής Ένωσης για τη χρονική περίοδο 2005-2022. Τα δεδομένα συλλέχθηκαν από την βάση δεδομένων FRED (Federal Reserve Economic Data), η οποία παρέχει πραγματικά τριμηνιαία μακροοικονομικά δεδομένα. Προκειμένου να διασφαλιστεί η εγκυρότητα των αποτελεσμάτων, εφαρμόστηκε αρχικά έλεγχος στασιμότητας (ADF test) σε όλες τις χρονοσειρές με αυτόματη επιλογή υστερήσεων σύμφωνα με τον κανόνα $k = \lfloor (n-1)^{(1/3)} \rfloor$, ο οποίος για το συγκεκριμένο πλήθος παρατηρήσεων επέστρεψε 4 υστερήσεις. Δεδομένου ότι διαπιστώθηκε μη στασιμότητα, οι συσχετίσεις υπολογίστηκαν στις δεύτερες διαφορές, ώστε να διασφαλιστεί η εγκυρότητα των αποτελεσμάτων. Εφόσον διαπιστώθηκε στασιμότητα στις δεύτερες διαφορές, πραγματοποιήθηκε ο υπολογισμός συντελεστών Pearson. Στην εικόνα 1 παρουσιάζονται τα αποτελέσματα όπως υπολογίστηκαν στο περιβάλλον της R με τη χρήση των πακέτων hmisc (Frank E, 2025) PerformanceAnalytics (Peterson & Carl, 2022) για τις βασικές μακροοικονομικές μεταβλητές της ανάλυσης. Οι συσχετίσεις Pearson, Distance Correlation, Partial Correlation και Hoeffding's D εφαρμόστηκαν στις δεύτερες διαφορές, συνεπώς οι τιμές ADF παραμένουν κοινές για όλα τα μέτρα συσχέτισης και αναφέρονται μία φορά στον πίνακα ADF στην εικόνα 1α.

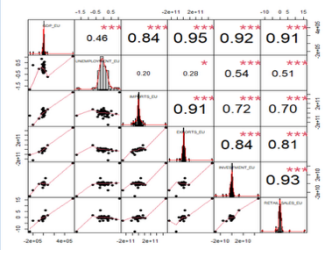
Εικόνα 1. *Υπολογισμός του συντελεστή συσχέτισης Pearson στη γλώσσα προγραμματισμού R (α) με το πακέτο Hmisc και (β) PerformanceAnalytics.*

```
> pearson_results
                GDP_EU UNEMPLOYMENT_EU IMPORTS_EU EXPORTS_EU INVESTMENT_EU
GDP_EU            1.00             0.46       0.84       0.95          0.92
UNEMPLOYMENT_EU   0.46             1.00       0.20       0.28          0.54
IMPORTS_EU        0.84             0.20       1.00       0.91          0.72
EXPORTS_EU        0.95             0.28       0.91       1.00          0.84
INVESTMENT_EU     0.92             0.54       0.72       0.84          1.00
RETAIL_SALES_EU   0.91             0.51       0.70       0.81          0.93
                RETAIL_SALES_EU
GDP_EU                     0.91
UNEMPLOYMENT_EU            0.51
IMPORTS_EU                 0.70
EXPORTS_EU                 0.81
INVESTMENT_EU              0.93
RETAIL_SALES_EU            1.00

n= 67

P
                GDP_EU UNEMPLOYMENT_EU IMPORTS_EU EXPORTS_EU INVESTMENT_EU
GDP_EU                      0.0001       0.0000     0.0000        0.0000
UNEMPLOYMENT_EU 0.0001                   0.1030     0.0231        0.0000
IMPORTS_EU      0.0000      0.1030                  0.0000        0.0000
EXPORTS_EU      0.0000      0.0231       0.0000                   0.0000
INVESTMENT_EU   0.0000      0.0000       0.0000     0.0000
RETAIL_SALES_EU 0.0000      0.0000       0.0000     0.0000        0.0000
                RETAIL_SALES_EU
GDP_EU                   0.0000
UNEMPLOYMENT_EU          0.0000
IMPORTS_EU               0.0000
EXPORTS_EU               0.0000
INVESTMENT_EU            0.0000
RETAIL_SALES_EU

> print(adf_second_diff_clean)
   ADF_stat p_value Lags_k          Variable
1 -5.718429    0.01      4            GDP_EU
2 -4.181770    0.01      4 UNEMPLOYMENT_EU
3 -4.574061    0.01      4       IMPORTS_EU
4 -5.149713    0.01      4       EXPORTS_EU
5 -5.910903    0.01      4    INVESTMENT_EU
6 -5.687676    0.01      4 RETAIL_SALES_EU
```



(α) *(β)*

Ο συντελεστής Biweight Midcorrelation (bicor) αποτελεί μια ανθεκτική εναλλακτική του Pearson, καθώς μειώνει την ευαισθησία σε ακραίες τιμές. Για την εφαρμογή του χρησιμοποιήθηκε το σύνολο δεδομένων "Wholesale customers" από το UCI Machine Learning Repository, που περιλαμβάνει δεδομένα για τις ετήσιες αγορές πελατών. Η ανάλυση πραγματοποιήθηκε στη γλώσσα προγραμματισμού R με τη χρήση του πακέτου Correlation, καθώς επιστρέφει τον συντελεστή Biweight Midcorrelation ανά ζεύγος μεταβλητών,

το 95% διάστημα εμπιστοσύνης και τον έλεγχο σημαντικότητας για κάθε συντελεστή. Ο μαθηματικός τύπος είναι:

$$\frac{\Sigma W_i^x w_i^y (x_i - M_x)(y_i + M_y)}{\sqrt{\Sigma (W_i^x (x_i - M_X) \cdot \Sigma \left(w_i^y (y - M_y)\right)^2}} \quad (1)$$

Όπου: x_i, y_i Οι παρατηρήσεις των μεταβλητών **X** και Y. M_x, M_y η διάμεσος των μεταβλητών **X**, και Y, αντί για μέσο όρο (λόγω ανθεκτικότητας σε ακραίες τιμές). w_i^x, w_i^y τα βάρη για κάθε παρατήρηση, τα οποία εξαρτώνται από την απόσταση του κάθε x_i ή y_i από τη διάμεσο τους.

Εικόνα 2. *Υπολογισμός του συντελεστή συσχέτισης Biweight Mid Correlation Coefficient με το πακέτο correlation.*

```
> bicor_result <- correlation(df, method = "biweight", p_adjust = "none")
> bicor_result
# Correlation Matrix (biweight-method)

Parameter1        |        Parameter2 |     r |          95% CI | t(438) |          p
-------------------------------------------------------------------------------------
Fresh             |             Milk | -0.08 | [-0.17,  0.02] |  -1.61 | 0.108
Fresh             |          Grocery | -0.10 | [-0.19,  0.00] |  -2.01 | 0.046*
Fresh             |           Frozen |  0.29 | [ 0.20,  0.37] |   6.27 | < .001***
Fresh             | Detergents_Paper | -0.07 | [-0.16,  0.02] |  -1.47 | 0.143
Fresh             |       Delicassen |  0.22 | [ 0.12,  0.30] |   4.61 | < .001***
Milk              |          Grocery |  0.72 | [ 0.67,  0.76] |  21.63 | < .001***
Milk              |           Frozen | -0.11 | [-0.20, -0.01] |  -2.23 | 0.026*
Milk              | Detergents_Paper |  0.44 | [ 0.36,  0.51] |  10.25 | < .001***
Milk              |       Delicassen |  0.28 | [ 0.19,  0.37] |   6.14 | < .001***
Grocery           |           Frozen | -0.14 | [-0.23, -0.05] |  -3.02 | 0.003**
Grocery           | Detergents_Paper |  0.57 | [ 0.50,  0.63] |  14.43 | < .001***
Grocery           |       Delicassen |  0.24 | [ 0.15,  0.33] |   5.24 | < .001***
Frozen            | Detergents_Paper | -0.12 | [-0.21, -0.03] |  -2.51 | 0.012*
Frozen            |       Delicassen |  0.14 | [ 0.04,  0.23] |   2.89 | 0.004**
Detergents_Paper  |       Delicassen |  0.17 | [ 0.07,  0.26] |   3.52 | < .001***

p-value adjustment method: none
Observations: 440
```

Ο συντελεστής συσχέτισης Percentage Bend αποτελεί μια ανθεκτική παραλλαγή του Pearson που μειώνει την επιρροή ακραίων τιμών μέσω ενός ποσοστιαίου ορίου καμπής. Στην ανάλυση χρησιμοποιήθηκαν οι ίδιες μακροοικονομικές μεταβλητές που χρησιμοποιήθηκαν για τον υπολογισμό του Biweight Midcorrelation. Με την αφαίρεση του 10% των χαμηλότερων και υψηλότερων τιμών (το ποσοστό 10% αποτελεί ενσωματωμένο χαρακτηριστικό της μεθόδου και δεν τροποποιείται) επιτυγχάνονται πιο αξιόπιστες εκτιμήσεις σε δεδομένα με ακραίες τιμές, αναδεικνύοντας τη σημασία διαφορετικών μεθοδολογικών προσεγγίσεων για την εκτίμηση γραμμικών συσχετίσεων. Η ανάλυση πραγματοποιήθηκε στη γλώσσα προγραμματισμού R με τη χρήση του πακέτου Correlation. Ο τύπος του συντελεστή Percentage Bend είναι:

$$r_{Pb} = \frac{\Sigma_{i-1}^{n} (x_i^* - \bar{x}^*)(y_i^* - \bar{y}^*)}{\sqrt{\Sigma_{i-1}^{n} (x_i^* - \bar{x}^*)^2} \sqrt{\Sigma_{i-1}^{n} (y_i^* - \bar{y}^*)^2}} \quad (2)$$

όπου x_i, y_i είναι οι παρατηρήσεις των μεταβλητών **X** και Y, ενώ $\bar{x}^*$, $\bar{y}^*$είναι οι μέσοι όροι των "καθαρών" τιμών x_i, y_i αντίστοιχα (όχι του αρχικού x_i ή y_i). Οι τιμές x_i, y_i προκύπτουν μέσω μιας διαδικασίας "bending" (κάμψης), η οποία αποκλείει ένα ποσοστό των πιο ακραίων τιμών και τροποποιεί τιμές ώστε να περιορίζεται η επιρροή των ακραίων τιμών.

Εικόνα 3. *Υπολογισμός του συντελεστή συσχέτισης Percentage Bend με το πακέτο correlation.*

```
> percentage_bend_results <- correlation(df, method = "percentage", p_adjust = "none")
> percentage_bend_results
# Correlation Matrix (percentage-method)

Parameter1        |        Parameter2 |     r |          95% CI | t(438) |          p
-------------------------------------------------------------------------------------
Fresh             |             Milk | -0.11 | [-0.20, -0.02] |  -2.34 | 0.020*
Fresh             |          Grocery | -0.14 | [-0.23, -0.05] |  -2.92 | 0.004**
Fresh             |           Frozen |  0.35 | [ 0.26,  0.43] |   7.73 | < .001***
Fresh             | Detergents_Paper | -0.26 | [-0.35, -0.17] |  -5.71 | < .001***
Fresh             |       DeliCassen |  0.23 | [ 0.14,  0.32] |   5.03 | < .001***
Milk              |          Grocery |  0.78 | [ 0.74,  0.81] |  26.13 | < .001***
Milk              |           Frozen | -0.10 | [-0.19, -0.01] |  -2.13 | 0.034*
Milk              | Detergents_Paper |  0.70 | [ 0.64,  0.74] |  20.29 | < .001***
Milk              |       DeliCassen |  0.34 | [ 0.26,  0.42] |   7.65 | < .001***
Grocery           |           Frozen | -0.17 | [-0.26, -0.08] |  -3.65 | < .001***
Grocery           | Detergents_Paper |  0.86 | [ 0.83,  0.88] |  34.63 | < .001***
Grocery           |       DeliCassen |  0.29 | [ 0.20,  0.37] |   6.30 | < .001***
Frozen            | Detergents_Paper | -0.26 | [-0.34, -0.17] |  -5.57 | < .001***
Frozen            |       DeliCassen |  0.22 | [ 0.13,  0.31] |   4.68 | < .001***
Detergents_Paper  |       DeliCassen |  0.16 | [ 0.07,  0.25] |   3.41 | < .001***

p-value adjustment method: none
Observations: 440
```

Ο συντελεστής μερικής συσχέτισης μετρά τη γραμμική σχέση μεταξύ δύο μεταβλητών, αφού αφαιρεθεί πρώτα η επίδραση μιας ή περισσότερων τρίτων μεταβλητών (Vinod, 2017). Ο συντελεστής μερικής συσχέτισης χρησιμοποιείται για την κατανόηση της «καθαρής» συσχέτισης μεταξύ δύο μεταβλητών, απομονώνοντας την επίδραση τρίτων μεταβλητών. Ο μαθηματικός τύπος είναι ο εξής:

$$r_{12;3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}} \tag{3}$$

Όπου: r_{12} είναι ο συντελεστής συσχέτισης μεταξύ των μεταβλητών X_1 και X_2. r_{13} είναι ο συντελεστής συσχέτισης μεταξύ X_1και της συγχέουσας μεταβλητής X_3 και r_{23} είναι ο συντελεστής συσχέτισης μεταξύ X_2 και X_3

Για τον υπολογισμό του συντελεστή μερικής συσχέτισης χρησιμοποιούνται οι ίδιες τριμηνιαίες μακροοικονομικές μεταβλητές που εξετάστηκαν και στον συντελεστή συσχέτισης Pearson, δηλαδή: Ακαθάριστο Εγχώριο Προϊόν, Ανεργία, Εισαγωγές, Εξαγωγές, Επενδύσεις και Όγκος Λιανικών Πωλήσεων της Ευρωπαϊκής Ένωσης. Η ανάλυση πραγματοποιήθηκε με τη χρήση του πακέτου psych όπου έχει εφαρμοστεί έλεγχος στασιμότητας (ADF test) στις δεύτερες διαφορές σε όλες τις χρονοσειρές. Τα αποτελέσματα παρουσιάζονται στις παρακάτω εικόνες.

Εικόνα 4. *Υπολογισμός συντελεστών μερικής συσχέτισης με το πακέτο psych.*

```
===== Πίνακας Partial Correlation (psych::partial.r) =====
> print(round(partial_cor_matrix, 4))
                GDP_EU UNEMPLOYMENT_EU IMPORTS_EU EXPORTS_EU INVESTMENT_EU
GDP_EU          1.0000           0.4386    -0.1918     0.8220        0.0907
UNEMPLOYMENT_EU 0.4386           1.0000     0.0650    -0.4801        0.2821
IMPORTS_EU      -0.1918          0.0650     1.0000     0.5936       -0.0734
EXPORTS_EU      0.8220           -0.4801    0.5936     1.0000        0.1400
INVESTMENT_EU   0.0907           0.2821     -0.0734    0.1400        1.0000
RETAIL_SALES_EU 0.5560           -0.1800    0.0928     -0.4023       0.5241
                RETAIL_SALES_EU
GDP_EU                   0.5560
UNEMPLOYMENT_EU         -0.1800
IMPORTS_EU               0.0928
EXPORTS_EU              -0.4023
INVESTMENT_EU            0.5241
RETAIL_SALES_EU          1.0000
```

Η συσχέτιση απόστασης (Distance Correlation) επιτρέπει την ανίχνευση τόσο γραμμικών όσο και μη γραμμικών σχέσεων (Székely et al., 2007). Ο μαθηματικός τύπος του συντελεστή απόστασης ορίζεται ως:

$$d\,Cov^2(X,Y) = \frac{1}{n^2}\sum_{ij-1}^{n} A_{ij}B_{ij} \quad (4)$$

Όπου: A_{ij} και B_{ij} είναι οι αποστάσεις των παρατηρήσεων X_i, X_j και Y_i, Y_j αντίστοιχα και n είναι το πλήθος των παρατηρήσεων. Οι ποσότητες αυτές προκύπτουν από τους πίνακες αποστάσεων των δύο μεταβλητών και υπολογίζονται μετά την εφαρμογή της διαδικασίας (double-centering). Ο συντελεστής υπολογίζεται για τις τριμηνιαίες μεταβλητές ΑΕΠ, Ανεργία, Εισαγωγές, Εξαγωγές, Επενδύσεις και Λιανικές Πωλήσεις. Ο υπολογισμός πραγματοποιήθηκε με το πακέτο energy στη γλώσσα προγραμματισμού R με τον διορθωμένο εκτιμητή του πακέτου, ενώ ο έλεγχος σημαντικότητας πραγματοποιήθηκε με 500 δοκιμές (permutations).

Εικόνα 5. *Υπολογισμός Distance Correlation με το πακέτο energy.*

```
===== Πίνακας Distance Correlation (2ες διαφορές) =====
> print(round(dcor_df, 4))
                GDP_EU UNEMPLOYMENT_EU IMPORTS_EU EXPORTS_EU INVESTMENT_EU
GDP_EU          1.0000          0.4565     0.7217     0.9046        0.8340
UNEMPLOYMENT_EU 0.4565          1.0000     0.3396     0.4351        0.4461
IMPORTS_EU      0.7217          0.3396     1.0000     0.7995        0.6038
EXPORTS_EU      0.9046          0.4351     0.7995     1.0000        0.7499
INVESTMENT_EU   0.8340          0.4461     0.6038     0.7499        1.0000
RETAIL_SALES_EU 0.8496          0.4577     0.6235     0.7530        0.8213
                RETAIL_SALES_EU
GDP_EU                   0.8496
UNEMPLOYMENT_EU          0.4577
IMPORTS_EU               0.6235
EXPORTS_EU               0.7530
INVESTMENT_EU            0.8213
RETAIL_SALES_EU          1.0000
```

Ο Μέγιστος Συντελεστής Πληροφορίας (Maximal Information Coefficient – MIC) είναι ένα μέτρο που μπορεί να ανιχνεύσει μη γραμμικές σχέσεις, που δεν μπορούν να εντοπιστούν από τα παραδοσιακά μέτρα συσχέτισης. Είναι βασισμένο στη θεωρία της πληροφορίας, και ανήκει στην οικογένεια στατιστικών MINE (Reshef et al., 2011). Ο μαθηματικός τύπος του είναι:

$$MIC(X,Y) = max\left(\frac{I(X;Y)}{log_2(min(|X|,|Y|))}\right) \quad (5)$$

Όπου: $I(X;Y)$ είναι η αμοιβαία πληροφορία μεταξύ των X και Y και $min(|X|,|Y|)$ είναι το ελάχιστο μέγεθος των δύο συνόλων παρατηρήσεων για τις μεταβλητές X και Y. Η ποσότητα $I(X;Y)$ μετρά την κοινή πληροφορία που μοιράζονται οι δύο μεταβλητές και μπορεί να ανιχνεύσει οποιοδήποτε είδος σχέσης μεταξύ τους, είτε γραμμική είτε μη γραμμική (Clark, 2013).

Τα δεδομένα είναι από το αρχείο “Wholesale customers” (UCI Repository). Η εφαρμογή έγινε στην R με τη χρήση του πακέτου Minerva για τη μελέτη συσχετίσεων ανάμεσα σε έξι ποσοτικές μεταβλητές με στόχο την ανάδειξη πιθανών σύνθετων σχέσεων που δεν εντοπίζονται με παραδοσιακά

μέτρα. Στο πακέτο Minerva χρησιμοποιήθηκε η προεπιλεγμένη παράμετρος α = 0.6, η οποία καθορίζει το μέγεθος του search-grid ως B(n) = $n^{0.6}$ (Reshef et al., 2011). Τα αποτελέσματα αποτυπώνονται στην εικόνα 6.

Εικόνα 6. *Υπολογισμός Maximal Information Coefficient με το πακέτο Minerva*

```
> library(minerva)
> mic_results <- mine(df_mic)
> mic_results
$MIC
                      Fresh      Milk   Grocery    Frozen Detergents_Paper Delicassen
Fresh             1.0000000 0.1697354 0.1741883 0.2662339        0.2328209  0.1755982
Milk              0.1697354 1.0000000 0.5840044 0.1714520        0.4806618  0.2828202
Grocery           0.1741883 0.5840044 1.0000000 0.2056580        0.6230436  0.2457021
Frozen            0.2662339 0.1714520 0.2056580 1.0000000        0.1799851  0.2018061
Detergents_Paper  0.2328209 0.4806618 0.6230436 0.1799851        1.0000000  0.1927976
Delicassen        0.1755982 0.2828202 0.2457021 0.2018061        0.1927976  1.0000000
```

Ο συντελεστής Spearman Rho (Spearman's rank correlation coefficient, Rs) αποτελεί μια μη παραμετρική εναλλακτική του Pearson, κατάλληλη για τη μέτρηση της μονοτονικής σχέσης μεταξύ δύο μεταβλητών που εκφράζονται με κατατάξεις (Clark, 2013). Ο μαθηματικός τύπος είναι:

$$r_s = 1 - \frac{6\Sigma d^2}{n(n^2-1)} \quad (6)$$

Όπου: r_s είναι ο συντελεστής Spearman, d είναι η διαφορά στις τάξεις για κάθε ζεύγος παρατηρήσεων (X,Y) και n είναι ο αριθμός των παρατηρήσεων.

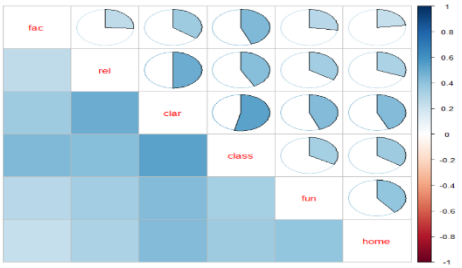
Ο συντελεστής Spearman's Rho υπολογίστηκε με τα δεδομένα του συνόλου "learning" από το πακέτο peopleanalyticsdata (McNulty, 2021) της R, το οποίο περιλαμβάνει αξιολογήσεις συμμετεχόντων σε εκπαιδευτικά προγράμματα. Εξετάστηκαν έξι διατακτικές μεταβλητές σε κλίμακα Likert (rel, fun, clar, home, class και fac) που αξιολογούν πλευρές του προγράμματος και της εκπαιδευτικής εμπειρίας. Τα αποτελέσματα μπορούμε να τα δούμε παρακάτω στην εικόνα 7.

Εικόνα 7. *Υπολογισμός συντελεστή συσχέτισης Spearman's Rho (α) με το πακέτο psych (β) με το πακέτο corrplot και (γ) με το πακέτο corrplot.*

```
> print(spearman_psych$r)
           rel       fun      clar      home     class       fac
rel  1.0000000 0.3272258 0.5025998 0.3186134 0.4189728 0.2459454
fun  0.3272258 1.0000000 0.4243673 0.3913999 0.3345627 0.2622015
clar 0.5025998 0.4243673 1.0000000 0.4347839 0.5398592 0.3441174
home 0.3186134 0.3913999 0.4347839 1.0000000 0.3539196 0.2288592
class 0.4189728 0.3345627 0.5398592 0.3539196 1.0000000 0.4358839
fac  0.2459454 0.2622015 0.3441174 0.2288592 0.4358839 1.0000000
>
> # 8. Εμφάνιση των p-values
> cat("\nP-values:\n")

P-values:
> print(spearman_psych$p)
             rel          fun          clar          home         class
rel   0.000000e+00 5.263942e-119 1.782699e-301 4.151343e-107 9.336508e-183
fun   8.773236e-120  0.000000e+00 2.038266e-204 1.024130e-164 2.190596e-112
clar  1.273356e-302 1.698558e-205  0.000000e+00 3.047937e-206  0.000000e+00
home  1.037836e-107 1.137922e-165 2.344567e-207  0.000000e+00 2.459505e-123
class 9.336508e-184 4.381193e-113  0.000000e+00 3.513579e-124  0.000000e+00
fac    1.412403e-66  7.764156e-75 3.824789e-130  2.815557e-54 3.669107e-198
             fac
rel   2.824805e-66
fun   2.329247e-74
clar  3.059832e-129
home  2.815557e-54
class 4.036017e-197
```

(α)



(β)



(γ)

Ο συντελεστής Shepherd's Pi αποτελεί μια παραλλαγή του Spearman's Rho, σχεδιασμένη να μειώνει την επίδραση των ακραίων τιμών. Βασίζεται σε κατατάξεις, αλλά ενσωματώνει μηχανισμό φιλτραρίσματος ακραίων τιμών outliers, προσφέροντας πιο αξιόπιστες εκτιμήσεις σε δεδομένα με αποκλίσεις. Ο μαθηματικός τύπος του συντελεστή συσχέτισης Shepherd's Pi είναι:

$$r_{\Pi} = \frac{\Sigma_{i-1}^{n} u_i v_i}{\sqrt{\Sigma_{i-1}^{n} u_i^2} \cdot \sqrt{\Sigma_{i}^{n} + v_i^2}} \tag{7}$$

Όπου: u_i και v_i είναι τα κατάλοιπα των μεταβλητών X και Y αντίστοιχα, όπως προκύπτουν μετά την εφαρμογή ανθεκτικής παλινδρόμησης (robust regression) που προηγείται του υπολογισμού του Shepherd's Pi. Τα κατάλοιπα αυτά έχουν φιλτραριστεί από ακραίες τιμές και χρησιμοποιούνται στη συνέχεια για τον υπολογισμό της συσχέτισης.

Για τον υπολογισμό του Shepherds Pi χρησιμοποιήθηκαν οι μεταβλητές (rel, fun, clar, home, class, fac) από το σύνολο δεδομένων του πακέτου peopleanalyticsdata χρησιμοποιώντας το πακέτο correlation (εικόνα 8). Η ανθεκτικότητα του Shepherd's Pi τον καθιστά κατάλληλο για αξιολόγηση κατατάξεων σε περιβάλλοντα με ακραίες τιμές.

Εικόνα 8. *Υπολογισμός του συντελεστή συσχέτισης Συντελεστής Συσχέτισης Shepherd's Pi με το πακέτο correlation.*

```
> result_shepherd <- correlation(df_selected, method = "shepherd", p_adjust = "none")
> result_shepherd
# Correlation Matrix (shepherd-method)

Parameter1 | Parameter2 |  rho |        95% CI |         S |          p
-----------------------------------------------------------------------
rel        |        fun | 0.36 | [0.34, 0.39] | 6.50e+09 | < .001***
rel        |       clar | 0.56 | [0.54, 0.59] | 4.36e+09 | < .001***
rel        |       home | 0.39 | [0.36, 0.41] | 6.24e+09 | < .001***
rel        |      class | 0.45 | [0.42, 0.48] | 5.30e+09 | < .001***
rel        |        fac | 0.29 | [0.26, 0.32] | 6.43e+09 | < .001***
fun        |       clar | 0.47 | [0.45, 0.50] | 5.35e+09 | < .001***
fun        |       home | 0.46 | [0.43, 0.48] | 5.51e+09 | < .001***
fun        |      class | 0.43 | [0.40, 0.45] | 5.59e+09 | < .001***
fun        |        fac | 0.33 | [0.30, 0.35] | 6.95e+09 | < .001***
clar       |       home | 0.50 | [0.48, 0.53] | 5.17e+09 | < .001***
clar       |      class | 0.55 | [0.53, 0.58] | 4.42e+09 | < .001***
clar       |        fac | 0.45 | [0.42, 0.47] | 5.30e+09 | < .001***
home       |      class | 0.44 | [0.41, 0.47] | 5.59e+09 | < .001***
home       |        fac | 0.23 | [0.20, 0.26] | 7.48e+09 | < .001***
class      |        fac | 0.46 | [0.43, 0.48] | 5.09e+09 | < .001***

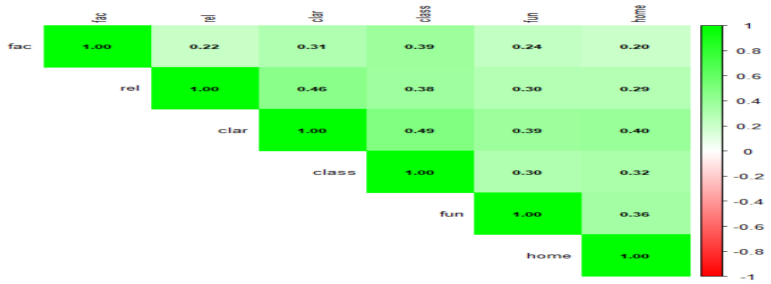
p-value adjustment method: none
Observations: 4120
```

Ο συντελεστής Kendall's Tau είναι κατάλληλος για περιπτώσεις όπου η σχέση μεταξύ δύο ποσοτικών μεταβλητών είναι μονοτονική μη γραμμική ή έχουμε διατακτικές μεταβλητές. (Lind et al., 2012). Χρησιμοποιήθηκαν οι διατακτικές μεταβλητές του συνόλου δεδομένων "learning" (rel, fun, clar, home, class, fac), ώστε να συγκριθούν τα αποτελέσματα με εκείνα των συντελεστών Spearman και Shepherd's Pi. Ο υπολογισμός έγινε μέσω της συνάρτησης cor() στην R με μέθοδο "kendall" (εικόνα 9).

Εικόνα 9. *Υπολογισμός Kendall's Tau Correlation (α) με το πακέτο psych και (β)με το πακέτο corrplot* .

```
> library(psych)
> df_selected <- learning[, c("rel", "fun", "clar", "home", "class", "fac")]
>
> kendall_psych <- psych::corr.test(df_selected, method = "kendall")
> kendall_psych
Call:psych::corr.test(x = df_selected, method = "kendall")
Correlation matrix
       rel   fun clar home class   fac
rel   1.00 0.30 0.46 0.29  0.38 0.22
fun   0.30 1.00 0.39 0.36  0.30 0.24
clar  0.46 0.39 1.00 0.40  0.49 0.31
home  0.29 0.36 0.40 1.00  0.32 0.20
class 0.38 0.30 0.49 0.32  1.00 0.39
fac   0.22 0.24 0.31 0.20  0.39 1.00
Sample Size
       rel   fun clar home class   fac
rel   4850 4764 4751 4543  4332  4764
fun   4764 4788 4715 4528  4305  4706
clar  4751 4715 4756 4509  4301  4677
home  4543 4528 4509 4547  4200  4476
class 4332 4305 4301 4200  4336  4283
fac   4764 4706 4677 4476  4283  4774
Probability values (Entries above the diagonal are adjusted for multiple tests.)
       rel fun clar home class fac
rel      0   0    0    0     0   0
fun      0   0    0    0     0   0
clar     0   0    0    0     0   0
home     0   0    0    0     0   0
class    0   0    0    0     0   0
fac      0   0    0    0     0   0
```

(α)



(β)

Ο συντελεστής συσχέτισης Gamma είναι ένα μη παραμετρικό μέτρο που εφαρμόζεται σε διατακτικές μεταβλητές και είναι ιδιαίτερα κατάλληλος όταν τα δεδομένα δεν περιλαμβάνουν πολλές ισοβαθμίες, καθώς ο συγκεκριμένος συντελεστής συσχέτισης δεν λαμβάνει υπόψη τις ισοβαθμίες, γεγονός που μπορεί να υπερεκτιμήσει τη δύναμη της συσχέτισης. Χρησιμοποιήθηκε το ίδιο σύνολο δεδομένων "learning" (rel, fun, clar, home, class, fac), ώστε να διασφαλιστεί η συγκρισιμότητα με τους συντελεστές Spearman's Rho, Shepherd's Pi, Kendall's Tau. Ο υπολογισμός πραγματοποιήθηκε στη R με τη χρήση του πακέτου Correlation. Τα αποτελέσματα παρουσιάζονται στην εικόνα 10. Ο μαθηματικός τύπος του συντελεστή συσχέτισης Gamma είναι:

$$G = \frac{C-D}{C+D} \tag{8}$$

Όπου: C είναι ο αριθμός των συμφωνούντων ζευγών, δηλαδή οι κατατάξεις που κινούνται προς την ίδια κατεύθυνση, ενώ D είναι ο αριθμός των διαφωνούντων ζευγών, όπου οι κατατάξεις μεταβάλλονται αντίθετα.

Εικόνα 10. *Υπολογισμός του συντελεστή συσχέτισης Gamma με το πακέτο correlation.*

```
> result_gamma <- correlation(df_selected, method = "gamma", p_adjust = "none")
> result_gamma
# Correlation Matrix (gamma-method)

Parameter1 | Parameter2 |    r |        95% CI |     t |   df |          p
-----------------------------------------------------------------------
rel        |        fun | 0.45 | [0.42, 0.47] | 34.53 | 4782 | < .001***
rel        |       clar | 0.66 | [0.64, 0.67] | 59.75 | 4749 | < .001***
rel        |       home | 0.43 | [0.40, 0.45] | 31.95 | 4541 | < .001***
rel        |      class | 0.55 | [0.52, 0.57] | 42.90 | 4330 | < .001***
rel        |        fac | 0.33 | [0.30, 0.35] | 23.79 | 4762 | < .001***
fun        |       clar | 0.57 | [0.55, 0.59] | 47.15 | 4713 | < .001***
fun        |       home | 0.53 | [0.51, 0.55] | 41.68 | 4526 | < .001***
fun        |      class | 0.45 | [0.42, 0.47] | 32.65 | 4303 | < .001***
fun        |        fac | 0.35 | [0.33, 0.38] | 25.88 | 4704 | < .001***
clar       |       home | 0.57 | [0.55, 0.59] | 46.44 | 4507 | < .001***
clar       |      class | 0.69 | [0.67, 0.71] | 62.63 | 4299 | < .001***
clar       |        fac | 0.45 | [0.43, 0.47] | 34.60 | 4675 | < .001***
home       |      class | 0.46 | [0.44, 0.49] | 33.89 | 4198 | < .001***
home       |        fac | 0.30 | [0.28, 0.33] | 21.40 | 4474 | < .001***
class      |        fac | 0.56 | [0.54, 0.58] | 43.78 | 4281 | < .001***

p-value adjustment method: none
Observations: 4200-4784
```

Ο συντελεστής συσχέτισης Hoeffding's D χρησιμοποιείται για την ανίχνευση οποιασδήποτε μορφής εξάρτησης μεταξύ δύο μεταβλητών, ακόμα κι αν η σχέση τους δεν είναι ούτε γραμμική ούτε μονοτονική (Hoeffding, 1948). Στην παρούσα ανάλυση χρησιμοποιούμε τις ίδιες τριμηνιαίες μακροοικονομικές μεταβλητές με τον συντελεστή απόστασης και τον συντελεστή μερικής συσχέτισης (ΑΕΠ, Ανεργία, Εισαγωγές, Εξαγωγές, Επενδύσεις, Λιανικές Πωλήσεις), για να ελέγξουμε εάν υπάρχουν πιο σύνθετες σχέσεις μεταξύ τους. Πριν από τον υπολογισμό πραγματοποιήθηκε έλεγχος (ADF test) στις δεύτερες διαφορές. Στην εικόνα 11 παρουσιάζουμε τα αποτελέσματα με την χρήση του πακέτου Hmisc. Ο μαθηματικός τύπος του Hoeffding's D είναι:

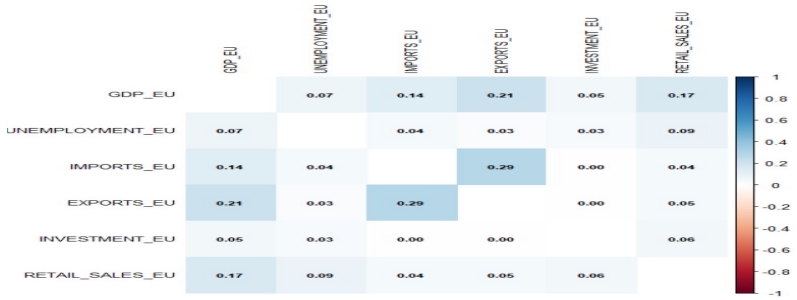
$$D = \frac{A-2(n-2)B+(n-2)(n-3)C}{n(n-1)(n-2)(n-3)(n-4)} \tag{9}$$

Όπου: n το μέγεθος του δείγματος, η ποσότητα A εκφράζει τον αριθμό των συμφωνούντων ζευγών ως προς τις κατατάξεις των δυο μεταβλητών, το B αντιστοιχεί στα ζεύγη όπου οι κατατάξεις κινούνται προς αντίθετες κατευθύνσεις, ενώ το C αφορά τις περιπτώσεις ισοβαθμιών ή κοινών τιμών. Οι τρεις ποσότητες χρησιμοποιούνται στην διαμόρφωση του δείκτη Hoeffding's D για τον εντοπισμό συσχετίσεων.

Εικόνα 11. *Υπολογισμός συντελεστή Hoeffding's D (α) με πακέτο Hmisc και (β) με το πακέτο Corrplot.*

```
----- Πίνακας Hoeffding's D (2ες διαφορές) -----
> print(round(hoeff_d_matrix, 4))
               GDP_EU UNEMPLOYMENT_EU IMPORTS_EU EXPORTS_EU INVESTMENT_EU
GDP_EU         1.0000          0.0019     0.0620     0.1943        0.0075
UNEMPLOYMENT_EU 0.0019         1.0000     0.0017     0.0187       -0.0058
IMPORTS_EU     0.0620          0.0017     1.0000     0.2282       -0.0062
EXPORTS_EU     0.1943          0.0187     0.2282     1.0000       -0.0057
INVESTMENT_EU  0.0075         -0.0058    -0.0062    -0.0057        1.0000
RETAIL_SALES_EU 0.0895        -0.0068     0.0146     0.0346        0.0180
               RETAIL_SALES_EU
GDP_EU                  0.0895
UNEMPLOYMENT_EU        -0.0068
IMPORTS_EU              0.0146
EXPORTS_EU              0.0346
INVESTMENT_EU           0.0180
RETAIL_SALES_EU         1.0000
```

(α)



(β)

Ο συντελεστής συσχέτισης Cramér's V χρησιμοποιείται για την ανάλυση της σχέσης μεταξύ κατηγορικών μεταβλητών (Asuero et al., 2006). Η κατασκευή του πίνακα συσχέτισης έγινε στο σύνολο δεδομένων "Bank Marketing" (μεταβλητές: job, marital, education, default, housing, loan, month), οι οποίες είναι κατηγορικές. Τα δεδομένα προέρχονται από το UC Irvine Repository. Τα αποτελέσματα αποτυπώνεται στην εικόνα 12 με την βοήθεια του πακέτου rcompanion . Ο μαθηματικός τύπος του Cramér's V είναι ο εξής:

$$V = \sqrt{\frac{x^2}{n*\min(k-1,r-1)}} \tag{10}$$

όπου x^2 είναι η τιμή του στατιστικού chi-square σε ένα πίνακα συνάφειας $r * k$, και n είναι το μέγεθος του δείγματος, k είναι ο αριθμός των στηλών του πίνακα συνάφειας και r είναι ο αριθμός των γραμμών του πίνακα συνάφειας.

Εικόνα 12. *Υπολογισμός Cramer's V με το πακέτο rcompanion.*

```
> cramers_v_df <- as.data.frame(cramers_v_matrix)
>
> cramers_v_df
              job marital education  default  housing     loan   month
job       1.00000 0.20600   0.45830 0.036530 0.281700 0.10650 0.11020
marital   0.20600 1.00000   0.12160 0.019230 0.020690 0.05194 0.07232
education 0.45830 0.12160   1.00000 0.015900 0.119300 0.08028 0.11010
default   0.03653 0.01923   0.01590 1.000000 0.006025 0.07723 0.05867
housing   0.28170 0.02069   0.11930 0.006025 1.000000 0.04132 0.50420
loan      0.10650 0.05194   0.08028 0.077230 0.041320 1.00000 0.18280
month     0.11020 0.07232   0.11010 0.058670 0.504200 0.18280 1.00000
```

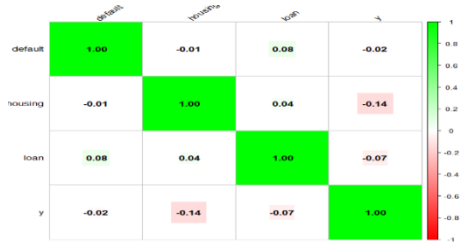
Ο συντελεστής συσχέτισης Φ (Phi Coefficient) εφαρμόζεται όταν και οι δύο μεταβλητές είναι δυαδικές. Στην παρούσα ανάλυση χρησιμοποιήθηκαν οι μεταβλητές default, housing, loan και y από το dataset "Bank Marketing", οι οποίες πληρούν αυτή την προϋπόθεση. Ο συντελεστής συσχέτισης υπολογίστηκε στη γλώσσα προγραμματισμού R με τη βοήθεια του πακέτου psych και τα αποτελέσματα φαίνονται στην εικόνα 13. Ο μαθηματικός τύπος είναι:

$$\Phi = \sqrt{\frac{x^2}{n}} \tag{11}$$

όπου: x^2 είναι η τιμή του στατιστικού chi-square και n είναι το συνολικό μέγεθος του δείγματος.

Εικόνα 13. *Υπολογισμός Phi Coefficient (α) με πακέτο psych και (β) με πακέτο corrplot.*

```
> phi_df <- as.data.frame(phi_matrix)
> phi_df
         default  housing     loan       y
default  1.00000 -0.00603  0.0772 -0.0224
housing -0.00603  1.00000  0.0413 -0.1390
loan     0.07720  0.04130  1.0000 -0.0682
y       -0.02240 -0.13900 -0.0682  1.0000
```



(α) *(β)*

Στην συνέχεια, εξετάζουμε συνοπτικά τέσσερις βασικές μετρικές για την ανίχνευση πολυσυγγραμμικότητας, χρησιμοποιώντας ένα σύνολο δεδομένων από το βιβλίο του γνωστού οικονομέτρη Damodar Gujarati (2003, σ. 290), που περιλαμβάνει τις μεταβλητές GNP, Housing Starts, Unemployment, Prime Rate και Line Gains για την περίοδο 1968-1983 (εικόνα 14).

Ο Συντελεστής Διόγκωσης της Διακύμανσης (VIF) αξιολογεί πόσο μπορεί να επηρεαστεί η διακύμανση της συσχέτισης της συγκεκριμένης μεταβλητής με τις υπόλοιπες μεταβλητές. Τιμές VIF μεγαλύτερες από 10 θεωρούνται υψηλή πολυσυγγραμμικότητα (Chan et al., 2022).

$$VIF(X_i) = \frac{1}{1-R_i^2} \tag{12}$$

όπου R_i^2 είναι ο συντελεστής προσδιορισμού που προκύπτει από την παλινδρόμηση της X_i ως συνάρτηση όλων των άλλων ανεξάρτητων μεταβλητών στο μοντέλο (Hanck et al., 2024).

Ο Συντελεστής ανοχής (Tolerance – TOL) είναι το αντίστροφο του Συντελεστή Διόγκωσης της Διακύμανσης (VIF) και μετρά το βαθμό στον οποίο μια ανεξάρτητη μεταβλητή δεν εξαρτάται από τις υπόλοιπες (Chan et al., 2022).

$$TC_{(i)} = \frac{1}{VIF} \tag{13}$$

όπου i υποδηλώνει τη συγκεκριμένη μεταβλητή για την οποία υπολογίζεται το VIF.

Εικόνα 14. *Υπολογισμός συντελεστή Variance Inflation Factor και υπολογισμός του συντελεστή ανοχής (Tolerance) με το πακέτο olsrr.*

```
> ols_vif_tol(model_gujarati)
      Variables  Tolerance       VIF
1           GNP 0.14481923  6.905160
2 Housing_Starts 0.23015249  4.344945
3  Unemployment 0.25202143  3.967917
4    Prime_Rate 0.06810576 14.683046
5    Line_Gains 0.18438282  5.423499
```

Παράλληλα υπολογίστηκαν οι ιδιοτιμές (eigenvalues) του πίνακα συσχέτισης των ανεξάρτητων μεταβλητών (εικόνα 15). Πολύ μικρές ιδιοτιμές υποδηλώνουν ύπαρξη πολυσυγγραμμικότητας, καθώς υποδηλώνουν σχεδόν γραμμική εξάρτηση μεταξύ των μεταβλητών (Chan et al., 2022). Το άθροισμα των ιδιοτιμών ισούται με τον αριθμό των μεταβλητών, επιτρέποντας τον υπολογισμό του ποσοστού διακύμανσης ανά συνιστώσα. Ο γενικός τύπος είναι:

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0 \tag{14}$$

Όπου: p αντιπροσωπεύει τον αριθμό ανεξάρτητων μεταβλητών, ενώ οι ιδιοτιμές του πίνακα συσχέτισης ικανοποιούν τη σχέση $\lambda_1 = \lambda_{max}$ και $\lambda_p = \lambda_{min}$, όπου λ_1 είναι η μέγιστη και λ_p η ελάχιστη ιδιοτιμή. Στην συνέχεια αναφερόμαστε στον Δείκτη Κατάστασης της εξίσωσης (15), όπου το μέγεθος κ υπολογίζεται ως την τετραγωνική ρίζα του λόγου της μέγιστης προς την ελάχιστη ιδιοτιμή ($CI = \kappa = \sqrt{\lambda_{max}/\lambda_{min}}$).

Εικόνα 15. *Υπολογισμός ιδιοτιμών με την συνάρτηση eigen() της R.*

```
> eig_vals <- eigen(cor_matrix)$values
> print(eig_vals)
[1] 3.71684128 0.71055658 0.35341055 0.17958812 0.03960348
```

Ο Δείκτης Κατάστασης (Condition Index) υπολογίζεται από τις ιδιοτιμές και χρησιμοποιείται για να ανιχνεύσει την πολυσυγγραμμικότητα ειδικά όταν οι ιδιοτιμές δείχνουν μεγάλο εύρος. Χρησιμοποιείται για την αξιολόγηση της σταθερότητας των παλινδρομήσεων σε περιπτώσεις όπου οι μεταβλητές έχουν υψηλό επίπεδο συσχέτισης (Salmerón et al., 2018). Ένας CI > 30 υποδηλώνει υψηλή πολυσυγγραμμικότητα. Ο μαθηματικός τύπος είναι:

$$\kappa = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}} \tag{15}$$

όπου λ_{max} η μέγιστη ιδιοτιμή είναι η μεγαλύτερη ιδιοτιμή που προκύπτει από τον πίνακα συσχέτισης ή συνδιακύμανσης. Αντιπροσωπεύει τη συνιστώσα που εξηγεί τη μεγαλύτερη ποσότητα διασποράς στα δεδομένα και λ_{min} η ελάχιστη ιδιοτιμή είναι η μικρότερη ιδιοτιμή (συνήθως κοντά στο μηδέν αν υπάρχει πολυδιάστατη συσχέτιση) και αντιπροσωπεύει τη συνιστώσα που εξηγεί τη λιγότερη ποσότητα διασποράς στα δεδομένα (εικόνα 16).

Εικόνα 16. *Υπολογισμός του Δείκτη κατάστασης.*

```
> X_vars <- df_model[, c("GNP", "Housing_Starts", "Unemployment", "Prime_Rate", "Line_G$
> cor_matrix <- cor(X_vars)
> eig_vals <- eigen(cor_matrix)$values
> condition_index <- sqrt(max(eig_vals) / eig_vals)
> condition_index
[1] 1.000000 2.287113 3.243002 4.549338 9.687692
```

Η Canonical Correlation Analysis (CCA), που εισήχθη από τον Hotelling το 1936, μελετά τη γραμμική σχέση μεταξύ δύο ομάδων μεταβλητών (Hotelling, 1936). Σκοπός του είναι η εύρεση ζευγών γραμμικών συνδυασμών που παρουσιάζουν τη μέγιστη συσχέτιση μεταξύ τους. Η μαθηματική του μορφή είναι:

$$U = a_1X_1 + a_2X_2 + \ldots + a_pX_p\ ,\quad V = b_1Y_1 + b_2Y_2 + \ldots + b_qY_q \tag{16}$$

Η ανάλυση κανονικής συσχέτισης εφαρμόστηκε στις μακροοικονομικές τριμηνιαίες μεταβλητές: ΑΕΠ, Ανεργία, Εισαγωγές και Εξαγωγές για την Ελλάδα (ομάδα 1) και την Ε.Ε. (ομάδα 2) για συνολικά n=67 παρατηρήσεις, όπου αρχικά προηγήθηκε έλεγχος στασιμότητας (ADF test) στις δεύτερες διαφορές. Ο υπολογισμός πραγματοποιήθηκε με χρήση του πακέτου CCP. Στην εικόνα 17 παρουσιάζονται τα αποτελέσματα. Οι τέσσερις κανονικές συσχετίσεις που προέκυψαν υποδεικνύουν υψηλή συμμεταβλητότητα μεταξύ των δύο ομάδων μεταβλητών. Η σημαντικότητα των κανονικών συνιστωσών ελέγχθηκε με το Wilks και την F- προσέγγιση του Rao. Οι υψηλές τιμές των πρώτων κανονικών συσχετίσεων οφείλονται εν μέρει στο γεγονός ότι οι ελληνικοί και οι ευρωπαϊκοί μακροοικονομικοί δείκτες εξελίσσονται με κοινές τάσεις και υψηλό βαθμό αλληλεξάρτησης, γεγονός που ενισχύει τεχνητά τη συσχέτιση μεταξύ των δύο ομάδων μεταβλητών.

Εικόνα 17. Αποτελέσματα Canonical Correlation Analysis μεταξύ των μεταβλητών της ομάδας 1 με το σύνολο των μεταβλητών της ομάδας 2 με το πακέτο ccp.

```
> print(rho)
[1] 0.87236655 0.45265374 0.12690878 0.05233513
> cat("\nΠλήθος παρατηρήσεων (n):", n, "\n")

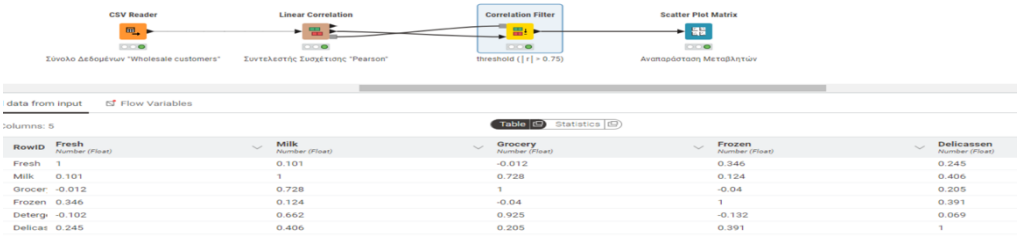
Πλήθος παρατηρήσεων (n): 67
> resWilks <- p.asym(rho, n, p, q, tstat = "Wilks")
Wilks' Lambda, using F-approximation (Rao's F):
      stat      approx df1      df2      p.value
1 to 4:  0.18643390 8.28566620  16 180.8856 9.1038229e-15
2 to 4:  0.78015561 1.74423331   9 146.1748 8.3964455e-02
3 to 4:  0.98111933 0.29081177   4 122.0000 8.8346624e-01
4 to 4:  0.99726103 0.17228223   1  62.0000 6.8128466e-01
> print(adf_diff2)
          Variable   ADF_stat   P_value Lags_k
1              GDP -6.144152 0.0100000      4
2     UNEMPLOYMENT -3.644996 0.0362911      4
3          IMPORTS -5.414547 0.0100000      4
4          EXPORTS -5.697447 0.0100000      4
5           GDP_EU -5.718429 0.0100000      4
6 UNEMPLOYMENT_EU -4.181770 0.0100000      4
7       IMPORTS_EU -4.574061 0.0100000      4
8       EXPORTS_EU -5.149713 0.0100000      4
```

4. ΥΠΟΛΟΓΙΣΤΙΚΗ ΠΡΟΣΕΓΓΙΣΗ ΤΗΣ ΑΝΑΛΥΣΗΣ ΣΥΣΧΕΤΙΣΗΣ ΜΕΣΩ ΤΗΣ ΠΛΑΤΦΟΡΜΑΣ ΑΝΑΛΥΤΙΚΗΣ ΔΕΔΟΜΕΝΩΝ KNIME.

Η πλατφόρμα αναλυτικής δεδομένων KNIME είναι μια δωρεάν, ανοιχτού κώδικα πλατφόρμα για ανάλυση δεδομένων, εξόρυξη γνώσης και μηχανική μάθηση. Αναπτύχθηκε αρχικά από το Πανεπιστήμιο του Κονσταντς (Γερμανία) και χρησιμοποιείται ευρέως τόσο στην έρευνα όσο και στον επιχειρηματικό χώρο. Κύρια χαρακτηριστικά της είναι το γραφικό περιβάλλον εργασίας με χρήση "κόμβων" (nodes) για κάθε στάδιο ανάλυσης, χωρίς ανάγκη προγραμματισμού. Αντί για γραφή κώδικα, χρησιμοποιείται drag-and-drop περιβάλλον εργασίας με κόμβους (nodes) που εκτελούν λειτουργίες όπως εισαγωγή, καθαρισμό, μετασχηματισμό και ανάλυση δεδομένων, μηχανική μάθηση και οπτικοποίηση (visual, low-code/no-code interface). Υπάρχει δυνατότητα ενσωμάτωσης Python, R, SQL, Java για πιο προχωρημένες αναλύσεις.

Η ανάλυση συσχέτισης μέσω της πλατφόρμας αναλυτικής δεδομένων KNIME επιβεβαιώνει την αξιοπιστία εναλλακτικών υπολογιστικών εργαλείων πέραν των υπολογιστικών εργαλείων με κώδικα. Ο υπολογισμός του συντελεστή συσχέτισης Pearson έγινε με τον κόμβο Linear Correlation, ενώ για τη μείωση της πολυσυγγραμμικότητας εφαρμόστηκε ο κόμβος Correlation Filter. Ο κόμβος Correlation Filter ρυθμίστηκε με την επιλογή (threshold) = $|r| > 0{,}75$, ώστε να αφαιρεθούν οι μεταβλητές με απόλυτη τιμή συσχέτισης μεγαλύτερη του 0,75. Σύμφωνα με τη σχετική βιβλιογραφία, η τιμή 0.75 θεωρείται ένδειξη ισχυρής γραμμικής συσχέτισης. Χρησιμοποιήθηκε το σύνολο δεδομένων Wholesale customers από το UCI Machine Learning Repository, με έξι ποσοτικές μεταβλητές. Μετά την εφαρμογή του κριτηρίου αφαιρέθηκε η μεταβλητή Detergents_Paper, καθώς παρουσίαζε πολύ υψηλή γραμμική συσχέτιση με τη μεταβλητή Grocery ($r = 0.925$) Η αφαίρεση της μεταβλητής οδήγησε στη μείωση των στηλών από 6 σε 5, διατηρώντας μόνο τις μεταβλητές που δεν εμφάνιζαν υπερβολικά υψηλή συσχέτιση με τις υπόλοιπες. Στην εικόνα 18 αποτυπώνονται τα σχετικά αποτελέσματα.

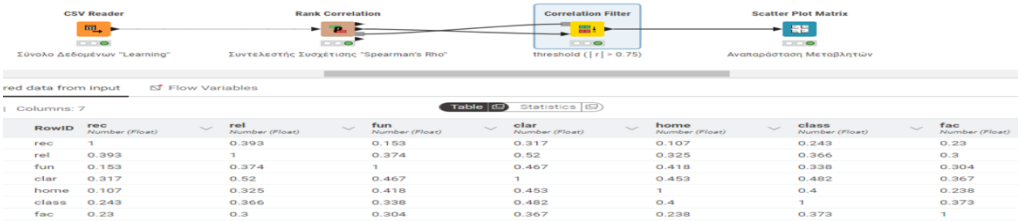
Εικόνα 18. *Εφαρμογή Pearson Correlation Coefficient με Correlation Filter (threshold 0.75) μέσω της πλατφόρμας αναλυτικής δεδομένων KNIME.*



RowID	Fresh Number (Float)	Milk Number (Float)	Grocery Number (Float)	Frozen Number (Float)	Delicassen Number (Float)
Fresh	1	0.101	-0.012	0.346	0.245
Milk	0.101	1	0.728	0.124	0.406
Grocer	-0.012	0.728	1	-0.04	0.205
Frozen	0.346	0.124	-0.04	1	0.391
Deterg	-0.102	0.662	0.925	-0.132	0.069
Delicas	0.245	0.406	0.205	0.391	1

Ο συντελεστής συσχέτισης Spearman' Rho υπολογίστηκε μέσω του κόμβου Rank Correlation χρησιμοποιώντας τις μεταβλητές από το Bank Marketing Dataset όπως και στη γλώσσα προγραμματισμού R (εικόνα 19). Η σύγκριση των αποτελεσμάτων επιβεβαιώνει την αξιοπιστία της μεθόδου σε διαφορετικά περιβάλλοντα. Στη εικόνα 19, προστέθηκε ο κόμβος Correlation Filter (με κριτήριο αποκλεισμού $|r| > 0{,}75$), όπου δεν υπήρξε απομάκρυνση καμιάς μεταβλητής, καθώς όλοι οι συντελεστές συσχέτισης έχουν τιμές μικρότερες από 0,75.

Εικόνα 19. *Εφαρμογή Spearman's Rho Correlation Coefficient με Correlation Filter (threshold 0.75) στη πλατφόρμα αναλυτικής δεδομένων KNIME.*



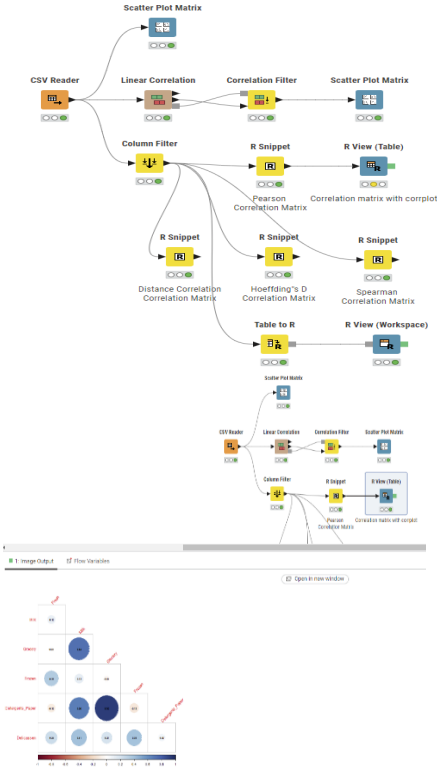
RowID	rec Number (Float)	rel Number (Float)	fun Number (Float)	clar Number (Float)	home Number (Float)	class Number (Float)	fac Number (Float)
rec	1	0.393	0.153	0.317	0.107	0.243	0.23
rel	0.393	1	0.374	0.52	0.325	0.366	0.3
fun	0.153	0.374	1	0.467	0.418	0.338	0.304
clar	0.317	0.52	0.467	1	0.453	0.482	0.367
home	0.107	0.325	0.418	0.453	1	0.4	0.238
class	0.243	0.366	0.338	0.482	0.4	1	0.373
fac	0.23	0.3	0.304	0.367	0.238	0.373	1

Ο συντελεστής συσχέτισης Kendall's Tau υπολογίστηκε στην πλατφόρμα KNIME με τον κόμβο Rank Correlation, χρησιμοποιώντας μεταβλητές σε κλίμακα Likert από το σύνολο δεδομένων "learning". Η ανάλυση επιβεβαιώνει ότι KNIME αποδίδει αντίστοιχα αποτελέσματα με τη γλώσσα προγραμματισμού R. Ο συντελεστής συσχέτισης Gamma μπορεί επίσης να υπολογιστεί με τον κόμβο Rank Correlation, χρησιμοποιώντας τις ίδιες διατακτικές μεταβλητές από το σύνολο δεδομένων "learning".

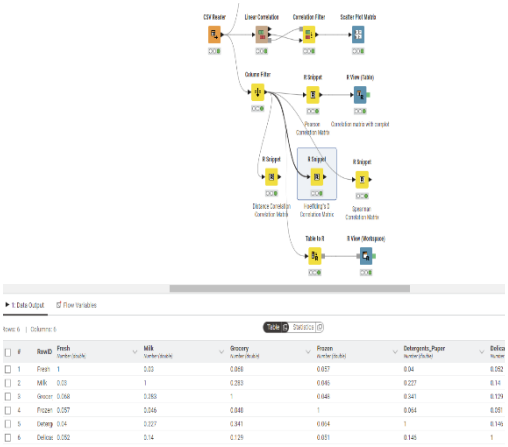
5. ΑΠΟΤΕΛΕΣΜΑΤΑ ΕΝΣΩΜΑΤΩΣΗΣ ΤΗΣ ΓΛΩΣΣΑΣ ΠΡΟΓΡΑΜΜΑΤΙΣΜΟΥ R ΣΤΗΝ ΠΛΑΤΦΟΡΜΑ ΑΝΑΛΥΤΙΚΗΣ ΔΕΔΟΜΕΝΩΝ KNIME.

Για την εφαρμογή εξειδικευμένων μέτρων συσχέτισης στο περιβάλλον της πλατφόρμας αναλυτικής δεδομένων KNIME αξιοποιήθηκε η ενσωμάτωση της γλώσσας προγραμματισμού R. Στην εικόνα 20 βλέπουμε ότι μέσω των κόμβων R Snippet και R View, υπολογίστηκαν μέτρα όπως ο Hoeffding's D και η Distance Correlation, τα οποία δεν υποστηρίζονται απευθείας από την πλατφόρμα.

Εικόνα 20. *(α) Πλήρης ροή εργασιών στο KNIME με ανάλυση συσχετίσεων με ενσωμάτωση της R (β) Δημιουργία πίνακα συσχέτισης (Pearson coefficients) με corrplot package (γ) Δημιουργία πίνακα συσχέτισης (Hoeffding's D coefficient) (δ) Δημιουργία πίνακα συσχέτισης (Distance correlation).*



(α) *(β)*



(γ)



(δ)

6. ΣΥΜΠΕΡΑΣΜΑΤΑ

Η παρούσα εργασία είχε ως στόχο τη διερεύνηση και σύγκριση των μέτρων συσχέτισης, εφαρμόζοντας τεχνικές και εργαλεία αναλυτικής δεδομένων. Μέσα από την παρουσίαση, αναδείχθηκε η σημασία της κατάλληλης επιλογής συντελεστή συσχέτισης, ανάλογα με τη φύση των δεδομένων και τη μορφή της σχέσης μεταξύ των μεταβλητών. Τα ευρέως χρησιμοποιούμενα στατιστικά λογισμικά, όπως το SPSS (IBM SPSS Statistics v.29) και το MINITAB (Minitab Statistical Software v.21), παρέχουν περιορισμένες δυνατότητες ως προς τον υπολογισμό των μέτρων συσχέτισης. Οι διαθέσιμοι συντελεστές περιορίζονται στους πλέον κλασικούς: Pearson's r, Spearman's Rho, Kendall's Tau, Cramer's V, Phi (Φ), καθώς και σε μετρικές που αφορούν την πολυσυγγραμμικότητα, όπως οι VIF, Tolerance, Eigenvalues και ο Condition Index. Ωστόσο, η ανάγκη για ανίχνευση πιο σύνθετων ή μη γραμμικών σχέσεων οδηγεί στην αναγκαιότητα χρήσης προηγμένων συντελεστών που δεν υποστηρίζονται από τα παραπάνω λογισμικά. Συντελεστές όπως ο Distance Correlation, ο Maximal Information Coefficient (MIC) και ο Hoeffding's D, οι οποίοι επιτρέπουν την ανίχνευση πολύπλοκων ή μη γραμμικών εξαρτήσεων που δεν μπορούν να αποτυπωθούν μέσω του συντελεστή συσχέτισης των παραδοσιακών συντελεστών Pearson's r, Spearman Rho, Kendall's Tau. Οι πιο εξειδικευμένοι συντελεστές αυτοί που προσεγγίστηκαν υπολογιστικά στην εργασία, αποδείχθηκαν πολύ σημαντικοί καθώς μπορούν να εντοπίζουν μη

γραμμικές σχέσεις, ιδιαίτερα όταν τα δεδομένα παρουσιάζουν αποκλίσεις από την γραμμικότητα.

Με τη χρήση εργαλείων της αναλυτικής δεδομένων, η ανάλυση συσχέτισης εμπλουτίζεται σημαντικά με τη δυνατότητα εύληπτων απεικονίσεων και γραφημάτων. Η γλώσσα προγραμματισμού R προσφέρει πρόσβαση σε ένα ευρύ φάσμα πακέτων για εξειδικευμένες μεθόδους. Η πλατφόρμα αναλυτικής δεδομένων KNIME, πέρα από τη φιλικότητα του οπτικού προγραμματισμού, επιτρέπει την ενσωμάτωση της R για την υλοποίηση πιο προχωρημένων αναλύσεων. Ο συνδυασμός των δύο εργαλείων απέδειξε πως είναι μια δυνατή ευέλικτη, διαφανής και επιστημονικά τεκμηριωμένη προσέγγιση στην ανάλυση συσχέτισης.

ABSTRACT

Correlation analysis encompasses a wide range of techniques and measures, which depend on the type of data being examined. In addition to classical correlation coefficients, such as Pearson, Spearman's Rho, Kendall's Tau, and Hoeffding's D new metrics like Distance Correlation and the Maximal Information Coefficient (MIC) have been developed to analyze complex and nonlinear relationships.

With the rise of open-source programming software, correlation techniques have evolved computationally, allowing for clear presentation of correlation matrices, multivariate analysis, and data visualization. In this study, results are presented using the R programming language, and computational approaches to correlation are also explored within the KNIME analytics platform through visual programming. The integration of R into KNIME is also employed to produce combined outputs.

Keywords: correlation analysis, R statistical programming language, visual programming, KNIME analytics platform.

ΑΝΑΦΟΡΕΣ

Albanese, D., Filosi, M., Visintainer, R., Riccadonna, S., Jurman, G., & Furlanello, C. (2013). Minerva and minepy: A C engine for the MINE suite and its R, Python and MATLAB wrappers. *Bioinformatics*, *29*(3). https://doi.org/10.1093/bioinformatics/bts707

Asuero, A. G., Sayago, A., & González, A. G. (2006). The correlation coefficient: An overview. In *Critical Reviews in Analytical Chemistry* (Vol. 36, Issue 1, pp. 41–59). https://doi.org/10.1080/10408340500526766

Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Meinl, T., Ohl, P., Thiel, K., & Wiswedel, B. (2007). *KNIME-The Konstanz Information Miner Version 2.0 and beyond*. http://labs.knime.org

Chan, J. Y. Le, Leow, S. M. H., Bea, K. T., Cheng, W. K., Phoong, S. W., Hong, Z. W., & Chen, Y. L. (2022). Mitigating the Multicollinearity Problem and Its Machine Learning Approach: A Review. In *Mathematics* (Vol. 10, Issue 8). MDPI. https://doi.org/10.3390/math10081283

Clark, M. (2013). *A Comparison of Correlation Measures*. Center for Social Research, University of Notre Dame. https://www3.nd.edu/~mclark19/learn/PDFs/ComparisonCorrelation.pdf

Frank E. (2025). *Hmisc: Harrell Miscellaneous*. (Version 5.2-3) [R package manual]. CRAN https://hbiostat.org/R/Hmisc/

Halkos, G. E., & Tsilika, K. D. (2018). Programming Correlation Criteria with free CAS Software. *Computational Economics*, *52*(1), 299–311. https://doi.org/10.1007/s10614-016-9604-1

Halkos, G., Nomikos, S., & Tsilika, K. (2022). Evidence for Novel Structures Relating CSR Reporting and Economic Welfare: Environmental Sustainability—A Continent-Level Analysis. *Computational Economics*, *59*(2), 415–444. https://doi.org/10.1007/s10614-020-10091-5

Hanck, C., Arnold, M., Gerber, A., & Schmelzer, M. (2024). *Introduction to Econometrics with R*.

Hebbali, A. (2025) olsrr: Tools for Building OLS Regression Models. R package version 0.6.1. Available at: https://CRAN.R-project.org/package=olsrr

Hoeffding, W. (1948). A Non-Parametric Test of Independence. In *Source: The Annals of Mathematical Statistics* (Vol. 19, Issue 4).

Hotelling, H. (1936). *Relations between two sets of variables*. 1–58.

Lind, D. A., Marchal, W. G., & Wathen, S. Adam. (2012). *Statistical techniques in business & economics*. McGraw-Hill / Irwin.

Makowski, D., Ben-Shachar, M., Patil, I., & Lüdecke, D. (2020). Methods and Algorithms for Correlation Analysis in R. *Journal of Open-Source Software*, *5*(51). https://doi.org/10.21105/joss.02306

Mangiafico, S.S. (2025). *{rcompanion}: Functions to Support Extension Education Program Evaluation*. https://CRAN.R-project.org/package=rcompanion

McNulty K. (2021). *Package "peopleanalyticsdata" Title Data Sets for Keith McNulty's Handbook of Regression Modeling in People Analytics*. http://peopleanalytics-regression-book.org

Menzel, U. (2022) CCP: Significance Tests for Canonical Correlation Analysis, Version 1.2. https://cran.r-project.org/package=CCP

Peterson, B. G., & Carl, P. (2022). PerformanceAnalytics: Econometric Tools for Performance and Risk Analysis. R package version 2.0.4. *R Package Version*, *1*(3). https://cran.r-project.org/package=PerformanceAnalytics

R Core Team. (2025). *_R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.*

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., Mcvean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M., Pardis, ‡, & Sabeti, C. (2011). *Detecting Novel Associations in Large Data Sets*. https://www.science.org

Revelle, W. (2025). *Package "psych" Title Procedures for Psychological, Psychometric, and Personality Research*. https://personality-project.org/r/psych/

Rizzo, M. L., & Székely, G. J. (2019). energy: E-Statistics: Multivariate Inference via the Energy of Data. In *R package version*. https://CRAN.R-project.org/package=energy

Salmerón, R., García, C. B., & García, J. (2018). Variance Inflation Factor and Condition Number in multiple linear regression. *Journal of Statistical Computation and Simulation*, *88*(12), 2365–2384. https://doi.org/10.1080/00949655.2018.1463376

Székely, G. J., Rizzo, M. L., & Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *Annals of Statistics*, *35*(6), 2769–2794. https://doi.org/10.1214/009053607000000505

Tinbergen, J. (1947). The Use of Correlation Analysis in Economic Research. In *Årg* (Vol. 49).

Trapletti, A. and Hornik, K. (2024). *tseries: Time Series Analysis and Computational Finance*. https://orcid.org/0000-0003-4198-9911

Vinod, H. D. (2017). Generalized correlation and kernel causality with applications in development economics. *Communications in Statistics: Simulation and Computation*, *46*(6), 4513–4534. https://doi.org/10.1080/03610918.2015.1122048

Wei, T., Simko, V., Levy, M., Xie, Y., Jin, Y. J., & Zemla, J. (2017). Visualization of a Correlation Matrix. R package "corrplot". *Statistician*, *56*. https://cran.r-project.org/package=corrplot

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 37ου Πανελληνίου & 3ου Διεθνούς Συνεδρίου Στατιστικής (2025), σελ.177-197

ΔΙΕΡΕΥΝΗΣΗ ΥΠΕΡΠΑΡΑΜΕΤΡΩΝ ΣΕ ΜΟΝΤΕΛΑ RESERVOIR COMPUTING

Ε. Παπαδοπούλου[1], Δ. Κουγιουμτζής[1]
[1]Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης, Πανεπιστημιούπολη, 54124, Θεσσαλονίκη
epapador@ece.auth.gr, dkugiu@auth.gr

ΠΕΡΙΛΗΨΗ

Ο πρόσφατα ανεπτυγμένος αλγόριθμος του Reservoir Computing έχει αποδειχθεί ότι είναι μία πολλά υποσχόμενη και αποτελεσματική προσέγγιση στο πεδίο της πρόβλεψης χρονοσειρών. Υπό αυτό το πλαίσιο, η ρύθμιση του μεγάλου αριθμού υπερπαραμέτρων που διαθέτει, καθίσταται ιδιαίτερα σημαντική διαδικασία καθώς μπορεί να επηρεάσει την απόδοση του. Η παρούσα εργασία αξιολογεί και συγκρίνει την παραδοσιακή μέθοδο της Μπεϋζιανής βελτιστοποίησης των υπερπαραμέτρων του Reservoir Computing με βιο-εμπνευσμένους μετα-ευρετικούς αλγορίθμους. Απώτερος στόχος αποτελεί η βελτίωση της προβλεπτικής ικανότητας του μοντέλου σε διαφορετικά συστήματα πολυμεταβλητών χρονοσειρών, τόσο ντετερμινιστικά όσο και στοχαστικά.

Λέξεις Κλειδιά: Μηχανική μάθηση, Πρόβλεψη χρονοσειρών, Reservoir Computing, Πολυμεταβλητές Χρονοσειρές, Βελτιστοποίηση υπερπαραμέτρων, Μετα-ευρετικοί αλγόριθμοι, Μπεϋζιανή βελτιστοποίηση.

1. ΕΙΣΑΓΩΓΗ

Η πρόβλεψη πολυμεταβλητών χρονοσειρών αποτελεί ένα ιδιαίτερα απαιτητικό έργο καθώς προϋποθέτει εκμάθηση της εξάρτησης των μελλοντικών τιμών καθεμίας μεταβλητής από τις προηγούμενές της αλλά και τη διερεύνηση ύπαρξης πιθανής αλληλεπίδρασης μεταξύ των μεταβλητών (Han et al., 2023; Nguyen-Thai et al., 2024). Μαθηματικά, μία πολυμεταβλητή χρονοσειρά μήκους N συμβολίζεται ως: $\{\boldsymbol{x}_t\}_{t=1}^{N} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_N\}$, όπου $\boldsymbol{x}_t = \{x_{1,t}, x_{2,t}, \dots, x_{K,t}\}$ είναι το διάνυσμα με τις τιμές των K μεταβλητών κατά τη χρονική στιγμή t.

Μία μέθοδος που χρησιμοποιείται ευρέως εξαιτίας της ευκολίας και της ευελιξίας της για την πραγματοποίηση προβλέψεων για πολλά βήματα μπροστά στο χρόνο σε χρονοσειρές είναι η αναδρομική (Sorjamaa et al., 2007). Σε πρώτο στάδιο εκπαιδεύεται ένα μοντέλο έτσι ώστε να πραγματοποιεί προβλέψεις για 1 βήμα μπροστά για όλες τις μεταβλητές, παράγοντας το διάνυσμα $\widehat{\boldsymbol{x}}_{t+1}$ δεδομένου ότι οι παρατηρήσεις ως τη χρονική στιγμή t είναι γνωστές. Σε επόμενο στάδιο, οι συγκεκριμένες εκτιμηθείσες τιμές αναδιαμορφώνονται κατάλληλα ώστε να συμπεριληφθούν στο διάνυσμα εισόδου

για την επακόλουθη πρόβλεψη των παρατηρήσεων των μεταγενέστερων χρονικών στιγμών. Έτσι, γίνεται αντιληπτό ότι αξιοποιείται επαναληπτικά το ίδιο εκπαιδευμένο μοντέλο σε κάθε περίπτωση, όποιος και αν είναι ο ορίζοντας πρόβλεψης.

Η ύπαρξη μη-γραμμικών συσχετίσεων μεταξύ των μεταβλητών σε πολλές πραγματικές χρονοσειρές έχει απομακρύνει τα τελευταία χρόνια τους ερευνητές από τη χρήση παραδοσιακών στατιστικών μοντέλων, όπως τα διανυσματικά αυτοπαλίνδρομα μοντέλα (Vector Autoregressive models, VAR) για την πρόβλεψη χρονοσειρών, με τα νευρωνικά δίκτυα να κερδίζουν έδαφος (Tealab, 2018). Ειδικά τα αναδρομικά νευρωνικά δίκτυα (recursive neural networks, RNN) με την ικανότητα μνήμης του πρόσφατου παρελθόντος τους έχουν αναδειχτεί ως ιδανικά για προβλήματα πρόβλεψης χρονοσειρών (Sharma et al., 2024). Στην οικογένεια των RNN ανήκει και το υπολογιστικό μοντέλο μηχανικής μάθησης Reservoir Computing (RC). Η αρχιτεκτονική του RC, όπως και των μοντέλων μηχανικής μάθησης γενικότερα, καθορίζεται από παραμέτρους που ονομάζονται υπερπαράμετροι. Οι υπερπαράμετροι δεν εκτιμώνται κατά την προσαρμογή του μοντέλου αλλά καθορίζονται από πριν, είτε αυθαίρετα και εμπειρικά (Pathak et al., 2018; Guo et al., 2026) ή από κάποια διερευνητική διαδικασία (Zhu et al., 2025). Οι τιμές των υπερπαραμέτρων είναι καθοριστικές για την απόδοση του μοντέλου.

Η παρούσα εργασία πραγματεύεται τη βελτιστοποίηση των υπερπαραμέτρων του RC με τρεις μεθόδους: τον γενετικό αλγόριθμο (Francescomarino et al., 2018), το σμήνος σωματιδίων (Chouikhi et al., 2016) και τη μπεϋζιανή βελτιστοποίηση (Alibrahim et al., 2021). Γίνεται αξιολόγηση των τριών αυτών μεθόδων μέσω προσομοιωτικής μελέτης με διάφορα δυναμικά συστήματα και στοχαστικές διαδικασίες. Για την αξιολόγηση και σύγκριση των τριών μεθόδων βελτιστοποίησης υπερπαραμέτρων του RC χρησιμοποιούνται και αναπτύσσονται κατάλληλοι στατιστικοί έλεγχοι.

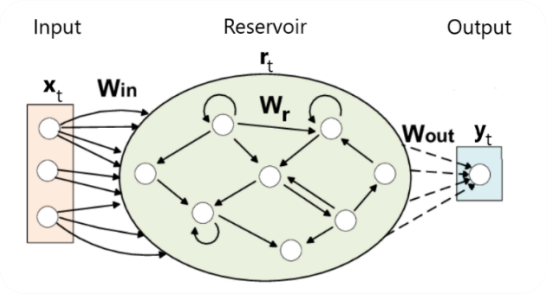
Η δομή της εργασίας είναι ως εξής. Στην ενότητα 2 παρουσιάζεται το RC, οι υπερπαράμετροι του και ο τρόπος λειτουργίας των τριών μεθόδων βελτιστοποίησης των υπερπαραμέτρων. Στη συνέχεια, η ενότητα 3 εστιάζει στην προσομοιωτική μελέτη, παρουσιάζοντας τα συστήματα που παράγουν τις χρονοσειρές, την προεπεξεργασία των δεδομένων και τους στατιστικούς ελέγχους που χρησιμοποιούνται για τη σύγκριση των μεθόδων, καθώς και τα αποτελέσματα των συγκρίσεων για κάθε ορίζοντα πρόβλεψης ξεχωριστά και συγκεντρωτικά για όλους τους ορίζοντες. Τα τελικά συμπεράσματα δίνονται συνοπτικά στην ενότητα 4.

2. ΘΕΩΡΗΤΙΚΟ ΠΛΑΙΣΙΟ

2.1 Το μοντέλο και οι υπερπαράμετροι του Reservoir Computing

Το μοντέλο του Reservoir Computing (RC) αποτελείται από τρία επίπεδα: της εισόδου, του reservoir και της εξόδου (Εικόνα 1).

Εικόνα 1. *Συνδεσιμότητα ενός Reservoir Computing μοντέλου. Προσαρμοσμένη εικόνα από (Skibinsky-Gitlin et al., 2018).*



Σε πρώτο επίπεδο, το διανυσματικό σήμα εισόδου $\boldsymbol{x_t}$ για μια δεδομένη χρονική στιγμή t διοχετεύεται στο κρυφό δυναμικό μέρος του δικτύου - το λεγόμενο στρώμα reservoir. Εκεί, συνδυάζεται με την προηγούμενη εσωτερική κατάσταση του reservoir $\boldsymbol{r_{t-1}}$, προκειμένου να παραχθεί η νέα κατάστασή του για τη χρονική στιγμή t. Η διαδικασία αυτή περιγράφεται από την ακόλουθη εξίσωση:

$$\boldsymbol{r_t} = (1-a)\cdot \boldsymbol{r_{t-1}} + a\cdot tanh(W_{in}\cdot \boldsymbol{x_t} + W_r \cdot \boldsymbol{r_{t-1}}), \quad \boldsymbol{r_0} = \mathbf{0}. \tag{1}$$

Αν θεωρήσουμε ότι το πλήθος των στοιχείων στα στρώματα εισόδου και reservoir είναι N^{in} και N^{r} αντίστοιχα, τότε ισχύει ότι $\boldsymbol{x_t} \in \mathbb{R}^{\mathrm{N}^{in}}$ και $\boldsymbol{r_t} \in \mathbb{R}^{\boldsymbol{N^r}}$. Οι πίνακες βαρών $W_{in} \in \mathbb{R}^{\mathrm{N}^r x \mathrm{N}^{in}}$ και $W_r \in \mathbb{R}^{\mathrm{N}^r x \mathrm{N}^r}$ αρχικοποιούνται με τυχαίες τιμές κατά την έναρξη της διαδικασίας της εκπαίδευσης και παραμένουν αμετάβλητοι καθ' όλη την υπόλοιπη διάρκειά της. Αξίζει να σημειωθεί ότι αυτοί οι πίνακες είναι αραιοί, με τα ποσοστά των μη-μηδενικών στοιχείων τους να αποτελούν υπερπαραμέτρους που ρυθμίζονται από τον χρήστη. Δεδομένου ότι η κατάσταση του reservoir για $t = 0$ καθορίζεται να είναι μηδενική, δημιουργείται ο κίνδυνος οι αρχικές καταστάσεις του συστήματος να επηρεάσουν δυσανάλογα τη δυναμική του reservoir, παράγοντας μη αντιπροσωπευτικές καταστάσεις της μακροπρόθεσμης συμπεριφοράς του στις επόμενες χρονικές στιγμές (Picco et al., 2025). Για το λόγο αυτό, τις περισσότερες φορές απομακρύνεται ένα πλήθος αρχικών χρονικών στιγμών, το λεγόμενο warmup ή washout, στις οποίες οι καταστάσεις του reservoir δεν λαμβάνονται υπόψη στον υπολογισμό της εξόδου και την εκπαίδευση των βαρών (Stepney, 2024).

Η παράμετρος a στην εξίσωση (1) είναι γνωστή ως ρυθμός διαρροής και ρυθμίζει τον βαθμό με τον οποίο η νέα κατάσταση του reservoir επηρεάζεται από την προηγούμενη (Viehweg et al., 2023). Όσο μεγαλύτερη είναι η τιμή του (π.χ. κοντά στο 1), τόσο λιγότερη μνήμη διατηρεί το reservoir από το παρελθόν (Dale et al., 2021) και τόσο πιο έντονα ανταποκρίνεται στα τρέχοντα δεδομένα εισόδου. Μία επιπρόσθετη κρίσιμη υπερπαράμετρος είναι η φασματική ακτίνα ρ του πίνακα W_r, η οποία αντιστοιχεί στη μέγιστη ιδιοτιμή του πίνακα κατά απόλυτη τιμή (Lukoševičius et al., 2009). Καθορίζοντας μία τιμή γι' αυτήν, αναπροσαρμόζονται οι τιμές του W_r μέσω του λόγου ρ/ρ_0, όπου ρ_0 είναι η αρχική φασματική ακτίνα του πίνακα. Κλιμάκωση εφαρμόζεται και στον πίνακα W_{in}, μέσω ενός συντελεστή που πολλαπλασιάζεται με τις μη-μηδενικές τιμές του πίνακα και ορίζει ο χρήστης. Αυτή η διαδικασία επιτρέπει

τον έλεγχο του βαθμού μη-γραμμικότητας στο στρώμα του reservoir (Bianchi et al., 2021).

Σε επόμενο στάδιο, η κατάσταση του reservoir $\boldsymbol{r}_t$ μετασχηματίζεται στην έξοδο του νευρωνικού δικτύου για τη χρονική στιγμή t, μέσω μιας απλής γραμμικής σχέσης. Η πρόβλεψη $\boldsymbol{y}_t = \hat{\boldsymbol{x}}_{t+1}$ δίνεται από τον τύπο:

$$\boldsymbol{y}_t = W_{out} \cdot \boldsymbol{r}_t \,. \tag{2}$$

Ο πίνακας $W_{out} \in \mathbb{R}^{\mathrm{N}^{out} x \mathrm{N}^r}$ σχετίζεται με τη σύνδεση των ενώσεων του reservoir με τις $\mathrm{N}^{out} = K$ σε πλήθος ενώσεις του στρώματος εξόδου. Περιλαμβάνει τα βάρη που αποτελούν τις παραμέτρους του μοντέλου οι οποίες προτιμάται να βελτιστοποιούνται κατά τη διάρκεια του σταδίου εκπαίδευσης με τη μέθοδο της ridge παλινδρόμησης, χωρίς οπισθοδιάδοση του σφάλματος κατά μήκος του χρόνου όπως γίνεται στα παραδοσιακά αναδρομικά νευρωνικά δίκτυα. Η ridge παλινδρόμηση εφαρμόζεται προκειμένου να αποφευχθεί το φαινόμενο της υπερπροσαρμογής, προσθέτοντας έναν επιπλέον όρο στη συνάρτηση σφάλματος που βοηθά στο να διατηρηθεί σε χαμηλά επίπεδα η νόρμα των βαρών. Με αυτόν τον τρόπο, δεν παράγονται πολύ μεγάλες τιμές βαρών που θα καταστούσαν το μοντέλο ιδιαίτερα πολύπλοκο χωρίς την ικανότητα γενίκευσης σε νέα δεδομένα (Wyffels et al., 2008).

Τελικά, το σύνολο των υπερπαραμέτρων που θα πρέπει να ρυθμιστούν από τον χρήστη είναι: 1) το μέγεθος N^r του reservoir, 2) η φασματική ακτίνα ρ του πίνακα W_r, 3) ο ρυθμός διαρροής a, 4) η κλιμάκωση των τιμών του πίνακα W_{in}, 5) ο συντελεστής της ridge παλινδρόμησης, 6) το ποσοστό μη-μηδενικών τιμών του πίνακα W_{in}, 7) το ποσοστό μη-μηδενικών τιμών του πίνακα W_r και 8) το πλήθος των αρχικών χρονικών στιγμών που αγνοούνται washout (Hinaut et al., 2021).

2.2 Γενετικός Αλγόριθμος

Ο γενετικός αλγόριθμος (Genetic Algorithm / GA) ανήκει στους εξελικτικούς αλγορίθμους) εμπνευσμένος από τη διαδικασία φυσικής επιλογής και εξέλιξης του Δαρβίνου (Simon, 2013; Loussaief et al., 2018). Εφαρμόζεται για την επίλυση προβλημάτων βελτιστοποίησης, ιδιαίτερα όταν οι χώροι αναζήτησης λύσεων είναι μεγάλοι (Erden, 2023). Στόχος είναι η εύρεση της βέλτιστης ή μίας σχεδόν βέλτιστης λύσης (Francescomarino et al., 2018). Ακολουθεί η περιγραφή του αλγορίθμου για συγκεκριμένες τιμές υπερπαραμέτρων, όπως έχει υλοποιηθεί στην παρούσα εργασία.

Κάθε γενιά αποτελείται από έναν πληθυσμό χρωμοσωμάτων-ατόμων ο οποίος στην πραγματικότητα είναι ένα σύνολο πιθανών λύσεων. Κάθε λύση αντιπροσωπεύει ένα σημείο στον χώρο αναζήτησης και αναπαρίσταται από ένα διάνυσμα 8 πραγματικών τιμών, όσοι και οι υπερπαράμετροι του RC, που ονομάζονται γονίδια. Ο γενετικός αλγόριθμος ξεκινά με την τυχαία αρχικοποίηση του πληθυσμού με 20 άτομα. Έπειτα, αξιολογείται η αποδοτικότητα κάθε ατόμου μέσω μίας συνάρτησης καταλληλότητας (fitness function). Στο πλαίσιο της παρούσας μελέτης ως δείκτης αξιολόγησης εκλαμβάνεται η συνάρτηση σφάλματος, με την καταλληλότητα να συσχετίζεται αντιστρόφως ανάλογα με το σφάλμα - δηλαδή, όσο μικρότερο είναι το σφάλμα, τόσο υψηλότερη είναι η καταλληλότητα του ατόμου.

Βάσει των τιμών αυτών, πραγματοποιείται μία διαδικασία επιλογής των πιο κατάλληλων ατόμων που θα αποτελέσουν τους γονείς (Mitchell, 1995). Αυτή

επιλέγεται να είναι η μέθοδος τουρνουά κατά την οποία επιλέγονται τυχαία 3 άτομα του πληθυσμού, το καταλληλότερο από τα οποία θα αποτελέσει τον γονέα. Έπειτα, με τη διασταύρωση, κάθε ζευγάρι γονέων συνδυάζεται κατάλληλα ο ένας με τον άλλον, ανταλλάσσοντας γονίδια μεταξύ τους έτσι ώστε να παράξουν τους λεγόμενους απογόνους, δηλαδή νέες λύσεις. Υπάρχουν διάφορα είδη διασταύρωσης, με την ομοιόμορφη να είναι αυτή που χρησιμοποιείται στη συγκεκριμένη εργασία. Σύμφωνα με αυτήν, αποφασίζεται, για κάθε απόγονο, με πιθανότητα διασταύρωσης ίση με 0.5, το εάν κάθε γονίδιο του θα προέρχεται από τον πρώτο ή από τον δεύτερο γονέα. Επόμενο βήμα είναι η μετάλλαξη των απογόνων η οποία συμβάλλει στην εξερεύνηση νέων περιοχών λύσεων. Κατά τη διάρκειά της τυχαίας μετάλλαξης, για κάθε απόγονο επιλέγεται με πιθανότητα 0.2 ένα σύνολο γονιδίων στα οποία ανατίθεται μία τυχαία τιμή από το σύνολο των επιτρεπόμενων τιμών τους.

Τελικά, σχηματίζεται ένα σύνολο λύσεων, ίδιου πλήθους με το αρχικό, που θα αποτελέσει τον πληθυσμό της επόμενης γενιάς. Η διαδικασία αξιολόγησης και εξέλιξης του νέου πληθυσμού επαναλαμβάνεται μέχρι να ικανοποιηθεί το κριτήριο τερματισμού του αλγορίθμου που είναι η ολοκλήρωση 50 γενεών, οπότε και εξάγεται το καταλληλότερο άτομο μέχρι εκείνη τη στιγμή (Alibrahim et al., 2021). Αυτό θα περιλαμβάνει τις καλύτερες τιμές υπερπαραμέτρων.

Συγκεντρωτικά, η λογική του γενετικού αλγορίθμου συνοψίζεται στο ότι τα άτομα με υψηλότερη τιμή συνάρτησης καταλληλότητας (fitness) έχουν μεγαλύτερη πιθανότητα να επιλεγούν για αναπαραγωγή (Zhong et al., 2018). Η διασταύρωση μεταξύ αυτών των «καλύτερων» λύσεων έχει ως στόχο τη δημιουργία απογόνων, κάποιοι από τους οποίους είναι πιθανό να κληρονομήσουν τα ευνοϊκά χαρακτηριστικά των γονέων τους. Παρ' ότι δεν είναι πάντα εξασφαλισμένη η παραγωγή βελτιωμένων λύσεων, η εισαγωγή στοχαστικών παραγόντων, όπως η μετάλλαξη, αυξάνει τις πιθανότητες εντοπισμού καλύτερων ατόμων μετά από την ολοκλήρωση κάθε γενιάς (Gad, 2024).

2.3 Σμήνος Σωματιδίων

Η βελτιστοποίηση με σμήνος σωματιδίων (Particle Swarm Optimization / PSO) είναι εμπνευσμένη από την κοινωνική συμπεριφορά των σμηνών πτηνών. Ανταλλάσσεται πληροφορία μεταξύ των μελών του σμήνους για τις καλύτερες θέσεις στις οποίες έχουν βρεθεί και μετά υπάρχει ανάλογη προσαρμογή της κατεύθυνσης και της ταχύτητάς τους (Fan et al., 2011). Ο αλγόριθμος εφαρμόζεται ευρέως σε πολύπλοκα μη-γραμμικά προβλήματα βελτιστοποίησης εξαιτίας του απλού και κατανοητού εννοιολογικού πλαισίου που τη διέπει (Kachitvichyanukul, 2012). Πετυχαίνει γρήγορη σύγκλιση σε καλές λύσεις, απαιτώντας από τον χρήστη τη ρύθμιση σχετικά λίγων υπερπαραμέτρων (Chouikhi et al., 2016) και πλεονεκτεί σε προβλήματα όπου οι μεταβλητές λαμβάνουν πραγματικές τιμές σε συνεχή διαστήματα (Wu et al., 2023).

Στη συγκεκριμένη μεθοδολογία ο πληθυσμός είναι ένα σμήνος σωματιδίων, σε καθένα από τα οποία αντιστοιχίζεται ένα ζευγάρι διανυσμάτων θέσης και ταχύτητας. Δεδομένου ότι το πλήθος των υπερπαραμέτρων προς βελτιστοποίηση είναι οκτώ,

θεωρούμε ότι κάθε σωματίδιο κινείται σε οκταδιάστατο χώρο. Έτσι γίνεται αντιληπτό ότι μία δεδομένη θέση ενός σωματιδίου υποδηλώνει την ανάθεση συγκεκριμένων τιμών για τις οκτώ υπερπαραμέτρους στο Reservoir Computing.

Ο αλγόριθμος ξεκινά με τη δημιουργία ενός αρχικού σμήνους 20 σωματιδίων με τυχαία θέση και ταχύτητα για καθένα από αυτά. Στη συνέχεια, αξιολογείται η καταλληλότητα κάθε σωματιδίου λαμβάνοντας υπόψη το πόσο ελαχιστοποιείται η τιμή της συνάρτησης σφάλματος όπως και στο γενετικό αλγόριθμο. Με τη βοήθεια των υπολογισμένων αυτών τιμών, εντοπίζεται η προσωπική καλύτερη θέση για κάθε σωματίδιο ξεχωριστά (personal best/pbest) αλλά και η ολική βέλτιστη θέση για όλο το σμήνος (global best/gbest). Σε επόμενο βήμα, ενημερώνονται η ταχύτητα και η θέση κάθε σωματιδίου σύμφωνα με τους τύπους (3) και (4) στη γενική περίπτωση, παράγοντας πρακτικά ένα νέο σμήνος για την επόμενη επανάληψη. Σε αυτό το σημείο χρονικά σηματοδοτείται το τέλος μίας επανάληψης, με τη διαδικασία να επαναλαμβάνεται, στη συγκεκριμένη μελέτη, συνολικά 50 φορές (ίδιο κριτήριο τερματισμού με τον γενετικό αλγόριθμο).

Οι ακόλουθοι τύποι αφορούν τον καθορισμό της ταχύτητας $\boldsymbol{v}_i{}^{t+1}$ και της θέσης $\boldsymbol{x}_i{}^{t+1}$ που θα έχει το σωματίδιο i κατά την επανάληψη $t+1$:

$$\boldsymbol{v}_i{}^{t+1} = w^{t+1} \cdot \boldsymbol{v}_i{}^{t} + c_1 \cdot \boldsymbol{r}_1 \odot \left(\boldsymbol{pbest}_i{}^{t} - \boldsymbol{x}_i{}^{t}\right) + c_2 \cdot \boldsymbol{r}_2 \odot (\boldsymbol{gbest}^{t} - \boldsymbol{x}_i{}^{t}) \quad (3)$$

$$\boldsymbol{x}_i{}^{t+1} = \boldsymbol{x}_i{}^{t} + \boldsymbol{v}_i{}^{t+1} \quad (4)$$

Γίνεται εύκολα αντιληπτό ότι με $\boldsymbol{pbest}_i{}^{t}$ συμβολίζεται η καλύτερη θέση στην οποία έχει βρεθεί το σωματίδιο i μέχρι και την επανάληψη t, ενώ με $\boldsymbol{gbest}^{t}$ η ολική βέλτιστη θέση που έχει εντοπιστεί και αφορά όλα τα σωματίδια του σμήνους μέχρι την ίδια επανάληψη. Οι συντελεστές c_1 και c_2 καλούνται γνωστική και κοινωνική παράμετρος και ρυθμίζονται να είναι ίσοι με 0.1 και 0.2 αντίστοιχα. Ο πρώτος εκφράζει τον βαθμό επίδρασης της νέας ταχύτητας κάθε σωματιδίου από την προσωπική εμπειρία του μέχρι εκείνη τη στιγμή, ενώ ο δεύτερος από την εμπειρία ολόκληρου του σμήνους. Τα διανύσματα $\boldsymbol{r}_1$ και $\boldsymbol{r}_2$ που εμφανίζονται στη σχέση (3) και πολλαπλασιάζονται στοιχείο προς στοιχείο (γινόμενο Hadamard που συμβολίζεται ως $\odot$) με τα διανύσματα απόστασης από τις βέλτιστες θέσεις, έχουν ως συνιστώσες τυχαίους αριθμούς από την ομοιόμορφη κατανομή στο διάστημα [0, 1] (Liu et al., 2018).

Επιπλέον, στην εξίσωση (3) εμφανίζεται ένας συντελεστής αδράνειας w ο οποίος αντισταθμίζει την νέα ταχύτητα κάθε σωματιδίου σε σχέση με την προηγούμενη ταχύτητά του. Η τιμή του μειώνεται γραμμικά καθώς περνούν οι επαναλήψεις σύμφωνα με τον τύπο (5), με τη μέγιστη και ελάχιστη τιμή του να καθορίζονται από τον χρήστη.

$$w^{t} = w_{max} - \frac{(w_{max} - w_{min}) \cdot t}{T} \quad (5)$$

Η μεταβλητή T παριστάνει το συνολικό πλήθος των επαναλήψεων του αλγορίθμου, με τη w_{max} να τίθεται ίση με 0.9 και τη w_{min} ίση με 0.4.

2.4 Μπεϋζιανή βελτιστοποίηση

Η Μπεϋζιανή βελτιστοποίηση (Bayesian Optimization / BO) στηρίζει τη λειτουργία της σε δύο βασικά στοιχεία: ένα πιθανοτικό προσεγγιστικό μοντέλο και μία συνάρτηση γνωστή ως συνάρτηση απόκτησης (Alibrahim et al., 2021). Πιο συγκεκριμένα, υποθέτει ότι η συνάρτηση σφάλματος για την οποία αναζητείται η ελάχιστη τιμή δεν είναι γνωστή. Οπότε εφαρμόζει μία μεθοδολογία προσέγγισης της που βασίζεται στην υπόθεση ότι αυτή η συνάρτηση προέρχεται από μία κατανομή πιθανότητας σε έναν χώρο συναρτήσεων (Yperman et al., 2017).

Στο ξεκίνημα της διαδικασίας επιλέγεται τυχαία ένα πεπερασμένο πλήθος σημείων $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n$, καθένα από τα οποία είναι ένας συνδυασμός τιμών των οκτώ υπερπαραμέτρων. Οι τιμές της συνάρτησης σφάλματος f σε καθένα από αυτά τα σημεία θεωρούνται τυχαίες μεταβλητές που ακολουθούν μία πολυμεταβλητή Γκαουσιανή κατανομή, δηλαδή:

$$f(\boldsymbol{x}_{1:n}) \sim N\big(\mu(\boldsymbol{x}_{1:n}), K(\boldsymbol{x}_{1:n}, \boldsymbol{x}_{1:n})\big). \tag{6}$$

Έτσι αποκτάται μία εκ των προτέρων κατανομή για τις τιμές $f(\boldsymbol{x}_1), f(\boldsymbol{x}_2), \dots, f(\boldsymbol{x}_n)$ με μέση τιμή το διάνυσμα στήλη $\mu(\boldsymbol{x}_{1:n}) = [\mu(\boldsymbol{x}_1), \dots, \mu(\boldsymbol{x}_n)]'$ και πίνακα διασπορών συνδιασπορών τον $K(\boldsymbol{x}_{1:n}, \boldsymbol{x}_{1:n}) = [k(\boldsymbol{x}_1, \boldsymbol{x}_1), \dots, k(\boldsymbol{x}_1, \boldsymbol{x}_n); \dots; k(\boldsymbol{x}_n, \boldsymbol{x}_1), \dots, k(\boldsymbol{x}_n, \boldsymbol{x}_n)]$ (Frazier, 2018). Κάθε συνιστώσα του διανύσματος $\mu(\boldsymbol{x}_{1:n})$, έστω η $\mu(\boldsymbol{x}_i)$, είναι η μέση τιμή των τιμών των συναρτήσεων στο σημείο $\boldsymbol{x}_i$ και συνήθως εκλαμβάνεται να είναι ίση με μηδέν. Ως $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ παριστάνεται η συνάρτηση πυρήνα για τα σημεία $\boldsymbol{x}_i$ και $\boldsymbol{x}_j$ και φανερώνει τη συσχέτιση μεταξύ διαφορετικών σημείων του χώρου αναζήτησης των υπερπαραμέτρων.

Στη συνέχεια, λαμβάνονται οι πραγματικές τιμές της συνάρτησης σφάλματος σε αυτά τα σημεία $f(\boldsymbol{x}_{1:n}) = [f(\boldsymbol{x}_1), \dots, f(\boldsymbol{x}_n)]'$. Με τη βοήθειά τους μπορεί να εκτιμηθεί σε οποιοδήποτε νέο σημείο $\boldsymbol{x}$ η δεσμευμένη κατανομή $f(\boldsymbol{x}) | f(\boldsymbol{x}_{1:n})$ ως εξής (Brochu et al., 2024):

$$f(\boldsymbol{x}) \mid f(\boldsymbol{x}_{1:n}), \boldsymbol{x} \sim N\big(\mu(\boldsymbol{x}), \sigma^2(\boldsymbol{x})\big), \tag{7}$$

όπου

$$\mu(\boldsymbol{x}) = \mathbf{k}' \cdot K(\boldsymbol{x}_{1:n}, \boldsymbol{x}_{1:n})^{-1} \cdot f(\boldsymbol{x}_{1:n}) \tag{8}$$

$$\sigma^2(\boldsymbol{x}) = k(\boldsymbol{x}, \boldsymbol{x}) - \mathbf{k}' \cdot K(\boldsymbol{x}_{1:n}, \boldsymbol{x}_{1:n})^{-1} \cdot \mathbf{k} \tag{9}$$

$$\mathbf{k} = [k(\boldsymbol{x}_1, \boldsymbol{x}), \dots, k(\boldsymbol{x}_n, \boldsymbol{x})]'. \tag{10}$$

Η μεθοδολογία, αποσκοπώντας στην επιλογή κατάλληλου νέου σημείου του χώρου προς αξιολόγηση, αξιοποιεί τη συνάρτηση αναμενόμενης βελτίωσης (Expected Improvement / EI). Αυτή είναι μία από τις πιο διαδεδομένες συναρτήσεις απόκτησης καθώς επιτυγχάνει ισορροπία ανάμεσα στην εκμετάλλευση (αναζήτηση κοντινών σημείων σε εκείνο που έχει τη μικρότερη μέση τιμή στη συνάρτηση σφάλματος) και την εξερεύνηση (αναζήτηση σε περιοχές με μεγάλη αβεβαιότητα) (Garnett, 2015). Εκφράζει την αναμενόμενη βελτίωση στην τιμή της συνάρτησης σφάλματος σε περίπτωση αξιολόγησης ενός νέου σημείου $\boldsymbol{x}$. Δίνεται από τον τύπο:

$$EI(\boldsymbol{x}) = \begin{cases} \big(f(\boldsymbol{x}^+) - \mu(\boldsymbol{x})\big) \cdot \Phi(Z) + \sigma(\boldsymbol{x}) \cdot \varphi(Z), & \text{εάν } \sigma(\boldsymbol{x}) > 0 \\ 0, & \text{εάν } \sigma(\boldsymbol{x}) = 0 \end{cases}, \tag{11}$$

όπου $Z = \frac{f(\boldsymbol{x}^+) - \mu(\boldsymbol{x})}{\sigma(\boldsymbol{x})}$ και $f(\boldsymbol{x}^+) = min_{\boldsymbol{x}_i}\big(f(\boldsymbol{x}_i)\big)$. Με $\Phi(\cdot)$ και $\varphi(\cdot)$ συμβολίζονται οι συναρτήσεις αθροιστικής κατανομής και πυκνότητας πιθανότητας της τυπικής κανονικής κατανομής αντίστοιχα. Τελικά, το νέο σημείο $\boldsymbol{x}_{n+1}$ που επιλέγεται είναι αυτό που ικανοποιεί την ακόλουθη σχέση:

$$\boldsymbol{x}_{n+1} = argmax_{\boldsymbol{x}}\big(EI(\boldsymbol{x})\big) \tag{12}$$

και μαζί με την πραγματική τιμή της συνάρτησης σφάλματος σε αυτό, $f(\boldsymbol{x}_{n+1})$, προστίθενται στο σύνολο των σημείων που έχουν ήδη αξιολογηθεί. Με αυτόν τον τρόπο ολοκληρώνεται η πρώτη φάση λειτουργίας του αλγορίθμου και ακολουθεί επανάληψη της ίδιας διαδικασίας μέχρι να ολοκληρωθεί το πλήθος των 50 αξιολογήσεων, οπότε και να εξαχθεί το σημείο με τη μικρότερη παρατηρούμενη τιμή σφάλματος.

3. ΠΡΟΣΟΜΟΙΩΤΙΚΗ ΜΕΛΕΤΗ

3.1 Συστήματα χρονοσειρών

Τα συστήματα που παράγουν τις χρονοσειρές στις οποίες είναι βασισμένη η παρούσα μελέτη πρόβλεψης είναι τόσο αιτιοκρατικά όσο και στοχαστικά.

Ως αιτιοκρατικό δυναμικό σύστημα διακριτού χρόνου χρησιμοποιείται το coupled Hénon σύστημα με τοπολογία ανοικτού δακτυλίου. Στο σύστημα αυτό, κάθε κόμβος του δακτυλίου αντιστοιχεί σε μία απεικόνιση Hénon, η οποία αλληλεπιδρά με τις γειτονικές της. Ο μαθηματικός ορισμός του συστήματος για K συζευγμένες απεικονίσεις δίνεται από τις εξής εξισώσεις:

$$\begin{cases} x_{1,t} = 1.4 - {x_{1,t-1}}^2 + 0.3 \cdot x_{1,t-2} \\ x_{j,t} = 1.4 - \left[0.5 \cdot c \cdot \left(x_{j-1,t-1} + x_{j+1,t-1}\right) + (1-c) \cdot x_{j,t-1}\right]^2 + 0.3 \cdot x_{j,t-2} \\ j = 2, \dots, K-1 \\ x_{K,t} = 1.4 - {x_{K,t-1}}^2 + 0.3 \cdot x_{K,t-2} \end{cases} \tag{13}$$

Η μεταβλητή c εκφράζει τον βαθμό σύζευξης κάθε μεταβλητής, εκτός της πρώτης και της τελευταίας, με τις γειτονικές της στον δακτύλιο, δηλαδή την εξάρτηση της στην προηγούμενη και στην επόμενή της. .

Εξετάστηκαν διάφορες τιμές για την ένταση σύζευξης c, το πλήθος K των μεταβλητών και το μήκος N της χρονοσειράς. Αναλυτικότερα, τα σύνολα τιμών τους ήταν: $c \in \{0,\ 0.1,\ 0.2,\ 0.4\}$, $K \in \{5,\ 20\}$ και $N \in \{256,\ 1024\}$. Για καθέναν από τους $4 \cdot 2 \cdot 2 = 16$ δυνατούς συνδυασμούς των (c, K, N) παράγονται 100 χρονοσειρές ρυθμίζοντας κάθε φορά διαφορετικές αρχικές συνθήκες, εξασφαλίζοντας, έτσι, τη στατιστική αξιοπιστία των επακόλουθων συμπερασμάτων.

Όσον αφορά τα στοχαστικά συστήματα, εφαρμόστηκε ένα σύστημα χαμηλής διάστασης ($K = 5$) και ένα άλλο μεγαλύτερης ($K = 20$). Αυτά περιγράφονται με τους τύπους (14) και (15). Σε καθένα από αυτά, το μήκος N των χρονοσειρών μεταβαλλόταν μεταξύ των τιμών 256 και 1024, και το πλήθος των πραγματοποιήσεων για κάθε συνδυασμό K και N ήταν και πάλι 100. Για $K = 5$, το στοχαστικό σύστημα είναι γραμμικό και διαμορφώνεται ως εξής (Schelter et al., 2006):

$$\begin{cases} x_{1,t} = 0.4 \cdot x_{1,t-1} - 0.5 \cdot x_{1,t-2} + 0.4 \cdot x_{5,t-1} + \eta_{1,t} \\ x_{2,t} = 0.4 \cdot x_{2,t-1} - 0.3 \cdot x_{1,t-4} + 0.4 \cdot x_{5,t-2} + \eta_{2,t} \\ x_{3,t} = 0.5 \cdot x_{3,t-1} - 0.7 \cdot x_{3,t-2} - 0.3 \cdot x_{5,t-3} + \eta_{3,t} \\ x_{4,t} = 0.8 \cdot x_{4,t-3} + 0.4 \cdot x_{1,t-2} + 0.3 \cdot x_{2,t-2} + \eta_{4,t} \\ x_{5,t} = 0.7 \cdot x_{5,t-1} - 0.5 \cdot x_{5,t-2} - 0.4 \cdot x_{4,t-1} + \eta_{5,t} \end{cases} \quad (14)$$

Οι συνιστώσες λευκού θορύβου $\eta_{i.t}, i = 1,2, \dots, 5$ που εμφανίζονται είναι ανεξάρτητες μεταξύ τους και ακολουθούν κανονική κατανομή με μέση τιμή 0 και τυπική απόκλιση 1.

Το δεύτερο στοχαστικό σύστημα έχει διάσταση $K = 20$ και περιλαμβάνει τόσο γραμμικούς όσο και μη-γραμμικούς πολυωνυμικούς όρους (τετραγωνικούς όρους καθώς και όρους αλληλεπίδρασης μεταξύ μεταβλητών με διαφορά μίας υστέρησης). Ορίζεται μέσω των παρακάτω εξισώσεων (Fotiadis et al., 2024), όπου οι στοχαστικοί όροι $e_{j,t}, j = 1,2, \dots, 20$ έχουν ίδια χαρακτηριστικά με τις συνιστώσες $\eta_{i.t}$:

$$\begin{cases} x_{1,t} = 0.8 \cdot x_{1,t-1} - 0.1 \cdot x_{3,t-2} - 0.3 \cdot x_{19,t-2} + e_{1,t} \\ x_{2,t} = 0.8 \cdot x_{2,t-1} - 0.2 \cdot x_{1,t-2} \cdot x_{15,t-1} + e_{2,t} \\ x_{3,t} = 0.8 \cdot x_{3,t-1} + e_{3,t} \\ x_{4,t} = 0.8 \cdot x_{4,t-1} - 0.2 \cdot x_{5,t-2} \cdot x_{7,t-1} + e_{4,t} \\ x_{5,t} = 0.8 \cdot x_{5,t-1} - 0.15 \cdot x_{3,t-1}^2 + e_{5,t} \\ x_{6,t} = 0.8 \cdot x_{6,t-1} - 0.4 \cdot x_{3,t-3} + 0.1 \cdot x_{10,t-1}^2 + e_{6,t} \\ x_{7,t} = 0.8 \cdot x_{7,t-1} + e_{7,t} \\ x_{8,t} = -0.6 \cdot x_{8,t-1} + 0.1 \cdot x_{11,t-3} \cdot x_{19,t-2} + e_{8,t} \\ x_{9,t} = 0.8 \cdot x_{9,t-1} - 0.1 \cdot x_{20,t-1} + e_{9,t} \\ x_{10,t} = 0.8 \cdot x_{10,t-1} + e_{10,t} \\ x_{11,t} = 0.8 \cdot x_{11,t-1} + 0.1 \cdot x_{20,t-1} - 0.1 \cdot x_{4,t-2}^2 + e_{11,t} \\ x_{12,t} = 0.8 \cdot x_{12,t-1} + e_{12,t} \\ x_{13,t} = 0.8 \cdot x_{13,t-1} + e_{13,t} \\ x_{14,t} = 0.8 \cdot x_{14,t-1} - 0.2 \cdot x_{4,t-3}^2 + e_{14,t} \\ x_{15,t} = 0.8 \cdot x_{15,t-1} + e_{15,t} \\ x_{16,t} = 0.8 \cdot x_{16,t-1} + e_{16,t} \\ x_{17,t} = 0.8 \cdot x_{17,t-1} - 0.1 \cdot x_{7,t-2}^2 + e_{17,t} \\ x_{18,t} = 0.8 \cdot x_{18,t-1} - 0.3 \cdot x_{7,t-2} + 0.5 \cdot x_{3,t-2} \cdot x_{20,t-1} + e_{18,t} \\ x_{19,t} = 0.8 \cdot x_{19,t-1} + e_{19,t} \\ x_{20,t} = 0.8 \cdot x_{20,t-1} + e_{20,t} \end{cases} \quad (15)$$

3.2 Υπολογιστικός σχεδιασμός προσομοίωσης

Μετά από την παραγωγή των παρατηρήσεων μίας πολυμεταβλητής χρονοσειράς μέσω των εξισώσεων των συστημάτων της προηγούμενης υποενότητας, καθορίζονται οι είσοδοι και οι επιθυμητοί έξοδοι του μοντέλου για το στάδιο της εκπαίδευσης. Συγκεκριμένα, για δεδομένη χρονική στιγμή δημιουργείται ένα διάνυσμα εισόδου με τις τιμές όλων των μεταβλητών εντός ενός παραθύρου m προηγούμενων χρονικών στιγμών, ενώ ως έξοδο τίθεται το διάνυσμα τιμών όλων των μεταβλητών στην επόμενη χρονική στιγμή. Η επιλογή του m βασίζεται στη μέγιστη υστέρηση που εμφανίζεται στις εξισώσεις των συστημάτων, διασφαλίζοντας την πλήρη κάλυψη της δυναμικής τους. Μία ενδεχόμενη υποεκτίμηση του οδηγεί σε απώλεια κρίσιμων χρονικών

συσχετίσεων, ενώ μία υπερεκτίμηση του αυξάνει τη διαστατικότητα και τον κίνδυνο υπερπροσαρμογής.

Τελικά για τη χρονική στιγμή t, η είσοδος $\mathbf{X}_t$ και η έξοδος $\boldsymbol{Y}_t$ του μοντέλου δίνονται όπως παρακάτω:

$$\mathbf{X}_t = [\mathbf{x}_t{}', \mathbf{x}'_{t-1}, \dots, \mathbf{x}'_{t-m+1}] \tag{16}$$

$$\boldsymbol{Y}_t = \boldsymbol{x}'_{t+1} \tag{17}$$

Σειρά έχει ο διαμερισμός σε τρία υποσύνολα: το σύνολο εκπαίδευσης (60% των αρχικών παρατηρήσεων), το σύνολο επικύρωσης (επόμενο 20%) και το σύνολο ελέγχου (τελευταίο 20%). Κατά την εκπαίδευση, οι παράμετροι του μοντέλου προσαρμόζονται με στόχο την όσο το δυνατόν πιο αποτελεσματική εκμάθηση των επιθυμητών εξόδων. Το υποσύνολο επικύρωσης αξιοποιείται για την επιλογή των βέλτιστων τιμών των υπερπαραμέτρων χωρίς να γίνεται περαιτέρω εκπαίδευση, επιτρέποντας στον χρήστη την αξιολόγηση της απόδοσης του μοντέλου σε νέα άγνωστα δεδομένα. Τέλος, με τη βοήθεια του συνόλου ελέγχου εκτιμάται η προβλεπτική του ικανότητα σε δεδομένα που δεν έχει προηγουμένως επεξεργαστεί.

Τα δεδομένα, στη συνέχεια, μετασχηματίζονται μέσω της κανονικοποίησης ελαχίστου-μεγίστου έτσι ώστε να βρίσκονται στο διάστημα μεταξύ -1 και 1. Υποθέτοντας ότι οι x_{min} και x_{max} αντιπροσωπεύουν την ελάχιστη και μέγιστη τιμή μίας μεταβλητής αντίστοιχα πριν την κανονικοποίηση, η παλιά τιμή x μετατρέπεται στην νέα κανονικοποιημένη τιμή x_{new} σύμφωνα με τον τύπο (18):

$$x_{new} = 2 \cdot \frac{x - x_{min}}{x_{max} - x_{min}} - 1, \tag{18}$$

Ο μετασχηματισμός αυτός αποτελεί επέκταση της κλασικής κανονικοποίησης στο διάστημα [0,1] και τοποθετεί τις τιμές καθεμίας μεταβλητής σε ένα συμμετρικό διάστημα γύρω από το 0, χωρίς να αλλοιώνει τη μεταξύ τους σχέση. Εναλλακτικά, θα μπορούσε να έχει χρησιμοποιηθεί κάποια άλλη τεχνική κλιμάκωσης των χαρακτηριστικών, όπως η z-score κανονικοποίηση (Pranolo et al., 2024). Με αυτήν, η νέα μεταβλητή θα είχε μέση τιμή 0 και διασπορά 1.

Ο χώρος αναζήτησης για τις οκτώ υπερπαραμέτρους του μοντέλου RC, όπως αυτές περιγράφηκαν στην ενότητα 2.1, παρουσιάζεται στον Πίνακα 1 και είναι ο ίδιος για τις τρεις μεθόδους βελτιστοποίησης.

Πίνακας 1. *Χώρος αναζήτησης για τις υπερπαραμέτρους του μοντέλου RC.*

Υπερπαράμετρος	**Χώρος αναζήτησης**	**Υπερπαράμετρος**	**Χώρος αναζήτησης**
Μέγεθος N^r του reservoir	50, 51, …, 500	Συντελεστής της ridge παλινδρόμησης	[0, 1]
Φασματική ακτίνα ρ του πίνακα W_r	[10^{-1}, 2]	Ποσοστό μη-μηδενικών τιμών του πίνακα W_{in}	[0, 1]
Ρυθμός διαρροής a	[0, 1]	Ποσοστό μη-μηδενικών τιμών του πίνακα W_r	[0, 1]
Κλιμάκωση των τιμών του πίνακα W_{in}	[-1, 1]	Πλήθος αρχικών χρονικών στιγμών που αγνοούνται washout	[0,1,…, 10% του μεγέθους του συνόλου εκπαίδευσης]

Τελικά, επιλέγονται εκείνες οι τιμές υπερπαραμέτρων που ελαχιστοποιούν τον μέσο όρο των επιμέρους τιμών του στατιστικού του σφάλματος για τους δέκα ορίζοντες πρόβλεψης στο σύνολο επικύρωσης. Το στατιστικό του σφάλματος για έναν ορίζοντα πρόβλεψης h ορίζεται σύμφωνα με τον ακόλουθο τύπο:

$$NRMSE(h) = \frac{\sqrt{det\left(cov\left(\boldsymbol{y}_{true}(h) - \boldsymbol{y}_{pred}(h)\right)\right)}}{\sqrt{det\left(cov(\boldsymbol{y}_{true}(h))\right)}}, \quad h = 1,2, \dots ,10. \tag{19}$$

Ο συμβολισμός $cov(.)$ υποδηλώνει τον πίνακα των διασπορών-συνδιασπορών. Στον αριθμητή του κλάσματος ο συγκεκριμένος πίνακας υπολογίζεται για τα σφάλματα που παρουσιάζονται λόγω της διαφοράς των πραγματικών τιμών των K μεταβλητών ($\boldsymbol{y}_{true}$) από τις εκτιμώμενες του μοντέλου ($\boldsymbol{y}_{pred}$) για τον ορίζοντα πρόβλεψης h, ενώ στον παρονομαστή ο πίνακας αφορά αποκλειστικά τις πραγματικές τιμές. Όσο πιο κοντά στο 0 είναι η τιμή του στατιστικού, τόσο πιο αποτελεσματικά προβλέπει το μοντέλο για όλα τα βήματα μπροστά στο χρόνο, ενώ σε περίπτωση που ξεπεράσει το 1, μπορούμε να συμπεράνουμε ότι το μοντέλο κάνει χειρότερες προβλέψεις από αυτές που θα προέκυπταν χρησιμοποιώντας τη μέση τιμή. Η ύπαρξη του παρονομαστή, γενικά, κανονικοποιεί το στατιστικό του σφάλματος, επιτρέποντας τη σύγκριση της απόδοσης του μοντέλου σε διαφορετικούς χρονικούς ορίζοντες ή σε δεδομένα όπου ενδεχομένως υπάρχει αλλαγή στην κλίμακά τους.

3.3 Στατιστικοί έλεγχοι

Για τη διερεύνηση ύπαρξης στατιστικά σημαντικών διαφορών στην προβλεπτική ικανότητα του μοντέλου Reservoir Computing όταν είναι σχεδιασμένο με τιμές υπερπαραμέτρων που έχουν ρυθμιστεί με τις τρεις μεθόδους, χρησιμοποιήθηκαν τρεις στατιστικοί έλεγχοι. Αυτοί εφαρμόστηκαν στις χρονοσειρές σφαλμάτων που προέκυψαν στο σύνολο ελέγχου για κάθε ορίζοντα πρόβλεψης ξεχωριστά και για κάθε πραγματοποίηση των συστημάτων. Πιο επεξηγηματικά, το σφάλμα πρόβλεψης για h βήματα μπροστά στο χρόνο και για δεδομένη χρονική στιγμή t εκλαμβάνεται σύμφωνα με τη σχέση (20):

$$g(e_t(h)) = \frac{1}{K}\sum_{i=1}^{K}\left({e_{i,t}}^2(h)\right), \tag{20}$$

όπου K είναι το πλήθος των μεταβλητών της μελετώμενης χρονοσειράς και $e_{i,t}(h)$ είναι η διαφορά μεταξύ πραγματικής και εκτιμηθείσας τιμής της μεταβλητής i για την χρονική στιγμή $t+h$, δηλαδή $e_{i,t}(h) = x_{i,t+h} - \hat{x}_{i,t+h}$.

Αναλυτικότερα, σε πρώτο στάδιο χρησιμοποιήθηκε ο έλεγχος Friedman ο οποίος είναι μη-παραμετρικός στατιστικός έλεγχος και εξετάζει αν υπάρχουν στατιστικά σημαντικές διαφορές στην απόδοση μεθόδων όταν αυτές συγκρίνονται σε συγκεκριμένο σύνολο δεδομένων (Rizkallah, 2025). Ουσιαστικά, συγκρίνει τις κατατάξεις των μεθόδων (ranks), με τη μηδενική υπόθεση του ελέγχου (H_0) να

αναφέρει ότι δεν υπάρχει στατιστικά σημαντική διαφορά μεταξύ των εφαρμοσμένων προσεγγίσεων.

Μετά από την εκτέλεση του, αξιοποιήθηκε ο έλεγχος Diebold-Mariano (Harvey et al., 1997; Mariano et al., 2012) για ανά δύο συγκρίσεις των μεθόδων και μόνο για τις περιπτώσεις που με τον προηγούμενο έλεγχο είχε απορριφθεί η μηδενική υπόθεση. Με αυτόν τον τρόπο κάθε φορά που απαιτείται πραγματοποιούνται 3 συγκρίσεις. Η μηδενική υπόθεση σε αυτόν τον έλεγχο είναι ότι οι προβλέψεις που έχουν παραχθεί με δύο διαφορετικές μεθόδους είναι ισοδύναμες ως προς την ακρίβεια.

Τελευταίο βήμα για τη σύγκριση των προβλέψεων που λαμβάνονται με το Reservoir Computing έχοντας προηγουμένως βελτιστοποιήσει τις υπερπαραμέτρους του με διαφορετικές μεθόδους, είναι ο έλεγχος multi-horizon model confidence set (multi-horizon MCS) (Quaedvlieg, 2021). Αυτός ελέγχει την συνολική αποδοτικότητα κάθε μοντέλου στην παραγωγή προβλέψεων για πολλά βήματα μπροστά στο χρόνο, προσδιορίζοντας τελικά ένα σύνολο μοντέλων με τη μεγαλύτερη αποτελεσματικότητα για δεδομένο επίπεδο εμπιστοσύνης. Η βασική του διαφορά με τους προηγούμενους ελέγχους και ο λόγος συμπερίληψής του στη μελέτη είναι το γεγονός ότι δεν λαμβάνει υπόψη τις προβλέψεις για έναν μόνο ορίζοντα αλλά για όσους αποφασίσει ο χρήστης. Αν ένα μοντέλο αποκλειστεί από το τελικό σύνολο, σημαίνει ότι η απόδοσή του στην πρόβλεψη πολλών βημάτων μπροστά είναι στατιστικά χειρότερη από αυτή των υπολοίπων (Gong et al., 2025).

Η στάθμη στατιστικής σημαντικότητας σε όλους τους ελέγχους ορίστηκε να είναι ίση με $\alpha = 0.05$, με εξαίρεση τους ελέγχους Diebold-Mariano όπου με τη διόρθωση Bonferroni μετατράπηκε σε $\alpha = 0.05/3$, προκειμένου να αποφευχθεί ο κίνδυνος λανθασμένης απόρριψης της μηδενικής υπόθεσης σε κάποια από τις τρεις συγκρίσεις μεταξύ δύο μεθόδων-μοντέλων (Hooper, 2025).

3.4 Αποτελέσματα

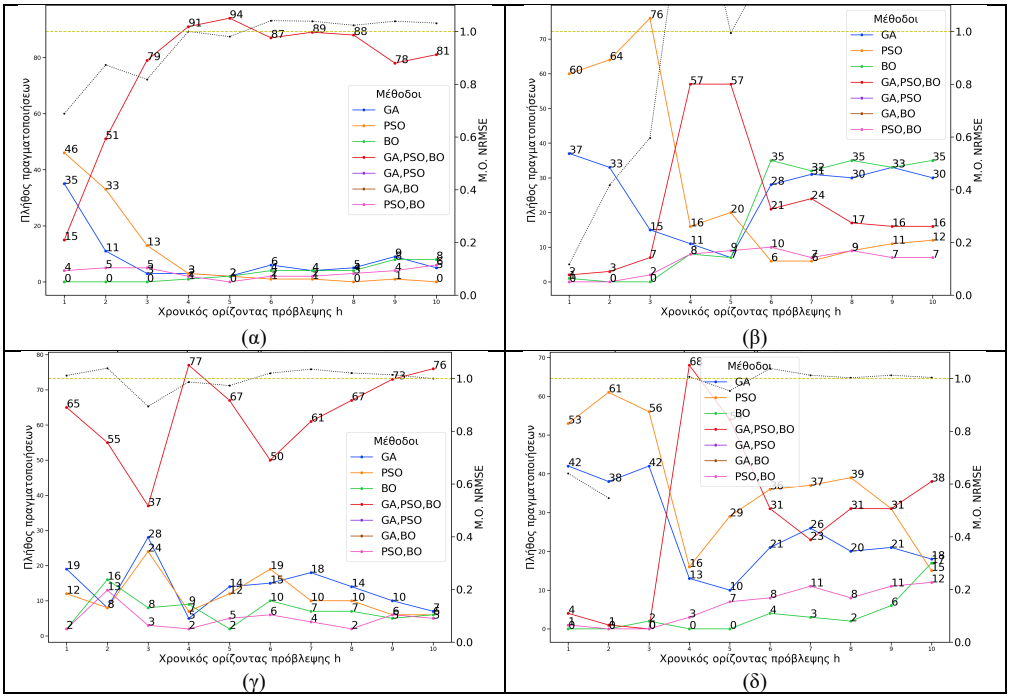
Τα αποτελέσματα των στατιστικών ελέγχων που αναλύθηκαν στην προηγούμενη υποενότητα συνοψίζονται στις ακόλουθες εικόνες. Η μελέτη κάθε συστήματος συνοδεύεται από τέσσερα διαγράμματα γραμμής στα οποία αλλάζει κάθε φορά ο συνδυασμός διάστασης-μήκους (K, N), ένα ραβδόγραμμα που σχετίζεται με τα αποτελέσματα του ελέγχου multi-horizon MCS και έναν πίνακα με τον χρόνο που απαιτήθηκε μέχρι την εύρεση των καλύτερων τιμών υπερπαραμέτρων με τις τρεις συγκρινόμενες μεθόδους.

Κάθε διάγραμμα γραμμής παρουσιάζει για κάθε ορίζοντα πρόβλεψης, το πλήθος των πραγματοποιήσεων από τις συνολικά 100 στις οποίες διαπιστώθηκε στατιστική ισοδυναμία τριών ή δύο μεθόδων, αλλά και ξεκάθαρη υπεροχή μόνο μίας μεθόδου. Σε εκείνες τις πραγματοποιήσεις όπου βρέθηκε ότι οι τρεις μέθοδοι είναι στατιστικά ισοδύναμες, υπολογίστηκε ο μέσος όρος του $NRMSE$ των τριών μεθόδων (δευτερεύων άξονας τιμών διαγραμμάτων). Στη συνέχεια, εντοπίστηκαν εκείνοι οι ορίζοντες όπου αυτός ο μέσος όρος ήταν μικρότερος του 1 και ήταν αυτοί που συμπεριλήφθηκαν για τον έλεγχο multi-horizon MCS. Η λογική πίσω από τη συγκεκριμένη μεθοδολογία είναι

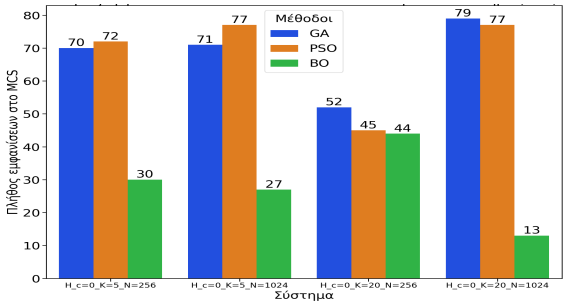
ο τελευταίος έλεγχος να επικεντρωθεί μόνο σε χρονικούς ορίζοντες όπου έχει νόημα η χρήση του μοντέλου για την πρόβλεψη των δεδομένων.

Οι Εικόνες 2-3 αφορούν το αιτιοκρατικό σύστημα με c=0. Μπορούμε να συμπεράνουμε ότι αυξάνοντας το πλήθος των παρατηρήσεων βελτιώνεται η απόδοση των μεθόδων GA και PSO. Στις περιπτώσεις που η διάσταση είναι μικρή (K=5), η PSO έχει σταθερά υψηλή απόδοση σε περισσότερες πραγματοποιήσεις συγκριτικά με τις υπόλοιπες μεθόδους, ενώ όταν η διάσταση αυξάνεται (K=20), τη θέση της παίρνει ο γενετικός αλγόριθμος. Η Μπεϋζιανή βελτιστοποίηση είναι η πιο αδύναμη μέθοδος ειδικά όταν ένα σύστημα έχει υψηλή διάσταση και μεγάλο μήκος.

Εικόνα 2. *Στατιστική σύγκριση μεθόδων για το αιτιοκρατικό σύστημα με c=0 και (α) (K,N)=(5,256), (β) (K,N)=(5,1024), (γ) (K,N)=(20,256) και (δ) (K,N)=(20,1024).*



Εικόνα 3. *Πλήθος εμφανίσεων κάθε μεθόδου στο MCS για το αιτιοκρατικό σύστημα (c=0).*

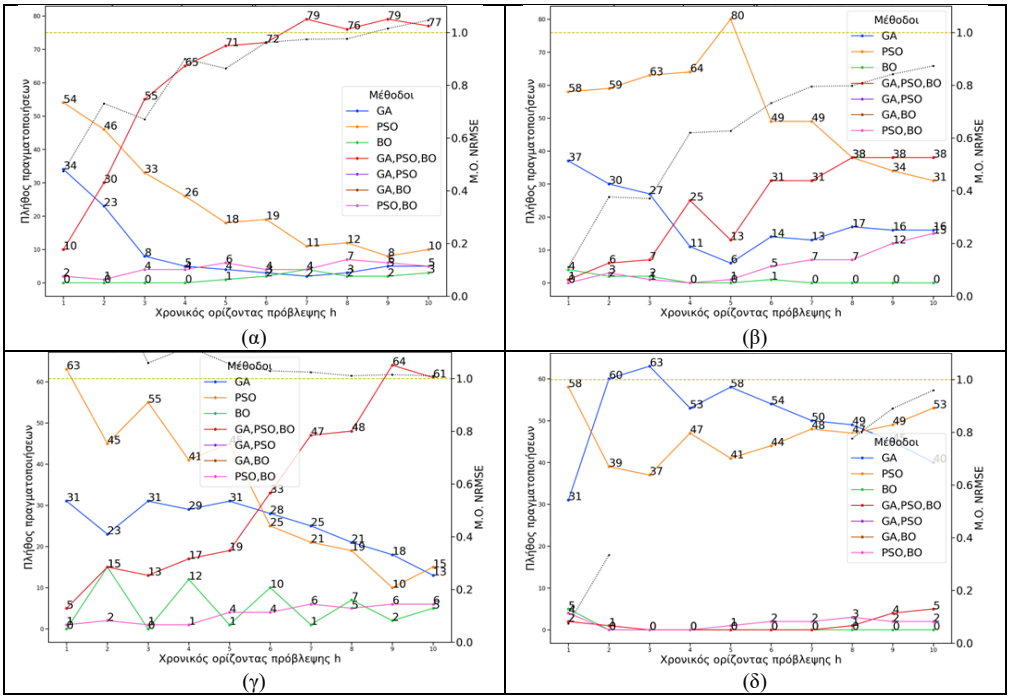


Πίνακας 2. *Μέσος όρος χρόνου βελτιστοποίησης (±1 SD) στο αιτιοκρατικό σύστημα (c=0).*

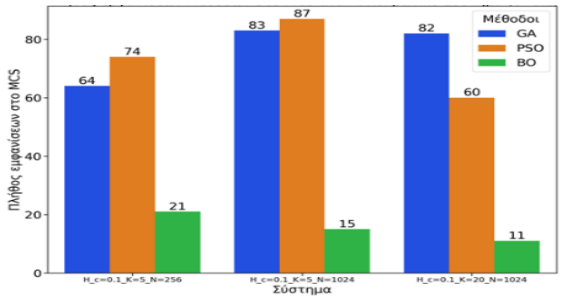
Σύστημα	Μέσος όρος χρόνου ± 1 τυπική απόκλιση		
	GA	PSO	BO
Διάσταση K=5, Μήκος N=256	9.9± 3.7min	3.2± 0.3min	0.8± 0.1min
Διάσταση K=5, Μήκος N=1024	18.1± 1.9min	9.6± 1.5min	1.9± 0.1min
Διάσταση K=20, Μήκος N=256	9.8± 3.8min	3.2± 0.3min	0.8± 0.1min
Διάσταση K=20, Μήκος N=1024	22.7± 4.4min	12.8± 1.2min	2± 0.1min

Ακολουθούν οι Εικόνες 4-5 και ο Πίνακας 3 για το αιτιοκρατικό σύστημα με c=0.1 από τα οποία μπορεί να διαπιστωθεί ότι η Μπεϋζιανή βελτιστοποίηση εντοπίζεται στο Multi Confidence Set τις λιγότερες φορές ανεξαρτήτως διάστασης και μήκους, με την αδυναμία της να αναδεικνύεται εντονότερα όταν το μήκος της χρονοσειράς είναι μεγάλο. Η βελτιστοποίηση με σμήνος σωματιδίων φαίνεται να είναι η πιο αξιόπιστη και σταθερά καλή μέθοδος σε όλα τα συστήματα, εκτός αυτού με υψηλή διάσταση και μεγάλο μήκος, όπου υπερισχύει ο γενετικός αλγόριθμος.

Εικόνα 4. *Στατιστική σύγκριση μεθόδων για το αιτιοκρατικό σύστημα με c=0.1 και (α) (K,N)=(5,256), (β) (K,N)=(5,1024), (γ) (K,N)=(20,256) και (δ) (K,N)=(20,1024).*



Εικόνα 5. *Πλήθος εμφανίσεων κάθε μεθόδου στο MCS για το αιτιοκρατικό σύστημα (c=0.1).*

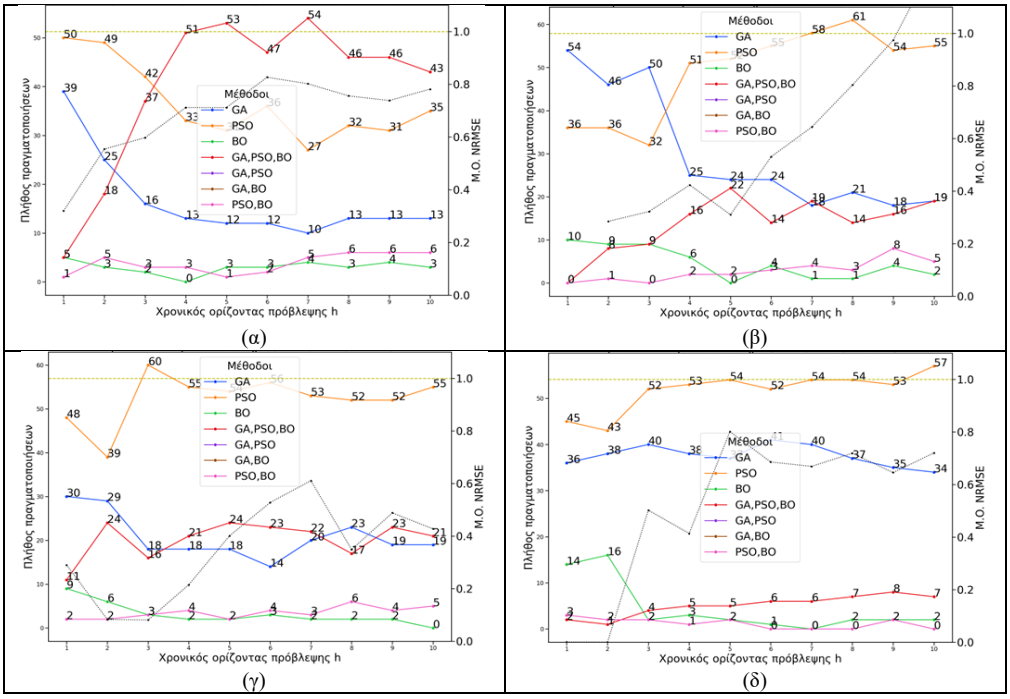


Πίνακας 3. *Μέσος όρος χρόνου βελτιστοποίησης (±1 SD) στο αιτιοκρατικό σύστημα (c=0.1).*

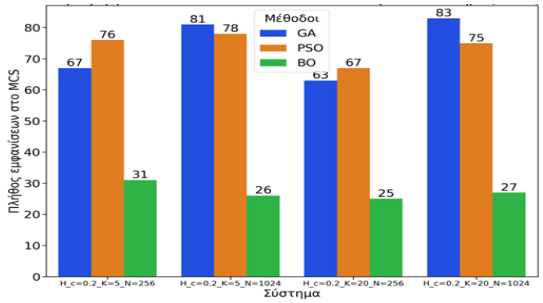
Σύστημα	Μέσος όρος χρόνου ± 1 τυπική απόκλιση		
	GA	PSO	BO
Διάσταση K=5, Μήκος N=256	10.1± 3.8min	3.1± 0.3min	0.8± 0.1min
Διάσταση K=5, Μήκος N=1024	18.3± 1.6min	9.7± 1.5min	1.9± 0.1min
Διάσταση K=20, Μήκος N=256	9.7± 3.5min	3.2± 0.3min	0.8± 0.1min
Διάσταση K=20, Μήκος N=1024	23.4± 3.8min	12.9± 1min	2± 0.1min

Στις Εικόνες 6-7 συγκρίνονται οι τρεις μέθοδοι για το αιτιοκρατικό σύστημα με c=0.2. Γίνεται αντιληπτό ότι αυξάνοντας το μήκος της χρονοσειράς, αυξάνεται το ποσοστό υπεροχής του GA σε όλους τους ορίζοντες, αυξάνεται ελάχιστα το ποσοστό υπεροχής της BO για χαμηλούς ορίζοντες πρόβλεψης (h=1,2), και μειώνεται το πλήθος των πραγματοποιήσεων όπου παρατηρείται ισοδυναμία των 3 μεθόδων. Το τελευταίο συμπέρασμα προκύπτει και με την αύξηση της διάστασης. Γενικά, οι GA και PSO εμφανίζονται ως ισοδύναμα οι πιο σταθερά καλές μέθοδοι στα συστήματα χαμηλής διάστασης-μεγάλου μήκους και υψηλής διάστασης-μικρού μήκους ενώ η BO ως η πιο αδύναμη μέθοδος σε όλα τα συστήματα.

Εικόνα 6. *Στατιστική σύγκριση μεθόδων για το αιτιοκρατικό σύστημα με c=0.2 και (α) (K,N)=(5,256), (β) (K,N)=(5,1024), (γ) (K,N)=(20,256) και (δ) (K,N)=(20,1024).*



Εικόνα 7. *Πλήθος εμφανίσεων κάθε μεθόδου στο MCS για το αιτιοκρατικό σύστημα (c=0.2).*

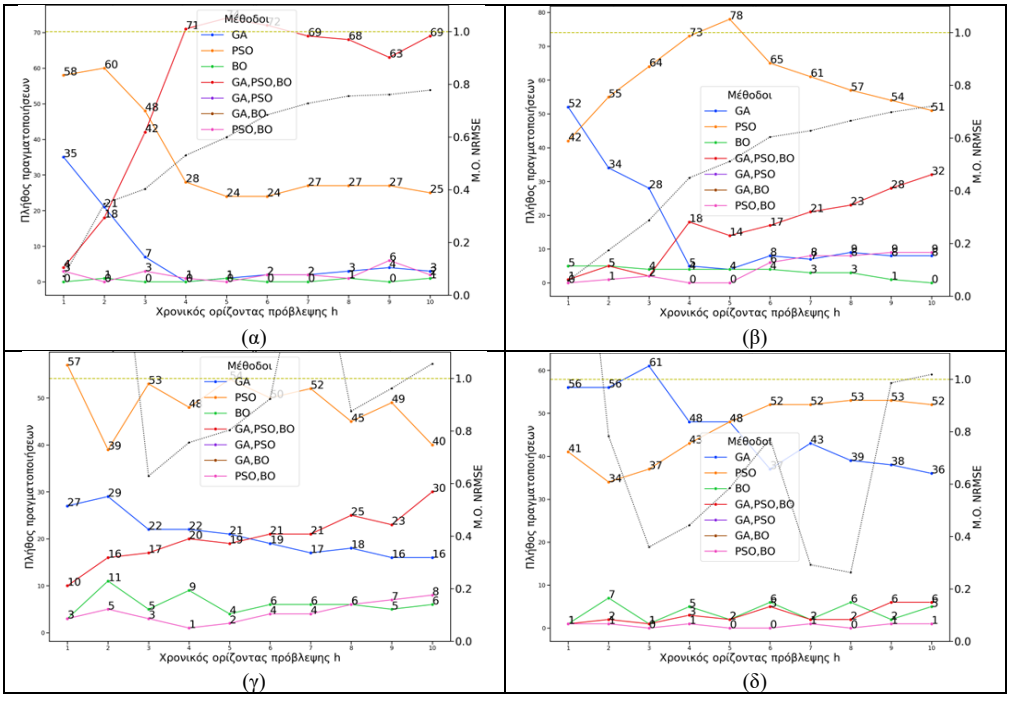


Πίνακας 4. *Μέσος όρος χρόνου βελτιστοποίησης (±1 SD) στο αιτιοκρατικό σύστημα (c=0.2).*

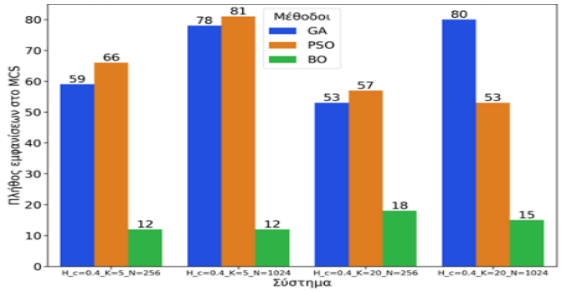
Σύστημα	Μέσος όρος χρόνου ± 1 τυπική απόκλιση		
	GA	PSO	BO
Διάσταση K=5, Μήκος N=256	9.7± 3.5min	3.2± 0.3min	0.8± 0.1min
Διάσταση K=5, Μήκος N=1024	17.7± 2.1min	9.1± 1.6min	1.9± 0.1min
Διάσταση K=20, Μήκος N=256	9.7± 3.9min	3.2± 0.3min	0.8± 0.1min
Διάσταση K=20, Μήκος N=1024	22.4± 3.6min	12.9± 1.3min	2± 0.1min

Οι Εικόνες 8-9 που ακολουθούν αναπαριστούν τα αποτελέσματα των ελέγχων για το αιτιοκρατικό σύστημα με c=0.4. Θέτοντας τις σε αντιπαράθεση με αυτές των προηγούμενων συστημάτων, παρατηρείται ότι με την αύξηση της έντασης σύζευξης c σε σύστημα μικρού μήκους και χαμηλής διάστασης, αυξάνεται και ο ορίζοντας πρόβλεψης μετά τον οποίο οι τρεις μέθοδοι ισοδυναμούν. Συνολικά, όταν c=0.4 η PSO έχει συχνότερη παρουσία στο σύνολο των καλύτερων μεθόδων σε όλα τα συστήματα εκτός αυτού που είναι το πιο πολύπλοκο (με υψηλή διάσταση και μεγάλο μήκος). Εκεί η GA είναι πιο αξιόπιστη. Η BO έχει την ίδια αδύναμη συμπεριφορά όπως και πριν.

Εικόνα 8. *Στατιστική σύγκριση μεθόδων για το αιτιοκρατικό σύστημα με c=0.4 και (α) (K,N)=(5,256), (β) (K,N)=(5,1024), (γ) (K,N)=(20,256) και (δ) (K,N)=(20,1024).*



Εικόνα 9. *Πλήθος εμφανίσεων κάθε μεθόδου στο MCS για το αιτιοκρατικό σύστημα (c=0.4).*

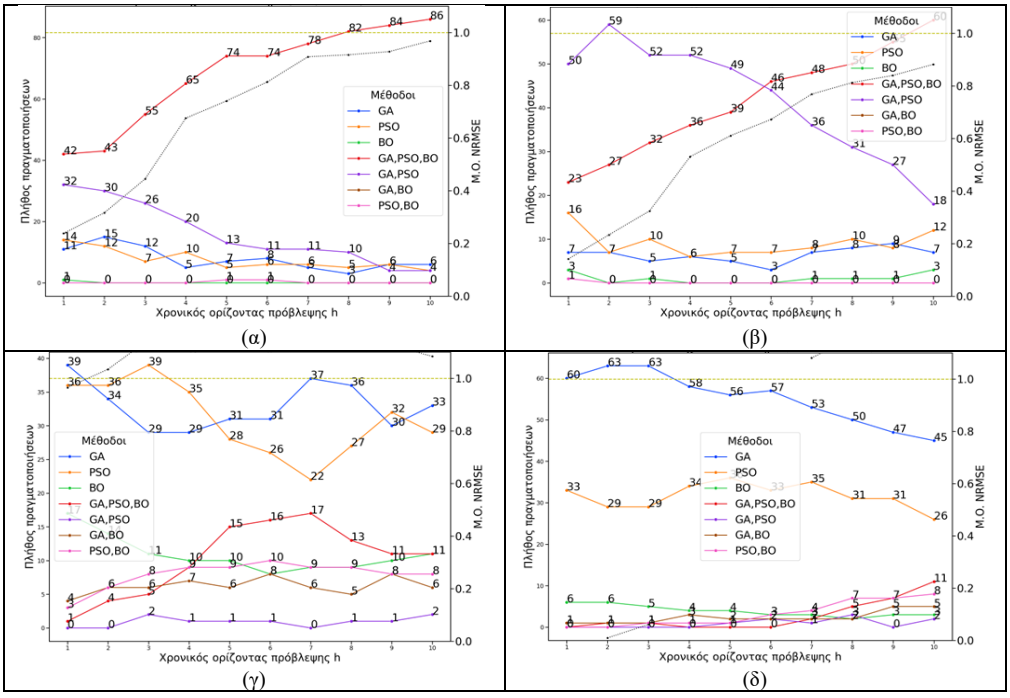


Πίνακας 5. *Μέσος όρος χρόνου βελτιστοποίησης (±1 SD) στο αιτιοκρατικό σύστημα (c=0.4).*

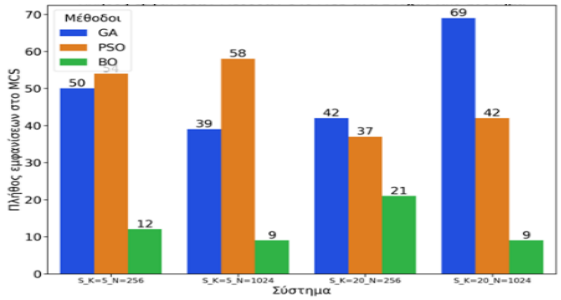
Σύστημα	Μέσος όρος χρόνου ± 1 τυπική απόκλιση		
	GA	PSO	BO
Διάσταση K=5, Μήκος N=256	9.5± 3.5min	3.2± 0.3min	0.8± 0.1min
Διάσταση K=5, Μήκος N=1024	18.3± 1.7min	9± 1.7min	1.9± 0.1min
Διάσταση K=20, Μήκος N=256	9.7± 3.6min	3.1± 0.3min	0.8± 0.1min
Διάσταση K=20, Μήκος N=1024	23.2± 3.5min	13± 1.2min	2± 0.1min

Τέλος, οι Εικόνες 10-11 και ο Πίνακας 6 σχετίζονται με το στοχαστικό σύστημα. Σε στοχαστικό σύστημα μικρού μήκους (N=256) και χαμηλής διάστασης (K=5), ανιχνεύεται να κυριαρχεί η ισοδυναμία των 3 μεθόδων σε καθέναν από τους 10 ορίζοντες πρόβλεψης. Η PSO έχει συχνότερη παρουσία στο σύνολο των καλύτερων μεθόδων στα συστήματα χαμηλής διάστασης, ενώ η GA είναι πιο αξιόπιστη για στοχαστικά συστήματα υψηλής διάστασης, ειδικά μεγάλου μήκους. Η BO εξακολουθεί και είναι η λιγότερο αποτελεσματική μέθοδος.

Εικόνα 10. *Στατιστική σύγκριση μεθόδων για το στοχαστικό σύστημα με (α) (K,N)=(5,256), (β) (K,N)=(5,1024), (γ) (K,N)=(20,256) και (δ) (K,N)=(20,1024).*



Εικόνα 11. *Πλήθος εμφανίσεων κάθε μεθόδου στο MCS για το στοχαστικό σύστημα.*



Πίνακας 6. *Μέσος όρος χρόνου βελτιστοποίησης (±1 SD) στο στοχαστικό σύστημα.*

Σύστημα	Μέσος όρος χρόνου ± 1 τυπική απόκλιση		
	GA	PSO	BO
Διάσταση K=5, Μήκος N=256	7.4± 0.4min	3.1± 0.3min	0.9± 0.1min
Διάσταση K=5, Μήκος N=1024	18± 1.8min	8.8± 1.6min	2± 0.1min
Διάσταση K=20, Μήκος N=256	7.5± 0.4min	3.2± 0.3min	0.9± 0.1min
Διάσταση K=20, Μήκος N=1024	25.4± 1.3min	12.8± 1.2min	2± 0.2min

Εστιάζοντας στους πίνακες 2-6 είναι σαφές ότι ανεξάρτητα από το σύστημα, ο ελάχιστος χρόνος καταγράφεται από την Μπεϋζιανή βελτιστοποίηση, ακολουθούμενη από τη μέθοδο PSO και τελευταία από τη GA.

4. ΣΥΜΠΕΡΑΣΜΑΤΑ

Στην εργασία αυτή έγινε μια συστηματική προσομοιωτική μελέτη για να αξιολογηθούν τρεις μέθοδοι βελτιστοποίησης των οκτώ υπερπαραμέτρων του μοντέλου Reservoir Computing: ο γενετικός αλγόριθμος, το σμήνος σωματιδίων και η μπεϋζιανή βελτιστοποίηση. Συνδυάζοντας όλα τα αποτελέσματα σε στοχαστικά και αιτιοκρακτικά συστήματα διαφορετικού μήκους και διάστασης συνάγεται το συμπέρασμα ότι μεταξύ των τριών μεθόδων, προτιμότερη είναι η βελτιστοποίηση με σμήνος σωματιδίων καθώς έχει τα πλεονεκτήματα του μικρού απαιτούμενου χρόνου υλοποίησης και της απλότητας στη λειτουργία και την παραμετροποίηση. Βέβαια, παίζει σημαντικό ρόλο και το ίδιο το σύστημα δεδομένου ότι ο γενετικός αλγόριθμος φαίνεται να υπερτερεί από άποψη αποτελεσματικότητας όταν το σύστημα, αιτιοκρατικό ή στοχαστικό, έχει μεγάλο μήκος και υψηλή διάσταση.

ABSTRACT

The recently developed Reservoir Computing algorithm has been proven to be a promising and effective approach in the field of time series forecasting. In this context, the tuning of its hyperparameters becomes a particularly important task, as it can significantly impact its performance. The present study evaluates and compares the traditional method of Bayesian optimization for tuning the hyperparameters of Reservoir Computing with bio-inspired metaheuristic algorithms. The main objective is to enhance the model's predictive capability across both deterministic and stochastic multivariate systems.

ΑΝΑΦΟΡΕΣ

Alibrahim, H. and Ludwig, S. A. (2021). Hyperparameter Optimization: Comparing Genetic Algorithm against Grid Search and Bayesian Optimization. In 2021 IEEE Congress on Evolutionary Computation (CEC), Kraków, Poland, 28 June–1 July 2021; pp. 1551–1559.

Bianchi, F. M., Scardapane, S., Løkse, S. and Jenssen, R. (2021). Reservoir Computing Approaches for Representation and Classification of Multivariate Time Series. *IEEE Transactions on Neural Networks and Learning Systems,* **32**.

Brochu, E., Cora, V. M. and de Freitas, N. (2024). A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning. *arXiv,* arXiv:1012.2599.

Chouikhi, N., Fdhila, R., Ammar, B., Rokbani, N. and Alimi, A. M. (2016). Single- and multi-objective particle swarm optimization of reservoir structure in Echo State Network. In Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, Canada, 24-29 July 2016; pp. 440-447.

Dale, M., O'Keefe, S., Sebald, A., Stepney, S. and Trefzer, M.A. (2021). Reservoir computing quality: connectivity and topology. *Natural Computing,* **20**.

Erden, C. (2023). Genetic algorithm-based hyperparameter optimization of deep learning models for $PM_{2.5}$ time-series prediction. *International Journal of Environmental Science and Technology*, **20**.

Fan, J. and Han, M. (2011). Online designed of Echo State Network based on Particle Swarm Optimization for system identification. In Proceedings of the 2011 2nd International Conference on Intelligent Control and Information Processing, Harbin, China, 25-28 July 2011; pp. 559-563.

Fotiadis, A., Vlachos, I. and Kugiumtzis, D. (2024). The causality measure of partial mutual information from mixed embedding (PMIME) revisited. *Chaos*, **34**.

Francescomarino, C. D., Dumas, M., Federici, M., Ghidini, C., Maggi, F. M., Rizzi, W., Simonetto, L. (2018). Genetic Algorithms for Hyperparameter Optimization in Predictive Business Process. *Information Systems,* **74**, 67-83.

Frazier, P. I. (2018). A Tutorial on Bayesian Optimization. *arXiv*, arXiv:1807.02811v1.

Gad, A. F. (2024). PyGAD: an intuitive genetic algorithm Python library. *Multimedia Tools and Applications*, **83**, 58029-58042.

Garnett, R. (2015, March 16). *Lecture 12: Bayesian Optimization* [Lecture notes]. Computer Science & Engineering, Washington University. https://www.cse.wustl.edu/~garnett/cse515t/spring_2015/files/lecture_notes/12.pdf .

Gong, H., Xing, H. and Wang, Q. (2025). Enhanced forecasting method for realized volatility of energy futures prices: A secondary decomposition-based deep learning model. *Engineering Applications of Artificial Intelligence,* **146**.

Guo, X., Liang, R., Xu, S., Dong, L. and Liu, Y. (2026). An investigation of echo state network for EEG-based emotion recognition with deep neural networks. *Biomedical Signal Processing and Control*, **111**.

Han, L., Ye, H. J. and Zhan, D. C. (2023). The Capacity and Robustness Trade-off: Revisiting the Channel Independent Strategy for Multivariate Time Series Forecasting. *arXiv,* arXiv:2304.05206.

Harvey, D., Leybourne, S. and Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, **13**.

Hinaut, X. and Trouvain, N. (2021). Which Hype for My New Task? Hints and Random Search for Echo State Networks Hyperparameters. In: Farkaš, I., Masulli, P., Otte, S., Wermter, S.

(eds) Artificial Neural Networks and Machine Learning – ICANN 2021. ICANN 2021. (pp. 83-97). Cham: Springer.
Hooper, R. (2025). To adjust, or not to adjust, for multiple comparisons. *Journal of Clinical Epidemiology,***180**.
Kachitvichyanukul, V. (2012). Comparison of Three Evolutionary Algorithms: GA, PSO, and DE. *Industrial Engineering & Management Systems,* **11**, 215-223.
Loussaief, S. and Abdelkrim, A. (2018). Convolutional Neural Network Hyper-Parameters Optimization based on Genetic Algorithms. *(IJACSA) International Journal of Advanced Computer Science and Applications,* **9**.
Lukoševičius, M. and Jaeger, H. (2009). Reservoir computing approaches to recurrent neural network training. *Computer Science Review*, **3**.
Mariano, R.S. and Preve, D. (2012). Statistical tests for multiple forecast comparison. *Journal of Econometrics*, **169**.
Mitchell, M. (1995), Genetic algorithms: An overview. *Complexity*, **1**, 31-39.
Nguyen-Thai, B., Le, V., Tieu, ND.T., Tran, T., Venkatesh, S. and Ramzan, N. (2024). Learning evolving relations for multivariate time series forecasting. *Applied Intelligence,* **54**, 3918-3932.
Pathak, J., Hunt, B., Girvan, M., Lu, Z. and Ott, E. (2018). Model-Free Prediction of Large Spatiotemporally Chaotic Systems from Data: A Reservoir Computing Approach. *Physical Review Letters*, **120**.
Picco, E., Jaurigue, L., Lüdge, K. and Massar, S. (2025). Efficient optimisation of physical reservoir computers using only a delayed input. *Communications Engineering*, **4**.
Pranolo, A., Setyaputri, F., Paramarta, A., Triono, A., Fadhilla, A., Akbari, A., Utama, A., Wibawa, A. and Uriu, W. (2024). Enhanced Multivariate Time Series Analysis Using LSTM: A Comparative Study of Min-Max and Z-Score Normalization Techniques. *ILKOM Jurnal Ilmiah,* **16**, 210-220.
Quaedvlieg, R. (2021). Multi-Horizon Forecast Comparison. *Journal of Business & Economic Statistics*, **39**, 40-53.
Rizkallah, L.W. (2025). Enhancing the performance of gradient boosting trees on regression problems. *Journal of Big Data,* **12**.
Schelter, B., Winterhalder, M., Hellwig, B., Guschlbauer, B., Lücking, C.H. and Timmer, J. (2006). Direct or indirect? Graphical models for neural oscillators. *Journal of Physiology-Paris*, **99**.
Sharma, D.K., Varshney, R.P., Agarwal, S., Alhussan, A.A. and Abdallah, H.A. (2024). Developing a multivariate time series forecasting framework based on stacked autoencoders and multi-phase feature. *Heliyon,* **10**.
Simon, D. (2013). *Evolutionary optimization algorithms*. New Jersey: John Wiley & Sons.
Skibinsky-Gitlin, E.S., Alomar, M.L., Frasser, C.F., Canals, V., Isern, E., Roca, M. and Rosselló, J.L. (2018). Cyclic Reservoir Computing with FPGA Devices for Efficient Channel Equalization. In Proceedings of the Artificial Intelligence and Soft Computing: 17th International Conference, Zakopane, Poland, 3–7 June 2018; pp. 226–234.
Sorjamaa, A., Hao, J., Reyhani, N., Ji, Y. and Lendasse, A. (2007). Methodology for long-term prediction of time series. *Neurocomputing,* **70**, 2861-2869.
Stepney, S. (2024). Physical reservoir computing: a tutorial. *Natural Computing,* **23**, 665-685.
Tealab, A. (2018). Time series forecasting using artificial neural networks methodologies: A systematic review. *Future Computing and Informatics Journal*, **3**, 334-340.
Viehweg, J., Worthmann, K. and Mäder, P. (2023). Parameterizing echo state networks for multi-step time series prediction. *Neurocomputing,* **522**, 214-228.

Wu, G., Zhang, J. and Xue, H. (2023). Long-Term Prediction of Hydrometeorological Time Series Using a PSO-Based Combined Model Composed of EEMD and LSTM. *Sustainability*, **15**.

Wyffels, F., Schrauwen, B. and Stroobandt, D. (2008). Stable Output Feedback in Reservoir Computing Using Ridge Regression. In Proceedings of the 18th International Conference on Artificial Neural Networks, Part I, Prague Czech Republic, 3–6 September 2008; pp. 808–817.

Yperman, J. and Becker, T. (2017). Bayesian optimization of hyper-parameters in reservoir computing. *arXiv,* arXiv:1611.05193.

Zhong, S., Xie, X., Lin, L. and Wang, F. (2017). Genetic algorithm optimized double-reservoir echo state network for multi-regime time series prediction. *Neurocomputing*, **238**.

Zhu, Y., Liu, Y., Wang, N., Zhang, Z. and Li, Y. (2025). Real-time Error Compensation Transfer Learning with Echo State Networks for Enhanced Wind Power Prediction. *Applied Energy*, **379**.

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 37ου Πανελλήνιου& 3ου Διεθνούς Συνεδρίου Στατιστικής (2025), σελ.199-213

ΚΑΤΑΝΟΜΗ POISSON: ΕΝΑ ΕΝΑΛΛΑΚΤΙΚΟ ΣΤΑΤΙΣΤΙΚΟ ΜΟΝΤΕΛΟ ΓΙΑ ΤΗΝ ΠΡΟΒΛΕΨΗ ΤΩΝ ΠΟΔΟΣΦΑΙΡΙΚΩΝ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

Σιόλου Ευδοξία, Συκιανάκης Νίκος, Χαλικιάς Μιλτιάδης
Τμήμα Λογιστικής και Χρηματοοικονομικής,
Πανεπιστήμιο Δυτικής Αττικής
esiolou@yahoo.gr, niksyk@uniwa.gr,mchalik@uniwa.gr

ΠΕΡΙΛΗΨΗ

Η κατανομή Poisson μπορεί να χρησιμοποιηθεί για την εκτίμηση των αποτελεσμάτων των ποδοσφαιρικών αγώνων χρησιμοποιώντας για παράμετρο λ το μέσο αριθμό των τερμάτων από προηγούμενους αγώνες. Στην παρούσα εργασία παρουσιάζεται ένα νέο μοντέλο πρόβλεψης για την πρόβλεψη των αποτελεσμάτων σε ποδοσφαιρικούς αγώνες, όπου για την εκτίμηση της παραμέτρου λ χρησιμοποιούνται όχι μόνο τα τέρματα, αλλά και το πλήθος των τελικών προσπαθειών της κάθε ομάδας για να πετύχει τέρμα. Σύμφωνα με τα ευρήματα της εργασίας, το νέο μοντέλο προβλέπει καλύτερα τα ακριβή σκορ σε διάφορες περιπτώσεις.

Λέξεις Κλειδιά: Μοντέλα Πρόβλεψης, Κατανομή Poisson

1. ΕΙΣΑΓΩΓΗ

Ο αθλητισμός και ιδιαίτερα το ποδόσφαιρο προσελκύει τεράστιο ενδιαφέρον παγκοσμίως. Ειδικότερα η πρόβλεψη αποτελεσμάτων έχει ενδιαφέρον και λόγω της ύπαρξης στοιχημάτων. Ακόμη, η πρόβλεψη αποτελεσμάτων αθλητικών αγώνων χρησιμοποιείται πολλές φορές και στη διδασκαλία της στατιστικής λόγω του ενδιαφέροντος που συγκεντρώνει από τους φοιτητές.

Ειδικότερα, ο Chalikias (2009) προκειμένου να βελτιώσει το επίπεδο κατανόησης της διωνυμικής κατανομής των φοιτητών, χρησιμοποίησε ένα παράδειγμα πρόβλεψης του νικητή του 49ου Παγκόσμιου Πρωταθλήματος της Διεθνούς Ομοσπονδίας Σκοποβολής στο διπλό τραπ που διεξήχθη το 2006 στο Ζάγκρεμπ της Κροατίας. Επίσης, ο Tijms (2015) τόνισε την αξία χρήσης πραγματικών δεδομένων για τη διδασκαλία της Μπεϋζιανής στατιστικής συμπερασματολογίας. Ειδικότερα, μελέτησε εάν είναι αληθείς οι ισχυρισμοί ότι τα αποτελέσματα της κλήρωσης για τα προημιτελικά στο Ευρωπαϊκό Πρωτάθλημα Ποδοσφαίρου 2013 ήταν προσυνεννοημένα. Επιπλέον, οι Williams και Williams (2016), οι οποίοι κατά τη διδασκαλία της Στατιστικής χρησιμοποίησαν πραγματικά δεδομένα για να συγκρίνουν

την επίδοση του Michael Jordan με την επίδοση του LeBron James, επισήμαναν ότι η χρήση ελκυστικών παραδειγμάτων που συνδέονται άμεσα με τα ενδιαφέροντα των φοιτητών οδηγούν σε βαθύτερη κατανόηση στατιστικών εννοιών.

Η κατανομή Poisson και οι έλεγχοι καλής προσαρμογής χρησιμοποιούνται ευρέως στον τομέα των αθλητικών δεδομένων. Ειδικότερα, ο Keller (1994) έδειξε ότι η κατανομή Poisson μπορεί να χρησιμοποιηθεί για την πρόβλεψη των αποτελεσμάτων σε αγώνες μπέιζμπολ και ποδοσφαίρου. Επιπλέον, οι Karlis και Ntzoufras (2003), οι οποίοι αντικατέστησαν δύο ανεξάρτητες κατανομές Poisson με ένα διμεταβλητό μοντέλο Poisson κατάφεραν να βελτιώσουν την προσαρμογή του μοντέλου και την πρόβλεψη του αριθμού των ισοπαλιών σε ποδοσφαιρικούς αγώνες, χρησιμοποιώντας σύνολα δεδομένων από το ποδόσφαιρο και την υδατοσφαίριση. Επιπρόσθετα, ο Chu (2003) χρησιμοποίησε δεδομένα από τα αποτελέσματα των αγώνων του Παγκόσμιου Κυπέλλου Ποδοσφαίρου για τη διδασκαλία της κατανομής Poisson σε φοιτητές. Ειδικότερα, χρησιμοποίησε την κατανομή Poisson με παράμετρο λ το μέσο αριθμό των τερμάτων που σημειώθηκαν στα 90 λεπτά της τυπικής διάρκειας των 232 ποδοσφαιρικών αγώνων του Παγκόσμιου Κυπέλλου Ποδοσφαίρου, ενώ για να εξετάσει την καλή προσαρμογή του μοντέλου χρησιμοποίησε τον έλεγχο X^2.

Σκοπός της παρούσας εργασίας είναι η βελτίωση του μοντέλου πρόβλεψης του Chu (2003) για την εκτίμηση πιθανοτήτων αποτελεσμάτων (σκορ) ποδοσφαιρικών αγώνων.

2. ΕΚΤΙΜΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ ΠΟΔΟΣΦΑΙΡΙΚΩΝ ΑΓΩΝΩΝ ΜΕΣΩ ΤΗΣ ΚΑΤΑΝΟΜΗΣ POISSON_ΠΑΡΟΥΣΙΑΣΗ ΕΝΟΣ ΝΕΟΥ ΜΟΝΤΕΛΟΥ ΠΡΟΒΛΕΨΗΣ

Στους ποδοσφαιρικούς αγώνες το τελικό σκορ είναι συνήθως συνάρτηση των τελικών προσπαθειών των ομάδων για να πετύχουν γκολ, καθώς οι ομάδες με τις περισσότερες τελικές προσπάθειες τείνουν να πετυχαίνουν περισσότερα γκολ (Lago-Ballesteros & Lago-Peñas, 2010). Ως εκ τούτου, για την εκτίμηση του σκορ των ποδοσφαιρικών αγώνων προτείνεται να χρησιμοποιηθεί η κατανομή Poisson με την εκτίμηση της παραμέτρου λ να βασίζεται όχι μόνο στα σκορ των προηγούμενων αγώνων, αλλά και στις τελικές προσπάθειες κάθε ομάδας να πετύχει γκολ.

Ειδικότερα, για την εκτίμηση της παραμέτρου λ, λαμβάνοντας υπόψη τόσο τα γκολ όσο και τις τελικές προσπάθειες, προτείνεται αρχικά να εκτιμώνται οι παράμετροι αποτελεσματικότητας c_1 και c_2, της γηπεδούχου και της φιλοξενούμενης ομάδας αντίστοιχα. Όπου το c_1, ορίζεται ως ο μέσος όρος των πηλίκων «γκολ/τελικές προσπάθειες» για τις ομάδες που παίζουν στο γήπεδό τους και αντίστοιχα το c_2 ορίζεται ως ο μέσος όρος των πηλίκων «γκολ/τελικές προσπάθειες» για τις φιλοξενούμενες ομάδες. Αξίζει να σημειωθεί ότι προϋπόθεση καλής προσαρμογής εφαρμογής της μεθοδολογίας είναι και οι δύο αντίπαλοι να πραγματοποιήσουν τουλάχιστον μία τελική προσπάθεια (έστω και ανεπιτυχή). Η προϋπόθεση αυτή μπορούμε να πούμε με σχετική ασφάλεια ότι τηρείται όταν οι ομάδες δεν έχουν πολύ μεγάλη διαφορά δυναμικότητας π.χ. ανήκουν στην ίδια κατηγορία όπως στην εφαρμογή που παρουσιάζουμε. Στη

συνέχεια, οι τιμές αυτές θα χρησιμοποιούνται ως παράμετροι που θα πολλαπλασιάζονται με τις τελικές προσπάθειες κάθε ομάδας. Στο τελικό βήμα του προτεινόμενου αλγορίθμου γίνεται ο προσδιορισμός των νέων παραμέτρων λ_1' και λ_2', οι οποίες ισούνται με το μέσο όρο των προηγούμενων τιμών που αφορούν τη γηπεδούχο και τη φιλοξενούμενη ομάδα, αντίστοιχα. Συγκεκριμένα, το λ_1' είναι ο μέσος όρος των εκτιμώμενων γκολ της γηπεδούχου ομάδας και το λ_2' είναι ο μέσος όρος των εκτιμώμενων γκολ της φιλοξενούμενης ομάδας. Επομένως, η εκτίμηση των πιθανοτήτων για τη γηπεδούχο αλλά και τη φιλοξενούμενη ομάδα να πετύχουν συγκεκριμένο αριθμό τερμάτων γίνεται χρησιμοποιώντας την κατανομή Poisson με παραμέτρους λ_1'και λ_2'αντίστοιχα. Για τον έλεγχο καλής προσαρμογής του μοντέλου χρησιμοποιείται ο έλεγχος X^2.

3. ΕΦΑΡΜΟΓΗ ΚΑΙ ΣΥΓΚΡΙΣΗ ΤΩΝ ΜΟΝΤΕΛΩΝ ΠΡΟΒΛΕΨΗΣ ΣΕ ΔΕΔΟΜΕΝΑ ΤΟΥ ΕΛΛΗΝΙΚΟΥ ΠΡΩΤΑΘΛΗΜΑΤΟΣ ΠΟΔΟΣΦΑΙΡΟΥ.

Στην παρούσα εργασία χρησιμοποιήθηκαν τα αποτελέσματα όλων των αγωνιστικών του ελληνικού πρωταθλήματος ποδοσφαίρου για την περίοδο 2014 - 2015. Στο ελληνικό επαγγελματικό πρωτάθλημα ποδοσφαίρου τη συγκεκριμένη χρονική περίοδο αγωνίστηκαν 17 σύλλογοι (Πίνακας 1), οι οποίοι συμμετείχαν σε μια σειρά 32 αγώνων εντός και εκτός έδρας σε δύο γύρους. Η σεζόν διήρκησε από τον Αύγουστο έως τον Μάιο.

Πίνακας 1. *Ομάδες ελληνικού πρωταθλήματος ποδοσφαίρου*

Ομάδα	Συμβολισμός
Αστέρας Τρίπολης	O1
Ατρόμητος	O2
Εργοτέλης	O3
Καλλονή	O4
Κέρκυρα	O5
Λεβαδιακός	O6
Ο.Φ.Η.	O7
Ολυμπιακός	O8
Π.Α.Ο.Κ.	O9
Παναθηναϊκός	O10
Παναιτολικός	O11
Πανιώνιος	O12
Πανθρακικός	O13
Πας Γιάννινα	O14
Πλατανιάς	O15
Σκόντα Ξάνθη	O16
Βέροια	O17

Αρχικά, χρησιμοποιήθηκαν τα αποτελέσματα μόνο του πρώτου γύρου, για τη γηπεδούχο και τη φιλοξενούμενη ομάδα, τα οποία παρουσιάζονται στον πίνακα 2:

Πίνακας2.*Αποτελέσματα ελληνικού πρωταθλήματος ποδοσφαίρου*

Αγωνιστική	Γηπεδούχος Ομάδα	Φιλοξενούμενη Ομάδα	Αποτέλεσμα		Αγωνιστική	ΓηπεδούχοςΟμάδα	Φιλοξενούμενη Ομάδα	Αποτέλεσμα		Αγωνιστική	ΓηπεδούχοςΟμάδα	Φιλοξενούμενη Ομάδα	Αποτέλεσμα	
1η	Ο7	Ο11	1	0	7η	Ο3	Ο8	2	3	13η	Ο17	Ο15	2	0
	Ο5	Ο13	2	1		Ο10	Ο7	1	2		Ο5	Ο2	1	1
	Ο9	Ο4	1	1		Ο9	Ο2	2	1		Ο8	Ο14	2	2
	Ο14	Ο1	3	1		Ο1	Ο11	1	1		Ο16	Ο9	4	2
	Ο2	Ο15	1	0		Ο12	Ο14	0	1		Ο11	Ο12	2	1
	Ο17	Ο16	3	2		Ο17	Ο6	2	1		Ο7	Ο6	0	2
	Ο6	Ο10	1	1		Ο15	Ο16	1	0		Ο1	Ο10	1	1
	Ο12	Ο3	2	1		Ο4	Ο5	1	0		Ο13	Ο3	0	0
2η	Ο13	Ο14	1	1	8η	Ο8	Ο10	1	0	14η	Ο12	Ο7	2	0
	Ο4	Ο3	2	0		Ο9	Ο17	4	1		Ο1	Ο17	2	0
	Ο11	Ο8	1	1		Ο7	Ο1	2	3		Ο3	Ο16	0	2
	Ο15	Ο9	0	4		Ο14	Ο4	0	0		Ο14	Ο2	1	0
	Ο16	Ο5	0	0		Ο2	Ο3	1	1		Ο9	Ο5	2	1
	Ο1	Ο6	2	1		Ο11	Ο13	3	1		Ο10	Ο13	4	0
	Ο10	Ο12	2	1		Ο6	Ο12	1	0		Ο4	Ο11	0	0
	Ο7	Ο17	0	1		Ο5	Ο15	1	2		Ο6	Ο8	1	2
3η	Ο17	Ο5	2	1	9η	Ο1	Ο8	0	0	15η	Ο5	Ο3	3	1
	Ο4	Ο10	1	0		Ο13	Ο7	2	0		Ο13	Ο1	1	1
	Ο12	Ο1	2	1		Ο3	Ο9	0	2		Ο11	Ο15	1	0
	Ο6	Ο10	1	1		Ο15	Ο14	0	0		Ο7	Ο4	1	1
	Ο14	Ο16	2	2		Ο17	Ο12	2	2		Ο2	Ο6	1	0
	Ο2	Ο11	0	0		Ο11	Ο16	5	2		Ο8	Ο12	2	0
	Ο8	Ο7	3	0		Ο4	Ο6	0	0		Ο14	Ο9	3	0
	Ο3	Ο15	0	3		Ο10	Ο2	2	0		Ο16	Ο10	4	2
4η	Ο8	Ο17	3	0	10η	Ο3	Ο17	2	2	16η	Ο13	Ο17	1	1
	Ο15	Ο10	2	3		Ο5	Ο11	1	2		Ο15	Ο7	1	0
	Ο11	Ο9	0	1		Ο16	Ο7	2	1		Ο4	Ο8	0	5
	Ο1	Ο4	1	0		Ο8	Ο13	5	1		Ο12	Ο2	2	2
	Ο5	Ο14	2	0		Ο6	Ο15	0	0		Ο1	Ο16	2	1

	O16	O6	2	0		O12	O4	0	2		O3	O14	1	0
	O13	O12	1	0		O9	O10	1	2		O6	O9	1	2
	O7	O2	1	0		O2	O1	4	3		O10	O5	2	0
5η	O2	O8	1	0	11η	O17	O4	1	1	17η	O17	O11	1	3
	O9	O7	4	0		O15	O12	3	0		O5	O1	1	0
	O15	O1	0	1		O1	O9	3	0		O9	O12	3	2
	O17	O14	2	0		O11	O14	0	0		O8	O15	2	1
	O12	O16	1	1		O13	O2	1	1		O2	O4	0	0
	O6	O5	2	3		O7	O5	3	2		O6	O3	1	1
	O3	O11	1	1		O10	O3	5	0		O14	O10	0	0
	O4	O13	2	0		O16	O8	1	3		O16	O13	0	0
6η	O8	O9	1	2	12η	O4	O15	0	0					
	O11	O10	0	1		O6	O11	0	0					
	O14	O6	0	4		O2	O16	2	0					
	O16	O4	2	1		O5	O8	0	4					
	O5	O12	1	0		O3	O1	0	1					
	O13	O15	1	0		O14	O7	3	0					
	O7	O3	1	0		O9	O13	3	2					
	O2	O17	0	0		O10	O17	2	1					

Για την πρόβλεψη των αποτελεσμάτων του δεύτερου γύρου του ελληνικού πρωταθλήματος ποδοσφαίρου, με ακρίβεια, χρησιμοποιήθηκαν τα παραπάνω δεδομένα με δύο τρόπους: α) Εφαρμόζοντας το μοντέλο πρόβλεψης που προτάθηκε από τον Chu (2003) και β) Εφαρμόζοντας το μοντέλο πρόβλεψης που προτείνεται στην παρούσα εργασία. Στη συνέχεια ακολουθεί η σύγκριση των δύο μοντέλων, η οποία βασίζεται στον έλεγχο καλής προσαρμογής X^2.

3.1 Εκτίμηση των αποτελεσμάτων του Ελληνικού Πρωταθλήματος Ποδοσφαίρου με το μοντέλο πρόβλεψης του Chu

Σύμφωνα με το μοντέλο πρόβλεψης του Chu (2003) η κατανομή Poisson με παράμετρο λ το μέσο αριθμό των τερμάτων που σημειώθηκαν κατά τη διάρκεια του πρώτου γύρου του ελληνικού πρωταθλήματος ποδοσφαίρου, μπορεί να χρησιμοποιηθεί για την εκτίμηση των πιθανοτήτων οι ομάδες να πετύχουν συγκεκριμένο αριθμό τερμάτων στο δεύτερο γύρο του ελληνικού πρωταθλήματος ποδοσφαίρου. Για την εκτίμηση του μέσου όρου των τερμάτων λ_1 της γηπεδούχου ομάδας και λ_2 της φιλοξενούμενης ομάδας, αντίστοιχα, χρησιμοποιήθηκαν τα δεδομένα που προέκυψαν από τις 17 αγωνιστικές του πρώτου γύρου του ελληνικού πρωταθλήματος (Πίνακας 1). Ειδικότερα, προέκυψε ότι:

$$\lambda_1=1,45588 \text{ και } \lambda_2=0,95588.$$

Στη συνέχεια, εκτιμήθηκαν οι πιθανότητες για τη γηπεδούχο ομάδα να πετύχει συγκεκριμένο αριθμό τερμάτων στον δεύτερο γύρο του ελληνικού πρωταθλήματος. Στον πίνακα 3 παρουσιάζονται όλα τα χρήσιμα στοιχεία για την πραγματοποίηση του ελέγχου καλής προσαρμογής X^2 για τη γηπεδούχο ομάδα στο δεύτερο γύρο του ελληνικού πρωταθλήματος.

Πίνακας 3. *Έλεγχος X^2 για τη γηπεδούχο ομάδα*

Αριθμός τερμάτων	f_i	p_i	$\theta_i = n*p_i$	$(f_i - \theta_i)^2/\theta_i$
0	29	0,233195062	31,71452842	0,232343501
1	41	0,339504027	46,17254764	0,579462266
2	34	0,247138561	33,61084433	0,004505752
3 ή περισσότερα	32	0,18016235	24,5020796	2,294450563
	136	1	136	3,110762082

Όπου f_i=παρατηρούμενη συχνότητα, δηλαδή ο αριθμός αγώνων που είχαμε το συγκεκριμένο αριθμό τερμάτων, θ_i=θεωρητική συχνότητα με βάση την κατανομή Poisson, p_i=η πιθανότητα επίτευξης i τερμάτων με βάση την κατανομή Poisson και $(f_i - \theta_i)^2/\theta_i$=ο όρος του αθροίσματος της στατιστικής συνάρτησης X^2

Αφού ολοκληρώθηκε η εκτίμηση των συχνοτήτων των τερμάτων και των πιθανοτήτων τους, εφαρμόστηκε ο έλεγχος X^2 καλής προσαρμογής προκειμένου να διαπιστωθεί εάν ο αριθμός των τερμάτων της γηπεδούχου ομάδας μπορεί να προσεγγιστεί από την κατανομή Poisson με παράμετρο λ_1.

Σύμφωνα με τα αποτελέσματα του ελέγχου καλής προσαρμογής X^2:

$$D^2 = \sum_{i=0}^{i=3} \frac{(f_i - \theta_i)^2}{\theta_i} \cong 3{,}11 < 7{,}815 = X_3^2$$

Επομένως, το προτεινόμενο μοντέλο μπορεί να χρησιμοποιηθεί για την πρόβλεψη των σκορ σε ποδοσφαιρικούς αγώνες για τη γηπεδούχο ομάδα.

Με ανάλογο τρόπο, εκτιμήθηκαν οι πιθανότητες για την φιλοξενούμενη ομάδα να πετύχει συγκεκριμένο αριθμό τερμάτων στον δεύτερο γύρο του ελληνικού πρωταθλήματος. Στον πίνακα 4, παρουσιάζονται όλα τα χρήσιμα στοιχεία για την πραγματοποίηση του ελέγχου καλής προσαρμογής X^2 για την φιλοξενούμενη ομάδα στο δεύτερο γύρο του ελληνικού πρωταθλήματος.

Πίνακας 4. *Έλεγχος X^2 για τη φιλοξενούμενη ομάδα*

Αριθμός τερμάτων	f_i	p_i	$\theta_i = n*p_i$	$(f_i - \theta_i)2/\theta_i$
0	62	0,384473659	52,2884176	1,80374234
1	42	0,367510681	49,98145262	1,274544506
2	21	0,175648055	23,88813546	0,349182818
3ή περισσότερα	11	0,072367605	9,84199428	0,136250562
	136	1	136	3,563720226

Σύμφωνα με τα αποτελέσματα του ελέγχου καλής προσαρμογής X^2:

$$D^2 = \sum_{i=0}^{i=3} \frac{(f_i - \theta_i)^2}{\theta_i} \cong 3{,}564 < 7{,}815 = X_3^2$$

Επομένως, το προτεινόμενο μοντέλο μπορεί να χρησιμοποιηθεί για την πρόβλεψη των σκορ σε ποδοσφαιρικούς αγώνες για τη φιλοξενούμενη ομάδα.

3.2 Εκτίμηση των αποτελεσμάτων του Ελληνικού Πρωταθλήματος Ποδοσφαίρου με το νέο/προτεινόμενο μοντέλο

Σύμφωνα με το προτεινόμενο μοντέλο πρόβλεψης, η κατανομή Poisson όπου ο υπολογισμός της παραμέτρου λ βασίζεται και στον αριθμό των τερμάτων που πετυχαίνει η κάθε ομάδα αλλά και στον αριθμό των προσπαθειών κάθε ομάδας για να πετύχει τέρμα, μπορεί να χρησιμοποιηθεί για την εκτίμηση των πιθανοτήτων οι ομάδες να πετύχουν συγκεκριμένο αριθμό τερμάτων στο δεύτερο γύρο του ελληνικού πρωταθλήματος ποδοσφαίρου. Για την εκτίμηση των παραμέτρων λ_1΄της γηπεδούχου ομάδας και λ_2΄της φιλοξενούμενης ομάδας, αντίστοιχα, χρησιμοποιήθηκαν τα δεδομένα που προέκυψαν από τις 17 αγωνιστικές του πρώτου γύρου του ελληνικού πρωταθλήματος, όπως παρουσιάζονται στον πίνακα 1, αλλά και ο αριθμός των προσπαθειών κάθε ομάδας για να πετύχει τέρμα στον πρώτο γύρο του ελληνικού πρωταθλήματος ποδοσφαίρου, όπως παρουσιάζονται στον πίνακα 5.

Πίνακας 5. *Αριθμός προσπαθειών κάθε ομάδας για να πετύχει τέρμα στον πρώτο γύρο του ελληνικού πρωταθλήματος ποδοσφαίρου*

Αγωνιστική	ΓηπεδούχοςΟμάδα	Φιλοξενούμενη Ομάδα	Αποτέλεσμα		Αγωνιστική	ΓηπεδούχοςΟμάδα	Φιλοξενούμενη Ομάδα	Αποτέλεσμα		Αγωνιστική	ΓηπεδούχοςΟμάδα	Φιλοξενούμενη Ομάδα	Αποτέλεσμα	
1η	O7	O11	16	3	7η	O3	O8	17	22	13η	O17	O15	13	14
	O5	O13	12	7		O10	O7	15	7		O5	O2	16	8
	O9	O4	16	8		O9	O2	12	2		O8	O14	24	7
	O14	O1	10	6		O1	O11	15	13		O16	O9	6	9
	O2	O15	14	10		O12	O14	12	7		O11	O12	17	13
	O17	O16	11	9		O17	O6	15	10		O7	O6	14	7
	O6	O10	5	16		O15	O16	10	12		O1	O10	17	9
	O12	O3	18	12		O4	O5	11	7		O13	O3	22	6
2η	O13	O14	18	11	8η	O8	O10	6	4	14η	O12	O7	8	2
	O4	O3	18	7		O9	O17	20	9		O1	O17	9	7
	O11	O8	6	11		O7	O1	6	12		O3	O16	6	12
	O15	O9	8	14		O14	O4	7	11		O14	O2	9	10

	O16	O5	10	8		O2	O3	7	10		O9	O5	13	14
	O1	O6	20	7		O11	O13	7	8		O10	O13	22	6
	O10	O12	18	7		O6	O12	17	8		O4	O11	11	6
	O7	O17	10	9		O5	O15	12	9		O6	O8	5	21
3η	O17	O5	12	10	9η	O1	O8	5	22	15η	O5	O3	12	14
	O4	O10	5	12		O13	O7	14	5		O13	O1	11	18
	O12	O1	13	9		O3	O9	7	10		O11	O15	21	7
	O6	O10	14	12		O15	O14	12	6		O7	O4	12	8
	O14	O16	15	7		O17	O12	13	8		O2	O6	13	8
	O2	O11	8	6		O11	O16	18	11		O8	O12	14	7
	O8	O7	23	5		O4	O6	11	10		O14	O9	13	6
	O3	O15	15	15		O10	O2	14	5		O16	O10	13	9
4η	O8	O17	13	5	10η	O3	O17	11	11	16η	O13	O17	9	8
	O15	O10	10	9		O5	O11	5	11		O15	O7	16	6
	O11	O9	8	9		O16	O7	9	6		O4	O8	5	19
	O1	O4	19	4		O8	O13	28	10		O12	O2	15	10
	O5	O14	6	10		O6	O15	13	6		O1	O16	8	7
	O16	O6	7	7		O12	O4	17	5		O3	O14	16	9
	O13	O12	10	5		O9	O10	8	9		O6	O9	20	7
	O7	O2	8	9		O2	O1	13	10		O10	O5	17	9
5η	O2	O8	6	8	11η	O17	O4	15	8	17η	O17	O11	12	5
	O9	O7	18	7		O15	O12	22	8		O5	O1	15	14
	O15	O1	11	8		O1	O9	10	6		O9	O12	7	11
	O17	O14	19	3		O11	O14	14	7		O8	O15	11	11
	O12	O16	23	10		O13	O2	3	12		O2	O4	14	9
	O6	O5	19	5		O7	O5	18	13		O6	O3	16	8
	O3	O11	16	7		O10	O3	14	4		O14	O10	10	6
	O4	O13	16	5		O16	O8	5	8		O16	O13	10	1
6η	O8	O9	14	5	12η	O4	O15	12	9					
	O11	O10	11	14		O6	O11	12	3					
	O14	O6	8	11		O2	O16	7	10					
	O16	O4	20	5		O5	O8	9	25					
	O5	O12	6	10		O3	O1	8	13					
	O13	O15	14	13		O14	O7	15	4					
	O7	O3	15	11		O9	O13	14	2					
	O2	O17	9	3		O10	O17	13	5					

Ακολούθησε ο υπολογισμός των παραμέτρων αποτελεσματικότητας c_1 για τη γηπεδούχο και c_2 τη φιλοξενούμενη ομάδα, που ισούνται με το μέσο όρο των πηλίκων

των διαιρέσεων του αριθμού των τερμάτων με τον αριθμό των προσπαθειών για κάθε ομάδα, που παρουσιάζονται στον πίνακα 6. Οι παράμετροι αποτελεσματικότητας είναι c_1=0,124156 για τη γηπεδούχο ομάδα και c_2=0,111998 για τη φιλοξενούμενη ομάδα, αντίστοιχα.

Πίνακας 6. *Πηλίκα διαιρέσεων του αριθμού των τερμάτων με τον αριθμό των προσπαθειών για κάθε ομάδα*

Αγωνιστική	Γηπεδούχος Ομάδα	Φιλοξενούμενη Ομάδα	Αγωνιστική	Γηπεδούχος Ομάδα	Φιλοξενούμενη Ομάδα	Αγωνιστική	Γηπεδούχος Ομάδα	Φιλοξενούμενη Ομάδα
1η	0,063	0	7η	0,118	0,136	13η	0,154	0
	0,167	0,143		0,067	0,286		0,063	0,125
	0,063	0,125		0,167	0,5		0,083	0,286
	0,3	0,167		0,067	0,077		0,667	0,222
	0,071	0		0	0,143		0,118	0,077
	0,273	0,222		0,133	0,1		0	0,286
	0,2	0,063		0,1	0		0,059	0,111
	0,111	0,083		0,091	0		0	0
2η	0,056	0,091	8η	0,167	0	14η	0,25	0
	0,111	0		0,2	0,111		0,222	0
	0,167	0,091		0,333	0,25		0	0,167
	0	0,286		0	0		0,111	0
	0	0		0,143	0,1		0,154	0,071
	0,1	0,143		0,429	0,125		0,182	0
	0,111	0,143		0,059	0		0	0
	0	0,111		0,083	0,222		0,2	0,095
3η	0,167	0,1	9η	0	0	15η	0,25	0,071
	0,2	0		0,143	0		0,091	0,056
	0,154	0,111		0	0,2		0,048	0
	0,071	0,083		0	0		0,083	0,125
	0,133	0,286		0,154	0,25		0,077	0
	0	0		0,278	0,182		0,143	0
	0,13	0		0	0		0,231	0
	0	0,2		0,143	0		0,308	0,222
4η	0,231	0	10η	0,182	0,182	16η	0,111	0,125
	0,2	0,333		0,2	0,182		0,063	0

	0	0,111		0,222	0,167		0	0,263
	0,053	0		0,179	0,1		0,133	0,2
	0,333	0		0	0		0,25	0,143
	0,286	0		0	0,4		0,063	0
	0,1	0		0,125	0,222		0,05	0,286
	0,125	0		0,308	0,3		0,118	0
5η	0,167	0	11η	0,067	0,125	17η	0,083	0,6
	0,222	0		0,136	0		0,067	0
	0	0,125		0,3	0		0,429	0,182
	0,105	0		0	0		0,182	0,091
	0,043	0,1		0,333	0,083		0	0
	0,105	0,6		0,167	0,154		0,063	0,125
	0,063	0,143		0,357	0		0	0
	0,125	0		0,2	0,375		0	0
6η	0,071	0,4	12η	0	0			
	0	0,071		0	0			
	0	0,364		0,286	0			
	0,1	0,2		0	0,16			
	0,167	0		0	0,077			
	0,071	0		0,2	0			
	0,067	0		0,214	1			
	0	0		0,154	0,2			

Στο επόμενο βήμα για τον υπολογισμό των παραμέτρων λ_1'και λ_2', οι παράμετροι αποτελεσματικότητας πολλαπλασιάστηκαν με τις τελικές προσπάθειες κάθε ομάδας. Τα γινόμενα των πολλαπλασιασμών παρουσιάζονται στον πίνακα 7.

Πίνακας 7. *Εκτιμώμενο σκορ με τη χρήση των τελικών προσπαθειών*

Αγωνιστική	Γηπεδούχος Ομάδα	Φιλοξενούμενη Ομάδα	Αγωνιστική	ΓηπεδούχοςΟμάδα	Φιλοξενούμενη Ομάδα	Αγωνιστική	ΓηπεδούχοςΟμάδα	Φιλοξενούμενη Ομάδα
$1^{η}$	1,987	0,336	$7^{η}$	2,111	2,464	13η	1,614	1,568
	1,49	0,784		1,862	0,784		1,987	0,896
	1,987	0,896		1,49	0,224		2,98	0,784
	1,242	0,672		1,862	1,456		0,745	1,008
	1,738	1,12		1,49	0,784		2,111	1,456
	1,366	1,008		1,862	1,12		1,738	0,784

	0,621	1,792		1,242	1,344		2,111	1,008
	2,235	1,344		1,366	0,784		2,732	0,672
2^{η}	2,235	1,232	8^{η}	0,745	0,448	14η	0,993	0,224
	2,235	0,784		2,483	1,008		1,117	0,784
	0,745	1,232		0,745	1,344		0,745	1,344
	0,993	1,568		0,869	1,232		1,117	1,12
	1,242	0,896		0,869	1,12		1,614	1,568
	2,483	0,784		0,869	0,896		2,732	0,672
	2,235	0,784		2,111	0,896		1,366	0,672
	1,242	1,008		1,49	1,008		0,621	2,352
3^{η}	1,49	1,12	9^{η}	0,621	2,464	15η	1,49	1,568
	0,621	1,344		1,738	0,56		1,366	2,016
	1,614	1,008		0,869	1,12		2,607	0,784
	1,738	1,344		1,49	0,672		1,49	0,896
	1,862	0,784		1,614	0,896		1,614	0,896
	0,993	0,672		2,235	1,232		1,738	0,784
	2,856	0,56		1,366	1,12		1,614	0,672
	1,862	1,68		1,738	0,56		1,614	1,008
4^{η}	1,614	0,56	10η	1,366	1,232	16η	1,117	0,896
	1,242	1,008		0,621	1,232		1,987	0,672
	0,993	1,008		1,117	0,672		0,621	2,128
	2,359	0,448		3,476	1,12		1,862	1,12
	0,745	1,12		1,614	0,672		0,993	0,784
	0,869	0,784		2,111	0,56		1,987	1,008
	1,242	0,56		0,993	1,008		2,483	0,784
	0,993	1,008		1,614	1,12		2,111	1,008
5^{η}	0,745	0,896	11η	1,862	0,896	17η	1,49	0,56
	2,235	0,784		2,732	0,896		1,862	1,568
	1,366	0,896		1,242	0,672		0,869	1,232
	2,359	0,336		1,738	0,784		1,366	1,232
	2,856	1,12		0,372	1,344		1,738	1,008
	2,359	0,56		2,235	1,456		1,987	0,896
	1,987	0,784		1,738	0,448		1,242	0,672
	1,987	0,56		0,621	0,896		1,242	0,112
6^{η}	1,738	0,56	12η	1,49	1,008			
	1,366	1,568		1,49	0,336			
	0,993	1,232		0,869	1,12			
	2,483	0,56		1,117	2,8			

	0,745	1,12		0,993	1,456
	1,738	1,456		1,862	0,448
	1,862	1,232		1,738	0,224
	1,117	0,336		1,614	0,56

Το τελικό βήμα του προτεινόμενου αλγορίθμου βασίζεται στον προσδιορισμό των παραμέτρων λ_1' και λ_2', οι οποίες ισούνται με το μέσο όρο των προηγούμενων τιμών που αφορούν τη γηπεδούχο και τη φιλοξενούμενη ομάδα, αντίστοιχα. Επομένως,

$$\lambda_1' = 1{,}571172 \text{ και } \lambda_2' = 0{,}992353$$

Στη συνέχεια, εκτιμήθηκαν οι πιθανότητες για τη γηπεδούχο ομάδα να πετύχει συγκεκριμένο αριθμό τερμάτων στο δεύτερο γύρο του ελληνικού πρωταθλήματος. Στον πίνακα 8 παρουσιάζονται όλα τα χρήσιμα στοιχεία για την πραγματοποίηση του ελέγχου καλής προσαρμογής X^2 για τη γηπεδούχο ομάδα στο δεύτερο γύρο του ελληνικού πρωταθλήματος (Πίνακας 8).

Πίνακας 8. *Έλεγχος X^2 για τη γηπεδούχο ομάδα*

Αριθμός τερμάτων	f_i	p_i	$\theta_i = n*p_i$	$(f_i - \theta_i)^2/\theta_i$
0	29	0,207801912	28,26106001	0,019321013
1	41	0,32649213	44,40292966	0,260792032
2	34	0,25648732	34,88227549	0,022315346
3 ή περισσότερα	32	0,209218638	28,45373477	0,441980541
	136	1	136	0,744408932

Σύμφωνα με τα αποτελέσματα του ελέγχου καλής προσαρμογής X^2:

$$D^2 = \sum_{i=0}^{i=3} \frac{(f_i - \theta_i)^2}{\theta_i} \cong 0{,}744 < 7{,}815 = X_3^2$$

Επομένως, το προτεινόμενο μοντέλο μπορεί να χρησιμοποιηθεί για την πρόβλεψη των σκορ σε ποδοσφαιρικούς αγώνες για τη γηπεδούχο ομάδα.

Ακολουθώντας την ίδια διαδικασία εκτιμήθηκαν οι πιθανότητες και για την φιλοξενούμενη ομάδα να πετύχει συγκεκριμένο αριθμό τερμάτων στον δεύτερο γύρο του ελληνικού πρωταθλήματος. Στον πίνακα 9 παρουσιάζονται όλα τα χρήσιμα στοιχεία για την πραγματοποίηση του ελέγχου καλής προσαρμογής X^2 για τη φιλοξενούμενη ομάδα, στο δεύτερο γύρο του ελληνικού πρωταθλήματος.

Πίνακας 9. *Έλεγχος X^2 για τη φιλοξενούμενη ομάδα*

Αριθμός τερμάτων	f_i	p_i	$\theta_i = n*p_i$	$(f_i - \theta_i)2/\theta_i$
0	62	0,370704511	50,4158135	2,661731858
1	42	0,367868622	50,03013252	1,288883821
2	21	0,182527213	24,82370101	0,588981046
3ή περισσότερα	11	0,078899654	10,73035294	0,006776062
	136	1	136	4,5463787

Σύμφωνα με τα αποτελέσματα του ελέγχου καλής προσαρμογής X^2:

$$D^2 = \sum_{i=0}^{i=3} \frac{(f_i - \theta_i)^2}{\theta_i} \cong 4{,}546 < 7{,}815 = X_3^2$$

Επομένως, το προτεινόμενο μοντέλο μπορεί να χρησιμοποιηθεί για την πρόβλεψη των σκορ σε ποδοσφαιρικούς αγώνες για τη φιλοξενούμενη ομάδα.

3.3 Σύγκριση των δύο μοντέλων πρόβλεψης

Για τη σύγκριση του μοντέλου πρόβλεψης του Chu (2003) με το προτεινόμενο μοντέλο πρόβλεψης, χρησιμοποιήθηκε ο έλεγχος καλής προσαρμογής X^2. Για τη γηπεδούχο ομάδα τα αποτελέσματα δίνουν μια ισχυρή και κρίσιμη διαφορά μεταξύ των δύο μοντέλων. Ειδικότερα, για τη γηπεδούχο ομάδα, η τιμή D^2 του προτεινόμενου μοντέλου είναι πολύ χαμηλότερη από την τιμή D^2 για μοντέλο του Chu. Για τη φιλοξενούμενη ομάδα, η τιμή D^2 του προτεινόμενου μοντέλου είναι υψηλότερη από την τιμή D^2 του μοντέλου του Chu. Ωστόσο, από τη συνολική σύγκριση των δύο μοντέλων, αθροίζοντας τα D^2, υπερτερεί το προτεινόμενο μοντέλο πρόβλεψης, όπως φαίνεται και στον πίνακα 10.

Πίνακας 10. *Σύγκριση των D^2 για τα δύο μοντέλα*

Μοντέλο του Chu	Γηπεδούχος ομάδα	3,11
	Φιλοξενούμενη ομάδα	3,564
	Άθροισμα	6,674
Προτεινόμενο μοντέλο	Γηπεδούχος ομάδα	0,744
	Φιλοξενούμενη ομάδα	4,546
	Άθροισμα	5,29

4. ΣΥΖΗΤΗΣΗ

Στην παρούσα εργασία παρουσιάστηκε και εφαρμόστηκε ένα νέο μοντέλο για την πρόβλεψη των αποτελεσμάτων των ποδοσφαιρικών αγώνων, το οποίο βασίζεται στην κατανομή Poisson. Σύμφωνα με τα ευρήματα της εργασίας η χρήση του αριθμού των τελικών προσπαθειών σε συνδυασμό με τον αριθμό των τερμάτων που σημείωσαν οι ομάδες, βελτίωσε την προσαρμογή του μοντέλου και την πρόβλεψη του σκορ/των αποτελεσμάτων για τους αγώνες ποδοσφαίρου. Επομένως, με χρήση περισσότερων εφαρμογών σε δεδομένα, συνιστάται να διερευνηθεί εάν το νέο μοντέλο εκτίμησης πιθανοτήτων μπορεί να αντικαταστήσει το υπάρχον μοντέλο του Chu και αν μπορεί να χρησιμοποιηθεί για την ακριβή εκτίμηση πιθανοτήτων αποτελεσμάτων και άλλων αθλημάτων.

Αξίζει να σημειωθεί ότι αν και στην πράξη η προϋπόθεση της ανεξαρτησίας των γεγονότων της κατανομής Poisson, στο ποδόσφαιρο συχνά δεν τηρείται καθώς η επίτευξη ενός τέρματος μπορεί να μεταβάλλει τη στρατηγική και τον ρυθμό παιχνιδιού, οδηγώντας σε μη σταθερό λ, ενώ ο αριθμός τερμάτων των δύο ομάδων μπορεί να είναι θετικά συσχετισμένος. Ωστόσο, η κατανομή Poisson είναι ο επικρατέστερος τρόπος εκτίμησης αποδόσεων σε ποδόσφαιρο και Πόλο όπως αναφέρετε στη βιβλιογραφία (Karlis και Ntzoufras, 2003; Chu, 2003; Chalikias et.al 2020)

ABSTRACT

Poisson distribution is one of the widely used tools that can be used to predict the probability of the exact scores in sporting events, such as football matches. The purpose of this paper is to suggest and analyze a new method to predict exact scores in football matches and to exhibit its results. While the original approach was based only on the goals by each team, in order to improve the approach, we used not only the goals scored by each team, but also the number of final attempts. The implementation of the new model on the results of the Greek football League revealed that the proposed model can be used to predict the exact score of football matches more accurately, than the previous one.

ΑΝΑΦΟΡΕΣ

Chalikias, M. S.(2009). The binomial distribution in shooting. *Teaching Statistics*,**31(3)**, 87-89.

Chalikias, M., Kossieri, E., & Lalou, P. (2020). Football matches: Decision making in betting. Teaching Statistics, **42(1)**, 4-9.

Chu, S. (2003).Using soccer goals to motivate the Poisson process. *INFORMS Transactions on Education,***3(2)**, 64-70.

Karlis, D. and Ntzoufras, I. (2003).Analysis of sports data by using bivariate Poisson models.*Journal of the Royal Statistical Society: Series D (The Statistician)*, **52(3)**, 381-393.

Keller, J. B. (1994). A characterization of the Poisson distribution and the probability of winning a game. *The American Statistician*, **48(4)**, 294-298.

Lago-Ballesteros, J. andLago-Peñas, C. (2010). Performance in team sports: Identifying the keys to success in soccer. *Journal of Humankinetics*, **25(1)**, 85-91.

Tijms, H. (2015). Teaching note–was the champions league draw rigged?.*Teaching Statistics*, **37(3)**, 104-105.

Williams, I. J. and Williams, K. K. (2016).Understanding summary statistics and graphical techniques to compare Michael Jordan versus LeBron James.*Teaching Statistics*, **38(3)**, 108-111.

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 37ου Πανελληνίου Συνεδρίου Στατιστικής (2025), σσ. 215–236

ΟΠΤΙΚΟΠΟΙΗΣΗ ΚΥΡΙΩΝ ΔΕΙΚΤΩΝ ΑΠΟΔΟΣΗΣ (KPIs) ΣΤΗΝ ΚΑΛΑΘΟΣΦΑΙΡΙΣΗ

Τσαφής Δημήτριος[1], Αυλογιάρης Γεώργιος[1]
[1] Τμήμα Στατιστικής, Πανεπιστήμιο Δυτικής Μακεδονίας, Ελλάδα
{dtsafis, gavlogiaris}@uowm.gr

ΠΕΡΙΛΗΨΗ

Το άθλημα της καλαθοσφαίρισης παρέχει στη σημερινή εποχή ένα διαρκώς αυξανόμενο όγκο δεδομένων, η διαχείριση των οποίων απαιτεί σύγχρονα αναλυτικά εργαλεία που υπερβαίνουν την απλή παράθεση αριθμητικών μεγεθών. Η παρούσα εργασία θέτει στο επίκεντρο την αξία της οπτικοποίησης δεδομένων, την οποία ανέδειξε ο Anscombe το 1973 με το περίφημο «κουαρτέτο» δεδομένων του.

Πέρα από την ανάδειξη της αναγκαιότητας της οπτικοποίησης των δεδομένων, κύριος σκοπός της εργασίας είναι η εφαρμογή σύγχρονων μεθόδων οπτικοποίησης παραδοσιακών και προηγμένων μετρικών ή κύριων δεικτών απόδοσης (KPI's) στην καλαθοσφαίριση. Η οπτικοποίηση των κύριων δεικτών απόδοσης (KPI's) στο άθλημα της καλαθοσφαίρισης (μπάσκετ) αποτελεί το εφαρμοσμένο μέρος της συγκεκριμένης εργασίας. Αξιοποιώντας πραγματικά δεδομένα από την EuroLeague για την περίοδο 2023-2024 και με τη χρήση της R και της Python, κατασκευάζονται μοντέρνα γραφήματα όπως shot charts, γραφήματα τεταρτημορίων, radial γραφήματα κλπ. Τα αποτελέσματα καταδεικνύουν ότι η οπτικοποίηση δεν αποτελεί απλώς ένα μέσο παρουσίασης, αλλά ένα απαραίτητο αναλυτικό εργαλείο για την κατανόηση της αποδοτικότητας των ομάδων σε άμυνα και επίθεση στο σύγχρονο μπάσκετ.

Λέξεις κλειδιά: Καλαθοσφαίριση, Κύριοι δείκτες απόδοσης, Οπτικοποίηση δεδομένων, EuroLeague 2023-2024

1. ΕΙΣΑΓΩΓΗ

Σκοπός της παρούσας εργασίας είναι, αφενός, η ανάδειξη της αξίας που κατέχει η οπτικοποίηση των δεδομένων και, αφετέρου, η παρουσίαση σύγχρονων μεθόδων γραφικής απεικόνισης κύριων δεικτών απόδοσης (KPI's) στο άθλημα της καλαθοσφαίρισης.

Η οπτικοποίηση δεδομένων είναι η γραφική απεικόνιση ή αναπαράσταση δεδομένων. Τα δεδομένα αποτελούν πρωτογενείς μετρήσεις (τιμές) που προκύπτουν από την παρατήρηση ενός φαινομένου. Με τον όρο γραφική αναπαράσταση δεδομένων εννοείται η χρήση τεχνικών για τη δημιουργία «εικόνων», όπως διαγραμμάτων, γραφημάτων, χαρτών κλπ. Με τη γραφική απεικόνιση επιτυγχάνεται αφενός η ανάδειξη συσχετίσεων

(ή μη) στα δεδομένα και αφετέρου αποκαλύπτονται πρότυπα ή τάσεις που - ενδεχομένως - υπάρχουν σε αυτά. Αυτό βοηθάει στη διάχυση πληροφοριών στο σύνολο των ενδιαφερομένων με αποτέλεσμα τη βελτίωση της διαδικασίας λήψης αποφάσεων.

Οι John Tukey [2], Edward Tufte [4] και F.J.Anscombe [1] είναι μερικοί από τους πολλούς στατιστικούς που τόνισαν τη σημασία της οπτικοποίησης δεδομένων. Συγκεκριμένα, ο Tukey (δημιουργός του box plot) υποστήριζε ότι το μεγαλύτερο μέρος μιας στατιστικής ανάλυσης θα έπρεπε να είναι οπτικό. Ο Tufte θεωρούσε ότι μια εικόνα είναι ισχυρότερη από τους αριθμούς και ότι κάθε γράφημα αφηγείται μια ιστορία για τα δεδομένα (data storytelling). Ο Anscombe μέσα από το περίφημο σύνολο δεδομένων του (Anscombe's Quartet) απέδειξε ότι τα στατιστικά στοιχεία - χωρίς τη γραφική τους αποτύπωση - μπορεί να μην είναι επαρκή και ορισμένες φορές να οδηγούν σε παραπλανήσεις.

Η εικόνα έχει δύναμη επειδή αποτυπώνεται ευκολότερα στη μνήμη. Προκαλεί την προσοχή και επιτρέπει την άμεση κατανόηση του φαινομένου που μελετάται. Στην ουσία ένα γράφημα «μετατρέπει» αριθμούς σε εικόνες, τις οποίες ο ανθρώπινος νους επεξεργάζεται πιο εύκολα [14]. Η χρήση ενός γραφήματος επιτρέπει τη σύγκριση δεδομένων, τη διαχρονική μελέτη ενός φαινομένου (π.χ χρονολογική σειρά), τον εντοπισμό ακραίων ή παράτυπων τιμών και την αποκάλυψη συσχετίσεων μεταξύ μεταβλητών. Αξίζει να σημειωθεί ότι η ραγδαία ανάπτυξη της πληροφορικής (και η δημιουργία λογισμικών γραφικών) δημιούργησε τεράστιες δυνατότητες σε ότι αφορά στη γραφική απεικόνιση δεδομένων.

Παρά τα πλεονεκτήματά της, η οπτικοποίηση δεδομένων ενέχει ορισμένους περιορισμούς, με κυριότερο την - ενδεχόμενη - απώλεια μέρους της πληροφορίας κατά τη διαδικασία της απεικόνισης. Είναι προφανές ότι σε ένα γράφημα δεν είναι εφικτό να αποτυπωθεί το σύνολο των πρωτογενών δεδομένων. Επιπρόσθετα, ορισμένα είδη γραφημάτων ενδέχεται να υποβαθμίζουν την αριθμητική ακρίβεια προς όφελος της οπτικής αντίληψης των τάσεων. Συχνά, η κακή εφαρμογή σχεδιαστικών πρακτικών, όπως η χρήση ακατάλληλων κλιμάκων μπορεί να οδηγήσει σε παρερμηνείες [4].

Στο άθλημα της καλαθοσφαίρισης τα πρωτογενή δεδομένα προέρχονται από την λεπτομερή καταγραφή γεγονότων κατά τη διάρκεια του παιχνιδιού όπως π.χ ένα άστοχο ή εύστοχο σουτ τριών πόντων, ένα κλέψιμο, ένα λάθος κλπ. Τα δεδομένα αυτά αποτελούν τη βάση για τον υπολογισμό προηγμένων μετρικών, που αξιοποιούνται πλέον ευρέως για τη βαθύτερη κατανόηση του αθλήματος. Στην παρούσα εργασία πραγματοποιείται μια ιστορική αναδρομή στην εξέλιξη της αναλυτικής καλαθοσφαίρισης (basketball analytics) από τη δεκαετία του 1980 έως σήμερα. Στη συνέχεια, με τη βοήθεια σύγχρονων εργαλείων οπτικοποίησης, αποτυπώνονται ορισμένες προηγμένες μετρικές που «αξιολογούν» την αποδοτικότητα των ομάδων τόσο στην επίθεση όσο και στην άμυνα.

2. ANSCOMBE´S QUARTET

2.1 Η εργασία του Anscombe

Το 1973 ο F.J Anscombe [1] δημοσίευσε στο περιοδικό «The american statistician» την εργασία του με τίτλο «Graphs in statistical analysis». Σε αυτή περιλαμβάνονταν 4 σύνολα δεδομένων (δομημένη συλλογή δεδομένων υπό μορφή πίνακα όπου οι γραμμές αντιπροσωπεύουν τις παρατηρήσεις και οι στήλες τις μεταβλητές), όπου κάθε σύνολο αποτελούνταν από 2 μεταβλητές X και Y (κάθε μεταβλητή από τις συνολικά 8 αποτελούνταν από n=11 παρατηρήσεις). Ακολουθεί ο πίνακας 1 που περιλαμβάνει τα 4 σύνολα δεδομένων.

Πίνακας 1: *Anscombe's Quartet*

Dataset 1		**Dataset 2**		**Dataset 3**		**Dataset 4**	
X_1	Y_1	X_2	Y_2	X_3	Y_3	X_4	Y_4
4	4.26	4	3.10	4	5.39	8	5.25
5	6.68	5	4.74	5	5.73	8	5.56
6	7.24	6	6.13	6	6.08	8	5.76
7	4.82	7	7.26	7	6.42	8	6.58
8	6.96	8	8.14	8	6.77	8	6.89
9	8.81	9	8.77	9	7.11	8	7.04
10	8.04	10	9.14	10	7.46	8	7.71
11	8.33	11	9.26	11	7.81	8	7.91
12	10.84	12	9.13	12	8.15	8	8.47
13	7.68	13	8.74	13	12.74	8	8.84
14	9.96	14	8.10	14	8.44	19	12.50

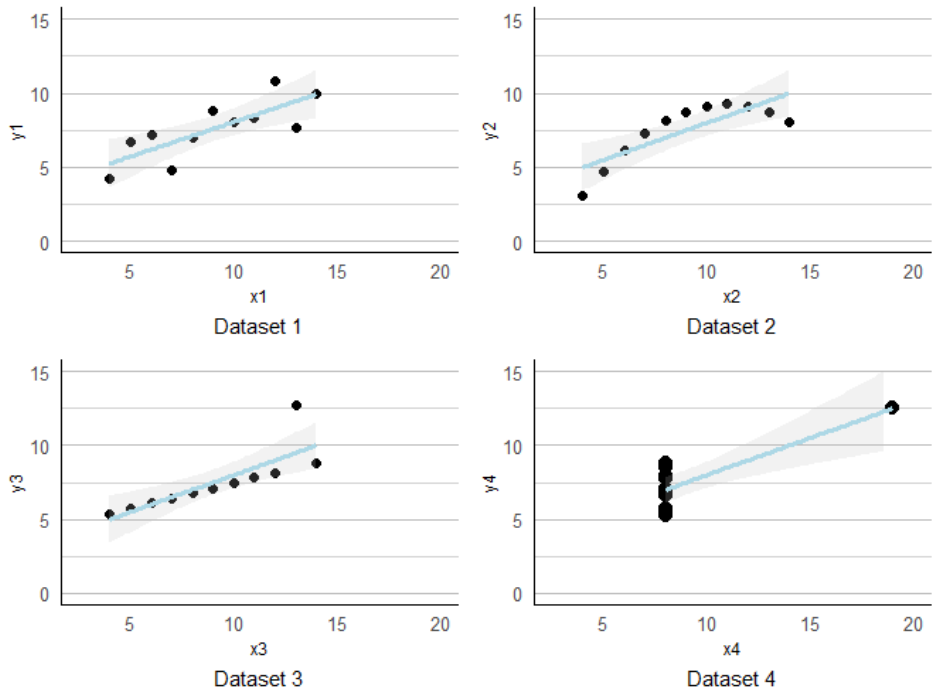
Ο Anscombe προχώρησε στον υπολογισμό ορισμένων βασικών στατιστικών μέτρων (βλ. πίνακα 2) και διαπίστωσε ότι τα 4 σύνολα δεδομένων είχαν:

- Ίδιους μέσους όρους και τυπικές αποκλίσεις για όλα τα X_i και Y_i (i=1,2,3,4)
- Ίδιους συντελεστές συσχέτισης για όλα τα ζευγάρια (X_i,Y_i)
- Ίδιες εξισώσεις γραμμικής παλινδρομήσης
- Ίδιους συντελεστές μεταβλητότητας R^2

Πίνακας 2: *Βασικά στατιστικά μέτρα και εξισώσεις παλινδρόμησης*

	Dataset 1	**Dataset 2**	**Dataset 3**	**Dataset 4**
n	11	11	11	11
Mean (X)	9.0	9.0	9.0	9.0
Mean (Y)	7.5	7.5	7.5	7.5
Standard Deviation (X)	3.31	3.31	3.31	3.31
Standard Deviation (Y)	2.03	2.03	2.03	2.03
Correlation (X, Y)	0.816	0.816	0.816	0.816
Regression Equation	$\hat{y} = 3.0 + 0.5x$	$\hat{y} = 3.0 + 0.5x$	$\hat{y} = 3.0 + 0.5x$	$\hat{y} = 3.0 + 0.5x$
Intercept (b_0)	3.0	3.0	3.0	3.0
Slope (b_1)	0.5	0.5	0.5	0.5
R^2	0.66	0.66	0.66	0.66

Από τα αποτελέσματα του πίνακα 2 κάποιος θα μπορούσε να ισχυριστεί ότι τα 4 σύνολα δεδομένων είναι πανομοιότυπα επειδή παράγουν τα ίδια αριθμητικά αποτελέσματα. Ο Anscombe στη συνέχεια προχώρησε στη γραφική απεικόνιση των 4 συνόλων, δημιουργώντας τα αντίστοιχα γραφήματα διασποράς (scatter plots). Ένα γράφημα διασποράς δείχνει αν και με ποιο τρόπο σχετίζονται 2 μεταβλητές. Στην ουσία φανερώνει το είδος της σχέσης που υπάρχει μεταξύ των μεταβλητών (π.χ γραμμική κλπ) και είναι ιδιαίτερα χρήσιμο στο να αποκαλύπτει ακραίες τιμές.



Graph 1-Visualization of Anscombe's Quartet

Στο σχήμα 1 περιλαμβάνονται τα 4 γραφήματα διασποράς. Σε κάθε νέφος σημείων ο Anscombe προσάρμοσε την αντίστοιχη ευθεία παλινδρόμησης, λαμβάνοντας ως εξαρτημένη μεταβλητή την Y και ως ανεξάρτητη την X. Από την μελέτη των γραφημάτων προέκυψαν τα ακόλουθα συμπεράσματα:

- Στο 1ο γράφημα διασποράς (dataset 1) παρατηρείται ότι η ευθεία παλινδρόμησης προσαρμόζεται πολύ καλά στα δεδομένα (γραμμική σχέση)
- Στο 2ο γράφημα διασποράς (dataset 2) το νέφος των σημείων δημιουργεί καμπύλη. Επομένως το μοντέλο δεν είναι γραμμικό
- Στο 3ο γράφημα διασποράς (dataset 3) η τιμή με συντεταγμένες (13,12.74) επηρεάζει της προσαρμογή της ευθείας παλινδρόμησης στα δεδομένα (ύπαρξη ακραίας τιμής)

- Στο 4ο γράφημα διασποράς (dataset 4) η τιμή με συντεταγμένες (19,12.50) καθορίζει δραστικά την προσαρμογή της ευθείας παλινδρόμησης στα δεδομένα (ύπαρξη σημαίνουσας τιμής).

Το σχήμα 1 καταδεικνύει ότι η οπτικοποίηση των 4 συνόλων δεδομένων αποκάλυψε την πραγματική τους δομή, η οποία παρέμενε «κρυφή» κατά την μαθηματική τους περιγραφή. Συνεπώς, η γραφική απεικόνιση καθίσταται σημαντική, καθώς συμβάλλει στην ανάδειξη της πραγματικής φύσης των δεδομένων και των μεταξύ τους συσχετίσεων. Ωστόσο, σύμφωνα με τον Anscombe, η στατιστική ανάλυση και η οπτικοποίηση θα πρέπει να προσεγγίζονται ως συμπληρωματικά εργαλεία και όχι ως ανεξάρτητες μεθοδολογίες, προκειμένου να διασφαλίζεται η πληρότητα και η εγκυρότητα της ανάλυσης των δεδομένων.

2.2 Συνήθεις πρακτικές οπτικοποίησης δεδομένων

Η οπτικοποίηση των δεδομένων αποτελεί καίριο βήμα για οποιαδήποτε στατιστική ανάλυση. Η συγκεκριμένη διαδικασία αποτελεί την πρώτη ενέργεια κάθε ερευνητή, δεδομένου ότι μια γραφική αναπαράσταση μπορεί να αποκαλύψει πρότυπα, μοτίβα ή ακραίες τιμές που δύσκολα γίνονται αντιληπτές από τη μελέτη των πρωτογενών δεδομένων ή κάποιων στατιστικών μέτρων (μέτρα θέσης ή διασποράς). Η κατασκευή ενός γραφήματος δεν είναι απλή διαδικασία. Θα πρέπει να ακολουθούνται κάποιοι κανόνες οι οποίοι ενισχύουν την αξιοπιστία και τη σαφήνεια μιας γραφικής παράστασης.

Κατ΄ αρχάς, ο κάθε ερευνητής οφείλει να ελέγξει τα δεδομένα που έχει στη διάθεσή του. Στη συνέχεια και ανάλογα με τον τύπο των δεδομένων που διαθέτει (π.χ ποσοτικά, ποιοτικά κλπ) επιλέγει το είδος του γραφήματος που θα κατασκευάσει. Ένας βασικός κανόνας για τη δημιουργία ενός γραφήματος είναι ότι αυτό θα πρέπει να είναι απλό, σαφές και να εξυπηρετεί το λόγο για τον οποίο κατασκευάστηκε [4]. Κάθε γράφημα θα πρέπει να έχει ορισμένα στοιχεία όπως τίτλο, άξονες, ετικέττες, κλίμακες και υπόμνημα. Ιδιαίτερη προσοχή απαιτείται στην επιλογή των κλιμάκων και στη βαθμονόμησή τους. Τα κομβικά - σημαντικά στοιχεία ενός γραφήματος θα πρέπει να επισημαίνονται. Επιπλέον, θα πρέπει να γίνεται αναφορά στην πηγή των δεδομένων. Τέλος, η χρήση ενός επεξηγηματικού κειμένου το οποίο θα πρέπει να είναι περιεκτικό (χρήση συντομογραφιών) και σύντομο μπορεί να είναι ιδιαίτερα επωφελές.

Η χρήση ακατάλληλων κλιμάκων οδηγεί σε παραπλανητικά γραφήματα. Τα τρισδιάστατα (3D) γραφήματα είναι συνήθως πολύπλοκα, προκαλούν «αλλοιώσεις» στο πραγματικό μέγεθος των στοιχείων και ως εκ τούτου μπορούν να οδηγήσουν σε παρερμηνείες. Σε ότι αφορά στο αισθητικό μέρος ενός γραφήματος, καλό είναι να αποφεύγονται τα έντονα χρώματα, οι πολλές γραμμές πλέγματος και τα εκτενή επεξηγηματικά κείμενα [13].

3. ΑΝΑΛΥΤΙΚΗ ΚΑΛΑΘΟΣΦΑΙΡΙΣΗΣ

Ο Bill James [3] - στα τέλη της δεκαετίας του 1970 - υπήρξε ο πρώτος που πρότεινε καινοτόμα στατιστικά μέτρα στο άθλημα του baseball. Εισήγαγε την επιστημονική α-

νάλυση στο συγκεκριμένο παιχνίδι με στόχο την αξιολόγηση ομάδων και παικτών με τρόπο αντικειμενικό (sabermetrics). Η εργασία του James επήρεασε βαθιά τον τρόπο με τον οποίο οι διοικήσεις των ομάδων προχωρούσαν στη στελέχωση των roster τους αλλά και τους προπονητές ως προς την τακτική προσέγγιση του παιχνιδιού. Ο James υπήρξε πρωτοπόρος και άνοιξε το δρόμο δίνοντας έμπνευση και σε άλλους ερευνητές να εφαρμόσουν τις ιδέες του και σε άλλα αθλήματα όπως το ποδόσφαιρο, το μπάσκετ, το βόλεϋ, το χόκεϊ επί πάγου κλπ.

Η αναλυτική καλαθοσφαίρισης (basketball analytics) εντάσσεται στο πεδίο της αναλυτικής αθλημάτων (sports analytics) [15] και γνωρίζει - την τελευταία εικοσαετία - ιδιαίτερη ανάπτυξη κυρίως στις Ηνωμένες Πολιτείες αλλά και στην Ευρώπη [8]. Η εφαρμογή υπολογιστικών μεθόδων επηρέασαν σε σημαντικό βαθμό το μπάσκετ, τόσο στο NBA [6] όσο και στις υπόλοιπες διοργανώσεις της FIBA. Βασικά στοιχεία του παιχνιδιού όπως ο ρυθμός (tempo), η στρατηγική στην επίθεση και στην άμυνα, η επιλογή των παικτών (ως προς συγκεκριμένα χαρακτηριστικά), ο τρόπος προπόνησης κλπ τροποποιήθηκαν εξαιτίας της αναλυτικής καλαθοσφαίρισης [6],[7],[11]. Στη συνέχεια ακολουθεί μια σύντομη παρουσίαση για το πώς η στατιστική και τα μαθηματικά επηρέασαν το παιχνίδι από τη δεκαετία του 1980 έως και σήμερα.

Σημαντική στιγμή στη δεκαετία του 1980 αποτέλεσε η εισαγωγή της γραμμής του τριπόντου (7.24 m) αρχικά στο NBA (1980) [9] και στη συνέχεια στα πρωταθλήματα και τις διεθνείς διοργανώσεις της FIBA (1984). Τα βασικά χαρακτηριστικά του παιχνιδιού στην άλλη πλευρά του Ατλαντικού ήταν ο υψηλός ρυθμός (κυρίως από τις ομάδες της Δύσης), με πίεση στη μπάλα και η γρήγορη μετάβαση στην επίθεση (καταστάσεις αιφνιδιασμού). Η ύπαρξη κυριαρχικών ψηλών (centers) έδινε στους προπονητές και την επιλογή για παιχνίδι μέσα από τη ρακέτα (post up). Στην Ευρώπη ο ρυθμός ήταν πιο χαμηλός (με εξαίρεση το ισπανικό πρωτάθλημα). Παρά την εισαγωγή του τριπόντου, κυριάρχησαν οι επιθέσεις με δίποντα μέσα από 1 εναντίον 1 δράσεις και το post up παιχνίδι (3PTA<10%). Σ΄ ότι αφορά στην αναλυτική καλαθοσφαίρισης - τη συγκεκριμένη δεκαετία - καταγράφονταν βασικά στατιστικά που αφορούσαν τους πόντους (PTS), τις ελεύθερες βολές (FT), τα δίποντα (2PT), τα τρίποντα (3PT), τα συνολικά rebounds (TREB) και τα λάθη (TOV) τόσο για τις ομάδες όσο και για τους παίκτες (boxscores). Σταδιακά προστέθηκαν οι τελικές πάσες (AST) και τα κοψίματα (BLK) ενώ τα rebounds (REB) διαχωρίστηκαν σε αμυντικά (DREB) και επιθετικά (OREB). Στα boxscores παρουσιάζονταν επίσης και τα ποσοστά ευστοχίας (FT%, 2PT%, 3PT%) ομάδων και παικτών. Ως δείκτης αξιολόγησης της αποδοτικότητας ενός παίκτη χρησιμοποιούνταν εκείνη την εποχή το σύστημα Tendex (κυρίως στην Ευρώπη) [5]. Ο δείκτης Tendex δινόταν από τον τύπο (όπου MFG=Missed Field Goals, MFT=Missed Free Throws και MIN=Minutes):

$$\text{Tendex} = \frac{(\text{PTS} + \text{REB} + \text{AST} + \text{STL} + \text{BLK}) - (\text{MFG} + \text{MFT} + \text{TOV})}{\text{MIN}} \quad (1)$$

Η δεκαετία του 1990 υπήρξε η χρυσή εποχή του σουτ μέσης απόστασης (midrange) και του παιχνιδιού μέσα από τη ρακέτα (inside game). Ήταν η περίοδος όπου κυριάρ-

χησαν ορισμένοι χαρισματικοί περιφερειακοί παίκτες με ιδιαίτερη ικανότητα στις 1 εναντίον 1 δράσεις (Jordan, Drexler, Hardaway κλπ). Κοντά στο καλάθι, υπήρξαν ψηλοί παίκτες (Olajuwon, Ewing, Malone, Robinson κλπ) που ήταν από τους καλύτερους όλων των εποχών, με ιδιαίτερη ικανότητα στο παιχνίδι με πλάτη (post up). Στον αμυντικό τομέα, οι ομάδες στο NBA συνέχισαν να στηρίζονται στην ένας εναντίον ενός άμυνα (man to man) με έμφαση στη σωματική επαφή και στη δύναμη (physical defense). Στις διοργανώσεις της FIBA - πέρα από το man to man - χρησιμοποιούνταν από τους προπονητές και οι άμυνες ζώνης (με παραλλαγές). Οι ομάδες - κυρίως στην Ευρώπη - άρχισαν να εκτελούν περισσότερα τρίποντα στα 90΄s (10%<3PT<15%). Στον τομέα της στατιστικής, ο Daryl Morey παρουσίασε ένα μοντέλο πρόβλεψης νίκης μιας ομάδας, που στηριζόταν στους συνολικούς πόντους που αυτή σκόραρε και δεχόταν (Winning percentage model) [9]. Επιπρόσθετα, άρχισαν να παρουσιάζονται σιγά σιγά και ορισμένοι δείκτες με σημαντικότερο το δείκτη assist to turnover ratio (AST/TOV).

Κατά τη δεκαετία του 2000, η σταδιακή αποχώρηση των κυριαρχικών ψηλών που σφράγισαν την προηγούμενη δεκαετία, υποχρέωσε τους προπονητές να αναπροσαρμόσουν τη στρατηγική τους κυρίως στον επιθετικό τομέα. Ως εκ τούτου, γεννήθηκε η pick and roll επιθετική δράση που επί της ουσίας ήταν η συνεργασία μεταξύ δύο παικτών (ενός χειριστή και ενός ψηλού) με στόχο να προκαλέσουν αμυντική ανισοροπία. Τη συγκεκριμένη περίοδο παρατηρήθηκε αύξηση στο σουτ τριών πόντων ως επιθετική επιλογή (15%<3PT<20%). Οι προπονητές άρχισαν να αναζητούν και να στηρίζονται σε περιφερειακούς οι οποίοι είχαν την ικανότητα να «διαβάσουν» την άμυνα (floor generals). Ορόσημο για τη συγκεκριμένη δεκαετία υπήρξε η έκδοση του βιβλίου του Dean Oliver με τίτλο Basketball on paper (2004)[6]. Με τη συγκεκριμένη εργασία ο Oliver εισήγαγε την επιστημονική μέθοδο στο άθλημα της καλαθοσφαίρισης. Η έννοια της κατοχής (possession) έγινε ζωτικής σημασίας και προτάθηκαν μετρικές (ανά κατοχή και ανά λεπτό) με τις οποίες επιτεύχθηκε η συγκρισιμότητα μεταξύ ομάδων και παικτών [6]. Προηγμένες μετρικές όπως τα four factors έγιναν το ευαγγέλιο προπονητών και παικτών προκειμένου να αποτιμάται η αποδοτικότητα (efficiency)[6]. Την περίοδο αυτή ο Hollinger παρουσίασε ένα νέο δείκτη με τον οποίο αξιολογούνταν η απόδοση ενός παίκτη. Το Player Efficiency Rating (PER) είχε το πλεονέκτημα ότι συνδύαζε πολλά πρωτογενή στατιστικά στοιχεία που κατέληγαν σε μια βαθμολογία. Βάσει της βαθμολογίας που προέκυπτε αξιολογούνταν η αποδοτικότητα του παίκτη [7].

Η δεκαετία του 2010 ήταν η περίοδος της συνειδητοποίησης της αξίας της αναλυτικής καλαθοσφαίρισης. Η διαπίστωση ότι ένα τρίποντο ισούται με ενάμισυ δίποντο, η έμφαση στη συχνότητα επισκέψεων στη γραμμή των ελευθέρων βολών και ότι η αποτελεσματικότητα των σουτ κοντά και πάνω από τη στεφάνη υπερβαίνει το 75% προκάλεσαν σημαντικές αλλαγές στο τρόπο παιχνιδιού [6],[9]. Το ποσοστό των επιθέσεων που κατέληγε σε τρίποντο ξεπερνούσε το 30% και σε ορισμένες περιπτώσεις πλησίαζε και το 35%. Η επιλογή το να επιτίθενται οι ομάδες είτε με τρίποντο είτε με δίποντο κοντά στο καλάθι (καρφώματα και layups) περιόρισε σε μεγάλο βαθμό τα σουτ μέσης απόστασης (midrange shots)[11]. Το pick and roll παιχνίδι βελτιώθηκε και οι προπονητές αναζητούσαν ψηλούς με ικανότητα στο μακρινό σουτ (stretch 4's και 5's) για να υποστηρίξουν

pick and pop δράσεις ώστε να «ανοίγει» το γήπεδο. Η ανάπτυξη του SportVU (tracking system) στα γήπεδα του NBA παρείχε στους αναλυτές, πληθώρα δεδομένων σχετικά με την κίνηση των παικτών εντός του γηπέδου κατά τη διάρκεια ενός αγώνα [12]. Η ανάλυση των χωρικών δεδομένων (tracking data) βοήθησε τους προπονητές να αναπτύξουν επιθέσεις με ροή (όπου και οι 5 παίκτες στο παρκέ κινούνταν σαν ένα) με διατήρηση σωστών αποστάσεων (spacing)[12]. Οι σωστές αποστάσεις και η συνεχής κίνηση στην επίθεση ανέπτυξαν τη συνεργασία μεταξύ των παικτών (multi-passing) και υποστήριξαν δράσεις με πολλά κοψίματα (multi-cutting) και πολλά screens. Τα προηγούμενα, αποτέλεσαν τις αρχές πάνω στις οποίες θεμελιώθηκε η επίθεση με συνεχή κίνηση (τόσο των 5 παικτών όσο και της μπάλας) (motion offense). Στο NBA κύριοι εκφραστές αυτής της επίθεσης υπήρξαν οι ομάδες του San Antonio και του Golden State. Σ΄ ότι αφορά στην άμυνα αναπτύχθηκαν τρόποι αντιμετώπισης του pick and roll (hedge out με όλες τις παραλλαγές του), close out άμυνα ενώ άρχισε να χρησιμοποιείται και η άμυνα με αλλαγές στα screens (switch defense). Στον τομέα της αναλυτικής - εκτός από τα προαναφερθεντα - παρουσιάστηκαν στατιστικά που μετρούσαν εξειδικευμένα στοιχεία του παιχνιδιού (π.χ playtype stats, hustle και clutch plays κλπ). Με τα χωρικά δεδομένα (tracking data) δόθηκε η ευχέρεια παραγωγής υψηλής ποιότητας γραφημάτων (π.χ shot charts, heatmaps κλπ) με τα οποία δίνονταν η δυνατότητα σε όλους τους ενδιαφερόμενους για οπτική «ανάλυση» της αποδοτικότητας μιας ομάδας ή παικτών π.χ στο σουτ από συγκεκριμένες θέσεις ή αποστάσεις κλπ.

Στη σημερινή εποχή το σύνολο των ομάδων του NBA έχουν δημιουργήσει τμήματα ανάλυσης δεδομένων στελεχωμένα με εξειδικευμένους επιστήμονες που μπορούν να διαχειριστούν και να αναλύσουν μεγάλα δεδομένα. Η αναλυτική καλαθοσφαίρισης χρησιμοποιείται ως κύριο εργαλείο στην ανίχνευση και ανάλυση του τρόπου παιχνιδιού της αντίπαλης ομάδας (scouting), στη στελέχωση των ομάδων, στη διαχείριση των τραυματισμών και στον τρόπο προπόνησης και εκγύμνασης των παικτών. Η χωρική ανάλυση (spatial analysis) είναι πολύ υψηλού επιπέδου και επικεντρώνεται σε λεπτομέρειες όπως η ανάλυση κίνησης των παικτών και της μπάλας κλπ [12]. Σε σχέση με το παιχνίδι παρατηρείται αύξηση του ρυθμού με έμφαση στην αποτελεσματική επιθετική και αμυντική μετάβαση (transition)[15]. Οι ομάδες επιτίθενται με τρίποντα (3PT>40%) και με δίποντα κοντά στο καλάθι και μέσα από το ζωγραφιστό (ρακέτα). Τα σουτ μέσης απόστασης συνεχίζουν να μην αποτελούν βασική επιλογή, αν και τα τελευταία 1-2 χρόνια υπάρχουν κάποιοι παίκτες (με γρήγορο 1ο βήμα και ικανότητα στο σουτ μετά από ντρίπλα) που επιλέγουν midrange εκτελέσεις. Στην Ευρώπη ως βασικός τρόπος επιθετικής λειτουργίας συνεχίζει να είναι το pick and roll. Ωστόσο αρκετές ομάδες υιοθετούν τις αρχές της motion offense και εντάσσουν δράσεις (plays) αυτής για να επιτεθούν στο αντίπαλο καλάθι (μέσα από spot up εκτελέσεις). Η τάση είναι ότι οι προπονητές προσπαθούν να δημιουργήσουν ομάδες εξίσου αποτελεσματικές τόσο στην άμυνα όσο και στην επίθεση (2-way ομάδες) με παίκτες που να παίζουν σε υψηλή ένταση, που να σουτάρουν καλά τη μπάλα (3PT>35%) και να είναι καλοί αμυντικοί (3 and D παίκτες)[20].

Η αύξηση του ρυθμού (παιχνίδια με περισσότερες κατοχές) που αναφέρθηκε και παραπάνω δημιουργεί νέες συνθήκες στο σύγχρονο μπάσκετ [20]. Ένα μπάσκετ που παί-

ζεται σε υψηλή ένταση, όπου κάθε κατοχή έχει αξία (value possession basketball) και κάθε αδυναμία τιμωρείται από τον αντίπαλο. Παράγοντες όπως η αθλητικότητα,η αντοχή και η ανθεκτικότητα είναι κομβικής σημασίας και θα είναι ακόμη περισσότερο στο μέλλον [20]. Αυτός είναι και ο λόγος που οι προπονητές προσπαθούν να στελεχώσουν τις ομάδες τους με παίκτες ταχυδυναμικούς και κυρίως ανθεκτικούς που να γνωρίζουν πολύ καλά τα βασικά (fundamentals) του αθλήματος σε άμυνα και επίθεση.

4. ΜΕΘΟΔΟΛΟΓΙΚΟ ΠΛΑΙΣΙΟ

4.1 Κύριοι δείκτες απόδοσης (KPI's)

Οι κύριοι δείκτες απόδοσης (KPI's) είναι αριθμοδείκτες που χρησιμοποιούνται για τη μέτρηση της απόδοσης ενός οργανισμού ή μιας διαδικασίας. Έχουν εφαρμογή σε πολλούς τομείς της ανθρώπινης δραστηριότητας, όπως τα χρηματοοικονομικά, η ελεγκτική και φυσικά η αναλυτική αθλημάτων. Στην καλαθοσφαίριση, οι κύριοι δείκτες απόδοσης (KPI's) ταυτίζονται τόσο με τα παραδοσιακές μετρικές όσο και με τα προηγμένες στατιστικές.

Ένας κύριος δείκτης απόδοσης (KPI) θα πρέπει να είναι SMART πράγμα που σημαίνει ότι θα πρέπει να έχει ορισμένα χαρακτηριστικά. Ειδικότερα ένας KPI θα πρέπει να είναι συγκεκριμένος (**S**pecific), να είναι μετρήσιμος (**M**easurable), να είναι ρεαλιστικός (**A**ttainable), να είναι συναφής (**R**elevant) και να είναι χρονικά προσδιορισμένος (**T**ime-based)[17]. Στην εφαρμογή που ακολουθεί παρουσιάζονται γραφήματα με κύριους δείκτες απόδοσης (KPI's) που αφορούν **ομάδες** καλαθοσφαίρισης. Οι μετρικές που περιγράφουν την απόδοση μιας ομάδας - συνήθως - κατανέμονται σε 8 κατηγορίες. Αυτές είναι: (α) τα four factors, (β) η κυκλοφορία της μπάλας (playmaking), (γ) οι κατοχές (possessions), (δ) η αποτελεσματικότητα στο σουτ (shooting efficiency), (ε) η αποδοτικότητα στην επίθεση (offensive efficiency), (στ) η αποδοτικότητα στην άμυνα (defensive efficiency), (ζ) το rebounding και (η) τα διάφορα στατιστικά (miscellaneous - π.χ οι πόντοι στον αιφνιδιασμό, οι πόντοι από λάθη, οι πόντοι από δεύτερες ευκαιρίες και οι πόντοι μέσα από τη ρακέτα).

4.2 Διοργάνωση EuroLeague

Η ευρωλίγκα αποτελεί το κορυφαίο ευρωπαϊκό πρωτάθλημα μπάσκετ σε επίπεδο συλλόγων. Αποτελεί τη μετεξέλιξη του πάλαι ποτέ κυπέλλου πρωταθλητριών (FIBA Champions Cup) που μετανομάστηκε σε FIBA EuroLeague το 1996. Η διαφωνία μεταξύ της FIBA και των διοικήσεων ιστορικών ομάδων έφερε - την περίοδο 2000-2001 - ρήξη που οδήγησε στη δημιουργία δύο παράλληλων διοργανώσεων. Της Suproleague που διοργανώθηκε από τη FIBA και της EuroLeague Basketball που διοργανώθηκε από τις ομάδες που «αποσχίστηκαν» από αυτή.

Την περίοδο 2001-2002 η ευρωλίγκα μετατράπηκε σε μια σχεδόν «κλειστή» διασυλλογική διοργάνωση καλαθοσφαίρισης ανδρών με τη συμμετοχή των μεγαλύτερων ευρωπαϊκών ομάδων. Την ίδια χρονιά η FIBA κατήργησε τη Suproleague. Η περίοδος

2016-2017 υπήρξε κομβικής σημασίας επειδή άλλαξε η δομή (format) της διοργάνωσης. Στη «νέα» ευρωλίγκα συμμετείχαν 16 ομάδες εκ των οποίων οι 11 είχαν κλειστά συμβόλαια. Η διοργάνωση μετατράπηκε σε πρωτάθλημα 30 αγωνιστικών (κανονική περίοδος) όπου κάθε ομάδα αντιμετώπιζε όλες τις υπόλοιπες (εντός και εκτός έδρας). Οι 8 πρώτες ομάδες της κανονικής περίοδου προκρίνονταν στα προημιτελικά (Playoffs) και οι 4 νικητές διεκδικούσαν το τρόπαιο στο Final 4. Την περίοδο 2019-2020 οι ομάδες αυξήθηκαν από 16 σε 18 και οι αγωνιστικές στην κανονική περίοδο έγιναν 34. Δυστυχώς η διοργάνωση δεν ολοκληρώθηκε τη συγκεκριμένη σεζόν λόγω της πανδημίας του κορωνοϊού. Την περίοδο 2021-2022 αποβλήθηκαν από τη διοργάνωση οι ρωσικές ομάδες εξαιτίας της εισβολής της Ρωσίας στην Ουκρανία.

Σήμερα η ευρωλίγκα θεωρείται η σπουδαιότερη διοργάνωση στο μπάσκετ σε επίπεδο συλλόγων παγκοσμίως μετά το NBA. Προσελκύει το ενδιαφέρον των φιλάθλων σε όλο τον κόσμο και οι αγώνες της έχουν υψηλή τηλεθέαση. Ενδεικτικά, αξίζει να αναφερθεί ότι την περίοδο 2023-2024 - κατά μέσο όρο - «κόβονταν» 11.000 εισιτήρια ανά παιχνίδι. Για την ίδια περίοδο υπολογίζεται ότι ο συνολικός κύκλος εργασιών της διοργάνωσης προσέγγισε το ποσό των 320 εκατομμυρίων ευρώ (χωρίς να λαμβάνονται υπόψιν τα budjet των ομάδων).

Η δομή (format) της διοργάνωσης την περίοδο 2023-2024 υπέστη μια μικρή αλλαγή με αποτέλεσμα αντί για 3 να υπάρχουν 4 διακριτές φάσεις:

- Κανονική περίοδος (Regular season) - Κάθε ομάδα από τις συνολικά 18 αντιμετώπισε όλες τις υπόλοιπες (από 2 φορές) σε ένα πρωτάθλημα 34 αγώνων (17 αγώνες εντός έδρας και 17 αγώνες εκτός έδρας)
- Play in φάση - Η 7η στην κατάταξη ομάδα στην κανονική περίοδο αντιμετώπισε σε αγώνα νοκ άουτ τη 10η. Η νικήτρια ομάδα προκρίθηκε στη φάση των Playoffs. Η ηττημένη ομάδα είχε μια 2η ευκαιρία να προκριθεί στα Playoffs αντιμετωπίζοντας το νικητή του ζευγαριού μεταξύ της ομάδας που τερμάτισε 8η στην κανονική περίοδο και αυτής που τερμάτισε 9η
- Φάση των προημιτελικών (Playoffs) - Οι 6 καλύτερες ομάδες της κανονικής περιόδου (1η έως 6η) μαζί με τις 2 ομάδες που προέκυψαν από την Play in φάση σχημάτισαν 4 ζευγάρια. Οι νικητές των 4 ζευγαριών μετά από μια σειρά έως και 5 αγώνων (best of five = 3 νίκες) προκρίθηκαν στην τελική φάση (Final 4)
- Τελική φάση (Final 4) - Οι 4 νικήτριες ομάδες των Playoffs συγκρότησαν τα 2 ζευγάρια των ημιτελικών. Οι νικητές των ημιτελικών αγωνίστηκαν στον μεγάλο τελικό και ο νικητής κατέκτησε το ευρωπαϊκό πρωτάθλημα.

4.3 Μεθοδολογία

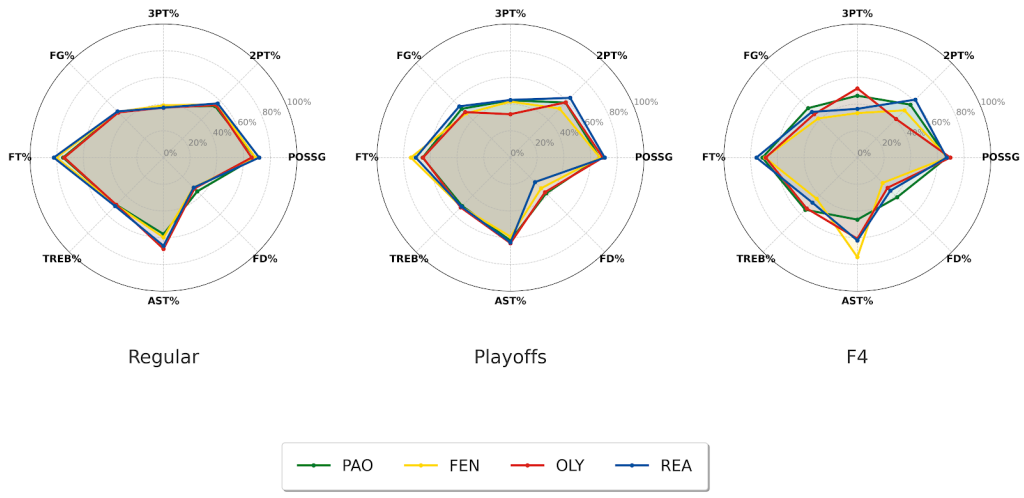
Για τη δημιουργία των γραφημάτων συμπεριλήφθηκαν οι 3 φάσεις (κανονική περίοδος, Playoffs και Final 4) που περιγράφηκαν παραπάνω, δεδομένου ότι τα Play in παιχνίδια (το πολύ 3) αποτελούσαν μια ενδιάμεση φάση προκειμένου να σχηματιστούν τα 4 ζευγάρια των προημιτελικών. Είναι προφανές ότι τα παιχνίδια στις φάσεις των Playoffs και του Final 4 θεωρούνται πολύ υψηλής πίεσης και έντασης για παίκτες και προπονητές.

Τα δεδομένα (data) για την περίοδο 2023-2024 αντλήθηκαν από την επίσημη ιστοσελίδα της Euroleague `https://www.euroleaguebasketball.net` και της ιστοσελίδας Kaggle `https://www.kaggle.com`. Τα γραφήματα κατασκευάστηκαν με τη χρήση της R [10], [19] (και σε δύο περιπτώσεις της Python [16]) και αφορούν κυρίως προηγμένες στατιστικές. Οι ομάδες που προκρίθηκαν στην τελική φάση (Final 4) της διοργάνωσης για την περίοδο 2023-2024 ήταν ο Παναθηναϊκός (PAO), η Ρεάλ Μαδρίτης (REA), ο Ολυμπιακός (OLY) και η Φενερμπαχτσέ (FEN). Για τις ανάγκες της στατιστικής ανάλυσης και της εφαρμογής που αναπτύχθηκε, χρησιμοποιήθηκε η διεθνής ορολογία για τις μετρικές απόδοσης, καθώς και τα προαναφερθέντα αρκτικόλεξα για τους 4 συλλόγους.

5. ΕΦΑΡΜΟΓΗ

5.1 Radial γράφημα

Τα βασικά ή παραδοσιακά στατιστικά (traditional stats) είναι τα πλέον συνήθη που χρησιμοποιούνται για να περιγράψουν έναν αγώνα μπάσκετ. Πρόκειται για τα ποσοστά ευστοχίας στις βολές (FT%), στα δίποντα (2PT%) και στα τρίποντα (3PT%). Τα δίποντα και τα τρίποντα αποτελούν τα σουτ εντός παιδιάς (FG%). Στο ακόλουθο σχήμα 2 παρουσιάζονται τα radial ή radar γραφήματα [16],[18] που αφορούν στις τέσσερις ομάδες PAO, REA, OLY και FEN που έφτασαν στο Final 4 της περιόδου 2023-2024. Κάθε radial γράφημα δίνει τις επιδόσεις κάθε ομάδας σε διαφορετική φάση διεξαγωγής της διοργάνωσης (το 1ο για την κανονική περίοδο, το 2ο για τα Playoffs και το 3ο για το Final 4).



Graph 2-Traditional KPI's for F4 teams 2023-2024

Ένα radial γράφημα προσομοιάζει με ιστό αράχνης στον οποίο υπάρχουν ακτίνες. Κάθε ακτίνα εκφράζει ένα δείκτη απόδοσης. Κατά μήκος της ακτίνας τοποθετούνται τιμές και δημιουργείται ένα πολύγωνο. Το συγκεκριμένο γράφημα (σχήμα 2) είναι χρήσιμο γιατί βοηθά στην οπτική σύγκριση μετρικών - στην περίπτωσή μας - τις επιδόσεις των PAO, REA, OLY και FEN ως προς τις 3 φάσεις της ευρωλίγκας.

Εκτός από τα βασικά στατιστικά στα παραπάνω γραφήματα παρουσιάζονται και ορισμένοι πολύ σημαντικοί δείκτες, όπως ο Total Rebounding Percentage (TREB%), ο Assist Percentage (AST%), ο Fouls Drawn Percentage (FDrawn%) και οι κατοχές ανά παιχνίδι (Possessions per game). Ο δείκτης FDrawn% εκφράζει το ποσοστό των κερδισμένων fouls ανά 100 προσπάθειες εντός παιδιάς. Η έννοια της κατοχής είναι κομβικής σημασίας για το άθλημα της καλαθοσφαίρισης. Ο αριθμός των κατοχών επηρεάζει το ρυθμό (tempo) του παιχνιδιού. Δίνεται από τον τύπο:

$$\text{Possessions} = \text{FGA} - \text{OREB} + \text{TOV} + 0.44 \times \text{FTA} \tag{2}$$

Από τη μελέτη των παραπάνω γραφημάτων προκύπτει ότι παρά το γεγονός ότι η REA υπήρξε η πιο συνεπής επιθετικά ομάδα (τόσο στην κανονική περίοδο όσο και στα Playoffs), ο PAO κατάφερε στο Final 4 να βελτιώσει την απόδοσή του σε κατηγορίες όπως τα συνολικά ριμπάουντ και τα κερδισμένα φάουλ με αποτέλεσμα να κερδίσει τη διοργάνωση.

5.2 Ραβδόγραμμα - Four factors

Οι τέσσερις παράγοντες (four factors) είναι δείκτες που ανέπτυξε ο Dean Oliver με στόχο τη στατιστική ανάλυση βασικών στοιχείων του παιχνιδιού που επηρεάζουν σημαντικά την απόδοση μιας ομάδας. Οι συγκεκριμένοι δείκτες περιγράφουν τόσο το επιθετικό όσο και το αμυντικό παιχνίδι μιας ομάδας. Αποτελούνται στη γενική τους μορφή από:

- Effective Field Goal Percentage (eFG%) που εκφράζει πόσο «εύκολα» μια ομάδα επιτυγχάνει καλάθι. Πρόκειται για ένα δείκτη που προσαρμόζει το ποσοστό ευστοχίας εντός παιδιάς λαμβάνοντας υπόψιν ότι τα τρίποντα έχουν μεγαλύτερη αξία από τα δίποντα. Δίνεται από τον τύπο:

$$\text{eFG\%} = \frac{\text{FGM} + 0.5 \times \text{3PM}}{\text{FGA}} \tag{3}$$

- Free Throw Attempt Rate (FTA Rate) που εκφράζει πόσο «εύκολα» μια ομάδα εκτελεί βολές. Ο δείκτης μετρά την ικανότητα μιας ομάδας να σουτάρει βολές σε σχέση με τον αριθμό των προσπαθειών εντός παιδιάς. Δίνεται από τον τύπο:

$$\text{FTA Rate} = \frac{\text{FTA}}{\text{FGA}} \tag{4}$$

- Rebound Percentage (REB%) που εκφράζει την ικανότητα μιας ομάδας να παίρνει rebounds. Π.χ το ποσοστό των επιθετικών rebounds (OREB%) που κερδίζει μια

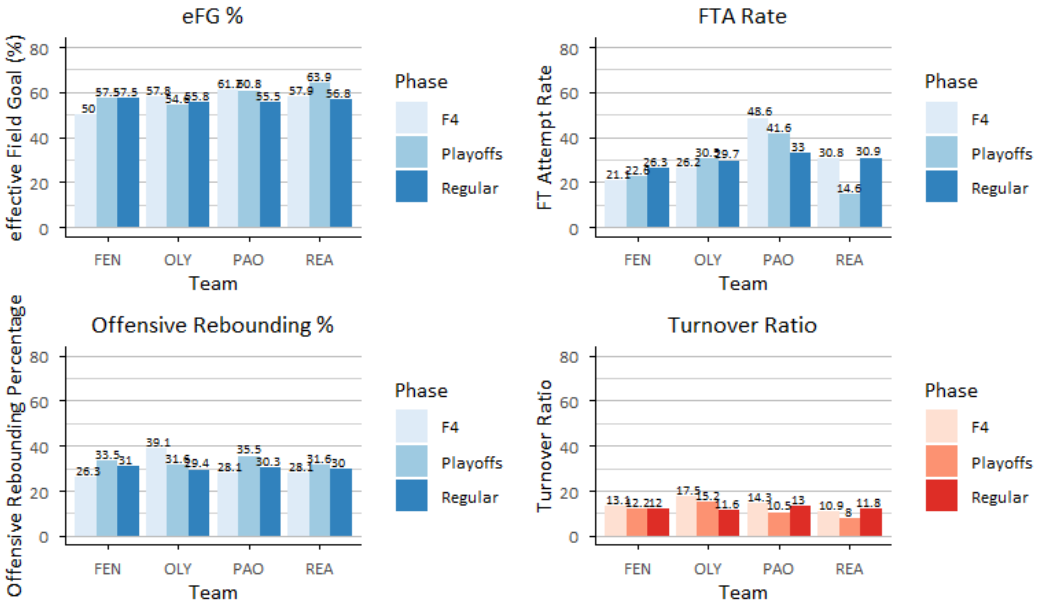
ομάδα δίνεται από τον τύπο:

$$\text{OREB\%} = \left(\frac{\text{OREB}}{\text{OREB} + \text{OppDREB}} \right) \times 100 \tag{5}$$

- Turnover Ratio (TOV%) που εκφράζει το πόσο «εύκολα» χάνει μια ομάδα την κατοχή. Πρόκειται στην ουσία για τον αριθμό των λαθών που πραγματοποιεί μια ομάδα ανά 100 κατοχές. Δίνεται από τον τύπο:

$$\text{TOV Ratio} = \frac{\text{TOV}}{\text{FGA} + (0.44 \times \text{FTA}) + \text{TOV}} \tag{6}$$

Για τους δείκτες eFG%, OREB%, FTA Rate είναι επιθυμητές υψηλές τιμές. Το αντίθετο ισχύει για το δείκτη TOV Ratio. Στα παρακάτω 4 ραβδογράμματα [19] (ένα για κάθε δείκτη - σχήμα 3) παρατηρείται ότι το eFG% του PAO βελτιώθηκε κατά τη διάρκεια των φάσεων, ειδικά στα Playoffs και το Final 4 (61%). Ο δείκτης FTA Rate παρουσίασε σημαντική αύξηση στο Final 4 (48%), γεγονός που δείχνει μεγάλη ικανότητα στο να κερδίζει μια ομάδα fouls. Η REA είχε μεγάλη άνοδο στο eFG% κατά τα Playoffs (64%) , αλλά σημείωσε μείωση στο Final 4 (58%). Τα λάθη της REA αυξήθηκαν σημαντικά στο Final 4 (11%). Ο OLY αύξησε το TOV Ratio σε όλες τις φάσεις, ιδιαίτερα στο Final 4 (17.5%), ενώ το eFG% του παρέμεινε σταθερό καθ' όλη τη διάρκεια της σεζόν. Το eFG% της FEN παρέμεινε σταθερό στα Playoffs, αλλά έπεσε σημαντικά στο final 4 (50%). Ο δείκτης FTA Rate μειώθηκε στα Playoffs (23%) και συνέχισε να μειώνεται και στο Final 4 (21%).



Graph 3-Offense Four Factors by Phase
EuroLeague 2023-2024 - F4 Teams

5.3 Bubble γράφημα

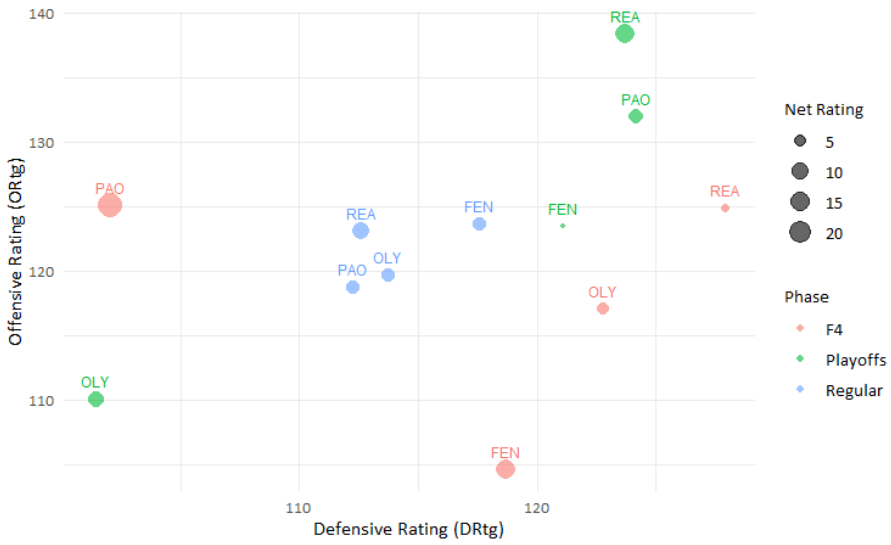
Η αποτελεσματικότητα στην επίθεση αποτυπώνεται από το δείκτη Offensive Rating (ORtg) που εκφράζει τους πόντους που επιτυγχάνει μια ομάδα ανά 100 κατοχές. Δίνεται από τον τύπο:

$$\text{ORtg} = \left(\frac{\text{Points}}{\text{Possessions}}\right) \times 100 \tag{7}$$

Αρκετοί ερευνητές χρησιμοποιούν και το δείκτη Points per Possession (PPP) που εκφράζει τους πόντους που σκοράρει μια ομάδα ανά μία κατοχή. Όσο μεγαλύτερες είναι οι τιμές των 2 δεικτών ORtg και PPP τόσο καλύτερη είναι η επιθετική λειτουργία μιας ομάδας. Αντίστοιχα, η αμυντική λειτουργία μιας ομάδας περιγράφεται από το δείκτη Defensive Rating (DRtg). Ο συγκεκριμένος δείκτης εκφράζει τον αριθμό των πόντων που δέχεται μια ομάδα ανά 100 κατοχές. Είναι φανερό ότι ομάδα αμύνεται καλά όταν ο δείκτης DRtg έχει μικρές τιμές. Δίνεται από τον τύπο:

$$\text{DRtg} = \left(\frac{\text{Points Allowed}}{\text{Possessions}}\right) \times 100 \tag{8}$$

Στο ακόλουθο bubble γράφημα [19] (σχήμα 4) απεικονίζεται γραφικά η επιθετική (ORtg - Κάθετος άξονας) και η αμυντική (DRtg - Οριζόντιος άξονας) λειτουργία των 4 ομάδων που έφτασαν μέχρι το Final 4 για την περίοδο 2023-2024 ως προς τις 3 φάσεις της διοργάνωσης. Στο γράφημα επεισέρχεται και η μεταβλητή Net Rating που αποτυπώνει την καθαρή απόδοση κάθε ομάδας. Όσο μεγαλύτερη είναι η φυσαλίδα (bubble) τόσο μεγαλύτερη είναι και η καθαρή απόδοση της εκάστοτε ομάδας.



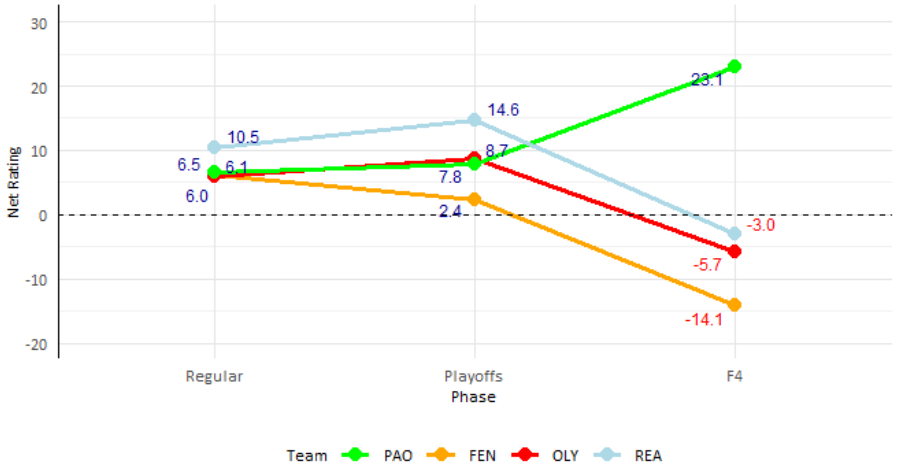
Graph 4-Offensive-Defensive Efficiency by Phase
EuroLeague 2023-2024 - F4 Teams

Ο PAO μείωσε εντυπωσιακά το δείκτη DRtg στο Final 4 (102.0) σε σχέση με την κανονική περίοδο (112.3), πράγμα που σημαίνει ότι βελτιώθηκε η αμυντική του απόδοση. Η REA διατήρησε υψηλό ORtg σε όλες τις φάσεις που δείχνει σταθερά καλή επιθετική λειτουργία. Ωστόσο, παρουσίασε μεγάλη αύξηση στο DRtg στο Final 4 (127.9) υποδεικνύοντας κακή αμυντική επίδοση. Ο OLY παρουσίασε επιθετικά προβλήματα στο Final 4 (ORtg: 117.0) με ταυτόχρονη αύξηση του DRtg που έδειξε ότι η συγκεκριμένη ομάδα απέτυχε να διατηρήσει την απόδοσή της σε υψηλά επίπεδα στα παιχνίδια υψηλής πίεσης. Τέλος, παρατηρώντας τους αντίστοιχους δείκτες για την FEN γίνεται σαφές ότι η συγκεκριμένη ομάδα είχε ένα άσχημο Final 4 τόσο επιθετικά όσο και αμυντικά.

5.4 Line plot - Γράφημα γραμμή

Η καθαρή απόδοση ή Net Rating (NetRtg) είναι μια μετρική που προκύπτει από τη διαφορά μεταξύ του δείκτη Offensive Rating και του δείκτη Defensive Rating. Η συγκεκριμένη μετρική παρέχει μια εικόνα της συνολικής απόδοσης μιας ομάδας λαμβάνοντας υπόψιν την επιθετική και αμυντική της αποτελεσματικότητα. Για την καθαρή απόδοση είναι επιθυμητές θετικές τιμές και ιδανικά όσο το δυνατόν μεγαλύτερες.

Στο γράφημα γραμμή [19] που ακολουθεί (σχήμα 5) παρουσιάζεται η καθαρή απόδοση (κάθετος άξονας) για κάθε ομάδα (γραμμή) ως προς τις 3 φάσεις της διοργάνωσης (οριζόντιος άξονας).



Graph 5-Net Rating Across Phases
EuroLeague 2023-2024-F4 Teams

Στην κανονική περίοδο παρατηρήθηκε ότι όλες οι ομάδες είχαν θετικές τιμές στην καθαρή απόδοση με τη REA να ξεχωρίζει έχοντας NetRtg 10.5. Ο PAO παρουσίασε εντυπωσιακή βελτίωση στο Final 4 με NetRtg 23.1. Αυτό δείχνει ότι τόσο στην επίθεση όσο και στην άμυνα του ήταν ιδιαίτερα αποτελεσματικός, γεγονός που εξηγεί και την

κατάκτηση του τίτλου. Η REA, ο OLY και η FEN αντιμετώπισαν δυσκολίες στο Final 4, με το NetRtg κάθε ομάδας να μειώνεται σημαντικά σε σχέση με τις προηγούμενες φάσεις και να παίρνει αρνητικές τιμές (-3.0, -5.7 και -14.1 αντίστοιχα).

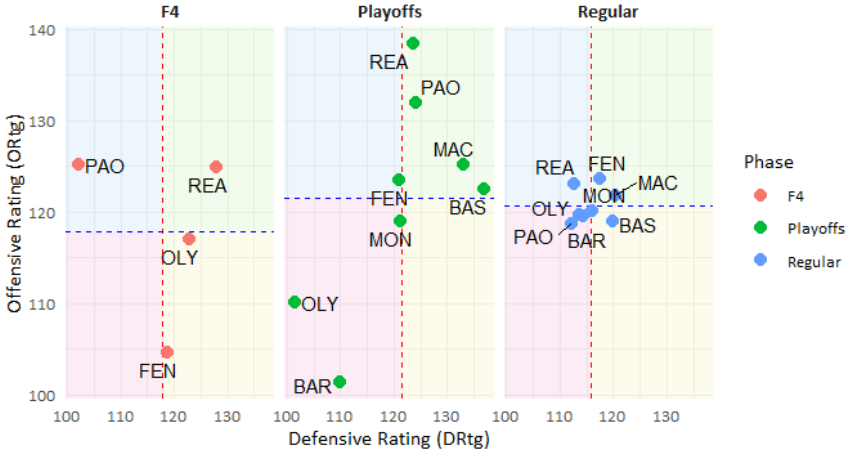
5.5 Γράφημα τεταρτημορίων

Το γράφημα τεταρτημορίων (σχήμα 6) αποτελείται από 3 γραφικές παραστάσεις (μία για κάθε φάση της διοργάνωσης) στις οποίες παρουσιάζονται η επιθετική (ORtg - Κάθετος άξονας) και η αμυντική (DRtg - Οριζόντιος άξονας) απόδοση των ομάδων. Για την κανονική περίοδο και τα Playoffs - πέρα από τις 4 ομάδες του Final 4 - αποτυπώνονται οι παραπάνω δείκτες και για τις 4 ομάδες που έφτασαν μέχρι τη φάση των Playoffs.

Κάθε γράφημα χωρίζεται σε 4 τεταρτημόρια με βάση τις μέσες τιμές των δεικτών ORtg και DRtg της αντίστοιχης φάσης (μπλε και κόκκινη διακεκομμένη γραμμή). Ειδικότερα, όσες ομάδες απεικονίζονται:

- στο 1ο τεταρτημόριο (με πράσινο φόντο) - έχουν καλή επιθετική αλλά κακή αμυντική λειτουργία,
- στο 2ο τεταρτημόριο (με μπλε φόντο) - έχουν καλή επιθετική και καλή αμυντική λειτουργία (2-way teams),
- στο 3ο τεταρτημόριο (με ροζ φόντο) - έχουν κακή επιθετική αλλά καλή αμυντική λειτουργία,
- στο 4ο τεταρτημόριο (με κίτρινο φόντο) - έχουν κακή επιθετική και κακή αμυντική λειτουργία

Το σύγχρονο μπάσκετ επιτάσσει αποτελεσματικότητα τόσο στην επίθεση όσο και στην άμυνα. Επομένως, στόχος των προπονητών είναι να δημιουργούν ομάδες που με την απόδοσή τους να τοποθετούνται στο 2ο τεταρτημόριο (με μπλε φόντο) των παραπάνω γραφημάτων. Με απλά λόγια να είναι 2-way ομάδες που θεωρούνται elite ομάδες.



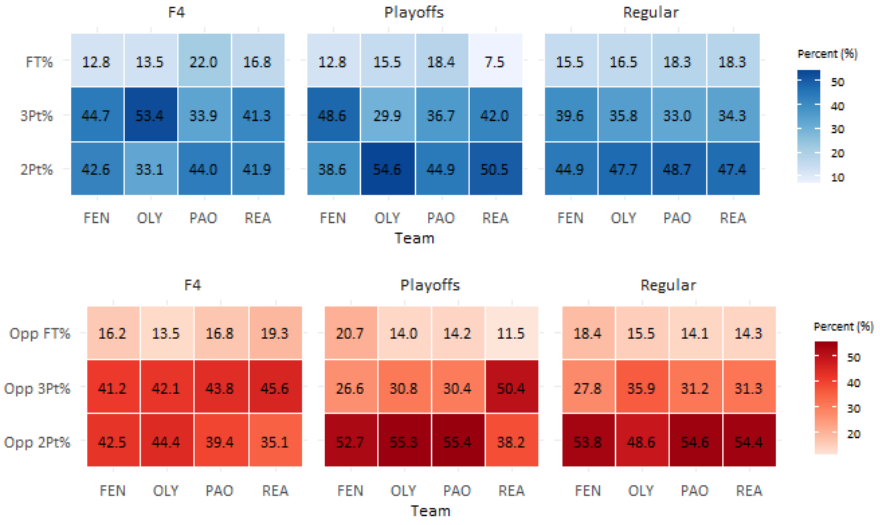
Graph 6-Quadrant Analysis for ORtg and DRtg across Phases
EuroLeague 2023-2024-F4 and Playoff Teams

Στην κανονική περίοδο, η REA ήταν η μοναδική ομάδα με 2-way χαρακτηριστικά (καλή επίθεση και καλή άμυνα). Στο Final 4, ο PAO είχε εντυπωσιακή αμυντική απόδοση και πολύ καλή επιθετική λειτουργία (elite ομάδα). Η REA ήταν πολύ δυνατή επιθετικά, αλλά αμυντικά δεν απέδωσε καλά. Ο OLY αλλά και η FEN είχαν μέτριες έως κακές εμφανίσεις τόσο στην επίθεση όσο και στην άμυνα στο Final 4.

5.6 Heatmap γράφημα

Ένα heatmap γράφημα [19] έχει τη μορφή πίνακα ο οποίος απεικονίζει δεδομένα με τη χρήση χρωμάτων με στόχο την οπτική «αποκάλυψη» μοτίβων ή προτύπων. Συνήθως, τα πιο έντονα (σκούρα) χρώματα υποδεικνύουν μεγαλύτερες τιμές. Σε αντίθεση με τα περισότερο ανοιχτά χρώματα που αποτυπώνουν μικρότερες τιμές.

Στο μπάσκετ οι πόντοι επιτυγχάνονται είτε με καλάθια εντός παιδιάς, δηλαδή με προσπάθειες 2 πόντων (δίποντα) και προσπάθειες 3 πόντων (τρίποντα), είτε με ελεύθερες βολές (σουτ από τη γραμμή των ελευθέρων βολών με κέρδος 1 πόντο ανά εύστοχη βολή). Στο ακόλουθο heatmap γράφημα (σχήμα 7) παρουσιάζεται η ποσοστιαία κατανομή των πόντων σε σχέση με τον τρόπο που επιτεύχθηκαν (FT, 2PT, 3PT) για τις 4 ομάδες (PAO, REA, OLY και FEN) ως προς τις 3 φάσεις της διοργάνωσης (σεζόν 2023-2024).



Graph 7-Total Points (%) distribution relative to FT%, 2Pt% and 3Pt%
EuroLeague 2023-2024-F4 teams

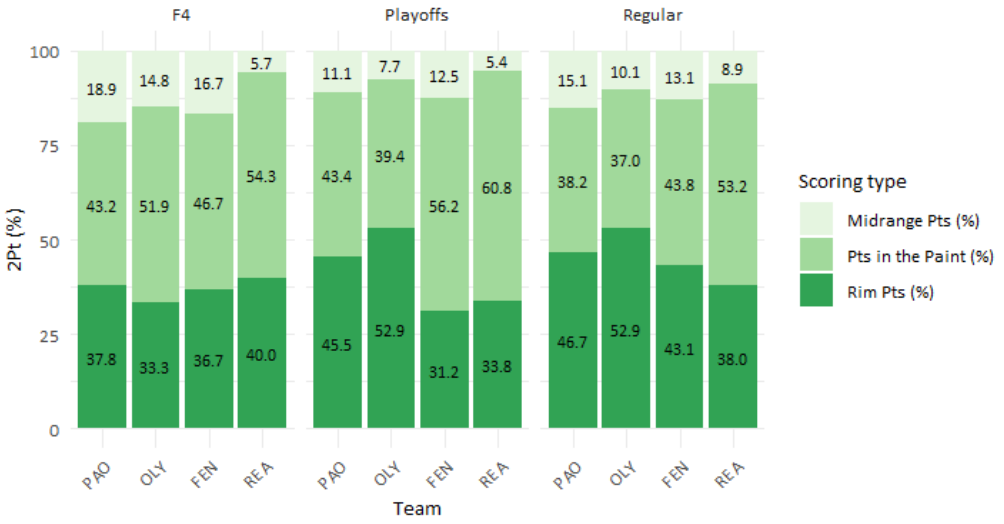
Ο PAO ήταν η ομάδα με το υψηλότερο ποσοστό πόντων από βολές (22.0%) μεταξύ των 4 ομάδων του Final 4. Ωστόσο υπήρξε και η ομάδα με τη χειρότερη περιφερειακή άμυνα στο Final 4 δεδομένου ότι δεχόταν το 43.8% των πόντων της από τρίποντα. Η REA - και στις 3 φάσεις της διοργάνωσης - σημείωνε τους περισσότερους πόντους της με δίποντα (Reg: 47%, Playoffs: 50.5% και F4: 42%). Στην κανονική περίοδο το 35%

με 40% των πόντων προέρχονταν από τρίποντα και για τις 4 ομάδες. Στο Final 4, ο OLY επιτύγχανε πάνω από το 53% των πόντων του με τρίποντα.

5.7 Stacked ραβδόγραμμα

Ένα stacked ραβδόγραμμα [19] είναι ένα γράφημα (σχήμα 8) που χρησιμοποιείται για να αποτυπώσει την ποσοστιαία κατανομή ενός συνόλου δεδομένων σε επιμέρους κατηγορίες. Κάθε ράβδος αθροίζει στο 100% και διαιρείται σε τμήματα (όσες και οι κατηγορίες). Κάθε τμήμα χρωματίζεται διαφορετικά και ανάλογα με την τιμή του (ποσοστό) διαμορφώνεται το μήκος αυτού.

Ενδιαφέρον παρουσιάζει ο τρόπος με τον οποίο οι 4 ομάδες σούταραν δίποντα κατά την περίοδο 2023-2024. Οι προσπάθειες για σουτ 2 πόντων εκδηλώνονται με 3 τρόπους. Με σουτ μέσης απόστασης (midrange), με σουτ μέσα από τη ρακέτα και με σουτ πάνω από τη στεφάνη (π.χ καρφώματα, lay ups).



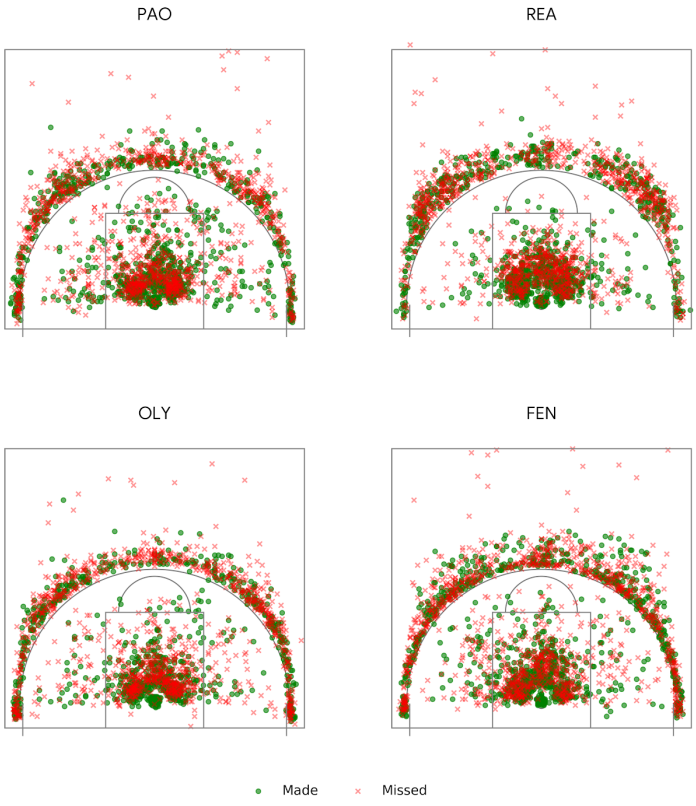
Graph 8-2Pts (%) distribution by scoring type across Phases
EuroLeague 2023-2024-F4 Teams

Ο PAO κατά τη διάρκεια της κανονικής περιόδου εκτέλεσε το 47% των διπόντων του πάνω από τη στεφάνη. Αξιοσημείωτη υπήρξε η επίδοση του στα midrange δίποντα (19%) στο Final 4. Η REA σε όλη τη σεζόν επέλεξε να εκτελεί δίποντα μέσα από τη ρακέτα και πάνω από τη στεφάνη. Ο OLY παρουσίασε μια ισορροπία στις εκτελέσεις διπόντων μέσα από το ζωγραφιστό σε όλη την περίοδο που άγγιξε το 53%.

Το παραπάνω γραφήμα αποτυπώνει καθαρά την τάση που υπάρχει στο σύγχρονο μπάσκετ σε σχέση με τα σουτ μέσης απόστασης (midrange). Τα συγκεκριμένα σουτ (από τα 4 έως 6 μέτρα) δεν αποτελούν πλέον βασική επιλογή στην επίθεση λόγω μικρής αποδοτικότητας σε σύγκριση με τα τρίποντα και τα δίποντα κοντά στο καλάθι.

5.8 Shot chart

Με το shot chart [16],[18] (σχήμα 9) αποτυπώνονται οπτικά οι προσπάθειες (σουτ) που πραγματοποιεί μια ομάδα ή ένας παίκτης σε ένα αγώνα ή σε μια ολόκληρη σεζόν. Ακολουθούν τέσσερα shot charts για τον PAO, τη REA, τον OLY και τη FEN για το σύνολο της περιόδου 2023-2024.



Graph 9-Made and Missed shots for F4 teams

EuroLeague 2023-2024

Είναι φανερό ότι τα σουτ μέσης απόστασης (midrange) δεν αποτέλεσαν κύρια επιθετική επιλογή για το σύνολο των ομάδων και κυρίως για τη REA. Οι εκτελέσεις έξω από τα 6.75 m κυριάρχησαν (σχεδόν το 45% των επιθέσων εκδηλώθηκαν με τρίποντα κυρίως από τις 45° και corner 3's) όπως και τα επιτυχημένα δίποντα μέσα από τη ρακέτα και πάνω από τη στεφάνη (βλ συνδυαστικά και σχήμα 8). Ο OLY υπήρξε η ομάδα με ξεκάθαρη στόχευση στα corner 3's σουτ τριών πόντων. Η REA φαίνεται ότι υπήρξε η πιο επιθετική ομάδα με τις περισσότερες προσπάθειες (πιο «γεμάτο» γράφημα)την περίοδο 2023-2024..

6. ΣΥΜΠΕΡΑΣΜΑΤΑ

Η οπτικοποίηση δεδομένων συνιστά σημαντικό παράγοντα στη διαδικασία της ανάλυσης, καθώς μετασχηματίζει μεγάλο όγκο δεδομένων σε κατανοητή πληροφορία. Μέσω της γραφικής αναπαράστασης, αναδεικνύονται κρίσιμα μοτίβα και τάσεις των δεδομένων και ως εκ τούτου ενισχύεται η αξιοπιστία και η εγκυρότητα των συμπερασμάτων. Ωστόσο, επισημαίνεται ότι η στατιστική επεξεργασία και η οπτικοποίηση των δεδομένων οφείλουν να λειτουργούν συνεργατικά και όχι ως μεμονωμένες μεθοδολογίες.

Η συνεισφορά της στατιστικής επιστήμης στην ανάλυση και κατανόηση του σύγχρονου μπάσκετ είναι αναμφισβήτητη, καθώς προσφέρει μια αντικειμενική βάση για τη βελτιστοποίηση της απόδοσης παικτών και ομάδων σε άμυνα και επίθεση. Ένα σημαντικό εργαλείο προς αυτή την κατεύθυνση είναι η χρήση μοντέρνων μεθόδων οπτικοποίησης δεδομένων και μετρικών. Οι προηγμένες μετρικές και τα σύγχρονα γραφήματα όπως τα heatmaps, τα γραφήματα τεταρτημορίων και τα shot charts ενισχύουν τη γνώση μας γύρω από το άθλημα και παρέχουν τα εφόδια για την καλύτερη και επιστημονικά τεκμηριωμένη προσέγγιση του παιχνιδιού. Ωστόσο, η αναλυτική δεδομένων (δείκτες, μοντέλα, γραφήματα) έχει πολύτιμο μεν, αλλά συμπληρωματικό ρόλο που υποστηρίζει τη διαδικασία λήψης αποφάσεων διοίκησης, προπονητών και παικτών σε ένα περιβάλλον πολύ έντονου ανταγωνισμού.

ABSTRACT

Modern basketball provides an ever-increasing volume of data, the management of which requires advanced analytical tools that transcend the mere presentation of numerical figures. This paper focuses on the value of data visualization, as highlighted by Anscombe in 1973 through his famous ”quartet” of data.

Beyond emphasizing the necessity of data visualization, the primary objective of this study is the application of modern visualization methods for both traditional and advanced metrics, or Key Performance Indicators (KPIs), in basketball. The visualization of these KPIs in the sport of basketball constitutes the empirical part of this work. Utilizing real-world data from the EuroLeague for the 2023-2024 season and employing R and Python, modern visualizations are constructed, such as shot charts, quadrant analysis, and radial plots.The results demonstrate that visualization is not merely a means of presentation but an indispensable analytical tool for understanding team efficiency in both offense and defense within the context of modern basketball.

ΑΝΑΦΟΡΕΣ

[1] Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, **27**(1), 17–21.

[2] Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley.

[3] James, B. (1980). *The Bill James Baseball Abstract*. Ballantine Books.

[4] Tufte, E. R. (1983). *The Visual Display of Quantitative Information*. Graphics Press.

[5] Heeren, D. (1988). *The Basketball Abstract*. Special Press, South Bend, IN.

[6] Oliver, D. (2004). *Basketball on Paper: Rules and Tools for Performance Analysis* (1st ed.), Potomac Books, Va.

[7] Hollinger, J. (2005). *Pro Basketball Forecast 2005-06*. Potomac Books, Inc.

[8] Kubatko, J., Oliver, D., Pelton, K., & Rosenbaum, D. T. (2007). A starting point for analyzing basketball statistics. *Journal of Quantitative Analysis in Sports*, **3**(3), Article 1.

[9] Winston, W. L. (2009). *Mathletics: How gamblers, managers, and sports enthusiasts use mathematics in sports*. Princeton University Press, NJ.

[10] Field, A. (2013). *Discovering statistics using R*. SAGE Publications, UK.

[11] Shea, S. M., & Baker, C. E. (2013). *Basketball Analytics: Objective and Efficient Strategies for Understanding How Teams Win*. Advanced Metrics LLC, Lake St. Louis, MO.

[12] Shea, S. M., (2014). *Basketball Analytics: Spatial Tracking*. CreateSpace, Lake St. Louis, MO.

[13] Wilke, O. C. (2019). *Fundamentals of data visualization*. O'Reilly Media, Sebastopol, CA.

[14] Ware, C. (2020). *Information Visualization: Perception for Design* (4th ed.). Morgan Kaufmann.

[15] Zuccolotto, P., & Manisera, M. (2020). *Basketball data science with applications in R* (1st ed.), CRC Press, Boca Raton, FL.

[16] VanderPlas, J. (2021). *Python Data Science Handbook: Essential Tools for Working with Data* (2nd ed.). O'Reilly Media.

[17] Γεωργίου, Γ. & Μπούσιος, Θ. (2021). *Έλεγχοι Απόδοσης (Performance Audits)*. Εκδόσεις Παπαζήση, Αθήνα.

[18] McKinney, W. (2022). *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and Jupyter* (3rd ed.). O'Reilly Media.

[19] Kobacoff, R. (2024). *Modern data visualization with R*. CRC Press, Boca Raton, FL.

[20] Oliver, D. (2024). *Basketball Beyond Paper: Insights into the Game's Analytics Revolution*. University of Nebraska Press, Lincoln.

Εργασίες στα Αγγλικά

Papers in English

Greek Statistical Institute
Proceedings of the 37th Panhellenic and 3rd Intern. Statistics Conference (2025), pp. 239–256

ON THE NUMERICAL REPRESENTATION OF ORDERED FACTORS USED AS EXPLANATORY VARIABLES: A PROPOSAL

A. Azzalini
Department of Statistical Sciences, University of Padua, Italy
azzalini@stat.unipd.it

ABSTRACT
In a regression model and in its various generalizations, the inclusion of an ordered factor among the explanatory variables requires the conversion of its qualitative levels into one ore more numeric components of the linear predictor. After recalling the general background and the main methods commonly adopted to deal with this situation, we present an alternative approach based on the construction of numeric scores to be assigned to the levels of each factor under consideration. Of this formulation, we first summarize its initially published version and then put forward a more flexible version, based on monotonic splines. The working of the proposal is illustrated by some numerical examples.

Keywords: ordered factors, ordered categorical variables, explanatory variables, linear predictors, numeric scores, quantile functions, monotonic splines.

1. INTRODUCTION

1.1 The problem of interest

The theme of the present contribution is the handling of ordered categorical variables, when they play the role of explanatory variables in the linear predictor of a statistical model, such as a linear regression model, a generalized linear model (GLM) or some other generalization of the linear model; we shall generically refer to them as 'regression-type model'. When a categorical variable is used as an explanatory variable, it is commonly called a 'factor'; hence we shall be working with ordered factors.

For the sake of clarity, we state that the term 'ordered' applies to a categorical variable, hence to a factor too, when there exists a recognized ordering of its levels, either increasing or decreasing. An elementary example is provided by the levels 'very poor', 'poor', ..., 'excellent' of the factor 'product quality'. Data of this type can also arise from the grouping of the values of a continuous variable; for instance, the age of individuals could be grouped as $[0, 5), [5, 15), \ldots [60, 75), [75, \infty)$.

Since working with a regression-type model requires numerical ingredients, the factors under consideration must be transformed to some numeric form, one way or another. Note that this requirement holds equally for non-ordered factors too; in this case, the standard solution involves, for any given factor F with K levels, the introduction of $K-1$ dummy variables entering the linear predictor, unless the intercept term is missing in which case K dummy variables are required. For a detailed account of this topic, see for instance McCullagh and Nelder (1989, § 3.3).

Clearly, the above-recalled methods for non-ordered factors are unsuitable for our problem, since they would disregard a crucial piece of information. There exist methods in the literature specifically targeted to our problem, and they will be recalled shortly.

After this overview of the existing procedures, we move on to our main aim, namely the discussion of an alternative approach to the problem, which we argue is preferable in terms of conceptual clarity and simpler interpretability of the resulting model, while preserving adequacy of data fitting.

1.2 Classical and newer procedures

Although the literature dealing with categorical variables is immense, the portion of it dedicated to the ordinal case is only a fraction. Furthermore, most of the effort produced in the context of regression-type analysis has considered the case where a given ordered categorical variable represents the response variable, while much less attention has been placed on the role as an explanatory variable, i.e. a factor. This unbalance emerges clearly, for instance, in the comprehensive account by Agresti (2010).

Continuing with the notation introduced in Section 1.1, denote by F a categorical variable with levels $F_1, \ldots, F_K$, which without loss of generality we assume to be arranged in increasing order, and suppose we want to include F, after a suitable transformation, as an explanatory variable in a linear model or in some of the many generalizations of the linear model. The long-standing way for undertaking this operation is to associate the levels $F_1, \ldots, F_K$ to a set of numeric scores $x_1, \ldots, x_K$, such that $x_1 < \cdots < x_K$, and then use these values as if they were the actual observations.

In practical work, the traditional choice for the scores $x_1, \ldots, x_K$ is $1, \ldots, K$, or some linear transformation of this sequence. This option has a clear advantage of simplicity, but, on the other hand, there is some arbitrariness in assigning equally spaced numeric values to qualitative levels which may not necessarily be "equally spaced". Consider for instance a factor which represents the condition of a patient after undergoing some surgery, with levels 'ideal', 'minor complications', ..., 'severe complications', 'dead'; here, the spacing between F_{K-1} and F_K cannot convincingly be claimed to be same as the spacing between F_1 and F_2.

Consequently, Armitage (1955) and Graubard and Korn (1987) have, very appropriately, recommended to develop reasonable scores based on a substantive meaning of the levels, if possible. Unfortunately, this ideal course of action can hardly ever be pursued in real situations, and the basic scores (as we shall denote them) $1, \ldots, K$ are commonly

used. In fact, this choice is defended by some authors as a generally reasonable one, on the grounds of the principle that "everything is linear to a first approximation"; see Pasta (2009) and Williams (2020).

Alternatives to the basic scores $1, \ldots, K$ have been considered. A noteworthy instance is represented by the ridits, proposed by Bross (1958) and developed by various authors, notably Brockett (1981). The informative summary on various scoring systems provided in Section 2.1 of Agresti (2010) clarifies, among other things, the equivalence between ridits and midranks, another scoring system in use. Ridits have been widely used in applied work, especially in epidemiological studies. The method has, however, also drawn criticism from various authors, such as Mantel (1979) and Graubard and Korn (1987), who actually refer to the midranks, but this makes no difference. The detailed illustration in Agresti (2010, Example 2.1.2) highlights the inadequacies of a numerical outcome based ridits. Consequently, we shall not examine further this direction and return to out main path.

Even if one adopts equally spaced scores $1, \ldots, K$, the linearity of the effect of factor levels can be tested by decomposing the total variability into linear and residual effect, as in Armitage (1955). Moving further in this direction, one can introduce more than one numeric variable, so that the total variability due to F is decomposed into components associated to linear, quadratic and other polynomial terms, with components up to a maximal degree $K-1$. An early instance of this logic is provided by Winer (1962, pp. 70–77), in a numerical illustration involving the effect of a six-level factor representing 'complexity' of a certain visual display, whose effect is decomposed into linear, quadratic, cubic and residual components. Note that, in this example, the linear predictor of the model includes three components associated to F, with three corresponding columns of the design matrix.

A systematic formulation of this scheme has been presented by Chambers and Hastie (1993, p. 32) where is prescribed that 'ordered factors are coded so that individual coefficients represent orthogonal polynomials if the levels of the factor were actually equally spaced numeric values'. Qualitatively, we can draw a parallel between this procedure and the use of orthogonal polynomials for approximating an arbitrary function of a continuous explanatory variable.

The scheme just described is currently standard practice. In the R computing environment (R Core Team, 2025), ordered factors are handled in this way unless an alternative choice is explicitly made; see the documentation of the command `contr.poly` for details. The implication is that the introduction of factor F in a given model translates into $K-1$ components of the linear predictor corresponding to the linear, quadratic, cubic,... effect of F. In the subsequent analysis of the fitted model, a number of the higher-order polynomial components are often removed, as non-significant. Alternatively, a maximal order less than $K-1$ can be set in advance, provided that we are able to specify such an order.

Alternative approaches to deal with ordered factors exist. One such instance is represented by regularization methods, discussed extensively by Tutz and Gertheiss (2016);

see specifically their Section 3 about the use of ordered categorical variables as explanatory variables. Their main focus is, however, on variables selection and on clustering of categories. Since these targets are somewhat different from ours, we do not dwell on this direction.

As already stated, the current method of choice in statistical practice appears to be the one formalized by Chambers and Hastie (1993, p. 32) and implemented in R as the standard one. Therefore, this approach, later denoted the 'polynomial method' for brevity, will be taken as the reference method in the subsequent discussion of our alternative proposal.

2. A METHOD FOR SELECTING SCORES

2.1 Motivation and initial remarks

The starting point of our approach is the same of early authors, concisely expressed by Armitage (1955) who wrote "a natural procedure is to allot a score x_i to the i-th column ($x_1 < x_2 < \cdots < x_k$), and to perform some sort of regression analysis of p on x" in a problem where p represented the quantity of interest. For reasons already discussed in Section 1.2, the commonly adopted scores are those equally spaced, in spite of known limitations of this option. In fact, we have seen that this choice is still firmly advocated by a number of statisticians, especially applied ones.

We agree with Armitage (1955) and with Graubard and Korn (1987) that it would be best "if possible, develop reasonable column scores based on the substantive meaning of the column categories", but this route is seldom feasible in practice. Hence, we decide to proceed in another direction: in essence, the driving idea is to select scores which work well for fitting the data under consideration. Clearly, this bare idea must be better articulated; this is the target of the rest of the current section.

For expository reasons, we consider initially only a single ordered factor F, but the procedure to be introduced can be applied to several factors simultaneously, when appropriate. Also, again for expository convenience, we initially set out the formulation in the context of a GLM, but the methodology can be applied far more generally. In a GLM, as formulated by McCullagh and Nelder (1989, § 2.2), the connection between the response variable Y and one (or more) explanatory variable(s) is typically expressed in the form

$$\mathbb{E}\{Y\} = g(\eta), \qquad \eta = \beta_0 + \beta_1\, x + \cdots \tag{1}$$

where the link function g is monotonic differentiable, x is an explanatory variable, β_0 and β_1 are parameters of the model, and the dots ($\cdots$) indicate other possible components of the linear predictor η, each component being the product of a parameter and an explanatory variable. The 'typically' inserted above refers to the intercept β_0, which is typically present, and for simplicity of argument we assume this is indeed the case in our development. The remaining essential ingredient of a GLM is a distributional assumption about Y, but for the moment we are not concerned with it.

In the context of our formulation, x in Eq. (1) denotes, for any given observation, the numeric score assigned to the level of the factor F pertaining to that observation. Before tackling the question of the choice of of the scores $x_1, \ldots, x_K$ associated to the levels $F_1 \ldots, F_K$, we note some elementary but crucial facts about the possible choices.

(i) It is implicit in the adoption of the qualification 'ordered factor' that the scores are ordered, at least weakly; otherwise the term 'ordered' would become meaningless. Most of what follows assumes strict inequalities in the ordering.

(ii) A linear transformation of the scores has no effect on the model fitting. If a set of x scores is changed to $a + b\,x$ for any a, b with $b \neq 0$, this changes $\beta_0 + \beta_1\,x$ to $\beta_0 + a\beta_1 + b\,\beta_1\,x$, which only amounts to a change of parameterization of η.

(iii) There is no loss of generality in considering increasing values $x_1 < \cdots < x_K$ since the case of decreasing levels $F_1 \ldots, F_K$ can be dealt with by a negative β_1.

(iv) By the same reason leading to (ii), we see that an ordering of a two-level factors is effectively void; hence, without loss of generality, we assume that $K \geq 3$.

In certain applications, it may be reasonable to think that the ordering among the levels $F_1, \ldots, F_K$ is not necessarily strict, and some levels have in fact an equivalent effect on the response variable. The implication is that a weak ordering of the scores should be considered, such that $x_1 \leq \cdots \leq x_K$. While this variant form of the problem is meaningful, it represents a somewhat special case. In the present contribution, we focus on the more common situation where strict ordering of the levels holds.

2.2 Scores obtained from quantile functions

The broad scheme delineated in the previous subsection must be converted into an operational procedure by further specification. The basic step of our formulation is to think of a candidate set of scores $(x_1, \ldots, x_K)$ as the values taken on by some 'appropriate function' evaluated at the points $U = (1, \ldots, K)$ or at some linear transformation of U, with the condition that the named function is monotonically increasing, so that $x_1 < \cdots < x_K$. This view on the genesis of the scores entails no restriction on their feasible values, apart from the restriction of strict ordering introduced stated in the final paragraph of Subsection 2.1.

In fact, there is no restriction even in considering this 'appropriate function' to be the strictly monotonic quantile function $s(\cdot)$ of a continuous random variable, since any strictly monotonic function over the interval $(0, 1)$ constitutes a valid quantile function and *vice versa*, under the above restriction of strict monotonicity. For our purposes, this quantile function must be evaluated at the set $u = (K + 1)^{-1}U$, and $x = s(u)$ provides a vector of candidate scores. The name 's' reminds us that the function produces scores.

The view just introduced is one adopted in the initial version of the proposed methodology, which was presented by Azzalini (2023). To make the proposal operational, one needs to indicate some criterion for choosing the quantile function $s(\cdot)$. A reasonable strategy is to select a parametric family of probability distributions, hence of quantile functions, with high degree of flexibility, and then choose a suitable member of this

family. Since this step does introduce a restriction on the set of quantile functions under consideration, it is appropriate to consider a parametric family with high degree of flexibility.

In summary, we are now working with a parametric quantile function $s(k;\theta)$ regulated by a (vector) parameter θ. For any given choice of θ, this function maps the factor levels to real numbers, written as

$$s(k;\theta) : F_k \to x_k, \qquad k \in \{1, \ldots, K\}\,, \tag{2}$$

so that the values of the explanatory variable x in Eq. (1) are generated.

Usually, a four-parameter family of distributions is considered sufficiently flexible for most practical problems; the Pearson's four-parameter class of distributions is a notable case in point. However, for our problem, location and scale can be fixed at some preassigned values, such as 0 and 1, say, because of remark (ii) in Subsection 2.1. Hence, in the following, we shall focus on quantile functions with two shape-regulating parameters, and fixed location and scale.

Another consideration is that it is computationally convenient to work with probability distributions whose quantile function has a simple explicit expression. This requirement is naturally satisfied by probability distributions which are generated via a transformation of a normally distributed variable, since their quantiles are obtained by a corresponding transformation of the normal quantiles, which are readily available. Popular instances of this sort are the Johnson's S_U distribution (Johnson, 1949), the Tukey's g-and-h distribution described by Hoaglin (1984), and the sinh–arcsinh distribution (Jones and Pewsey, 2009). All these families feature two shape parameters; other basic features relevant for usage in the present formulation are summarized by Azzalini (2023).

Suppose that we have now adopted a certain family of distributions whose quantile function $s(\cdot;\theta)$ is regulated by a certain parameter θ. The remaining key component of our formulation still to be specified is a criterion for selecting the values of θ best suited to fit the data under consideration.

A natural choice for such criterion is to adopt the same criterion used for fitting the regression-type model. In our reference case of a GLM model, the standard criterion for estimating the vector β of parameters is the (residual) deviance function $Q(\beta)$, say, whose minimum value $Q(\hat\beta)$ identifies the estimates $\hat\beta$. When the scores are not fixed, but regulated by θ, there will be one such deviance value $Q_s(\hat\beta_\theta;\theta)$ for any named θ, with associated estimates $\hat\beta_\theta$; the s subscript in Q_s is a reminder of the role of the $s(k,\theta)$ function. Minimization of $Q_s(\hat\beta_\theta;\theta)$ with respect to θ over its feasible range delivers an optimal $Q_s(\hat\beta_{\hat\theta};\hat\theta)$ with associated $\hat\theta$ and $\hat\beta_{\hat\theta}$. With models of different type, the fitting criterion must be adapted correspondingly. For instance, in the case of a classical linear model, the usual target criterion is the residual sum of squares, or some monotonically increasing transformation of it, such as the residual standard deviation, although the deviance is technically equivalent.

Minimization of $Q(\beta)$ requires numerical methods, almost invariably. Minimization of $Q_s(\hat{\beta}_\theta;\theta)$ involves another layer of complexity and recourse to numerical methods is inevitable. This situation involves numerous computations of the factor scores, namely one set of scores for each choice of θ examined in the search process. This reinforces the above-stated preference for families of distributions whose quantile function is easily computable.

Each of above-recalled parametric families of probability distributions features two parameters for shape regulation. Although this option allows ample flexibility of the quantiles and correspondingly of the factor scores, presumably suitable for most practical situations, problems exist where an additional degree of flexibility is appropriate.

The natural resort to meet this requirement is to work with even more flexible parametric families of distributions. A well-known instance of this type is provided by finite mixtures of normal distributions, just to mention the most common option, but other alternatives exist; see for instance the formulation of Ma and Genton (2004) and more generally the wide stream of literature reviewed by Azzalini (2022). The problem with these options is the numerical computation of the quantiles, which are not available in an explicit form, hence involving an increased computational burden.

2.3 Scores obtained from monotonic splines

Given the closing remarks of the previous subsection, we take a different course of action when a high level of flexibility is required for the factor scores, that is, we now change the type of function s in Eq. (2).

Splines are a fairly natural option to consider, since they represent a relatively simple yet effective way for introducing highly flexible functions. In our setting, we need to work with monotonic functions, but this requirement is readily accomplished by using appropriate types of splines proposed in the literature. In reality, a spline function is typically defined over a continuous domain, while we are mostly concerned with its behaviour at $k \in \{1,\dots,K\}$. However, for simplicity, we shall use the term 'spline' also when we refer to its restriction over a discrete set.

The procedure described next is based on the preliminary scheme presented by Azzalini (2024a). Since location and scale of the scores can be chosen arbitrarily, we choose to set the conditions

$$s(1,\theta) = 1, \qquad s(K,\theta) = K \tag{3}$$

for the spline $s(k;\theta)$.

Next, select the number m, with $1 \le m \le K-2$, of internal knots having abscissae $t_1,\dots,t_m$, and corresponding ordinates $y_1,\dots y_m$, under the conditions

$$1 < t_1 < \cdots < t_m < K, \qquad 1 < y_1 < \cdots < y_m < K \tag{4}$$

but otherwise arbitrary. From the abscissae of the internal knots and the corresponding ordinates, we form the vector of $2\,m$ parameters

$$\theta = (t_1,\dots,t_m,y_1,\dots y_m)\,. \tag{5}$$

Specification of θ, plus the condition in Eq. (3), must be supplied to an algorithm for constructing a monotonic spline which interpolates the points

$$(1, 1),\ (t_1, y_1),\ \ldots,\ (t_m, y_m),\ (K, K)\,.$$

In the numerical work to be described shortly, we have considered the methods proposed by Fritsch and Carlson (1980) and by Hyman (1983) for constructing a monotonic cubic spline. In our experience, there was no appreciable difference between these two procedures. Finally, evaluation of this spline at $k = 1, \ldots, K$ provides the scores $x_1, \ldots, x_K$, as prescribed by Eq. (2).

For any given real problem, we must find an appropriate vector θ, which depends on the data under consideration. This task is completely analogous to the one described in Subsection 2.2, the only change being in the type of scoring function $s(\cdot; \theta)$. Hence, the numerical target is still minimization of $Q_s(\hat{\beta}_\theta; \theta)$ with respect to θ, or a suitable analogue of $Q_s(\hat{\beta}_\theta; \theta)$ if the fitted model is not of GLM type.

From a computational viewpoint, the constraints of Eq. (4) require some care. To perform optimization of $Q_s(\hat{\beta}_\theta; \theta)$, working parameters free of constraints have been adopted in place of θ. The reason is that, for an optimization task, unconstrained parameters are more convenient to work with than constrained parameters. At the end of the optimization stage, the optimized working parameters are mapped back to a vector $\hat{\theta}$ satisfying the inequalities of Eq. (4).

Notice that, while from an operational viewpoint the nature of $s(\cdot; \theta)$ considered here appears different from the one of Subsection 2.2 based on quantile functions, the two schemes share common traits from a formal viewpoint. The spline functions employed here are monotonically increasing functions and the domain set of the t values in Eq. (4) and Eq. (5) can be simply mapped from the $(1, K)$ to the $(0, 1)$ interval. This means that, up to an elementary domain mapping, these $s(\cdot; \theta)$ functions are legitimate quantile functions too.

3. NUMERICAL ILLUSTRATIONS

We now illustrate the working of the proposed methodology with a few numerical examples, starting with a very simple case and then moving to a more elaborate one. The computations have been accomplished with the help of the R package `smof` which implements the proposed methodology (Azzalini, 2024b).

3.1 Malformation data

A dataset with very simple structure, suitable for an introductory illustration, is the one presented in Section 3 of Graubard and Korn (1987), originated from a prospective study of maternal drinking and congenital malformations. Women completed a questionnaire early in their pregnancy concerning alcohol use in the first trimester, expressed as the average number of drinks per day; subsequently, data were recorded on

their pregnancy outcomes. For convenience, Table (1) reproduces the data on congenital sex organ malformations cross-classified by maternal alcohol consumption.

Table 1: *Malformation data: number of events classified by alcohol consumption of the mother*

Malformation	Alcohol consumption				
	0	<1	1–2	3–5	≥ 6
Absent	17066	14464	788	126	37
Present	48	38	5	1	1

The natural course of action is to model the presence of malformation as a binomial random variable whose probability of event depends on alcohol consumption via a logistic regression model. The explanatory variable 'alcohol' is numeric in principle, but it is recorded in a way which requires to treat it as an ordered categorial variable. In the R computing environment, fitting a logistic model with the `ordered` explanatory variable 'alcohol' leads to the fitted model summarized in the top part of Table (2), unless we explicitly require a different action. Contrasts based on orthogonal polynomials up to the maximal feasible order are introduced, that is, order four in this case since there are five levels in factor 'alcohol'.

The summary numerical components of the fitted model so constructed are displayed in the top part of Table (2). Labels 'alcohol.L' to 'alcohol^4' denote the explanatory variables associated to the polynomials components, from linear to quartic. Since the polynomial degree is the maximal possible, the model is saturated, with zero residual deviance.

As there is clear evidence of overfitting, the linear predictor must be reduced. Among the possible strategies to pursue this reduction, we adopt the one of step-down regression, also called backwards regression. This proceeds by sequential elimination of the highest-order non-significant polynomial coefficient and refitting the reduced model, until a significant highest-order coefficient is encountered. In the similar context of classical regression with normally distributed observations, Anderson (1962) has proved an optimality property of this procedure.

After elimination of the non-significant (at any meaningful significance level) fourth-order term, the model is refitted, but the new cubic term is still non-significant and further reduction is necessary; details about the cubic fitting are omitted. Finally, a second-order polynomial with significant quadratic coefficient at 5% level is obtained, and this model is then retained; its summary ingredients are displayed in the bottom part of Table (2).

Next, we compare the above outcome with the one of the methodology proposed here, starting with the variant form based on the quantile function of the g-and-h distribution. The fitting procedure described on Subsection 2.2 leads to parameters $(g = 1.855, h = 0)$ with associated u-level quantiles $(-0.45, -0.30, 0.00, 0.66, 2.70)$. However, to ease comparison with the basic scores $(1, \ldots, 5)$, the raw quantiles are linearly

Table 2: *Malformation data: fitting the probability of malformation using polynomial contrasts*

Fourth-order polynomial

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-5.06	0.30	-16.76	0.000
alcohol.L	1.78	0.72	2.46	0.014
alcohol.Q	0.52	0.66	0.79	0.432
alcohol.C	0.02	0.72	0.02	0.982
alcohol^4	0.39	0.60	0.66	0.512

Residual deviance: 0 on 0 degrees of freedom

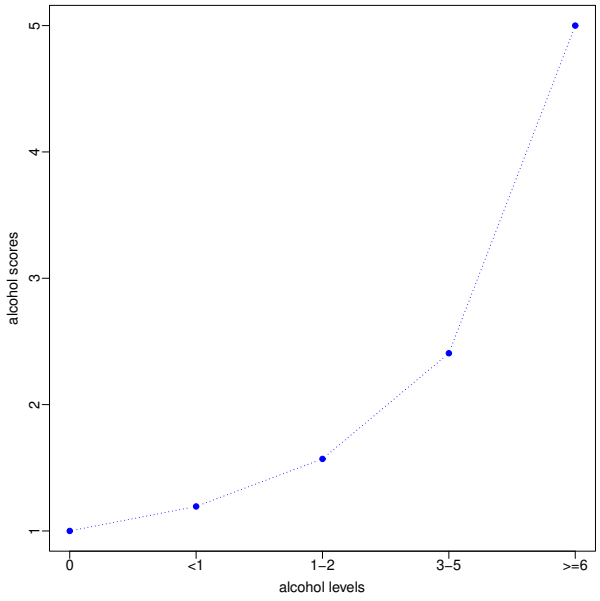
Second-order polynomial

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-5.07	0.28	-18.41	0.000
alcohol.L	1.94	0.65	2.99	0.003
alcohol.Q	0.76	0.36	2.10	0.036

Residual deviance: 0.949 on 2 degrees of freedom

transformed to match the range $(1, 5)$, leading to score values $(1, 1.19, 1.57.2.41, 5)$. Figure (1) displays these numeric scores, effectively comparing them with the basic ones, $(1, \ldots, 5)$, since the original qualitative levels are placed at equally spaced positions on the horizontal axis.

Figure 1: *Malformation data: numeric scores assigned to the levels of factor alcohol*



To any given observation of factor 'alcohol', we associate the pertaining numeric score to the named observation. This process generates a new explanatory variable, denoted 'alcohol.score' in Table (3) which summarizes the fitting of the logistic model which employs the new variable 'alcohol.score' in place of polynomial contrasts. Since the linear predictor features only one component now, apart from the constant, there are 3 degrees of freedom for the residual deviance. This deviance is now somewhat larger than the one in the bottom part of Table (2), but it is still well below its expected value, given the degrees of freedom, hence providing a reassuring indication of the model adequacy.

Table 3: *Malformation data: fitting the probability of malformation using numeric scores*

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-6.55	0.29	-22.79	0.000
alcohol.score	0.61	0.23	2.66	0.008

Residual deviance: 2.03 on 3 degrees of freedom

Some remarks are appropriate. Differently from the approach involving polynomial components, we do not need to go through a model reduction process, like the one leading to the final part of Table (2). The proposed procedure delivers the single explanatory variable 'alcohol.score' to be used directly in the linear predictor, with simplification of the model building and its interpretation. In some complex cases, the model simplification can be more substantial than in the present simple example. Another relevant product of the procedure is a set of scores $(1, 1.19, 1.57, 2.41, 5)$ which intend to quantify the "real" spacings among the qualitative levels $(0, <1, 1\text{–}2, 3\text{–}5, \geq 6)$. This quantification can be of interest also independently of the statistical analysis.

As stated earlier, this numerical work has been based on the quantiles of the g-and-h distribution. There is no compelling reason for this choice; it is simply the default distribution chosen by the R function `smof` which was the key tool for our numerical work. When alternative feasible distributions have been considered, namely the S_U and the sinh-arcsinh distributions, the outcome was very much the same.

3.2 Diamond pricing data

The next example is somewhat more substantial as for number of variables, and it also exhibits a more elaborate data pattern. Various packages within the R environment, such as `ggplot2`, include the dataset 'diamonds' which comprises 10 characteristics of 53,940 round cut diamonds. The variable of interest here is the price of a diamond (expressed in USD), which is connected to some determinants, often collectively denoted as "the 4C's", described as follows:

'carat'	denotes weight (1 carat equals 0.2 grams);
'clarity'	indicates how clear the diamond is,
	with levels I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best);
'color'	of the diamond, from D (best) to J (worst);
'cut'	indicates the quality of the cut,
	with levels Fair, Good, Very Good, Premium, Ideal.

Among these, only 'carat' is of numeric type; all the others are of ordinal type. The purpose of the exercise is to find a simple rule for predicting the monetary value of a diamond from its 4C's variables.

Given the large number of cases in the dataset and the purely illustrative aim of the present demonstration, we examined a small subset of the data, namely those with row number $1, 21, 41, \ldots$, yielding 2697 cases. A smaller subset of the 'diamonds' dataset has been used for similar purposes by Azzalini (2023), selecting one case every one hundred. The target and the pattern of the statistical analysis are largely the same in the two situations, but the present larger number of data points calls for a somewhat more elaborate treatment of the qualitative factors, suitable for illustrating the variant method of Subsection 2.3.

Similarly to Azzalini (2023), a classical linear model has been formulated initially, taking the square-root of the price as the response variable and the 4C's variables as explanatory. First, the factors effects has been coded using polynomial contrasts with maximal degree, given the number of factor levels, namely 7, 6, 4 for clarity, color and cut, respectively. Next, the non-significant polynomial terms have been dropped, following the step-down procedure already employed for the malformation data. Removal of four clear outliers, followed by model re-fitting, leads to the summary values in Table (4). Although the cubic component of 'color' is not significant, it has been retained because the quartic component is, in order to follow a hierarchical building scheme. Inspection of graphical model diagnostics indicates the presence of heavier-than-normal tails, but we decided not to elaborate in this direction, since our purpose is to illustrate the handling of the ordered factors.

There are various indications that the effect of the factor levels, for all three factors under consideration, is fairly complex; one such indication comes from inspection of the estimates in Table (4). Therefore, for the construction of scoring of the three factors, we employ the more versatile version of the proposed methodology based on monotonic splines. Note that, since in this case there are three factors in the model, the scheme described in Subsection 2.3 must be replicated three times in parallel. After some preliminary exploration, we have chosen to the values $(2, 1, 1)$ for the number m of internal knots of the three factors, implying a total of $4{+}2{+}2{=}8$ parameters. For selecting the optimal set of parameters in a linear model context, it is natural to take the residual sum of squares as the target criterion, or equivalently the residual standard deviation.

The outcome of the parameter optimization process leads to the model summarized in Figure (2) and Table (5). The standard deviation reported at the bottom of Table (5) is essentially the same of the matching value in Table (4). The three panels of Figure (2)

Table 4: *Diamond pricing data: linear model of square-root price using polynomial contrasts*

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.25	0.33	0.76	0.450
carat	65.54	0.29	224.07	0.000
clarity.L	25.33	0.75	33.86	0.000
clarity.Q	-12.06	0.70	-17.34	0.000
clarity.C	5.48	0.60	9.12	0.000
clarity^4	-1.65	0.46	-3.59	0.000
clarity^5	1.67	0.38	4.39	0.000
color.L	-13.35	0.42	-32.18	0.000
color.Q	-4.01	0.38	-10.63	0.000
color.C	0.03	0.35	0.07	0.942
color^4	1.35	0.32	4.17	0.000
cut.L	3.51	0.47	7.42	0.000
cut.Q	-1.32	0.40	-3.28	0.001
cut.C	0.91	0.34	2.72	0.007

Residual standard deviation: 6.19 on 2679 degrees of freedom

display how the various levels of the factors are numerically spaced; in other words, they provide an indication about the "real" separation of the levels. The corresponding explanatory variables appearing in Table (5) are associated to a single estimate, instead of the multitude of estimates in Table (4), providing an easier-to-read summary message.

Table 5: *Diamond pricing data: linear model of square-root price using spline-based scores*

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-22.57	0.95	-23.78	0.000
carat	65.58	0.29	227.65	0.000
clarity.score	4.51	0.10	44.22	0.000
color.score	-2.49	0.07	-33.68	0.000
cut.score	1.27	0.16	8.16	0.000

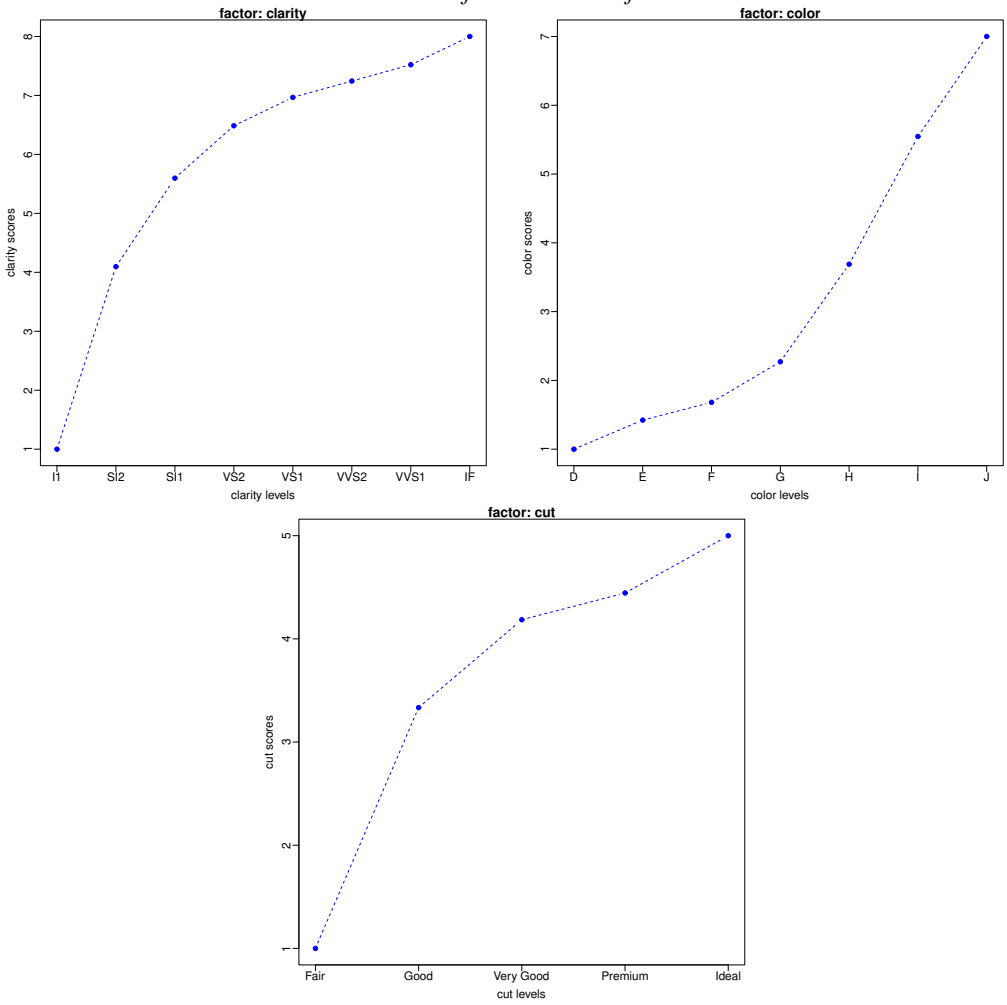
Residual standard deviation: 6.20 on 2688 degrees of freedom

4. OUT-OF-SAMPLE PERFORMANCE

The aim of the present section is to examine some inferential aspects of the proposed methodology, in comparison with the classical one based on polynomial contrasts. The discussion will make use the diamond pricing data of Subsection 3.2 as the reference case; these data are well suited to the purpose, for reasons to be explained shortly.

To compare the performance of the two competing modelling procedures, a simple indicator is represented by the standard deviation of the residual error. This ingredient

Figure 2: *Diamond pricing data: numeric scores assigned to the levels of three ordered factors*



is nearly the same for Table (4) and Table (5), as already pointed out. One can, however, remark that this indication is somewhat simplistic, considering how we have formulated the linear models summarized in those two tables.

Recall that, in general, the validity of the properties of inferential methods associated to a linear model (such as p-values, for instance) are valid under the assumption, among others, that the explanatory variables are determined before observing the data, but this requirement does not hold true for either of the two settings. In the model involving spline-based scores, three of the explanatory variables appearing in Table (5) are constructed from the data themselves, and the standard errors of the associated estimates do not take into account the variability inherent in the construction of the scores. On the other hand, the explanatory variables of Table (4) have been chosen via a selection

process which depends on the data too, hence running into problems similar to those of the other method. Our specific procedure for selecting the polynomial degrees was of step-down type, but the above remark would apply also to other selection strategies.

This sort of issues arise commonly in model building of some complexity, and are not specifically linked to the present setting.

To compare the merit of competing procedures with respect to their predictive performance, the usual procedure is to perform an out-of-sample comparison, hence using different data with the same structure, not the training set used to estimate the parameters and other tuning elements of the models. Recall that, of the full diamond pricing data, we have used only the subset with units numbered $1, 21, 41, \ldots$; hence a similar separate subset can be extracted by selecting, for instance, the units numbered $2, 22, 42, \ldots$; this yields a test set of $N = 2697$ cases.

Denote by z_i the vector of explanatory variables of the ith unit of the test set, for $i = 1, \ldots, N$, and by y_i the corresponding value of the response variable, after square-root transformation of the price. By applying the linear predictor and the parameter estimates of Table (4) to z_i, we obtain the predicted value of y_i, denoted $\hat{y}_i^P$, where the superscript P is a reminder of the polynomial-based approach. A similar operation is performed on the same data using the scoring system represented graphically in Figure (2) and the parameters of Table (5), yielding predicted values $\hat{y}_i^S$, for $i = 1, \ldots, N$, where the superscript S is a reminder of the score-based approach. Finally, we obtain two vectors of prediction errors, e^P and e^S, with components

$$e_i^P = y_i - \hat{y}_i^P \,, \qquad e_i^S = y_i - \hat{y}_i^S \,,$$

for $i = 1, \ldots, N$, on which we base the comparison of the two methods.

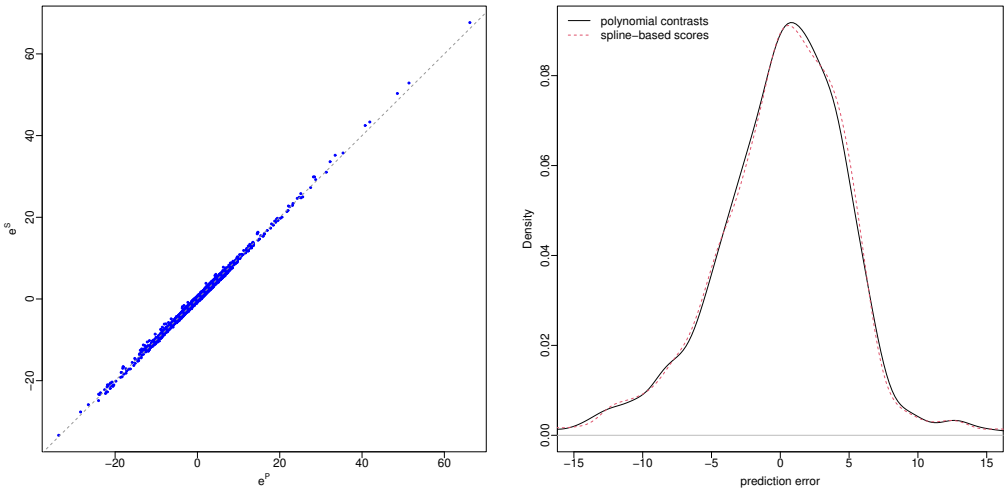
The left panel of Figure (3) displays the scatter plot of the points (e_i^P, e_i^S) with superimposed the identity line. This simple graph yields a clear indication that the two error vectors are essentially coincident. A closer look at the behaviour of prediction errors is offered by the right panel of Figure 3, which displays non-parametric kernel estimates of the density functions of error variable e^P and e^S, represented graphically by the continuous and the dashed curve, respectively. The two curves are barely distinguishable; they follow each other very closely, even in the fine details. This is the reason why the horizontal axis of the right panel had to be substantially restricted with respect to the range on the left panel; otherwise, the two densities would have been visually undistinguishable.

To summarize, there is convincing evidence that, when employed for prediction of future diamond pricing, the two models are essentially equivalent.

5. CONCLUDING REMARKS

The motivation of the proposal is to accomplish the aim expressed by various authors, starting from early ones, of creating a 'reasonable' numeric scores to be attached to the factor levels. This scheme was denoted as 'natural' by Armitage (1955).

Figure 3: *Diamond pricing data: comparing the prediction errors of two competing methods*



The proposed procedure does not attempt to improve the data fitting with respect to the method based on polynomial contrasts. Such an improvement is not even possible, if one considers polynomials of maximal degree, which can achieve perfect fitting with a saturated model. This situation is exemplified in the top component of Table (2) which has zero residual deviance. However, saturated models are not often used in practice.

In the illustrative cases which we have considered here, as well as in other situations not reported here, the fitting quality of the two competing methods turned out to be largely equivalent, and often essentially the same.

The key motivation of the proposed procedure is in the natural logic of the formulation and the ease of interpretation of the outcome. The first stage of the procedure generates a set of of numeric scores for each factor under consideration, such as those represented graphically in Figures (1) and (2). These scores are of interest even outside statistical inference, as a direct indication of the 'real' spacings between successive factor levels. In the second stage of the procedure, these scores allow to generate summary tables of the fitted model, involving a single parameter to express the effect of each factor, instead of the multitude of parameters appearing for instance in Table (4).

One could remark that the entries of, say, Table (4) could be used to produce numerical values with the role like those pictured in Figure (2). While this is technically true, the logical path seems somewhat tortuous and outside the logic of initial formulation; in fact, it is not commonly considered.

Finally, we underline that the proposed procedure can be applied beyond linear and generalized linear models, provided the model of interest features a linear predictor. In fact, the R package `smof` employed in the numerical work of Section 3 can apply the present methodology to proportional hazard models, and others.

6. ACKNOWLEDGEMENTS

I gratefully acknowledge the generous support of the Cacoullos family fund for participating in the 37th Panhellenic and 3rd International Statistics Conference held in Larissa, in April 2025. It was thanks to this participation that eventually the present contribution was produced. Additinal thanks are due to a reviewer for drawing my attention to a few oversights in the original submission.

ΠΕΡΙΛΗΨΗ

Αυτή είναι η περίληψη που ενσωματώνει η Εκδοτική Επιτροπή.

The paper should contain an abstract in Greek which will be prepared by the Editorial Committee.

REFERENCES

Agresti, A. (2010). *Analysis of Ordinal Categorical Data*. J. Wiley and Sons, second edition.

Anderson, T. W. (1962). The choice of the degree of a polynomial regression as a multiple decision problem. *Ann. Math. Statist.*, **33**, 255–265.

Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics*, **11**, 375–386.

Azzalini, A. (2022). An overview on the progeny of the skew-normal family — a personal perspective. *J. Multiv. Anal.*, **188**, 104851. Available online 13 October 2021.

Azzalini, A. (2023). On the use of ordered factors as explanatory variables. *Stat*, **12**, e624.

Azzalini, A. (2024a). On the use of splines for representing ordered factors. `arXiv:2406.15933`.

Azzalini, A. (2024b). *`smof`: Scoring Methodology for Ordered Factors (version 1.2.2)*. `https://cran.r-project.org/package=smof`.

Brockett, P. L. (1981). A note on the numerical assignment of scores to ranked categorical data. *J. Math. Sociology*, **8**, 91–101.

Bross, I. D. T. (1958). How to use ridit analysis. *Biometrics*, **14**, 18–38.

Chambers, J. M. and Hastie, T. J. (1993). Statistical models. In J. M. Chambers and T. J. Hastie (Eds.), *Statistical Models in S* chapter 2. Wadsworth and Brooks/Cole.

Fritsch, F. N. and Carlson, R. E. (1980). Monotone piecewise cubic interpolation. *SIAM J. Numer. Anal.*, **17**, 238–246.

Graubard, B. I. and Korn, E. L. (1987). Choice of column scores for testing independence in ordered $2 \times K$ contingency tables. *Biometrics*, **43**, 471–476.

Hoaglin, D. C. (1984). Summarizing shape numerically: the g-and-h distributions. In D. C. Hoaglin, F. Mosteller, and J. W. Tukey (Eds.), *Exploring Data Tables, Trends, and Shapes* chapter 11. J. Wiley and Sons.

Hyman, J. M. (1983). Accurate monotonicity preserving cubic interpolation. *SIAM J. Sci. Stat. Comput.*, **4**, 645–654.

Johnson, N. L. (1949). Systems of frequency curves generated by methods of translation. *Biometrika*, **36**, 146–176.

Jones, M. C. and Pewsey, A. (2009). Sinh-arcsinh distributions. *Biometrika*, **96**, 761–780.

Ma, Y. and Genton, M. G. (2004). Flexible class of skew-symmetric distributions. *Scand. J. Statist.*, **31**, 459–468.

Mantel, N. (1979). Ridit analysis and related ranking procedures—use at your own risk. *Amer. J. Epidemiol.*, **109**, 25–29.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. London: Chapman and Hall, 2nd edition.

Pasta, D. J. (2009). Learning when to be discrete: continuous vs. categorical predictors. In *SAS Global Forum 2009* (Paper 248–2009).

R Core Team (2025). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Tutz, G. and Gertheiss, J. (2016). Regularized regression for categorical data. *Stat. Modelling*, **16**, 161–200.

Williams, R. A. (2020). Ordinal independent variables. In P. Atkinson, S. Delamont, A. Cernat, J. W. Sakshaug, and R. A. Williams (Eds.), *Linear Regression*. SAGE Research Methods Foundations.

Winer, B. J. (1962). *Statistical Principles in Experimental Design*. McGraw-Hill.

Greek Statistical Institute
Proc. of the 37th Panhellenic & 3rd Intern. Statistics Conference (2025), pp. 257–276

ON THE ESTIMATION OF COMPLEX STATISTICS COMBINING DIFFERENT SURVEYS

Vasilis Chasiotis, Dimitris Karlis
Department of Statistics, Athens University of Economics and Business, Greece
{chasiotisv, karlis}@aueb.gr

ABSTRACT

We employ the alignment method to combine information from two surveys for estimating complex statistics. Deriving alignment weights for complex statistics is challenging due to their non-linear nature. To address this, we propose aligning the linearized variable for the complex statistic of the target variable. This approach also allows us to use alignment weights derived from the linearized variable associated with one complex statistic to estimate other complex statistics. Linearized variables have been used for variance estimation, enabling us to estimate the variance of the combined estimate of complex statistics. Based on the current literature, the variance of a complex statistic estimate is typically computed by using the residuals from the regression of the linearized variable on the auxiliary variables used in the alignment. We propose estimating the variance of a complex statistic using a new linearized variable derived from the alignment weights. These innovations could lead to improvements in computational efficiency and, in some cases, the quality and reliability of the estimators. The proposed approaches are evaluated through a simulation study.

Keywords: Official statistics, variance reduction, alignment weights, multiple surveys, pooling information.

1. INTRODUCTION

The need to investigate the possibility of integrating different surveys to improve efficiency and reduce costs has been widely recognized (Hidiroglou, 2001; Wu, 2004; Ybarra and Lohr, 2008; Merkouris, 2010a; Guandalini and Tillé, 2017). Combining samples from different surveys, or even different samples from the same survey, can result in a more representative sample of the target population, enabling more precise estimation of parameters.

A method of pooling information from different surveys is through alignment, which involves modifying the weights of each survey when two separate surveys have been conducted within the same population and share common variables. The goal is to ensure that both surveys produce equal estimates for these common variables. Zieschang

(1990) introduced a methodology for composite estimation based on an extended regression constraint system. Renssen and Nieuwenbroek (1997) extended the idea of Zieschang (1990) and proposed a class of composite regression estimators. In the two-step method proposed by Renssen and Nieuwenbroek (1997), separate initial estimates are first calculated for the common variables within each sample. These individual estimates are then combined to derive a common estimate, incorporating information from both samples. Each sample then treats the common variable as an auxiliary variable, with the common estimate used as the known population total. Using this approach, new estimates for all variables are calculated. Merkouris (2004) proposed a one-step method that requires all microdata from both samples. In this method, alignment weights are calculated based on a regression (or equivalently, calibration) procedure applied to the combined data. Specifically, the method modifies Zieschang's approach to account for differences in effective sample sizes between the two surveys. This modification retains the practical advantages of Zieschang's method while improving the efficiency of the derived composite regression estimators. Kim and Rao (2012) proposed a model-assisted method to combine data from two independent surveys. In this approach, a large sample from one survey provides only auxiliary information, while a smaller sample from another survey provides information on both the variables of interest and the auxiliary variables. Dong et al. (2014) introduced a new method for combining multiple surveys that adjusts for the complex sampling design features in each survey. Merkouris (2015) developed an efficient estimation method for matrix survey sampling with overlapping subsets of questions, leveraging correlations among items surveyed in various subsamples to improve the precision of survey estimates. More recently, Berger and Kabzińska (2020) developed a design-based empirical likelihood approach for aligning information from multiple surveys and estimating complex parameters defined by estimating equations. Elliott et al. (2018) provided a brief review of methodologies for combining estimates from multiple surveys, and Yang and Kim (2020) offered a comprehensive overview of data integration in survey sampling.

Complex statistics refer to quantitative measures that involve intricate mathematical expressions, such as non-linear relationships among multiple variables. Examples of such statistics include the median, the quintile share ratio S80/S20, the Gini coefficient, the at-risk-of-poverty rate, and the relative median at-risk-of-poverty gap, all of which are useful in social statistics. Due to the non-linearity of complex statistics, their variance can be estimated by applying the method of Deville (1999), which is based on a linearized variable related to the complex statistic under consideration. The variance of the linearized variable under a sampling design provides an approximation of the variance of the complex statistic. There are many references on variance estimation by linearization (Binder and S., 1995; Binder, 1996; Demnati and Rao, 2004, 2010; Vallée and Tillé, 2019). The derivation techniques differ significantly; however, all these methods yield approximately equal results. Additionally, Berger et al. (2009) proposed an estimator for the variance of survey estimates calibrated on estimated auxiliary totals, with an application to the regression composite estimator and the extended regression

estimator.

The alignment method of Merkouris (2004) can, as of now, only be applied using totals. Consequently, its application to complex statistics is not straightforward, as deriving the alignment weights involves solving a cumbersome non-linear optimization problem. This aspect highlights the challenges associated with complex statistics. In this paper, we first propose aligning the linearized variable for the complex statistic of the survey variable of interest, also referred to as the target variable. This approach enables the derivation of alignment weights for complex statistics of the target variable, thereby facilitating the pooling of information from two different surveys. As a result, the proposed method eliminates the need to use the total of the target variable to derive the alignment weights, allowing instead for the use of another complex statistic. Furthermore, the proposed approach makes it possible to use alignment weights derived from the linearized variable associated with one complex statistic to estimate other complex statistics. Additionally, we propose estimating the variance of a complex statistic using a new linearized variable derived from the alignment weights. This approach avoids relying on residuals from the estimated regression of the linearized variable on the auxiliary variables used in the alignment. As a result, it enhances computational efficiency while sometimes improving the quality and reliability of the estimators.

We should note that we introduce an approach for obtaining alignment weights for complex statistics of the target variable, whereas Lesage (2011) focused on using complex statistics of auxiliary variables for calibration, employing the method proposed by Deville and Särndal (1992), which relies on calibration equations. However, the complex statistics discussed in Section 3. are not straightforward to derive using such equations. Fortunately, the linearized variables associated with these statistics are available. Recently, there has been growing attention in the literature on deriving linearized variables for complex statistics, such as the Bonferroni inequality index (Dong et al., 2021). This makes our proposed method particularly attractive. Additionally, the approach of Berger and Kabzińska (2020) could be applied using linearized variables. However, we apply the alignment method of Merkouris (2004) using an adjusted composite regression estimator, which is likely to be more precise than the aligned empirical likelihood estimator proposed by Berger and Kabzińska (2020) when the second-order inclusion probabilities are available.

The remainder of this paper proceeds as follows. Section 2. describes variance estimation for complex statistics using their associated linearized variables, the general regression estimator via calibration, and the alignment method proposed by Merkouris (2004). Section 3. presents the linearized variables for the Laeken indicators from the EU-SILC survey, which are the focus of this paper. In Section 4., we describe our proposed method for estimating complex statistics by pooling information from different surveys through alignment. Section 5. provides simulation results based on synthetic data, while concluding remarks are given in Section 6..

2. Background

In this section, we describe the generalized linearization technique, employed for variance estimation of complex statistics, as proposed by Deville (1999). This technique forms the foundation for our derivations. We also present the derivation of the general regression estimator via calibration for a single sample. Furthermore, we outline the alignment method introduced by Merkouris (2004), which is applied in Section 4. to pool information from two different surveys for estimating complex statistics.

We consider a finite population $U = \{1, \ldots, k, \ldots, N\}$, from which a probability sample S ($S \subseteq U$) of size n is drawn according to a sampling design, $p(\cdot)$, that is $p(S)$ is the probability that S is selected. The first – and second – order inclusion probabilities $\pi_k = \Pr(k \in S)$ and $\pi_{kl} = \Pr(k \in S \,\&\, l \in S)$ are assumed to be strongly positive. Let $\mathbf{w}$ be the $N \times 1$ sampling weight vector with the k-th entry defined as $w_k = (1/\pi_k)\, I(k \in S)$, where I denotes the indicator variable, and let $\mathbf{Y}$ be the $N \times 1$ population vector of a target variable y. The Horvitz-Thompson estimator of the population total of y, $t_y = \mathbf{Y}^{\mathrm{T}}\mathbf{1}$, where $\mathbf{1}$ is the $N \times 1$ unit vector, is given by $\widehat{Y} = \mathbf{Y}^{\mathrm{T}}\mathbf{w}$.

2.1 Variance estimation for complex statistics

Poverty and inequality indicators are examples of complex statistics due to their non-linear nature. These statistics cannot be expressed as standard functions of totals (i.e., functions that are continuously differentiable up to the second order). For instance, the Gini coefficient is a rank statistic, while other indicators are based on quantile statistics. As noted by Osier (2009), their variance cannot be estimated using Taylor linearization. Instead, the generalized linearization method is necessary (Deville, 1999; Demnati and Rao, 2004; Osier, 2009).

In this paper, we employ the generalized linearization technique, proposed by Deville (1999), which is based on the concept of the influence function, to estimate the variance of complex statistics such as the Laeken indicators in the European Statistics on Income and Living Conditions (EU-SILC) survey. In the generalized linearization technique, a complex estimator is approximated by a simpler, linear estimator using the concept of the influence function, which was initially developed in the field of robust statistics (Hampel, 1974). The core idea is to derive a linearized variable that corresponds to the complex statistic under consideration. This approach simplifies the calculation of the estimator's variance, as the variance of the linear approximation can be used to approximate the variance of the non-linear indicator.

The underlying concept of the generalized linearization technique is to examine the impact of unit k on a population parameter of interest θ by introducing an infinitesimal change in the weight assigned to that unit. The parameter θ is represented as a functional $\theta = T(M)$, where M is a measure that assigns a mass of 1, $M(k) = 1$, exclusively at points on the continuum corresponding to units $k \in U$. By specializing the general

measure M into a discrete measure, the functional T, originally defined on a continuum, is converted into a discrete functional, similar to the way t_y is defined. The influence function of T, also referred to as the linearized variable, is defined as

$$z_k = I[T(M)]_k = \lim_{t \to 0} \frac{T(M + t\, d_k) - T(M)}{t},$$

for all $k \in U$, where d_k represents the Dirac measure for the k-th unit, meaning $d_k(i) = 1$ if $i = k$ and 0 otherwise. However, the values of z_k are unknown because they depend on population-level information that is typically unavailable. In practice, since data is only accessible from a sample S, Deville (1999) introduced the concept of an empirical influence function, denoted as $\widehat{z}_k$. This is derived by evaluating the above limit using differential calculus and substituting unknown population quantities with their sample-based estimates. Deville (1999) justified this approach by demonstrating that

$$T(\widehat{M}) - T(M) \approx \sum_{k \in S} z_k w_k - \sum_{k \in U} z_k, \tag{1}$$

where $\widehat{M}$ is the measure assigning a mass w_k to each $k \in S$. The values of z_k can be estimated from the sample using the plug-in estimator $\widehat{z}_k = I[T(\widehat{M})]_k$.

Let $\widehat{\theta}$ be an estimator of θ. A key result, under the asymptotic conditions outlined by Deville (1999), is that when the sample is "sufficiently large", the variance (V) of the estimated total of the linearized variable $\widehat{z}_k$ approximates the variance of the complex statistic $\widehat{\theta}$, that is $V(\sum_{k \in S} z_k w_k) \approx V(\widehat{\theta})$. Consequently, for any sampling design, the variance of a complex statistic $\widehat{\theta}$ can be estimated. The variance of $\hat{\theta}$ can be estimated using the first – and second – order inclusion probabilities. Defining

$$\tilde{\Delta}_{kl} = (\pi_{kl} - \pi_k \pi_l)\, / \pi_{kl}, \tag{2}$$

the variance of $\widehat{\theta}$ can be estimated using the Horvitz-Thompson variance estimator:

$$\widehat{V}(\widehat{\theta}) = \sum_{k \in S} \sum_{l \in S} \tilde{\Delta}_{kl} \frac{\widehat{z}_k \widehat{z}_l}{\pi_k \pi_l}. \tag{3}$$

This approach establishes a framework for deriving alignment weights for complex statistics and approximating the variance of combined estimates of complex statistics, as further elaborated in Section 4..

2.2 The general regression estimator by calibration

In surveys, auxiliary information is commonly used in various ways to improve survey estimates. For instance, the general regression estimator for a finite population total or mean relies on a vector of auxiliary variables with a known population total. A

calibration estimator employs calibrated (or alignment) weights that are as close as possible to the original sampling design weights, according to a specified distance measure, while also satisfying a set of constraints known as the calibration equations.

Let $\mathbf{X}$ be the $N \times 1$ population vector of an auxiliary variable x, whose total $t_x = \mathbf{X}^{\mathrm{T}}\mathbf{1}$ is assumed to be known. Let $\mathbf{\Pi}$ be the $N \times N$ diagonal "weighting" matrix whose k-th diagonal element is w_k/q_k, where q_k is a positive constant. For simplicity, we use the subscript S to denote the subvectors and submatrices corresponding to sample S. The $n \times 1$ alignment weight vector $\boldsymbol{a}$, whose k-th element is denoted by a_k, is obtained minimizing the generalized least squares distance $(\boldsymbol{a} - \mathbf{w}_S)^{\mathrm{T}} \mathbf{\Pi}_S^{-1} (\boldsymbol{a} - \mathbf{w}_S)$ under the constraint $\mathbf{X}_S^{\mathrm{T}}\boldsymbol{a} = t_x$. This procedure generates the alignment weight vector $\boldsymbol{a}$ given by

$$\boldsymbol{a} = \mathbf{w}_S + \mathbf{\Pi}_S\mathbf{X}_S \left(\mathbf{X}_S^{\mathrm{T}}\mathbf{\Pi}_S\mathbf{X}_S\right)^{-1} \left(t_x - \mathbf{X}_S^{\mathrm{T}}\mathbf{w}_S\right). \tag{4}$$

The calibration estimator of the total t_y is obtained as

$$\mathbf{Y}_S^{\mathrm{T}}\boldsymbol{a} = \mathbf{Y}_S^{\mathrm{T}}\mathbf{w}_S + \mathbf{Y}_S^{\mathrm{T}}\mathbf{\Pi}_S\mathbf{X}_S \left(\mathbf{X}_S^{\mathrm{T}}\mathbf{\Pi}_S\mathbf{X}_S\right)^{-1} \left(t_x - \mathbf{X}_S^{\mathrm{T}}\mathbf{w}_S\right),$$

that can take the form of a generalized regression estimator,

$$\widehat{Y^R} = \widehat{Y} + \widehat{\beta}\left(t_x - \widehat{X}\right) = \widehat{\beta}t_x + \left(\mathbf{Y}_S - \mathbf{X}_S\widehat{\beta}\right)^{\mathrm{T}} \mathbf{w}_S, \tag{5}$$

where $\widehat{X} = \mathbf{X}_S^{\mathrm{T}}\mathbf{w}_S$ is the Horvitz-Thomspon estimator of t_x and $\widehat{\beta} = \mathbf{Y}_S^{\mathrm{T}}\mathbf{\Pi}_S\mathbf{X}_S \left(\mathbf{X}_S^{\mathrm{T}}\mathbf{\Pi}_S\mathbf{X}_S\right)^{-1}$ is the sample regression coefficient.

In (5), the vector $\left(\mathbf{Y}_S - \mathbf{X}_S\widehat{\beta}\right)^{\mathrm{T}}$ contains the sample regression residuals, the term $\left(\mathbf{Y}_S - \mathbf{X}_S\widehat{\beta}\right)^{\mathrm{T}} \mathbf{w}_S$ is the sum of weighted sample regression residuals, and $\widehat{\beta}$ minimizes the quadratic form $\left(\mathbf{Y}_S - \mathbf{X}_S\widehat{\beta}\right)^{\mathrm{T}} \mathbf{\Pi}_S \left(\mathbf{Y}_S - \mathbf{X}_S\widehat{\beta}\right)$.

The asymptotic variance (AV) of the generalized regression estimator $\widehat{Y^R}$ is

$$AV(\widehat{Y^R}) = (\mathbf{Y} - \mathbf{X}\beta)^{\mathrm{T}} \tilde{\mathbf{\Pi}} (\mathbf{Y} - \mathbf{X}\beta), \tag{6}$$

where $\mathbf{E} = \mathbf{Y} - \mathbf{X}\beta$ are the residuals and $\tilde{\mathbf{\Pi}}$ is a $N \times N$ matrix whose kl-th element is $(\pi_{kl} - \pi_k\pi_l)/\pi_k\pi_l$, $(\pi_{kk} \equiv \pi_k)$, with β satisfying the normal equation

$$\beta\left(\mathbf{X}^{\mathrm{T}}\mathbf{\Lambda}\mathbf{X}\right) = \mathbf{Y}^{\mathrm{T}}\mathbf{\Lambda}\mathbf{X}, \tag{7}$$

where $\mathbf{\Lambda}$ is a $N \times N$ diagonal matrix whose k-th diagonal element is $1/q_k$. Clearly, β minimizes the weighted least squares expression

$$\mathbf{\Lambda} (\mathbf{Y} - \mathbf{X}\beta)^{\mathrm{T}} (\mathbf{Y} - \mathbf{X}\beta). \tag{8}$$

The residuals $\mathbf{E}$ cannot be used to estimate (6), since β is unknown. An estimator, $\widehat{\beta_a}$, is obtained since the expression in (8) is the unknown population total of the fixed

quantities $\mathbf{\Lambda}\left(\mathbf{Y}-\mathbf{X}\beta\right)^{\mathrm{T}}\left(\mathbf{Y}-\mathbf{X}\beta\right)$. Let $\mathbf{\Pi}^{a}$ be a $N \times N$ diagonal matrix whose k-th diagonal element is a_k/q_k. The calibrated weights estimator of the total in (8) is $\mathbf{\Pi}_S^a\left(\mathbf{Y}_S-\mathbf{X}_S\beta\right)^{\mathrm{T}}\left(\mathbf{Y}_S-\mathbf{X}_S\beta\right)$, which is minimized by $\widehat{\beta_a}$ satisfying the sample-based normal equation $\widehat{\beta_a}\left(\mathbf{X}_S^{\mathrm{T}}\mathbf{\Pi}_S^a\mathbf{X}_S\right)=\mathbf{Y}_S^{\mathrm{T}}\mathbf{\Pi}_S^a\mathbf{X}_S$. Therefore, the sample-based residuals now can be calculated as $\mathbf{e}=\mathbf{Y}_S-\mathbf{X}_S\widehat{\beta_a}$. Let $\mathbf{R}^w$ and $\mathbf{R}^a$ be the $N \times N$ matrices whose k-th diagonal elements are w_k and a_k, respectively, and $\tilde{\mathbf{\Pi}}^{\circ}$ is a $N \times N$ matrix whose kl-th element is equal to $\tilde{\Delta}_{kl}$ given in (2). Deville and Särndal (1992) advocated the variance estimator of $\widehat{Y_R}$ that is given by

$$\widehat{V}(\widehat{Y^R})=\left(\mathbf{Y}_S-\mathbf{X}_S\widehat{\beta_a}\right)^{\mathrm{T}}\mathbf{R}_S^a\tilde{\mathbf{\Pi}}_S^{\circ}\left(\mathbf{Y}_S-\mathbf{X}_S\widehat{\beta_a}\right). \tag{9}$$

The alignment weights a_k in (9) are used to weight the residuals $\mathbf{e}$. From a purely design-based perspective, one could simply use $\mathbf{R}_S^w$ instead of $\mathbf{R}_S^a$, which corresponds to the standard design weights w_k, to weight $\mathbf{e}$, resulting in a design-consistent variance estimator. However, when both model-based and design-based properties are taken into account, there is a compelling reason to favor $\mathbf{R}_S^a$ over $\mathbf{R}_S^w$.

2.3 Pooling information from different surveys

Pooling information from two or more independent surveys is a common challenge in survey sampling. When done appropriately, pooling can reduce costs and enhance the efficiency of sample surveys, making it a topic of significant interest. Pooling information can be derived either from independent surveys targeting the same population or from two independent samples within the same survey. The concept of alignment suggests that if two sample surveys share common variables, it is desirable for both surveys to produce nearly identical estimates for the population totals of these variables. These shared variables, referred to as common variables, are observed in both surveys, but their corresponding population totals are unknown. Examples of such variables often include household characteristics, such as size or composition.

For each survey, it is typically required that known population totals of auxiliary variables be accurately reproduced by weighting-type estimators. Auxiliary variables, such as sex, age, marital status, or region, have known population totals, allowing their estimates from the sample to be adjusted to match these known frequencies. This requirement is always satisfied when auxiliary variables are included as regressors in the general regression estimator. For common variables, population totals are estimated by pooling data from both surveys, and these common variables are simultaneously used as additional regressors. The weights derived from this adjusted regression estimator, known as alignment weights, ensure that the known population totals (or frequencies) of the auxiliary variables are reproduced, while also providing consistent estimates for the population totals of the common variables. These weights are termed reproductive with respect to the auxiliary variables and consistent with respect to the common variables.

In summary, the adjusted regression estimator leverages alignment weights to balance the reproduction of known auxiliary variable totals and the consistent estimation of common variable totals, thereby enhancing the accuracy and efficiency of pooled survey data. The use of alignment weights derived from simpler statistics, such as totals, offers several interesting properties (Zieschang, 1990). These include optimality in the sense that they align with known estimators, improved interpretability, relatively straightforward derivation, and the ability to correct for potential issues within each sample.

In this subsection, the subscript i is used to index vectors and matrices associated with different samples drawn from the same population. Additionally, the subscripts S_1 and S_2 denote the subvectors and submatrices corresponding to samples S_1 and S_2, respectively. We assume that the samples are independent and that the probability of overlapping units between them is negligible. The standard convention of setting $q_k = 1$ for all k is followed throughout.

We consider two independent surveys of the same target population, or two independent subsamples, S_1 and S_2, of the same survey. Recall that $\mathbf{Y}$ is the $N \times 1$ population vector of a target variable y, which is considered common to the two surveys. Let $\mathbf{X}_1$ and $\mathbf{X}_2$ be the $N \times 1$ population vectors of two auxiliary variables x_1 and x_2, respectively, which may or may not be common. The totals of x_1 and x_2, that is $t_{x_1} = \mathbf{X}_1^{\mathrm{T}}\mathbf{1}$ and $t_{x_2} = \mathbf{X}_2^{\mathrm{T}}\mathbf{1}$, respectively, are assumed to be known. Moreover, let $\mathbf{w}_i$ be the $N \times 1$ sampling weight vector based on sample S_i, $i = 1, 2$. Assuming that the probability sample S_i has size n_i, let $\boldsymbol{a}_i$ be the $n_i \times 1$ alignment weight vector based on sample S_i.

Let $\mathbf{w} = \left(\mathbf{w}_1^{\mathrm{T}}, \mathbf{w}_2^{\mathrm{T}}\right)^{\mathrm{T}}$ and $\boldsymbol{a} = \left(\boldsymbol{a}_1^{\mathrm{T}}, \boldsymbol{a}_2^{\mathrm{T}}\right)^{\mathrm{T}}$. In the alignment method proposed by Merkouris (2004), the alignment weight vectors $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ are obtained minimizing the generalized least squares distance $(\boldsymbol{a} - \mathbf{w}_S)^{\mathrm{T}} \boldsymbol{\Pi}_S^{-1} (\boldsymbol{a} - \mathbf{w}_S)$, where $\mathbf{w}_S = \left(\mathbf{w}_{S_1}^{\mathrm{T}}, \mathbf{w}_{S_2}^{\mathrm{T}}\right)^{\mathrm{T}}$ and $\boldsymbol{\Pi}_S = \mathrm{diag}(\boldsymbol{\Pi}_{S_1}, \boldsymbol{\Pi}_{S_2})$, under the constraints $\widehat{Y_1^a} = \mathbf{Y}_{S_1}^{\mathrm{T}}\boldsymbol{a}_1 = \mathbf{Y}_{S_2}^{\mathrm{T}}\boldsymbol{a}_2 = \widehat{Y_2^a}$, $\widehat{X_1^a} = \mathbf{X}_{S_1}^{\mathrm{T}}\boldsymbol{a}_1 = t_{x_1}$ and $\widehat{X_2^a} = \mathbf{X}_{S_2}^{\mathrm{T}}\boldsymbol{a}_2 = t_{x_2}$.

The alignment weight vectors, $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$, are chosen not only to calibrate the auxiliary variables to their known totals but also to ensure that common target variables yield aligned (identical) estimates of the total across samples. Notably, the alignment weights, together with the data from a single sample, contain all available information about the common variables across samples. The advantage of this approach is that it can enhance the estimates of variables that are not common to both samples.

The vector $\boldsymbol{a}$ of the alignment weights is $\boldsymbol{a} = \mathbf{w}_S + \boldsymbol{\Pi}_S\mathbf{N}\left(\mathbf{N}^{\mathrm{T}}\boldsymbol{\Pi}_S\mathbf{N}\right)^{-1}\left(\mathbf{t}_x - \mathbf{N}^{\mathrm{T}}\mathbf{w}_S\right)$, where $\mathbf{N} = \begin{pmatrix} \mathbf{X}_{S_1} & 0 & \mathbf{Y}_{S_1} \\ 0 & \mathbf{X}_{S_2} & -\mathbf{Y}_{S_2} \end{pmatrix}$ and $\mathbf{t}_x = \begin{pmatrix} t_{x_1} & t_{x_2} & 0 \end{pmatrix}^{\mathrm{T}}$.

We consider the independent generalized regression estimators $\widehat{Y_i^R} = \widehat{Y}_i + \widehat{\beta}_i\left(t_{x_i} - \widehat{X_i}\right)$ of the total t_y, where $\widehat{Y}_i = \mathbf{Y}_{S_i}^{\mathrm{T}}\mathbf{w}_{S_i}$ and $\widehat{X_i} = \mathbf{X}_{S_i}^{\mathrm{T}}\mathbf{w}_{S_i}$ are the Horvitz-Thomspon estimators of t_y and t_{x_i} based on the sample S_i, $i = 1, 2$, and $\widehat{\beta}_i = \mathbf{Y}_{S_i}^{\mathrm{T}}\boldsymbol{\Pi}_{S_i}\mathbf{X}_{S_i}\left(\mathbf{X}_{S_i}^{\mathrm{T}}\boldsymbol{\Pi}_{S_i}\mathbf{X}_{S_i}\right)^{-1}$.

Also, let $\widehat{\gamma} = \phi \mathbf{Y}_{S_2}^{\mathrm{T}} \boldsymbol{\Xi}_2 \mathbf{Y}_{S_2} \left[(1-\phi)\, \mathbf{Y}_{S_1}^{\mathrm{T}} \boldsymbol{\Xi}_1 \mathbf{Y}_{S_1} + \phi \mathbf{Y}_{S_2}^{\mathrm{T}} \boldsymbol{\Xi}_2 \mathbf{Y}_{S_2}\right]^{-1}$, where ϕ is chosen as $\phi = (n_1/d_1)/[(n_1/d_1) + (n_2/d_2)]$ and $\boldsymbol{\Xi}_i = \boldsymbol{\Pi}_{S_i} \left[\mathbf{I} - \mathbf{X}_{S_i} \left(\mathbf{X}_{S_i}^{\mathrm{T}} \boldsymbol{\Pi}_{S_i} \mathbf{X}_{S_i}\right)^{-1} \mathbf{X}_{S_i}^{\mathrm{T}} \boldsymbol{\Pi}_{S_i}\right]$. Here, d_i represents the design effect associated with the sample S_i, $i = 1, 2$.

Therefore, we obtain the estimators $\widehat{Y_1^a} = \mathbf{Y}_{S_1}^{\mathrm{T}} \boldsymbol{a}_1 = \widehat{Y_1^R} + (1-\widehat{\gamma}) \left(\widehat{Y_2^R} - \widehat{Y_1^R}\right)$ and $\widehat{Y_2^a} = \mathbf{Y}_{S_2}^{\mathrm{T}} \boldsymbol{a}_2 = \widehat{Y_2^R} - \widehat{\gamma} \left(\widehat{Y_2^R} - \widehat{Y_1^R}\right)$, such that $\widehat{Y_1^a} = \widehat{Y_2^a} = \widehat{\gamma} \widehat{Y_1^R} + (1-\widehat{\gamma})\, \widehat{Y_2^R}$.

The asymptotic variance estimator of $\widehat{Y_1^a}$ $\left(\text{and } \widehat{Y_2^a}\right)$ is given by $\widehat{AV}\left(\widehat{Y_1^a}\right) = \boldsymbol{\Psi} \mathbf{R}_S^w \tilde{\boldsymbol{\Pi}}_S^{\circ} \boldsymbol{\Psi}^{\mathrm{T}}$, where $\boldsymbol{\Psi} = \left[\widehat{\gamma} \left(\mathbf{Y}_{S_1}^{\mathrm{T}} - \widehat{\beta}_1 \mathbf{X}_{S_1}^{\mathrm{T}}\right), (1-\widehat{\gamma}) \left(\mathbf{Y}_{S_2}^{\mathrm{T}} - \widehat{\beta}_2 \mathbf{X}_{S_2}^{\mathrm{T}}\right)\right]$.

3. Complex statistics

In this paper, we evaluate our proposed method using the Laeken indicators from the EU-SILC survey as examples of complex statistics. Therefore, in this section, we provide details about some of these indicators and their corresponding linearized variables.

Let y_k be a given characteristic of a target variable y associated with the k-th unit. Recall that the Horvitz-Thompson estimator of the population total of y, $t_y = \mathbf{Y}^{\mathrm{T}} \mathbf{1}$, is given by $\widehat{Y} = \mathbf{Y}^{\mathrm{T}} \mathbf{w}$. Similarly, the estimator of the population size is given by $\widehat{N} = \mathbf{w}^{\mathrm{T}} \mathbf{1}$. It is important to note that one can use the alignment weight vector $\boldsymbol{a}$ instead of the sampling weight vector $\mathbf{w}$.

3.1 Median

The median is regarded as a more suitable measure of location than the mean when analyzing variables of interest, such as income and expenditure, that exhibit highly skewed distributions. This is because the median is less sensitive to outliers compared to the mean. For this reason, the median is widely used in most household wealth surveys, including the Household Finance and Consumption Survey (HFCS) conducted by the European Central Bank across Eurozone countries.

Let the data y_i be sorted in ascending order. Then the estimated ξ-quantile is

$$\widehat{Q_\xi} = \begin{cases} \frac{1}{2}(y_j + y_{j+1}) & \text{if } \sum_{i=1}^{j} w_i = \xi \widehat{N} \\ y_{j+1} & \text{if } \sum_{i=1}^{j} w_i < \xi \widehat{N} < \sum_{i=1}^{j+1} w_i \end{cases}.$$

This is the standard quantile estimator, but it incorporates weights for each observation.

For the median, the linearized variable z_k is estimated from a sample S using its plug-in estimator:

$$\widehat{z_k^m} = -\frac{1}{\widehat{f(Q_{0.5})}} \frac{1}{\widehat{N}} [I(y_k \leq \widehat{Q_{0.5}}) - 0.5], \tag{10}$$

where $\widehat{f(\widehat{Q_{0.5}})}$ is the estimated density at $\widehat{Q_{0.5}}$. This density estimator can be kernel-based (Graf and Tillé, 2014), that is $\widehat{f(x)} = \frac{1}{h\widehat{N}} \sum_{k \in S} w_k K\left(\frac{x-y_k}{h}\right)$, where h is the bandwidth that determines the amount of smoothing applied to the data. It is obtained from the data as $\widehat{h} = \widehat{\sigma}\widehat{N}^{-0.2}$, where $\widehat{\sigma}$ is an estimator of the variance of the variable y. A natural choice for the kernel is the standard normal distribution, i.e., $K(\cdot)$ is the probability density function of the standard normal distribution. More details can be found in Graf and Tillé (2014).

Thus, to estimate the variance of the median, one must first calculate the linearized variable given in (10) and then apply (3) to obtain the variance estimate.

3.2 Quintile share ratio S80/S20

The quintile share ratio S80/S20 (QSR) has become a widely used measure of social inequality and cohesion (Atkinson et al., 2004). The QSR is defined as the ratio of the total income earned by the richest 20% of the population relative to that earned by the poorest 20%. It can be expressed in terms of quantile shares, where a quantile share represents the proportion of total income earned by all units up to a given quantile.

For the QSR, the linearized variable z_k is estimated from a sample S by its plug-in estimator:

$$\begin{aligned} \widehat{z_k^{QSR}} = & \frac{y_k - 0.8\widehat{Q_{0.8}} + (\widehat{Q_{0.8}} - y_k)I(y_k \leq \widehat{Q_{0.8}})}{\widehat{Y_{0.2}}} \\ & - \frac{(\widehat{Y} - \widehat{Y_{0.8}})[0.2\widehat{Q_{0.2}} - (\widehat{Q_{0.2}} - y_k)I(y_k \leq \widehat{Q_{0.2}})]}{(\widehat{Y_{0.2}})^2}. \end{aligned} \tag{11}$$

Thus, to estimate the variance of the QSR, one must first calculate the linearized variable given in (11) and then apply (3) to obtain the variance estimate.

3.3 Gini coefficient

The Gini coefficient is the most widely used inequality index, frequently employed to measure income inequality (Karsu and Morton, 2015). In various applications, the Gini coefficient defines equity as the degree of fairness in the distribution of impacts, recognizing disparities in outcomes across different populations (Kaplan et al., 2014).

The Gini coefficient can be estimated from a sample S by

$$\begin{aligned} \widehat{G} &= \frac{2}{\widehat{N}\widehat{Y}} \sum_{k \in S} w_k \widehat{N_k} y_k - \left(\frac{1}{\widehat{N}\widehat{Y}} \sum_{k \in S} w_k^2 y_k \right) \\ &= \frac{\sum_{k \in S} \sum_{l \in S} w_k w_l |y_k - y_l|}{2\widehat{N}\widehat{Y}}, \end{aligned}$$

where $\widehat{N_k} = \sum_{l \in S} w_l I(y_l \leq y_k)$. In the case of equal weights, the expression for the estimate of the Gini coefficient can be simplified.

For the Gini coefficient, the linearized variable z_k is estimated from a sample S by its plug-in estimator:

$$\widehat{z_k^G} = \frac{1}{\widehat{N}\widehat{Y}}\left[2\widehat{N_k}(y_k - \widehat{\widehat{Y}}) + \widehat{Y} - \widehat{N}y_k - \widehat{G}(\widehat{Y} + y_k\widehat{N})\right], \tag{12}$$

where $\widehat{\widehat{Y}} = \widehat{Y}/\widehat{N_k}$.

Thus, to estimate the variance of the Gini coefficient, one must first calculate the linearized variable given in (12) and then apply (3) to obtain the variance estimate.

3.4 Relative median at-risk-of-poverty gap

The relative median at-risk-of-poverty gap (RMPG) is calculated as the difference between the median equivalized disposable income of individuals below the at-risk-of-poverty threshold (ARPT) and the ARPT itself, expressed as a percentage of the ARPT. The ARPT is defined as 60% of the national median equivalized disposable income. The equivalized disposable income is defined as the total household disposable income divided by the equivalized household size.

The at-risk-of-poverty threshold (ARPT) can be estimated from a sample S by $\widehat{ARPT} = 0.6\widehat{Q_{0.5}}$. The at-risk-of-poverty rate (ARPR) is the proportion of persons with an equivalized disposable income below the ARPT. It takes values in the interval $[0, 1]$, as it is a proportion. Like the Gini coefficient and the QSR, the ARPR is scale-independent. The ARPR can be estimated from a sample S by $\widehat{ARPR} = \sum_{k \in S} w_k I(y_k \leq \widehat{ARPT})/\widehat{N}$ (Fabrizi et al., 2020).

The estimator of the median equivalized disposable income of individuals below the ARPT is denoted as $\widehat{m_p}$. It is estimated in the same way as any other quantile but using only the data points below the ARPT. Therefore, the RMPG can be estimated from a sample S by $\widehat{RMPG} = \left(\widehat{ARPT} - \widehat{m_p}\right)/\widehat{ARPT}$. It is clear that the estimator $\widehat{RMPG}$ depends on both the ARPT and the ARPR, as well as involves the estimator of the median equivalized disposable income of individuals below the ARPT.

For the ARPT, the linearized variable z_k is estimated from a sample S by its plug-in estimator:

$$\widehat{z_k^{ARPT}} = -\frac{0.6}{\widehat{f(\widehat{Q_{0.5}})}}\frac{1}{\widehat{N}}[I(y_k \leq \widehat{Q_{0.5}}) - 0.5].$$

For the ARPR, the linearized variable z_k is estimated from a sample S by its plug-in estimator:

$$\widehat{z_k^{ARPR}} = \frac{1}{\widehat{N}}(I(y_k \leq \widehat{ARPT}) - \widehat{ARPR}) + \widehat{f(\widehat{ARPT})}\widehat{z_k^{ARPT}},$$

and so, for the $\widehat{m_p}$, the linearized variable z_k is estimated from a sample S by its plug-in

estimator:

$$\widehat{z_k^{\widehat{m_p}}} = \frac{1}{\widehat{f(\widehat{m_p})}} \frac{\widehat{z_k^{ARPR}}}{2} - \frac{1}{\widehat{N}}[I(y_k \leq \widehat{m_p}) - F(\widehat{m_p})].$$

Therefore, for the RMPG (Osier, 2009; Graf and Tillé, 2014), the linearized variable z_k is estimated from a sample S by its plug-in estimator:

$$\widehat{z_k^{RMPG}} = \frac{\widehat{m_p}\widehat{z_k^{ARPT}} - \widehat{ARPT}\widehat{z_k^{\widehat{m_p}}}}{\widehat{ARPT}^2}. \tag{13}$$

Thus, to estimate the variance of the RMPG, one must first calculate the linearized variable given in (13) and then apply (3) to obtain the variance estimate.

4. Derivation of alignment weights using linearized variables

In this section, we describe the proposed approach for estimating complex statistics by pooling information from different surveys using the alignment method proposed by Merkouris (2004). Although our approach is implemented for the complex statistics discussed in Section 3., it can be readily applied to other complex statistics.

The main obstacle in applying the alignment method proposed by Merkouris (2004) is the non-linear form of complex statistics, as this method is built upon totals, ensuring that common target variables yield identical estimates of the total across samples. Therefore, we propose using the linearized variable for the complex statistic of the target variable to obtain the alignment weights. The key to implementing this method for the complex statistic lies in the result demonstrated by Deville (1999) in (1). Specifically, we recall that a population parameter of interest $\theta = T(M)$ can be expressed as

$$\widehat{\theta} - \theta \approx \sum_{k \in S} z_k w_k - \sum_{k \in U} z_k,$$

where θ represents a complex statistic in our case. Considering two samples, S_1 and S_2, along with their respective complex statistic estimates, $\widehat{\theta_1}$ and $\widehat{\theta_2}$, derived from them, the alignment method proposed by Merkouris (2004) aims to set $\widehat{\theta_1} = \widehat{\theta_2}$. Based on (1) and the fact that the quantities $\sum_{k \in U} z_k$ and θ are population values, it follows that

$$\sum_{k \in S_1} z_k w_k \approx \sum_{k \in S_2} z_k w_k,$$

which implies that the linearized variable for the complex statistic of the target variable must be aligned. Therefore, the approach proposed by Merkouris (2004) is applicable and can be implemented, as we are now dealing with totals. This allows us to derive alignment weights for the samples S_1 and S_2.

The proposed approach offers many new opportunities. First, we are now able to derive alignment weights for a variety of complex statistics of the target variable, provided their linearized variables can be obtained. This means we are no longer restricted to using only the total of the target variable to derive the alignment weights. Additionally, obtaining alignment weights directly for the complex statistic of interest, rather than equating the totals of the target variable across the two samples as in the alignment method of Merkouris (2004), provides a more targeted approach. This is because the alignment weights are tailored to the specific complex statistic, leading to more accurate and precise estimates for that statistic. Moreover, the results are easier to interpret, as the alignment weights are explicitly derived based on the complex statistic of interest. Finally, this procedure is computationally efficient, as it only requires knowledge of the linearized variable of the complex statistic of the target variable.

As described above, the proposed approach provides a way to avoid necessarily using the total of the target variable and instead allows the use of another complex statistic to derive the alignment weights. This is particularly advantageous when the focus is on estimating a specific complex statistic of the target variable. However, just as we can use the total of the target variable even when we are interested in estimating a complex statistic of it, the proposed approach enables the derivation of alignment weights based on a complex statistic for a target variable, even when the goal is to estimate another statistic for the same target variable. This flexibility contrasts with the earlier discussion but also highlights a key advantage of the new approach, that is the alignment weights can be derived without being restricted to using the total, even when the complex statistic of interest is not used in the alignment procedure. Instead, we can choose from a wide range of other complex statistics to use in the alignment procedure. This approach allows us to derive alignment weights once based on a complex statistic and then reuse them to estimate multiple statistics. Such a strategy offers benefits, especially when the estimated complex statistic is related to the one used in the alignment procedure. Furthermore, this method is particularly useful in cases where the linearized variable of the complex statistic we want to estimate is unknown. Additionally, it accommodates scenarios where an agency imposes the constraint that a specific set of alignment weights, derived from a particular complex statistic, must be used for all statistics being estimated. However, this approach is not without risks. Potential issues include bias in the estimates, inefficiency in the variance estimates, and challenges in interpreting the results. Despite these risks, the approach can be considered acceptable if the trade-offs are justified, particularly in the presence of constraints, such as the unavailability of the linearized variable for the complex statistic of interest or specific requirements for the alignment weights imposed by agency policies.

In Subsection 2.2, we described the derivation of the variance estimation for the generalized regression estimator using the residuals from the estimated regression of the target variable on the auxiliary variables used for alignment. Accordingly, the variance estimation for both $\widehat{\theta}_1$ and $\widehat{\theta}_2$ can be derived using (3), replacing $\widehat{z}_k$ with the corresponding residuals from the estimated regression of $\widehat{z}_k$ on the auxiliary variables used

for alignment. In (3), the sampling weights $\mathbf{w}_1$ and $\mathbf{w}_2$ are used to preserve design-based properties. However, if both model-based and design-based properties are considered, the alignment weights $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ can be used instead of $\mathbf{w}_1$ and $\mathbf{w}_2$.

Instead of using the residuals in (3), we propose using a new linearized variable, denoted as $z_{k,a}$, estimated using the alignment weights. Let $\widehat{z_{k,a}}$ be the plug-in estimator of $z_{k,a}$. For the proposed estimators obtained using $\widehat{z_{k,a}}$, denoted as $\widehat{\theta_{1,a}}$ and $\widehat{\theta_{2,a}}$, the variance estimates are given by

$$\widehat{V}(\widehat{\theta_{1,a}}) = \sum_{k\in S_1}\sum_{l\in S_1}\tilde{\Delta}_{kl}\frac{\widehat{z_{k,a}}\widehat{z_{l,a}}}{\pi_k\pi_l}, \tag{14}$$

and

$$\widehat{V}(\widehat{\theta_{2,a}}) = \sum_{k\in S_2}\sum_{l\in S_2}\tilde{\Delta}_{kl}\frac{\widehat{z_{k,a}}\widehat{z_{l,a}}}{\pi_k\pi_l}. \tag{15}$$

In (14) and (15), either the alignment weights or the sampling weights can be used, depending on the desired properties.

Let $\widehat{\theta}_c$ be the adjusted composite estimator of θ (Merkouris, 2004, 2010a,b), obtained by combining the estimators $\widehat{\theta_{1,a}}$ and $\widehat{\theta_{2,a}}$. Therefore, we have

$$\widehat{\theta}_c = \delta\widehat{\theta_{1,a}} + (1-\delta)\widehat{\theta_{2,a}}, \tag{16}$$

where $\delta \in [0,1]$. The estimated variance of $\widehat{\theta}_c$ is then given by

$$\widehat{V}(\widehat{\theta}_c) = \delta^2\widehat{V}(\widehat{\theta_{1,a}}) + (1-\delta)^2\widehat{V}(\widehat{\theta_{2,a}}).$$

A choice of δ that minimizes $\widehat{V}(\widehat{\theta}_c)$ (Merkouris, 2004) can be derived as a weighted average based on $\widehat{V}(\widehat{\theta_{1,a}})$ and $\widehat{V}(\widehat{\theta_{2,a}})$. Specifically,

$$\delta_c = \frac{\widehat{V}(\widehat{\theta_{2,a}})}{\widehat{V}(\widehat{\theta_{1,a}}) + \widehat{V}(\widehat{\theta_{2,a}})}. \tag{17}$$

The new linearized variable, estimated using the alignment weights, is built upon the alignment weights and reflects the impact of calibration on the variability of the estimated complex statistic. This is because the alignment weights redistribute the sampling weights to calibrate the auxiliary variables to their known totals and ensure that common target variables yield identical estimates of the total across samples. Moreover, using the alignment weights to construct the new linearized variable eliminates the need for additional regression steps required for estimating residuals. As a result, deriving the variance estimation of the estimated complex statistic using the new linearized variable is computationally more efficient than using residuals. However, we should consider a potential risk regarding the bias of the variance estimation of the estimated complex statistic, which must be weighed against the computational gains achieved.

In Algorithm 1, we outline the steps of our proposed method for estimating complex statistics by combining information either from independent surveys targeting the same population or from two independent samples within the same survey.

Algorithm 1

1: Obtain the plug-in estimator $\widehat{z_k}$ for the linearized variable z_k associated with the complex statistic under consideration, using the weights w_k derived according to the sampling scheme.
2: Apply the alignment method proposed by Merkouris (2004) using the plug-in estimator $\widehat{z_k}$ from Step 1 to obtain the alignment weights $\boldsymbol{a} = \left(\boldsymbol{a}_1^{\mathrm{T}}, \boldsymbol{a}_2^{\mathrm{T}}\right)^{\mathrm{T}}$.
3: Compute the estimators $\widehat{\theta_{1,a}}$ and $\widehat{\theta_{2,a}}$, using the alignment weights $\boldsymbol{a} = \left(\boldsymbol{a}_1^{\mathrm{T}}, \boldsymbol{a}_2^{\mathrm{T}}\right)^{\mathrm{T}}$ obtained in Step 2, as well as their variances using (14) and (15).
4: Obtain the weighted estimator $\widehat{\theta_c}$ for the considered complex statistic using δ_c in (16), based on the estimators $\widehat{\theta_{1,a}}$ and $\widehat{\theta_{2,a}}$ and their variances computed in Step 3.

5. Simulation study

The simulation study is based on synthetic population data generated from Austrian EU-SILC data for 2006, provided by Statistics Austria. The population data are simulated using the methodology described in Alfons et al. (2011) and implemented in the R package `simPopulation` (Alfons and Gábor, 2013).

The package `laeken` (Alfons and Templ, 2013) provides the synthetic example dataset named `eusilc`, which consists of $14,827$ individuals with 28 variables from $N = 6,000$ households. We consider the entire `eusilc` dataset as the population. The target variable is the equivalized disposable income (`eqIncome`). We also use the employee cash or near cash income (net) (`py010n`)

Before describing the simulation scenarios, we introduce some notation. Let N_r denote the number of households in the r-th federal state, such that $N = \sum_{r=1}^{R} N_r$. Additionally, let $n_{j,r}$ represent the number of selected households in the j-th sample from the r-th federal state, and let n_j denote the number of selected households in the j-th sample. Our results are based on a two-stage sampling design, where households are sampled first, and then all individuals from these households are included. This sampling scheme simplifies the calculation of the second-order inclusion probabilities for individuals in selected households, which is required to demonstrate the potential of the proposed methodology. We consider a simulation scenario (SS1), in which households are selected with equal probability, that is for $j = 1, 2$, we randomly sample n_j households from the N available. The first-order inclusion probability of a household (individual) is equal to n_j/N. The second-order inclusion probability for two individuals is equal to n_j/N if they belong to the same household and $\frac{n_j}{N} \cdot \frac{n_j - 1}{N - 1}$ if they belong to different households. The auxiliary variable is the `py010n`, which has the largest correlation (0.36) with `eqIncome` among all the remaining variables in the `eusilc` dataset.

We apply the proposed method described in Section 4. to the Laeken indicators provided in Section 3.. For the calculation of the estimators of these complex statistics, we consider three cases:

1. In the first case, abbreviated as "(sw)", we use sampling weights.
2. In the second case, abbreviated as "(oa)", we use alignment weights obtained by aligning the linearized variable for the corresponding complex statistic of the target variable.
3. In the third case, abbreviated as "(m)", we use alignment weights obtained by aligning the linearized variable for the median of the target variable.

In all cases, we combine the estimators from the two samples using δ_c as given in (17). The estimated variances used to derive δ_c are computed under two frameworks:

1. Design-based properties, abbreviated as "d".
2. Model-based and design-based properties, abbreviated as "m".

For the proposed approach to variance estimation, where we use a new linearized variable estimated from the alignment weights instead of residuals, the abbreviation "New" is adopted.

The simulation is repeated $1,000$ times. We evaluate our proposed approach on the aforementioned simulation scenarios reporting the Monte Carlo estimates of the standard errors (SEs) and the relative mean squared errors (RMSEs). Also, we report the Monte Carlo estimate of the bias of the variance estimator of $\widehat{\theta}_c$. We should note that for the estimation of the variance we used the formula provided in (3).

In Table 1 we provide the Monte Carlo estimates of SE and the RMSE for the median, the QSR, the Gini coefficient and the RMPG of the equivalized disposable income for the SS1 simulation scenario when $n_1 = 500$ and $n_2 = 1000$. The results show that all estimation methods produce very similar SEs and RMSEs across the different statistics, with only negligible differences observed between the variants. Notably, the "New" methods yield marginal improvements in RMSE for the Gini coefficient and RMPG, suggesting slight efficiency gains without compromising precision. Overall, the consistency across methods indicates robustness in the estimation of income inequality and poverty indicators under the given sampling design.

In Table 2 we provide the Monte Carlo estimates of the bias of the variance estimator of $\widehat{\theta}_c$ for the median, the QSR, the Gini coefficient and the RMPG of the equivalized disposable income for the case that $n_1 = 500$ and $n_2 = 1000$. The results indicate that the bias of the variance estimator is generally low for the QSR, Gini coefficient, and RMPG across all methods, with minimal variation between them. For the median, however, the "New" methods show higher bias compared to the others, reflecting an increase in estimated variance.

In conclusion, it seems that the preferable method depends on the complex statistic. First, (sw) consistently results in higher SE and RMSE compared to the other methods. The methods d (oa) and m (oa) are preferable for the median and the RMPG due to their lower bias, while for the QSR and the Gini coefficient, all methods perform similarly.

Table 1: *The Monte Carlo estimates of SE and the RMSE for the median, the QSR, the Gini coefficient and the RMPG of the equivalized disposable income for SS1 when* $n_1 = 500$ *and* $n_2 = 1000$*.*

		Median	QSR	Gini coefficient	RMPG
(sw)	SE	250.7214	0.1130	0.5573	1.4353
	RMSE	1.3432	2.9152	2.1385	7.5532
d (oa)	SE	238.0430	0.1128	0.5564	1.4451
	RMSE	1.2763	2.9106	2.1370	7.5986
m (oa)	SE	238.0050	0.1128	0.5570	1.4447
	RMSE	1.2756	2.9117	2.1395	7.5964
New d (oa)	SE	238.4102	0.1129	0.5560	1.4453
	RMSE	1.2791	2.9132	2.1355	7.6000
New m (oa)	SE	238.2264	0.1128	0.5563	1.4443
	RMSE	1.2774	2.9126	2.1369	7.5947
d (m)	SE	238.0430	0.1131	0.5564	1.4585
	RMSE	1.2763	2.9191	2.1370	7.6687
m (m)	SE	238.0050	0.1132	0.5572	1.4571
	RMSE	1.2756	2.9224	2.1404	7.6614
New d (m)	SE	238.4102	0.1131	0.5560	1.4586
	RMSE	1.2791	2.9200	2.1352	7.6693
New m (m)	SE	238.2264	0.1132	0.5565	1.4565
	RMSE	1.2774	2.9218	2.1374	7.6582

Table 2: *The Monte Carlo estimates of the bias of the variance estimator of* $\widehat{\theta}_c$ *for the median, the QSR, the Gini coefficient and the RMPG of the equivalized disposable income for SS1 when* $n_1 = 500$ *and* $n_2 = 1000$*.*

	Median	QSR	Gini coefficient	RMPG
(sw)	14649.89	-0.0028	-0.0977	-0.2546
d (oa)	13333.35	-0.0028	-0.0967	-0.2838
m (oa)	13356.44	-0.0028	-0.0974	-0.2821
New d (oa)	20621.83	-0.0028	-0.0961	-0.2791
New m (oa)	20729.12	-0.0027	-0.0965	-0.2757
d (m)	13333.35	-0.0028	-0.0968	-0.3226
m (m)	13356.44	-0.0029	-0.0977	-0.3179
New d (m)	20621.83	-0.0028	-0.0961	-0.3153
New m (m)	20729.12	-0.0028	-0.0966	-0.3083

6. Concluding remarks

In this paper, we extend the alignment method of Merkouris (2004), originally for totals, to complex statistics by aligning the linearized variable of the target statistic. This enables deriving alignment weights using multiple complex statistics, rather than relying solely on totals. The weights from one statistic's linearized variable can also estimate other statistics. We propose a new variance estimation method based on a linearized variable from alignment weights, avoiding the use of residuals. While illustrated with social equality indicators, the method applies to any complex statistic with an available linearized variable.

Simulation results show that alignment weights can reduce variance estimates, though the best method depends on the statistic analyzed. Although not always superior to residual-based variance estimation, the approach offers comparable performance and computational advantages by eliminating extra regression steps.

ΠΕΡΙΛΗΨΗ

Η σημασία της διερεύνησης πιθανών ενσωματώσεων μεταξύ ερευνών έχει αναγνωριστεί ευρέως, με σκοπό την ενίσχυση της αποτελεσματικότητας και τη μείωση του κόστους. Εφαρμόζουμε τη μέθοδο ευθυγράμμισης για τον συνδυασμό πληροφοριών από δύο έρευνες, με στόχο την εκτίμηση σύνθετων στατιστικών μεγεθών. Η εξαγωγή βαρών ευθυγράμμισης για σύνθετα στατιστικά μεγέθη είναι δύσκολη λόγω της μη γραμμικής φύσης τους. Για να αντιμετωπίσουμε αυτή την πρόκληση, προτείνουμε την ευθυγράμμιση της γραμμικοποιημένης μεταβλητής που αντιστοιχεί στο σύνθετο στατιστικό του υπό μελέτη χαρακτηριστικού. Αυτή η προσέγγιση μας επιτρέπει επίσης να χρησιμοποιήσουμε τα βάρη ευθυγράμμισης που έχουν εξαχθεί από τη γραμμικοποιημένη μεταβλητή που σχετίζεται με ένα σύνθετο στατιστικό, για την εκτίμηση άλλων σύνθετων στατιστικών. Οι γραμμικοποιημένες μεταβλητές έχουν χρησιμοποιηθεί για την εκτίμηση της διακύμανσης, επιτρέποντάς μας να εκτιμήσουμε τη διακύμανση του συνδυασμένου εκτιμητή σύνθετων στατιστικών. Σύμφωνα με την τρέχουσα βιβλιογραφία, η διακύμανση ενός εκτιμητή σύνθετου στατιστικού υπολογίζεται συνήθως χρησιμοποιώντας τα κατάλοιπα από την παλινδρόμηση της γραμμικοποιημένης μεταβλητής στις βοηθητικές μεταβλητές που χρησιμοποιήθηκαν στην ευθυγράμμιση. Προτείνουμε την εκτίμηση της διακύμανσης ενός σύνθετου στατιστικού χρησιμοποιώντας μια νέα γραμμικοποιημένη μεταβλητή που προκύπτει από τα βάρη ευθυγράμμισης. Αυτές οι καινοτομίες θα μπορούσαν να οδηγήσουν σε βελτιώσεις στην υπολογιστική αποτελεσματικότητα και, σε ορισμένες περιπτώσεις, στην ποιότητα και αξιοπιστία των εκτιμητών. Οι προτεινόμενες προσεγγίσεις αξιολογούνται μέσω μελέτης προσομοίωσης.

REFERENCES

Alfons, A. and Gábor, C. (2013). Simulation of synthetic populations for surveys based on sample. R package version 0.4.1. https://github.com/cran/simPopulation.

Alfons, A., Kraft, S., Templ, M., and Filzmoser, P. (2011). Simulation of close-to-reality population data for household surveys with application to EU-SILC. *Statistical Methods and Applications*, 20(3):383–407.

Alfons, A. and Templ, M. (2013). Estimation of social exclusion indicators from complex surveys: The R package `laeken`. *Journal of Statistical Software*, 54(15):1–25.

Atkinson, A. B., Marlier, E., and Nolan, B. (2004). Indicators and targets for social inclusion in the european union. *JCMS: Journal of Common Market Studies*, 42(1):47–75.

Berger, Y. G. and Kabzińska, E. (2020). Empirical likelihood approach for aligning information from multiple surveys. *International Statistical Review*, 88(1):54–74.

Berger, Y. G., Muñoz, J. F., and Rancourt, E. (2009). Variance estimation of survey estimates calibrated on estimated control totals—an application to the extended regression estimator and the regression composite estimator. *Computational Statistics & Data Analysis*, 53(7):2596–2604.

Binder, D. A. (1996). Linearization methods for single phase and two-phase samples: a cookbook approach. *Survey Methodology*, 22(1):17–22.

Binder, D. A. and S., K. M. (1995). Estimating some measures of income inequality from survey data: an application of the estimating equations approach. *Survey Methodology*, 21(2):137–145.

Demnati, A. and Rao, J. N. K. (2004). Linearization variance estimators for survey data. *Survey Methodology*, 30(1):17–26.

Demnati, A. and Rao, J. N. K. (2010). Linearization variance estimators for model parameters from complex survey data. *Survey Methodology*, 36(2):193–201.

Deville, J.-C. (1999). Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey Methodology*, 25(2):193–203.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418):376–382.

Dong, Q., Elliott, M. R., and Raghunathan, T. E. (2014). Combining information from multiple complex surveys. *Survey Methodology*, 40(2):347–354.

Dong, Z., Tillé, Y., Giorgi, G. M., and Guandalini, A. (2021). Linearization and variance estimation of the Bonferroni inequality index. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 184(3):1008–1029.

Elliott, M. R., Raghunathan, T. E., and Schenker, N. (2018). Combining estimates from multiple surveys. In *Wiley StatsRef: Statistics Reference Online*, pages 1–10. John Wiley & Sons, Ltd.

Fabrizi, E., Ferrante, M. R., and Trivisano, C. (2020). A functional approach to small area estimation of the relative median poverty gap. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 183(3):1273–1291.

Graf, E. and Tillé, Y. (2014). Variance estimation using linearization for poverty and social exclusion indicators. *Survey Methodology*, 40:61–79.

Guandalini, A. and Tillé, Y. (2017). Design-based estimators calibrated on estimated totals from multiple surveys. *International Statistical Review*, 85(2):250–269.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393.

Hidiroglou, M. (2001). Double sampling. *Survey Methodology*, 27:143–154.

Kaplan, S., Popoks, D., Prato, C. G., and Ceder, A. (2014). Using connectivity for measuring equity in transit provision. *Journal of Transport Geography*, 37:82–92.

Karsu, O. and Morton, A. (2015). Inequity averse optimization in operational research. *European Journal of Operational Research*, 245:343–359.

Kim, J. K. and Rao, J. N. (2012). Combining data from two independent surveys: a model-assisted approach. *Biometrika*, 99(1):85–100.

Lesage, E. (2011). The use of estimating equations to perform a calibration on complex parameters. *Survey Methodology*, 37(1):103–108.

Merkouris, T. (2004). Combining independent regression estimators from multiple surveys. *Journal of the American Statistical Association*, 99(468):1131–1139.

Merkouris, T. (2010a). Combining information from multiple surveys by using regression for efficient small domain estimation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 72(1):27–48.

Merkouris, T. (2010b). An estimation method for matrix survey sampling. In *Proceedings of the Section on Survey Research Methods: American Statistical Association*, pages 4880–4886, Vancouver, Canada.

Merkouris, T. (2015). An efficient estimation method for matrix survey sampling. *Survey Methodology*, 41(1):237–262.

Osier, G. (2009). Variance estimation for complex indicators of poverty and inequality using linearization techniques. *Survey Research Methods*, 3:167–195.

Renssen, R. and Nieuwenbroek, N. (1997). Combining independent regression estimators from multiple surveys. *Journal of the American Statistical Association*, 92:368–374.

Vallée, A.-A. and Tillé, Y. (2019). Linearisation for variance estimation by means of sampling indicators: Application to non-response. *International Statistical Review*, 87(2):1–15.

Wu, C. (2004). Combining information from multiple surveys through the empirical likelihood method. *Canadian Journal of Statistics*, 32:15–26.

Yang, S. and Kim, J. K. (2020). Statistical data integration in survey sampling: a review. *Japanese Journal of Statistics and Data Science*, 3:625–650.

Ybarra, L. and Lohr, S. (2008). Small area estimation when auxiliary information is measured with error. *Biometrika*, 95(4):919–931.

Zieschang, K. D. (1990). Sample weighting methods and estimation of totals in the consumer expenditures surveys. *Journal of the American Statistical Association*, 85:986–1001.

Greek Statistical Institute
Proceedings of the 37th Panhellenic & 3rd International Stat Conf (2025), pp.277-295

TWO MODIFIED CUSUM-TYPE CONTROL CHARTS FOR JOINT MONITORING OF ORIGIN AND SCALE OF A TWO-PARAMETER EXPONENTIAL DISTRIBUTION

K. G. Fountoukidis, D. L. Antzoulakos and A. C. Rakitzis
University of Piraeus, Department of Statistics and Insurance Science, 18534, Piraeus Greece
k.fountoukidis@unipi.gr, dantz@unipi.gr, arakitz@unipi.gr

ABSTRACT

The exponential distribution is the most widely used distribution for describing the time for an event to occur. However, in many cases the origin of the exponential distribution may be different from zero. For example, when products come with a warranty time, this assures that failures occur only after that time. In these cases, we often encounter a two-parameter (or shifted) exponential distribution to model the lifetime of the product. In this paper, we propose two new CUSUM-type control charts for the joint monitoring of the origin and scale parameters of the shifted exponential distribution based on their maximum likelihood estimators. A detailed description of the operation of the proposed charts is provided while empirical rules for selecting the values of their design parameters are given, as well. The in-control and out-of-control performance of the proposed charts is computed via Monte Carlo simulation. The results of the numerical study show that the proposed CUSUM charts significantly outperform the usual Shewhart-type control chart. Comparisons are also made against other competitive CUSUM-type charts. Finally, the implementation of the proposed control charts is exhibited via a numerical example.

Keywords: CUSUM chart, shifted exponential distribution, joint process monitoring, origin, scale.

1. INTRODUCTION

Statistical Process Control (SPC) utilizes statistical tools and techniques to monitor and control processes. Among these tools, the control chart is the primary tool designed to enable practitioners to identify processes operating out-of-control (OOC). The main aim of a control chart is the quick detection of assignable causes of variation, facilitating timely corrective actions to prevent

the production of large amounts of non-conforming products. While the usual assumption is that the monitored quality characteristic X (random variable, r.v.) follows a normal distribution, contemporary literature increasingly addresses scenarios where this assumption is violated, particularly in the context of process related to lifetime data.

Exponential distribution plays an important role in quality and reliability engineering (see e.g. [Xie et al. (2002)] and [Santiago and Smith (2013)]) where it is employed to model the lifetime of items with constant failure rate. It is well known that in some situations the exponential distribution has an origin parameter θ that is greater than zero, in addition to its scale parameter λ. For example, businesses very often offer warranty periods for their products to attract customers. These warranties provide assurance that the product will function for at least a specified period before any failure occurs. In these situations, it is better to adopt a shifted (or two-parameter) exponential distribution to model the products lifetime. Thus, it is very important to employ control charts to test whether the parameters of a shifted exponential distribution have been shifted. The shifted exponential distribution can be applied in many fields. For example, [Baten and Kamil (2009)] applied it to optimize total costs in managing hazardous items inventory. Furthermore, [Kao (2010)] found that the current needed to break down a wafer in acceptable tasting follow the shifted exponential distribution.

In general, two approaches may be used for monitoring the parameters of a shifted exponential distribution. With the first approach, two separate control charts may be used to monitor each one of the two parameters, while with the second approach a single control chart is utilized to simultaneously monitor the two parameters. The main advantage of the second approach is that it simplifies the complexity of using two separate charts to monitor origin and scale parameters. Also, it is less expensive and time consuming. In this paper, we focus on the second approach. For detailed overviews on single control charts see [Jalilibal et al. (2022)] and [Chen and Thaga (2006)].

[Chen and Cheng (1998)] introduced the Max chart, a Shewhart type control chart, that is the most prominent single control chart. [Chen et al. (2001)] incorporated the exponentially weighted moving average (EWMA) statistic in the Max chart and proposed the Max-EWMA chart which can detect small changes in process mean and/or variability. [Thaga (2010)] proposed the Max-CUSUM (Cumulative Sum) chart and the SS-CUSUM chart (Sum of Squares), which can detect small shifts in process mean and/or standard deviation. These

control charts were designed under the assumption that the quality characteristic follows a normal distribution. [Mukherjee et al. (2015)] introduced Max-type charts for monitoring the origin and the scale parameters of the shifted exponential distribution based on the maximum likelihood estimators (MLEs) of the two parameters. [Chong et al. (2021)] studied the case where the values of the shifted exponential process parameters are unknown, while [Mukherjee et al. (2022)] studied the effect of using different mappings and Minkowski distances in joint monitoring of TBE processes. [Huang et al. (2018)] introduced a CUSUM chart specifically designed for the simultaneous monitoring of the origin and scale parameters of a shifted exponential process. [Saha et al. (2022)] extended the work of [Chong et al. (2021)] by studying the performance of the respective CUSUM charts in the case of estimated process parameters. [Li et al. (2022)] proposed a generally weighted moving average control chart for monitoring shifted exponential processes in the case of measurement errors. Recently, [Baranwal et al. (2024)] proposed and studied the performance of EWMA, double-EWMA and triple-EWMA charts in the detection of shifts in the parameters of an exponential process, while [Antzoulakos et al. (2025)] proposed a run sum single control chart for detecting changes in the origin and the scale of the process.

In this study, we propose two modified CUSUM-type control charts for the joint monitoring of origin and scale parameters of a shifted exponential distribution. We investigate the performance of the proposed charts in a wide range of OOC conditions. Moreover, the performance of the proposed charts is compared with other competitive control charts. This paper is organized as follows: Section 2 presents the statistical framework and the design of the proposed control charts. The results of the numerical study are presented in Section 3 and a comparison study with other control charts is offered in Section 4. An illustrative example is given in Section 5, while conclusions are drawn in Section 6.

2. STATISTICAL FRAMEWORK

Assume that the quality characteristic X follows the shifted exponential distribution with origin parameter $\theta \geq 0$ and scale parameter $\lambda > 0$, that is $X \sim SE(\theta, \lambda)$. The probability density function of the shifted exponential distribution is given by

$$f(x; \theta, \lambda) = \frac{1}{\lambda} \exp\left(-\frac{x-\theta}{\lambda}\right), \quad x \geq \theta. \tag{1}$$

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a shifted exponential process with origin and scale parameters θ and λ, respectively. The process is in-control (IC) if $\theta = \theta_0$ and $\lambda = \lambda_0$. On the other hand, the OOC values of the parameters will be denoted

by $\theta = \theta_1$ and $\lambda = \lambda_1$. The MLEs of the origin and scale parameters are $\hat{\theta} = X_{(1)}$ and $\hat{\lambda} = \bar{X} - X_{(1)}$, respectively, and $\hat{\theta}$ and $\hat{\lambda}$ are independent (see [Lawless (1982)], [Johnson and Kotz (1970)]).

[Mukherjee et al. (2015)] defined the following two statistics:

$$E_1 = \frac{2n(\hat{\theta}-\theta_0)}{\lambda_0}, \quad E_2 = \frac{2}{\lambda_0}\sum_{i=1}^{n}\left(X_i - \hat{\theta}\right). \tag{2}$$

When the process is IC, then $E_1 \sim \chi^2_2$ and $E_2 \sim \chi^2_{2n-2}$, where χ^2_ν denotes the of chi-square distribution with ν degrees of freedom. Now, let:

$$W^{(1)} = \Phi^{-1}\{H_2(E_1)\}, \quad W^{(2)} = \Phi^{-1}\{H_{2n-2}(E_2)\}, \tag{3}$$

where $H_\nu(\cdot)$ denotes the cumulative distribution function (cdf) of χ^2_ν, and $\Phi(\cdot)$ is the cdf of the standard normal distribution. It follows from the probability integral transform that $H_2(E_1)$ and $H_{2n-2}(E_2)$ follow the uniform distribution over $(0,1)$ when the process is IC, while $W^{(1)}$ and $W^{(2)}$ follow the standard normal distribution. Furthermore, $W^{(1)}$ and $W^{(2)}$ are independent.

2.1 DESIGN OF THE MAX-SE-CUSUM CHART

Assume that the process characteristic of interest X follows the shifted exponential distribution $SE(\theta, \lambda)$. Let $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{in})$, $i = 1,2,\ldots,n$, be independent random samples of size n from the process. Let also $W_i^{(1)}$ and $W_i^{(2)}$ be the statistics defined in (3) based on the i^{th} sample. The construction of the Max-Shifted Exponential-Cumulative Sum (Max-SE-CUSUM) chart requires the following four statistics:

$$C_i^+ = \max\left[0, W_i^{(1)} - k + C_{i-1}^+\right], \tag{4}$$

$$C_i^- = \max\left[0, -W_i^{(1)} - k + C_{i-1}^-\right], \tag{5}$$

and

$$S_i^+ = \max\left[0, W_i^{(2)} - k + S_{i-1}^+\right], \tag{6}$$

$$S_i^- = \max\left[0, -W_i^{(2)} - k + S_{i-1}^-\right], \tag{7}$$

where $k > 0$ is the reference value of the chart, and C_0^+, C_0^-, S_0^+ and S_0^- are the starting values of the corresponding statistics with $C_0^+ = C_0^- = S_0^+ = S_0^- = 0$. Since $W_i^{(1)}$ and $W_i^{(2)}$ are both normally distributed we can combine the statistics C_i^+, C_i^-, S_i^+ and S_i^- to develop a new statistic defined by:

$$M_i = \max[C_i^+, C_i^-, S_i^+, S_i^-]. \tag{8}$$

The statistic M_i is always positive and indicates that the process origin has drifted away from θ_0 and/or the process scale has drifted away from λ_0, whenever its value exceeds a suitable upper control limit (decision interval) h_M. Small values of M_i indicate that the process is IC. Therefore, if $M_i > h_M$ then an OOC signal is obtained.

The charting procedure and operation of the Max-SE-CUSUM chart is as follows:

1. Specify the IC parameters θ_0 and λ_0, choose the sample size $n \geq 2$ and the desired IC average run length value ARL_0.
2. Given the reference value k, determine the value of the decision interval h_M.
3. For the i^{th} sample, $i = 1,2, \ldots$, compute $W_i^{(1)}$, $W_i^{(2)}$and the value of the statistic M_i.
4. Declare the process as OOC at the i^{th} sample if $M_i > h_M$.
5. Follow-up procedure: When the process is declared as OOC at the i^{th} sample, compare each of C_i^+, C_i^-, S_i^+ and S_i^- with h_M.

   (i) If $C_i^+ > h_M$, then we have indication for an increase in origin.
   (ii) If $C_i^- > h_M$, then we have indication for a decrease in origin.
   (iii) If $S_i^+ > h_M$, then we have indication for an increase in scale.
   (iv) If $S_i^- > h_M$, then we have indication for a decrease in scale.

   If $C_i^+, S_i^+ > h_M$, or $C_i^+, S_i^- > h_M$, or $C_i^-, S_i^+ > h_M$ or $C_i^-, S_i^- > h_M$, it should be safe to conclude that λ has shifted, and further investigation is required about the shift in origin θ. This is because the distribution of the origin estimator $\hat{\theta}$ depends on the scale parameter λ (see Mukherjee et al. (2015)). If $C_i^+, S_i^+ > h_M$, or $C_i^+, S_i^- > h_M$, or $C_i^-, S_i^+ > h_M$ or $C_i^-, S_i^- > h_M$

2.2 DESIGN OF THE SS-SE-CUSUM CHART

Consider the four CUSUM statistics defined by equations (4-7). The construction of the Sum of Squares-Shifted Exponential-Cumulative Sum (SS-SE-CUSUM) chart is based on the following statistics:

$$N_i = \max[C_i^+, C_i^-], \tag{9}$$

$$Q_i = \max[S_i^+, S_i^-], \tag{10}$$

and

$$SS_i = N_i^2 + Q_i^2. \tag{11}$$

The statistic in equation (11) defines a circle. Since both N_i and Q_i are nonnegative, it is sufficient to use only the first quadrant of the circle centered at (0, 0) to plot the pairs (Q_i, N_i) for each sample. A point failing outside a circle with radius $r = h_S$ ($SS_i > h_S$) indicates an OOC situation, otherwise the process is IC.

The charting procedure and operation of the SS-SE-CUSUM chart is as follows:

1. Specify the IC parameters θ_0 and λ_0, choose the sample size $n \geq 2$ and the desired IC average run length value ARL_0.
2. Given the reference value k, determine the value of the decision interval h_S.
3. For the i^{th} sample, $i = 1,2, \ldots$, compute N_i, Q_i and the value of the statistic SS_i.
4. Declare the process as OOC at the i^{th} sample if $SS_i > h_S$.
5. Follow-up procedure: Assume that the process is declared as OOC at the i^{th} sample.

   If $N_i > r$ and $Q_i \leq r$, then:

   (i) If $C_i^+ > r$, then we have indication for an increase in origin.
   (ii) If $C_i^- > r$, then we have indication for a decrease in origin.

   If $Q_i > r$ and $N \leq r$, then:

   (iii) If $S_i^+ > r$, then we have indication for an increase in scale.
   (iv) If $S_i^- > r$, then we have indication for a decrease in scale.

   If $N_i > r$ and $Q_i > r$, then

   (v) If $C_i^+, S_i^+ > r$, or $C_i^+, S_i^- > r$, or $C_i^-, S_i^+ > r$, or $C_i^-, S_i^- > r$, it should be safe to conclude that λ has shifted, but a further investigation is required about the shift in origin θ, since the distribution of the origin estimator $\hat{\theta}$ depends on the scale parameter λ (see Mukherjee et al. (2015)).

3. NUMERICAL STUDY

In this section we evaluate the performance of the proposed Max-SE-CUSUM and SS-SE-CUSUM control charts. We use Monte Carlo simulation with 10^4 runs to investigate the performance of the proposed charts in terms of ARL. Following [Huang et al. (2018)] we consider the cases $k \in \{0.5,1,1.5\}$, $ARL_0 = 500$ and $(\theta_0, \lambda_0) \in \{(0,1), (10,100)\}$.

In Table 1 we provide the values of the decision intervals of the proposed control charts when $n = 5$. From Table 1 we deduce that the values of the decision intervals decrease as k increases.

Table 1. *Decision intervals for* $n = 5$ *and* $ARL_0 = 500$.

	Max-SE-CUSUM	SS-SE-CUSUM
k	h_M	h_S

0.5	5.75	5.97
1	3.01	3.07
1.5	1.94	1.96

For the OOC values of θ_1 and λ_1 we choose $\theta_1 \in \{0, 0.25, 0.5, 1, 1.5\}$ and $\lambda_1 \in \{0.25, 0.5, 0.75, 0.95, 1, 1.25, 1.50, 1.75, 2\}$ when $(\theta_0, \lambda_0) = (0,1)$, while for $(\theta_0, \lambda_0) = (10,100)$ we choose $\theta_1 \in \{5, 7, 10, 13, 15, 25, 35\}$ and $\lambda_1 \in \{60, 80, 100, 120, 140, 200, 300\}$. The ARL performance of the proposed control charts is given in Tables 2 and 3. The gray-shaded cells show the IC performance of the charts. The boldfaced entries in Tables 2 and 3 indicate the minimum OOC ARL value. It follows from Table 2 that the SS-SE-CUSUM chart has slightly better performance than the Max-SE-CUSUM for almost all the considered shifts.

The results in Table 3 show that the SS-SE-CUSUM chart outperforms the Max-SE-CUSUM chart for most of the pairs (θ_1, λ_1). Specifically, the Max-SE-CUSUM chart has better performance for small increases in origin and small decreases in scale (e.g. for $\theta_1 \in (10,15]$ and $\lambda_1 = 80$), as well as in the cases that correspond to large increasing or decreasing shifts in the process mean (i.e. for $(\theta_1, \lambda_1) \in \{(5, 60), (13, 60), (15, 60), (25, 100), (35, 100)\}$.

Next, we examine the overall performance of the proposed charts. It is noted that for the derivation of the OOC ARL values given in Tables 2 and 3 we use specific OOC values of θ_1 and λ_1. However, in practice, it is difficult to know their true values. Therefore, it is more realistic to assume that θ and λ are r.v.'s and to use an overall performance measure to assess the performance of the control chart. A usual assumption made in this case is that θ and λ are independent with uniform distributions over the intervals $[\theta_{\min}, \theta_{\max}]$ and $[\lambda_{\min}, \lambda_{\max}]$, respectively. In this case, the overall performance of a control chart can be calculated by the expected average run length ($EARL$) (see [Mukherjee et al. (2022)]). The $EARL$ is calculated by the following formula

$$EARL = \frac{1}{(\lambda_{max} - \lambda_{min})(\theta_{max} - \theta_{min})} \int_{\theta_{min}}^{\theta_{max}} \int_{\lambda_{min}}^{\lambda_{max}} ARL(\theta, \lambda) d\lambda d\theta. \qquad (12)$$

The smaller the value of the $EARL$ for a control chart, the better its overall performance.

In the sequel, we consider 4 scenarios of possible shifts ranges. When $(\theta_0, \lambda_0) = (0,1)$, we consider $[\theta_{\min}, \theta_{\max}] = [0, 1.5]$ and $[\lambda_{\min}, \lambda_{\max}] = [0.25,\ 2]$ (Scenario 1). When $(\theta_0, \lambda_0) = (10,100)$ we consider the following three scenarios (see [Huang et al. (2018)]):

- $[\theta_{\min}, \theta_{\max}] = [10, 13]$, $[\lambda_{\min}, \lambda_{\max}] = [100, 130]$ (Scenario 2)
- $[\theta_{\min}, \theta_{\max}] = [7, 13]$, $[\lambda_{\min}, \lambda_{\max}] = [70, 130]$ (Scenario 3)
- $[\theta_{\min}, \theta_{\max}] = [7, 10]$, $[\lambda_{\min}, \lambda_{\max}] = [70, 100]$ (Scenario 4)

From Table 4 we deduce that, in terms of $EARL$, the SS-SE-CUSUM chart with $k = 0.5$ has slightly better performance than the Max-SE-CUSUM chart for all scenarios except from Scenario 3 in which the Max-SE-CUSUM chart has better performance.

The performance of a CUSUM chart depends on the value of the reference parameter k. Based on the results of the numerical study when $(\theta_0, \lambda_0) = (0,1)$ we suggest taking $k = 0.5$ for detecting small increases in process origin and/or scale, while for detecting moderate to large increases in origin regardless the shift in scale we suggest taking $k = 1.5$. In the case where $(\theta_0, \lambda_0) = (10, 100)$, we recommend using $k = 0.5$ for any shift in the process parameters. We note also that the best performance in terms of $EARL$ for both control charts results from the case $k = 0.5$.

Table 2. *ARL profiles for the Max-SE-CUSUM and SS-SE-CUSUM charts,* $(\theta_0, \lambda_0) = (0,1)$.

		Max-SE-CUSUM			**SS-SE-CUSUM**		
θ_1	λ_1	$k = 0.5$	$k = 1$	$k = 1.5$	$k = 0.5$	$k = 1$	$k = 1.5$
0.00	0.25	3.96	3.14	3.52	3.82	**3.07**	3.48
	0.50	8.79	11.87	33.10	**8.49**	11.52	32.96
	0.75	43.18	140.17	301.70	**42.07**	139.08	296.85
	1.00	501.07	500.33	501.34	500.95	503.68	499.47
	1.25	43.23	72.92	102.42	**41.56**	70.52	97.87
	1.50	13.77	17.59	25.49	**13.38**	16.74	24.52
	1.75	8.20	8.20	10.59	7.89	**7.82**	10.27
	2.00	6.00	5.25	6.01	5.77	**5.07**	5.86
0.25	0.25	3.97	**3.14**	3.55	4.04	3.17	3.57
	0.50	8.67	11.94	35.77	**7.75**	11.98	35.69
	0.75	10.91	33.92	321.33	**10.37**	32.50	325.37
	1.00	**9.12**	16.43	143.38	9.20	16.21	141.48
	1.25	7.80	10.32	36.19	**7.57**	9.99	34.23
	1.50	6.64	7.07	13.93	**6.14**	6.64	12.95
	1.75	5.65	5.20	7.36	5.08	**4.78**	6.84
	2.00	4.84	4.02	4.77	4.31	**3.72**	4.51
0.50	0.25	3.96	3.13	3.53	3.52	**2.89**	3.52
	0.50	5.49	5.02	12.77	4.83	**4.62**	11.80
	0.75	5.15	4.62	9.08	5.13	**4.58**	8.92
	1.00	4.82	**4.18**	6.66	4.89	4.19	6.63
	1.25	4.54	3.80	5.32	4.52	**3.77**	5.25
	1.50	4.27	3.46	4.32	4.12	**3.34**	4.24
	1.75	4.03	3.15	3.64	3.74	**2.96**	3.46
	2.00	3.76	2.85	3.08	3.41	**2.64**	2.86
1.00	0.25	2.99	2.13	1.96	2.83	1.99	**1.91**
	0.50	2.99	2.05	1.99	2.98	2.06	**1.98**
	0.75	2.99	2.02	1.98	2.99	2.06	**1.98**
	1.00	2.98	2.01	1.96	2.98	2.03	**1.95**
	1.25	2.95	1.99	1.91	2.95	2.00	**1.91**
	1.50	2.91	1.97	1.87	2.87	1.97	**1.83**
	1.75	2.85	1.95	1.80	2.75	1.92	**1.76**
	2.00	2.78	1.91	1.73	2.62	1.85	**1.66**
1.50	0.25	2.82	1.99	1.88	2.05	1.88	**1.48**
	0.50	2.48	1.99	1.71	2.45	1.97	**1.65**
	0.75	2.29	1.97	1.57	2.51	1.97	**1.58**
	1.00	2.19	1.94	1.47	2.40	1.94	**1.49**
	1.25	2.14	1.90	1.40	2.28	1.89	**1.40**
	1.50	2.10	1.85	1.34	2.17	1.81	**1.32**
	1.75	2.07	1.80	1.30	2.10	1.72	**1.25**
	2.00	2.06	1.74	1.26	2.06	1.62	**1.20**

Table 3. *ARL profiles for the Max-SE-CUSUM and SS-SE-CUSUM charts,* $(\theta_0, \lambda_0) = (10,100)$.

		Max-SE-CUSUM			**SS-SE-CUSUM**		
θ_1	λ_1	$k = 0.5$	$k = 1$	$k = 1.5$	$k = 0.5$	$k = 1$	$k = 1.5$
5	60	**2.86**	2.89	2.89	2.87	2.87	2.86
	80	3.69	3.71	3.72	**3.67**	3.68	3.72
	100	4.55	4.53	4.56	4.49	**4.46**	4.46
	120	5.19	5.18	5.29	5.23	5.21	**5.14**
	140	5.47	5.38	5.54	5.39	5.41	**5.38**
7	60	4.25	4.26	4.36	**4.21**	4.27	4.37
	80	5.73	5.71	5.77	**5.69**	5.77	5.81
	100	7.22	**6.93**	7.01	7.04	7.06	6.98
	120	8.01	8.20	8.12	8.05	8.14	**7.93**
	140	7.74	7.81	8.27	**7.49**	7.82	8.11
10	60	13.75	28.74	91.85	**13.56**	28.31	92.83
	80	75.96	225.82	393.33	**73.90**	219.73	401.95
	100	499.77	506.22	498.00	502.15	502.03	502.43
	120	65.69	106.58	142.50	**62.18**	102.14	138.66
	140	19.15	27.59	40.12	**18.43**	26.73	40.20
13	60	**14.25**	30.56	103.45	14.63	32.00	106.82
	80	**87.11**	287.92	552.56	90.93	302.96	573.33
	100	349.13	567.42	612.91	**347.37**	565.93	616.40
	120	50.29	95.69	139.25	**47.90**	94.25	137.49
	140	17.78	25.86	39.69	**16.56**	24.85	39.41
15	60	**14.16**	30.44	103.07	14.81	31.22	108.37
	80	**85.17**	279.76	539.72	88.09	295.12	565.00
	100	183.80	468.99	572.12	**175.83**	479.06	579.00
	120	38.49	88.33	133.16	**36.46**	84.70	131.44
	140	16.18	24.73	38.33	**15.11**	23.47	37.01
25	60	14.01	29.99	101.86	**13.68**	31.20	104.92
	80	24.18	215.83	499.88	**22.96**	211.13	519.61
	100	**17.64**	104.31	348.35	17.73	103.81	353.13
	120	13.24	34.30	91.21	**12.80**	32.34	87.83
	140	10.09	15.07	29.98	**9.30**	14.08	28.64
	200	5.42	4.61	5.48	4.87	**4.30**	5.13
35	60	10.85	27.64	103.42	**9.41**	25.89	103.56
	80	10.52	28.82	358.65	**10.29**	28.17	349.41
	100	**9.10**	16.23	141.42	9.22	16.29	139.52
	120	8.00	11.15	44.96	**7.89**	10.89	44.51
	200	4.85	4.06	4.81	4.29	**3.69**	4.54
	300	3.14	2.27	2.10	2.82	2.09	**1.99**

Table 4. *EARL profiles for the Max-SE-CUSUM and SS-SE-CUSUM charts,* $ARL_0 = 500, n = 5$

Chart	k	Scenario 1	Scenario 2	Scenario 3	Scenario 4
Max-SE-CUSUM	0.5	7.80	154.91	**103.64**	26.37
	1	14.024	225.15	169.50	33.80
	1.5	29.26	264.46	225.26	39.93
SS-SE-CUSUM	0.5	**7.59**	**154.79**	105.22	**23.22**
	1	13.90	225.24	173.47	28.72
	1.5	29.29	265.78	229.66	32.99

4. NUMERICAL COMPARISONS

In this section the Max-SE-CUSUM and SS-SE-CUSUM charts are compared, in terms of ARL, with the SEMLE-Max chart of [Mukherjee et al. (2015)] and the CUSUM-MLE-Max and CUSUM-LR charts of [Huang et al. (2018)]. We consider both the cases of fixed and random shifts in the OOC values of origin and scale. To ensure a fair comparison, all control charts have the same IC ARL value ($ARL_0 = 500$) and sample size ($n = 5$). For the reference value k we choose the values $0.5, 1, 1.5$. For θ_0 and λ_0 we assume, as in Section 3, that $(\theta_0,\ \lambda_0) \in \{(0,1), (10,100)\}$.

In Tables 4 and 5 we provide numerical comparisons, in terms of OOC ARL, for the five previously mentioned control charts. The entries in bold correspond to the minimum OOC ARL value for the specific pair (θ_1, λ_1). Due to space economy, we present results only for $\theta_1 \in \{0, 0.25, 0.5, 1, 1.5\}$ and $\lambda_1 \in \{0.5, 0.75, 1, 1.25, 1.50\}$ when $(\theta_0, \lambda_0) = (0,1)$. For $(\theta_0, \lambda_0) = (10,100)$ we consider the cases $\theta_1 \in \{5, 7, 10, 13, 15, 25, 35\}$ and $\lambda_1 = \{60, 80, 100, 120\}$. The results from Table 4 (case $(\theta_0, \lambda_0) = (0,1)$) show that the SS-SE-CUSUM chart has the best performance in most OOC cases except from $(\theta_1, \lambda_1) \in \{(0.25, 0.25), (0.5,0.25), (0.5,0.5), (1,0.25)\}$ where the CUSUM-LR chart is slightly better. Also, the Max-SE-CUSUM chart has slightly better performance only for $(\theta_1,\ \lambda_1) = (0.25, 1)$. From Table 5 (case $(\theta_0, \lambda_0) = (10,100)$) it can be seen that the SS-SE-CUSUM chart outperforms the other charts in most cases. The Max-SE-CUSUM chart appears to be better than the other charts for small increases in origin and small decreases in scale, and for large increases in origin only. The CUSUM-LR chart is the best chart only when $(\theta_1, \lambda_1) \in \{(13, 100), (35, 60)\}$.

Table 6 provides $EARL$ values for all the control charts examined for Scenarios 1-4 of Section 3. The numerical findings show that the SS-SE-CUSUM chart with $k = 0.5$ has the best performance for Scenarios 1, 2 and 4, while for Scenario 3 the best chart is the Max-SE-CUSUM chart.

Table 4. *ARL values of the Max-SE-CUSUM, SS-SE-CUSUM, CUSUM-MLE-Max and CUSUM-LR charts:* $(\theta_0, \lambda_0) = (0,1)$, $n = 5$ *and* $ARL_0 = 500$.

			Max-SE-CUSUM			SS-SE-CUSUM			CUSUM-MLE-Max			CUSUM-LR		
θ_1	λ_1	SEMLE-Max	$k = 0.5$	$k = 1$	$k = 1.5$	$k = 0.5$	$k = 1$	$k = 1.5$	$k = 0.5$	$k = 1$	$k = 1.5$	$k = 0.5$	$k = 1$	$k = 1.5$
0	0.25	17.24	3.96	3.14	3.52	3.82	**3.07**	3.48	4.26	6.66	13.53	4.83	4.45	4.77
	0.50	141.83	8.79	11.87	33.1	**8.49**	11.52	32.96	45.85	103.62	135.79	21.68	32.25	49.89
	0.75	463.71	43.18	140.17	301.70	**42.07**	139.08	296.85	455.01	455.12	458.25	187.2	247.3	288.7
	1.25	137.13	43.23	72.92	102.42	**41.56**	70.52	97.87	97.27	121.91	135.31	223.21	262.54	287.57
	1.50	41.81	13.77	17.59	25.49	**13.38**	16.74	24.52	24.45	33.76	39.70	55.66	72.92	88.64
0.25	0.25	17.82	3.97	3.14	3.55	4.04	3.17	3.57	4.62	7.32	14.23	3.41	2.97	**2.94**
	0.50	165.04	8.67	11.94	35.7	**7.75**	11.98	35.69	69.20	132.85	157.97	7.29	7.51	9.6
	0.75	602.11	10.91	33.92	321.33	**10.37**	32.50	325.37	484.54	605.26	609.06	12.83	17.49	33.34
	1	364.32	**9.12**	16.43	143.38	9.20	16.21	141.48	194.65	331.48	352.85	15.69	25.41	56.29
	1.25	92.53	7.80	10.32	36.19	**7.57**	9.99	34.23	37.66	72.37	88.66	13.55	19.22	36.75
	1.50	31.60	6.64	7.07	13.93	**6.14**	6.64	12.95	13.18	22.73	29.45	10.13	11.82	17.48
0.5	0.25	17.82	3.96	3.13	3.53	3.52	2.89	3.52	4.32	7.17	14.41	2.68	2.27	**2.18**
	0.50	164.12	5.49	5.02	12.77	4.83	4.62	11.80	14.06	113.89	154.68	4.35	**3.93**	4.14
	0.75	390.25	5.15	4.62	9.08	5.13	**4.58**	8.92	14.47	240.30	379.97	5.7	5.52	6.54
	1	141.06	4.82	4.18	6.66	4.89	**4.19**	6.63	9.90	71.57	132.93	6.17	6.14	7.61
	1.25	46.63	4.54	3.80	5.32	4.52	**3.77**	5.25	6.95	22.87	42.04	5.83	5.71	6.81
	1.50	19.92	4.27	3.46	4.32	4.12	**3.34**	4.24	5.21	10.33	17.53	5.15	4.88	5.46
1	0.25	17.81	2.99	2.13	1.96	2.83	1.99	1.91	2.69	2.82	10.99	1.96	1.66	**1.50**
	0.50	86.69	2.99	2.05	1.99	2.98	2.06	**1.98**	2.76	2.88	18.09	2.53	2.13	2.02
	0.75	30.66	2.99	2.02	1.98	2.99	2.06	**1.98**	2.59	2.62	8.18	2.86	2.43	2.32
	1	13.30	2.98	2.01	1.96	2.98	2.03	**1.95**	2.43	2.42	5.19	2.95	2.51	2.41
	1.25	7.75	2.95	1.99	1.91	2.95	2.00	**1.91**	2.30	2.25	3.94	2.88	2.46	2.35
	1.50	5.23	2.91	1.97	1.87	2.87	1.97	**1.83**	2.20	2.11	3.17	2.75	2.32	2.22

Table 5. *ARL values of the Max-SE-CUSUM, SS-SE-CUSUM, CUSUM-MLE-Max and CUSUM-LR charts:* $(\theta_0, \lambda_0) = (10,100)$, $n = 5$ *and* $ARL_0 = 500$.

			Max-SE-CUSUM			*SS*-SE-CUSUM			CUSUM-MLE-Max			CUSUM-LR		
θ_1	λ_1	SEMLE-Max	$k = 0.5$	$k = 1$	$k = 1.5$	$k = 0.5$	$k = 1$	$k = 1.5$	$k = 0.5$	$k = 1$	$k = 1.5$	$k = 0.5$	$k = 1$	$k = 1.5$
5	60	2.91	**2.86**	2.89	2.89	2.87	2.87	2.86	2.91	2.91	2.92	98.16	137.9	172.2
	80	3.70	3.69	3.71	3.72	**3.67**	3.68	3.72	3.81	3.68	3.71	618.5	518.3	495.7
	100	4.48	4.55	4.53	4.56	4.49	**4.46**	4.46	4.56	4.53	4.49	1153.1	804.8	701.3
	120	5.19	5.19	5.18	5.29	5.23	5.21	**5.14**	5.25	5.22	5.16	663.5	526.7	481.5
7	60	4.45	4.25	4.26	4.36	**4.21**	4.27	4.37	4.49	4.52	4.39	74.4	113.7	146.9
	80	5.79	5.73	5.71	5.77	**5.69**	5.77	5.81	5.70	5.84	5.74	446	431	431
	100	7.09	7.22	**6.93**	7.01	7.04	7.06	6.98	7.12	7.09	7.18	835.6	671.6	617.9
	120	8.17	8.01	8.20	8.12	8.05	8.14	**7.93**	8.10	8.08	8.18	475.1	437.2	424.1
10	60	247.64	13.75	28.74	91.85	**13.56**	28.31	92.83	138.79	215.02	244.78	51.12	83.73	117.5
	80	535.73	75.96	225.8224	393.33	**73.90**	219.73	401.95	557.07	541.24	539.78	226.8	321	350.4
	120	180.89	65.69	106.58	142.50	**62.18**	102.14	138.66	137.99	168.25	176.48	286.9	326.2	247.8
13	60	311.72	**14.25**	30.56	103.45	14.63	32.0	106.82	290.24	296.55	309.16	36.25	61.45	93.06
	80	795.24	**87.11**	287.92	552.56	90.93	302.96	573.33	1362.08	848.00	791.12	157.3	235.9	282.7
	100	632.46	349.13	567.42	612.91	347.37	565.93	616.40	903.12	662.72	629.02	**283.1**	369.8	405.6
	120	186.97	50.29	95.69	139.25	**47.90**	94.25	137.49	172.96	176.98	180.89	171.1	243.8	284.2
15	60	311.65	**14.16**	30.44	103.07	14.81	31.22	108.37	291.99	297.85	310.20	29.73	50.05	79.8
	80	787.72	**85.17**	279.76	539.72	88.09	295.12	565.0	1315.52	856.02	790.21	111	296.2	350.7
	100	608.94	183.8	468.99	572.12	**175.83**	479.06	579.0	824.77	638.59	604.20	192.5	188.5	243.1
	120	181.07	38.49	88.33	133.16	**36.46**	84.70	131.44	158.11	164.91	178.73	121.3	195.5	243.3
25	60	311.05	14.01	29.99	101.86	**13.68**	31.20	104.92	263.29	294.77	301.66	14.52	20.41	35.04
	80	736.28	24.18	215.83	499.88	**22.96**	211.13	519.61	959.47	762.12	735.46	28.34	57.03	104.32
	100	485.79	**17.64**	104.31	348.35	17.73	103.81	353.13	460.00	479.59	481.85	36.84	85.67	152.72
	120	150.39	13.24	34.30	91.21	**12.80**	32.34	87.83	97.89	130.83	146.13	29.62	61.64	107.34
35	60	309.67	10.85	27.64	103.42	9.41	25.89	103.56	200.63	283.75	300.91	**9.34**	10.75	16.05
	80	656.23	10.52	28.82	358.65	**10.29**	28.17	349.41	510.17	657.88	647.77	13.76	20.07	40.13
	100	364.32	**9.10**	16.23	141.42	9.22	16.29	139.52	189.64	321.97	354.29	15.73	25.58	56.66
	120	119.64	8.00	11.15	44.96	**7.89**	10.89	44.51	49.76	96.67	116.33	14.17	20.83	42.2

Table 6. *EARL values for the SEMLE-Max, Max-SE-CUSUM, SS-SE-CUSUM, CUSUM-MLE-Max and CUSU-LR charts:* $n = 5$, $ARL_0 = 500$.

Chart	k	Scenario 1	Scenario 2	Scenario 3	Scenario 4
SEMLE-Max		83.78	283.65	26.33	26.33
Max-SE-CUSUM	0.5	7.80	154.91	**103.64**	26.37
	1	14.024	225.15	169.50	33.80
	1.5	29.26	264.46	225.26	39.93
SS-SE-CUSUM	0.5	**7.59**	**154.79**	105.22	**23.22**
	1	13.90	225.24	173.47	28.72
	1.5	29.29	265.78	229.66	32.99
CUSUM-MLE-Max	0.5	47.72	336.46	400.34	26.60
	1	61.20	299.67	296.73	26.34
	1.5	77.25	299.04	282.97	26.30
CUSUM-LR	0.5	11.186	272.95	350.52	442.48
	1	14.40	327.89	373.72	435.05
	1.5	18.57	355.17	388.63	436.70

5. ILLUSTRATIVE EXAMPLE

In this section, we present an illustrative example for the practical implementation of the Max-SE-CUSUM and SS-SE-CUSUM charts using a dataset from [Mukherjee et al. (2015)]. The data set refers to mileage of military personnel carriers that failed in service and consists of 20 samples each of size $n = 5$. The first 10 samples come from an IC process modeled by a shifted exponential distribution with parameters $(\theta_0, \lambda_0) = (162, 836)$. The subsequent 10 samples come from an OOC process, modeled as a shifted exponential distribution with parameters $(\theta_1, \lambda_1) = (324, 209)$. It's worth noting that the mean level for the IC process is 998, while it drops to 533 in the OOC situation.

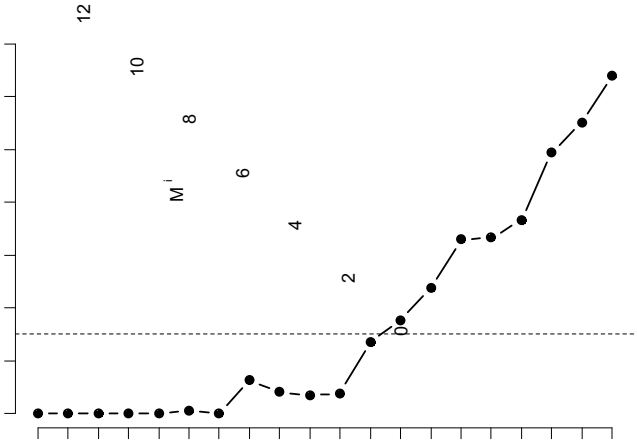
Let the desired IC ARL value be 500 ($ARL_0 = 500$). Using Table 1 and computing the OOC ARL values of the Max-SE-CUSUM and SS-SE-CUSUM charts for $(\theta_1, \lambda_1) = (324, 209)$, we conclude that the minimum values are 3.12 and 3.20, respectively, corresponding to $k = 1$, $h_M = 3.01$ and $h_S = 3.069$. The values of the statistics $W_i^{(1)}$, $W_i^{(2)}$, C_i^+, C_i^-, S_i^+, S_i^-, M_i, N_i, Q_i and SS_i for $i = 1,2, \ldots, 20$ are presented in Table 7.

Table 7. *Values of the statistics associated with the Max-SE-CUSUM and SS-SE-CUSUM charts for the mileage data set.*

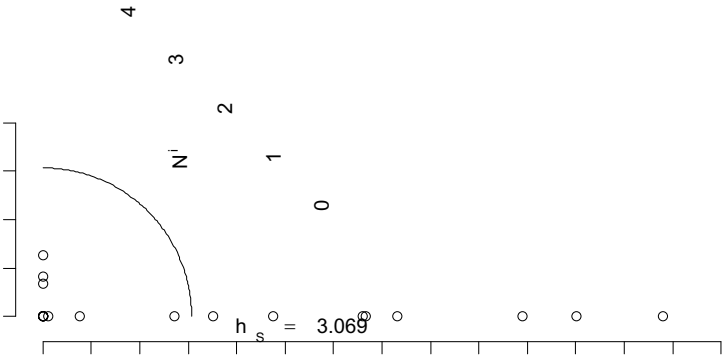
i	$W_i^{(1)}$	$W_i^{(2)}$	C_i^+	C_i^-	S_i^+	S_i^-	M_i	N_i	Q_i	SS_i
1	-0.0449	-0.1743	0	0	0	0	0	0	0	0
2	-0.3545	0.0953	0	0	0	0	0	0	0	0
3	0.6707	-0.4793	0	0	0	0	0	0	0	0
4	-0.2854	0.5694	0	0	0	0	0	0	0	0
5	-0.6210	-0.2359	0	0	0	0	0	0	0	0
6	0.9083	-1.1216	0	0	0	0.121	0.121	0	0.121	0.015
7	-0.2386	0.4226	0	0	0	0	0	0	0	0
8	-2.2606	-0.6507	0	1.260	0	0	1.260	1.260	0	1.588
9	-0.5566	-0.1010	0	0.817	0	0	0.817	0.817	0	0.667
10	-0.8643	0.8254	0	0.68	0	0	0.681	0.681	0	0.464
11	0.4214	-1.7684	0	0	0	0.768	0.768	0	0.768	0.590
12	0.4487	-2.9386	0	0	0	2.707	2.707	0	2.707	7.328
13	**0.3709**	**-1.8173**	**0**	**0**	**0**	**3.524**	**3.524**	**0**	**3.524**	**12.419**
14	0.5585	-2.2219	0	0	0	4.746	4.746	0	4.746	22.525
15	0.3766	-2.8488	0	0	0	6.595	6.595	0	6.595	43.494
16	0.4487	-1.0636	0	0	0	6.659	6.659	0	6.659	44.342
17	0.3363	-1.6514	0	0	0	7.310	7.310	0	7.310	53.436
18	0.6518	-3.5774	0	0	0	9.887	9.887	0	9.887	97.753
19	0.5433	-2.1330	0	0	0	11.020	11.020	0	11.020	121.440
20	0.3304	-2.7764	0	0	0	12.797	12.797	0	12.797	163.763

The Max-SE-CUSUM and SS-SE-CUSUM charts are presented in panels (a) and (b) of Figure 1, respectively. We observe that both charts produce an OOC signal for the first time at the 13th sample. By applying the follow-up procedure described in Section 2, we conclude that both charts show a decrease in process scale, since $S_{13}^{-} > 3.01$ for the Max-SE-CUSUM chart and $S_{13}^{-} > 3.069$ for the SS-SE-CUSUM chart.

Figure 1. *The proposed CUSUM charts for mileage data*



(a). Max-SE-CUSUM ($k = 1$)



(b). SS-SE-CUSUM ($k = 1$)

6. CONCLUSIONS

In this work, we proposed and studied two CUSUM single control charts, based on a Max-type statistic, for monitoring a shifted exponential process. The performance of the charts is evaluated in terms of ARL and $EARL$, which are computed via Monte Carlo simulation.

The numerical comparisons in terms of ARL revealed that the proposed charts are more powerful than the competitive charts SEMLE-Max, CUSUM-MLE-Max and CUSUM-LR in most OOC cases. Since there is no chart that outperforms the remaining ones for all the considered shifts in origin and scale, we suggest the use of the SS-SE-CUSUM chart for the detection of small magnitude shifts (either increasing or decreasing) in both parameters. Also, we suggest the use of the Max-SE-CUSUM chart for detecting small increases in origin and moderate to large decreases in scale.

In terms of $EARL$ the SS-SE-CUSUM chart with reference parameter $k = 0.5$ is the chart with the best overall performance for the considered scenarios except from the case where both parameters may shift to any direction; in this case the Max-SE-CUSUM chart has the best overall performance. Finaly, topics for future research consist of the case of SS-SE-CUSUM and Max-SE-CUSUM chart with estimated parameters, as well as the study of the respective EWMA-type charts.

ΠΕΡΙΛΗΨΗ

Η εκθετική κατανομή είναι η πιο διαδεδομένη κατανομή για την περιγραφή του χρόνου που απαιτείται για να συμβεί ένα ενδεχόμενο. Ωστόσο, σε πολλές περιπτώσεις η παράμετρος αρχής της μπορεί να είναι μεγαλύτερη από το μηδέν. Για παράδειγμα, τα προϊόντα συνήθως συνοδεύονται από χρόνο εγγύησης γεγονός που διασφαλίζει ότι αποτυχίες μπορούν να συμβούν μόνο μετά από αυτό το χρόνο. Σε αυτές τις περιπτώσεις χρησιμοποιείται συνήθως μια διπαραμετρική (ή μετατοπισμένη) εκθετική κατανομή για να μοντελοποιηθεί η διάρκεια ζωής του προϊόντος. Στην παρούσα εργασία, προτείνουμε δύο νέα διαγράμματα ελέγχου τύπου CUSUM για την από κοινού παρακολούθηση των παραμέτρων αρχής και κλίμακας της μετατοπισμένης εκθετικής κατανομής με βάση τους εκτιμητές μεγίστης πιθανοφάνειας των παραμέτρων. Παρουσιάζεται αναλυτικά η λειτουργία των προτεινόμενων διαγραμμάτων ελέγχου ενώ δίνονται και εμπειρικοί κανόνες για την επιλογή των τιμών των παραμέτρων τους. Για την αξιολόγηση της εντός και εκτός ελέγχου απόδοσης των προτεινόμενων διαγραμμάτων ελέγχου χρησιμοποιείται το μέσο μήκος ροής και το αναμενόμενο μέσο μήκος ροής. Ο υπολογισμός τους γίνεται μέσω προσομοίωσης Monte Carlo. Τα αποτελέσματα της αριθμητικής μελέτης δείχνουν ότι τα προτεινόμενα διαγράμματα CUSUM υπερτερούν έναντι των διαγραμμάτων τύπου Shewhart. Επιπρόσθετα, δίνονται αριθμητικές συγκρίσεις μεταξύ των προτεινόμενων διαγραμμάτων και άλλων ανταγωνιστικών διαγραμμάτων τύπου CUSUM. Τέλος, παρουσιάζεται μια εφαρμογή των προτεινόμενων διαγραμμάτων ελέγχου σε πραγματικά δεδομένα.

Acknowledgment

This work has been partly supported by the University of Piraeus Research Center.

REFERENCES

Antzoulakos, D.L., Fountoukidis, K.G., and Rakitzis, A.C. (2025) A Max-type run sum control chart for monitoring a shifted exponential process. *Quality and Reliability Engineering International (accepted)*. https://doi.org/10.1002/qre.3773

Baranwal, A., Kumar, N., Chatterjee, K., and Koukouvinos, C. (2024). A new EWMA chart for simultaneously monitoring the parameters of a shifted exponential distribution. *Journal of Applied Statistics*, **52**(1), 221–252.

Baten, M. A., and Kamil, A. A. (2009). Inventory management systems with hazardous items of two parameter exponential distribution". *Journal of Social Sciences*, **5**(3), 183-187.

Cheng, S.W., and Thaga, K. (2006). Single variables control charts: an overview. *Quality and Reliability Engineering International*, **22**, 811-820.

Cheng, S.W., Thaga, K. (2010). The Max-CUSUM Chart. In: Lenz, HJ., Wilrich, PT., Schmid, W. (eds) *Frontiers in Statistical Quality Control 9*. Physica-Verlag HD.

Chen, G., and Cheng, S.W. (1998). Max chart: combining X-bar chart and *S* chart. *Statistica Sinica*, **8**, 263-271.

Chen, G., Cheng, S.W., and Xie, H. (2001). Monitoring process mean and variability with one EWMA chart. *Journal of Quality Technology*, **33**, 223-233.

Chong, Z.L., Mukherjee, A., and Marozzi, M. (2021). Simultaneous monitoring of origin and scale of a shifted exponential process with unknown and estimated parameters. *Quality and Reliability Engineering International*, **37**, 242–261.

Huang, S., Mukherjee, A., and Yang, J. (2018). Two CUSUM schemes for simultaneous monitoring of parameters of a shifted exponential time to events. *Quality and Reliability Engineering International*, **34**, 1158-1173.

Jalilibal, Z., Amiri, A., and Khoo, M.B.C. (2022). A literature review on joint control schemes in statistical process monitoring. *Quality and Reliability Engineering International*, **38**, 3270-3289.

Johnson, N.L. and Kotz, S. (1970). *Distributions in Statistics, vol. 1: Continuous Univariate Distributions*. Boston, MA: Houghton Mifflin.

Kao, S. C., (2010). Normalization of the shifted exponential distribution for control chart construction". *Journal of Applied Statistics*, **37**(7), 1067-1087.

Lawless, J. (1982). *Statistical Models and Methods for Lifetime Data*. New York, NY: John Wiley & Sons.

Li, Q., Yang, J., Huang, S., and Zhao, Y. (2022). Generally weighted moving average control chart for monitoring two-parameter exponential distribution with measurement errors. *Computers & Industrial Engineering*, **165**, 107902.

Mukherjee, A., McCracken, A.K., and Chakraborti, S. (2015). Control charts for simultaneous monitoring of parameters of a shifted exponential distribution. *Journal of Quality Technology*, **47**, 176-192.

Mukherjee, A., Li, Q., and Song, Z. (2022). An assessment of the effect of using different mappings and Minkowski distances in joint monitoring of the time-between-event processes. *Journal of Computational and Applied Mathematics*, **404**, 113776.

Saha, S., Mukherjee, A., and Khoo, M.B.C. (2022). Two CUSUM schemes for simultaneous monitoring of unknown parameters of a shifted exponential process and its application in monitoring of call durations in telemarketing. *Quality Technology & Quantitative Management*, **19**(1), 113-137.

Santiago, E., and Smith, J. (2013). Control charts based on the exponential distribution: Adapting runs rules for the *t* chart. *Quality Engineering*, **25**(2), 85-96.

Xie M., Goh T.N., and Ranjan P. (2002). Some effective control chart procedures for reliability monitoring. *Reliability Engineering & System Safety*, **77**(2), 143-150.

Greek Statistical Institute
Proceedings of the 37th Panhellenic & 3d International Stat Conf (2025), pp.297-312

A TRIVARIATE COMPOSITIONAL FAY-HERRIOT MODEL FOR SMALL AREA ESTIMATION OF TREE DIAMETER DISTRIBUTION

Aristeidis Georgakis[1], Vasileios E. Papageorgiou[2], Georgios Stamatellos[3]
Aristotle University of Thessaloniki, PC-54124, Thessaloniki
School of Forestry and Natural Environment, Laboratory of Forest Biometry[1,3],
e-mail: arisgeorg@for.auth.gr[1], stamatel@for.auth.gr[3]
Department of Mathematics, e-mail: vpapageor@math.auth.gr[2]

ABSTRACT

Accurate small area estimation (SAE) of tree diameter distributions is crucial for forest management inventories, yet traditional models often require large sample sizes and are rarely applied to subpopulations. This study introduces the Trivariate Compositional Fay-Herriot (TriCompFH) model, which leverages compositional data and auxiliary satellite-derived and census variables to provide reliable small area estimates of tree diameter distributions. Unlike conventional univariate approaches, the TriCompFH model captures correlations among diameter classes, reducing estimation uncertainty. By integrating remote sensing data and clustering techniques, it enhances estimation in small areas with minimal field samples. The proposed methodology supports data-driven forest management by offering scalable, robust estimates of diameter distributions, facilitating sustainable decision-making.

Keywords: Small area estimation, Multivariate Fay-Herriot model, Compositional data, Remote sensing, Tree diameter distribution, Forest management, Correlation analysis

1. INTRODUCTION

Forest inventories provide spatially detailed assessments that are critical for effective forest management, offering insights into stand structure, productivity, and biodiversity. Among these, the diameter at breast height (DBH) distribution is a key forest structure measure for guiding decision-making, silvicultural treatments, stock/carbon estimation and sustainable planning. However, accurately estimating DBH distributions at fine spatial scales —such as individual forest management units (FMUs)— remains challenging when sample sizes are limited. Traditional methods, like the Weibull distribution, are commonly used to model DBH distributions at the stand level. Yet, these methods require a minimum number of sample plots per area to yield reliable estimates and become computationally impractical when applied

separately to many small areas. Consequently, Weibull-type models are often unsuitable for forest inventories with sparse data.

Small area estimation (SAE) techniques offer a robust alternative by leveraging auxiliary information to generate reliable estimates in domains (areas) with small sample sizes [Rao and Molina (2015); Γεωργάκης, (2024)]. Conventional SAE approaches could model each DBH class independently, where each class corresponds to a predefined diameter range. However, this approach ignores the inherent interdependence among classes, as the proportions of trees across DBH classes must collectively sum to one. Compositional data analysis —wherein data are represented as non-negative vectors summing to a whole [Coenders and Pawlowsky-Glahn (2020)]− provides a natural framework for modeling DBH distributions. Aitchison (1982) laid the foundation for such transformations, although applications in forestry have been limited. Previous studies have demonstrated the potential of compositional models in forestry, for example, in assessing species proportions [Huber et al. (2014)], long-term forest dynamics [Kobal et al. (2017)] and biomass estimation [Poudel (2016); Silva Pedro et al. (2017)], yet their use in small area DBH distribution estimation remains underexplored.

Modern SAE methods have evolved to incorporate multivariate and mixed-effects frameworks that jointly model correlated responses [Krause et al. (2022)]. Recent advances, including temporal SAE models [Benavent and Morales (2021); Georgakis et al. (2024a)] and approaches for compositional proportions [Franco and Maitra (2023)], underscore the potential of these techniques. In particular, Georgakis et al. (2024b) addressed the challenge of limited sample sizes in structurally complex, multilayered FMUs with multiple vertical strata (understory, midstory, and overstory) by integrating clustering with multivariate Fay-Herriot (MFH) models.

This study introduces the Trivariate linear mixed-effects Compositional Fay-Herriot (TriCompFH) [Esteban et al. (2020)], for compositional DBH class estimation in clustered FMUs. The model effectively handles small samples, captures inter-class correlations, enhances estimation accuracy, improves computational efficiency, and provides reliable uncertainty estimates via parametric bootstrap techniques. By bridging compositional data analysis with SAE, TriCompFH offers a novel and practical framework for DBH estimation in forest inventories, especially where Weibull-type models are infeasible.

2. MATERIALS AND METHODS

This section outlines the methodology used in the study. First, the study area and data collection are described, followed by the data preprocessing steps applied to auxiliary variables. Finally, the TriCompFH model is presented, focusing on the mathematical formulation of the multivariate Fay-Herriot model and its adaptation for compositional small-area estimation.

2.1 Study Area and Data

The study was conducted in the University Forest of Pertouli, located in Central Greece. The forest, covering 2,260 hectares, is dominated by the hybrid fir species *Abies borisii-regis Mattf.* and is divided into 174 FMUs. After excluding unmanaged FMUs and those with insufficient supporting data, 160 FMUs were included in the analysis. Data collection in 2018 involved a systematic sampling of 239 sample plots, each with an area of 0.1 hectares, corresponding to a sampling intensity of approximately 1%. Roughly half of the FMUs contained one sample plot, while the rest had two. A small number of FMUs had either three plots or none.

The effectiveness of model-based estimators in SAE relies on the availability of informative auxiliary variables. In this study, remote sensing and historical census data were incorporated to improve model-based estimations. Height-related measures were derived from a 2-meter canopy height model (CHM), which was generated from LiDAR-based digital terrain and surface models obtained from WorldView-2 satellite stereo-pairs data. Height variables included percentiles (h25, h50, h75, h90), measures of central tendency such as mean and mode, dispersion measures including standard deviation and coefficient of variation and shape characteristics of the height distribution such as skewness and L-moments (Table 1). Additionally, historical census data were integrated to enhance predictive accuracy. Tree density and volume estimates were obtained from past forest inventories, specifically ForestDensity97 (trees per hectare for all species), DenFir88 (fir tree density, trees per hectare), and ForestGSV97 (forest growing stock volume, m^3/ha) as illustrated in Table 1.

Table 1. *Height and Census Auxiliary Variables*

Data Type	Data Group	Descriptive statistics	Variable Abbreviations	Unit
Height	A	Quantiles	h25; h50; h75; h90	Meters (m)
Height	A	Central Tendency	hmean; hmode	m
Height	B	Dispersion Measures	hsd; hcv	m; Dimensionless
Height	B	Distribution's Shape	hLskew, L4	Dimensionless
Census	B	Central Tendency	ForestDensity97	Trees/hectare (ha)
Census	B	Central Tendency	ForestGSV97	m^3/ha
Census	B	Central Tendency	DenFir88	Trees /ha

To improve model robustness and flexibility all auxiliary variables were transformed using logarithmic and squared functions, allowing for alternative representations of the data that better capture their relationships with small-area estimates.

2.2 Data Preprocessing

This section provides a detailed description of the data preprocessing pipeline, which consists of the following steps: *a) Cluster Analysis:* Initially, FMUs were grouped based on three-dimensional vegetation height data and tree density statistics from previous forest censuses to provide enhanced correlation between response and auxiliary variables. To determine the optimal number of clusters, multiple clustering techniques were evaluated based on forest characteristics in small areas. The final clustering approach, as proposed by Georgakis et al. (2023), utilized the mean tree height as the grouping variable, investigating cluster formations ranging from 2 to 50 clusters of FMUs. The Euclidean distance metric was employed to construct the distance-dissimilarity matrix, while hierarchical single-link clustering was used for grouping. The Friedman and Rubin index [Friedman and Rubin (1967)] determined the optimal number of clusters, resulting in 32 groups, of which eight contained only a single FMU. The remaining 24 clusters (domains) had an average of 9.63 FMUs per cluster.

b) Response Variables: The selection of dependent variables is crucial for the accuracy and reliability of SAE. In this study, the dependent variables of interest were the tree density across four DBH classes, expressed as a proportion of total tree count per class. Increasing the number of DBH classes improves the precision of tree diameter distribution estimates across small areas. While the methodology can accommodate multiple dependent variables, their number must be carefully selected to balance model accuracy and computational feasibility. For this study, four DBH classes were defined, each comprising approximately 25% of the total tree population across the 239 sample plots (Table 2). The TriCompFH model was applied to predict three out of the four diameter classes, while the remaining class (c2) was inferred as the complement of the three estimated proportions [Esteban et al. (2020)]. The exclusion of c2 from direct estimation was based on its weaker correlation with auxiliary variables.

c) Estimation of Class Proportions (Percentages): Following DBH class definition, tree density proportions were estimated as follows: $\text{Proportion per diameter class} = \frac{\text{Total trees in class}}{\text{Total trees in domain}}$ similar to Egozcue and Pawlowsky-Glahn (2019). After grouping, each domain contained an average of 579 trees per hectare (ha), equating to approximately 145 trees/ha per class (Table 2).

Table 2. *Average number of trees per DBH (diameter at breast height) class based on 239 sampled plots (per hectare, ha)*

DBH classes (in cm)	c1 (Class I) [8-12]	c2 (Class II) [14-20]	c3 (Class III) [22-34]	c4 (Class IV) [>35-]	Sum
Trees/ha	164	148	155	112	579
Percentage (%)	28.33%	25.56%	26.77%	19.34%	100%

d*) Transformation and Selection of Covariates:* To improve model robustness, all auxiliary variables underwent logarithmic and squared transformations, providing

alternative representations that better capture relationships with small area estimates. The selection of auxiliary variables was performed using stepwise selection based on the Akaike Information Criterion (AIC) in a univariate empirical best linear unbiased prediction Fay-Herriot (EBLUP-FH) model. To avoid multicollinearity, we apply the methodology deployed by Georgakis et al. (2024b) for EBLUP-FH [Marhuenda et al. (2014)].

e) Smoothing of Covariances: To mitigate variance instability, a variance and covariance smoothing procedure was applied following Goerndt et al. (2011) and Georgakis et al. (2024a), correspondingly. Weighting was performed based on small area sample size, ensuring stability in both variances and covariances of predictor and response variables.

f) Outlier Detection and Correlation Analysis: A robust correlation analysis was conducted to ensure strong linear relationships between auxiliary and response variables. The analysis also identified three distinct categories of outliers, each adversely affecting model performance: i) Extreme response variable outliers: Observed in domains with only two sample plots, causing excessively high direct estimates and weakening predictor-response correlations. ii) Computationally problematic outliers: Resulted from extreme (smoothed) variances or covariances, leading to convergence issues. These cases primarily occurred in domains with only two sample plots. iii) Model assumption violations: Outliers exhibiting non-normal residuals or random area effects, again concentrated in domains with two sample plots. After removing 30 outliers, a final dataset of 209 sample plots (87.3% of the original data) was retained. Among the 24 clusters, 15 remained viable for small area modeling, with an average of 13.9 sampling units per group. Similarly, out of 152 FMUs, 131 (86.19%) were included in the final estimation. The remaining 13.81% of FMUs were addressed through alternative estimation methods.

g) Estimation of Small Areas Without a Sample: For FMUs lacking direct observations or classified as extreme outliers, a synthetic estimator was employed, leveraging information from similar areas. FMUs were grouped into two broad clusters using the Silhouette method [Rousseeuw (1987)] based on direct percentage estimates as clustering variables. The selection of only two large groups maximized coverage by avoiding leaving some small areas unassociated with sampled domains.

h) Transformation of Compositional Data: Following initial small area estimates, an additive log-ratio transformation (*alr*) was applied to direct proportion estimates [Aitchison (1982)] to facilitate the TriCompFH modeling.

i) Model Adaptation and Statistical Software: The TriCompFH model was implemented using EBLUP. Statistical analysis was conducted in R programming language utilizing the "sae.prop" package [Sholihin and Sumarni (2022)] and the function "mseFH.ns.mprop". This function provides parametric Bootstrap MSE estimates for EBLUPs in sampled areas. For unsampled FMUs, MSE estimates were

obtained using a customized method [Haris and Ubaidillah (2020)] within a Type-1 MFH [Anisa et al. (2014)].

2.3 The Trivariate Compositional Fay-Herriot Model (TriCompFH)

As we previously mentioned, the objective of the present paper is the estimation of the tree distribution based on their stem diameter, through a multivariate SAE methodology. More specifically, the examined dataset contains four tree diameter categories, whose proportions should be determined inside a total of D domains U_d, $d = 1, \dots, D$. In what follows, we describe the general structure of the TriCompFH model.

Let N be the size of the population under study, which is partitioned into D domains of size N_d, $d = 1, \dots, D$. The response variables are the domain proportions $\bar{Z}_{dr} = \frac{Z_{dr}}{N_d}$, where $Z_{dr} = \sum_{i \in U_d} z_{dri}$ are the respective counts for $d = 1, \dots, D$ and $r = 1, \dots, 4$. It is evident that obtaining the population characteristics is almost impossible, leading to the sample proportions $\bar{z}_{dr} = \frac{z_{dr}}{n_d}$, where $z_{dr} = \sum_{i \in s_d} z_{dri}$, while $z_{dri} = 1$, if the i-th tree from domain U_d belongs to the category r and $z_{dri} = 0$ otherwise. For any sample $s \subset U$ extracted from the population, s_d denotes the subsample from U_d of size n_d and w_{dj} represents the sampling weight of the unit j from the subsample s_d. Direct estimators of the population proportions and counts can be generated after sampling surveys, namely

$$\hat{\bar{Z}}_{dr}^{dir} = \frac{\hat{Z}_{dr}^{dir}}{\hat{N}_d^{dir}}\,,\ \hat{Z}_{dr}^{dir} = \sum_{i \in s_d} w_{di} z_{dri}\,,\ d = 1, \dots, D \text{ and } r = 1, \dots, 4\,, \qquad (1)$$

$\hat{N}_d^{dir} = \sum_{i \in s_d} w_{di}$, while the sampling weights w_{di} are generated from the inverses of the inclusion probabilities according to the utilized sampling design. Since the sample sizes n_{dr} are usually quite small, the direct estimates (1) derived from the sampling design result in large design-based variances. Hence, the implementation of model-based approaches for the estimation of the examined proportions is necessary.

Based on the above, we propose the utilization of the TriCompFH model, since the investigated response variables sum to unity, i.e. $\bar{z}_{d1} + \bar{z}_{d2} + \bar{z}_{d3} + \bar{z}_{d4} = 1, d = 1, \dots, D$. For the implementation of the multivariate FH model, a transformation is needed from the $(r-1)$ −dimensional composition space to $\mathbb{R}^{r-1}$. Namely, we employ the additive logistic transformation (*alr*), where

$$y_{dr} = alr(\bar{\mathbf{z}}_d) = \log\left(\frac{\bar{z}_{dr}}{\bar{z}_{d4}}\right) = \log\left(\frac{\bar{z}_{dr}}{1 - \bar{z}_{d1} - \bar{z}_{d2} - \bar{z}_{d3}}\right), \qquad (2)$$

for $r = 1,2,3$. Equivalent transformations, such as the isometric log-ratio (*ilr*) or the centered log-ratio (*clr*) transformation can be also used. In this analysis, the *alr*

transform of equation (2), is used according to the experimental results of Esteban et al. (2020).

The MFH represents a mixed linear model defined in two stages. First, the "sampling model" can be described as

$$\begin{pmatrix} y_{d1} \\ y_{d2} \\ y_{d3} \end{pmatrix} = \begin{pmatrix} m_{d1} \\ m_{d2} \\ m_{d3} \end{pmatrix} + \begin{pmatrix} \varepsilon_{d1} \\ \varepsilon_{d2} \\ \varepsilon_{d3} \end{pmatrix}, d = 1, \dots, D, \tag{3}$$

where the sampling errors follow a three-dimensional trivariate normal distribution, namely $\boldsymbol{\varepsilon}_d \sim N_3(\mathbf{0}, \boldsymbol{V}_{\varepsilon d})$. The elements of the 3×3 covariance matrix are considered known in the case of FH modeling, an assumption that is often unrealistic in real scenarios. As a result, matrix $\boldsymbol{V}_{\varepsilon d}$ is estimated from the covariance matrix $\widehat{\boldsymbol{cov}}(\bar{\boldsymbol{z}}_d)$ based on a sample drawing. The elements of this matrix are calculated according to

$$\widehat{cov}(\bar{z}_{dk}, \bar{z}_{dl}) = \frac{1}{(\hat{N}_d^{dir})^2} \sum_{j \in s_d} w_{dj}(w_{dj} - 1)(z_{dkj} - \bar{z}_{dk})(z_{dlj} - \bar{z}_{dl}).$$

where we have employed the notation $\bar{z}_{dk} = \hat{\bar{Z}}_{dk}^{dir}$ for simplicity. Vector $\boldsymbol{m}_d = (m_{d1}, m_{d2}, m_{d3})^T$, is linearly related to a set of auxiliary variables derived from older census or remote sensing data. Thus, we obtain the "linking model" that connects the components of $\boldsymbol{m}_d$ with the matrix of auxiliary variables $\boldsymbol{X}_d = diag(\boldsymbol{x}_{d1}, \boldsymbol{x}_{d2}, \boldsymbol{x}_{d3})$, through the vector of coefficients $\boldsymbol{b}_d = vec(\boldsymbol{b}_1, \boldsymbol{b}_2, \boldsymbol{b}_3)$. The linking model for the trivariate case is presented by

$$\begin{pmatrix} m_{d1} \\ m_{d2} \\ m_{d3} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_{d1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{x}_{d2} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{x}_{d3} \end{pmatrix} \begin{pmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \\ \boldsymbol{b}_3 \end{pmatrix} + \begin{pmatrix} u_{d1} \\ u_{d2} \\ u_{d3} \end{pmatrix}, \tag{4}$$

or

$$\begin{pmatrix} m_{d1} \\ m_{d2} \\ m_{d3} \end{pmatrix} = \begin{pmatrix} x_{d11} & \cdots & x_{d1p_1} & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & x_{d21} & \cdots & x_{d2p_2} & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & \cdots & 0 & x_{d31} & \cdots & x_{d3p_3} \end{pmatrix}_{3 \times p} \begin{pmatrix} b_{11} \\ \vdots \\ b_{1p_1} \\ b_{21} \\ \vdots \\ b_{2p_2} \\ b_{31} \\ \vdots \\ b_{3p_3} \end{pmatrix}_{p \times 1} + \begin{pmatrix} u_{d1} \\ u_{d2} \\ u_{d3} \end{pmatrix}, \tag{5}$$

for $d = 1, \dots, D$, where $p = p_1 + p_2 + p_3$, shows the number of the auxiliary variables used for m_{d1}, m_{d2} and m_{d3}. The vector of random effects also follows a normal (Gaussian) distribution of zero mean, namely $\boldsymbol{u}_d \sim N_3(\mathbf{0}, \boldsymbol{V}_{ud})$.

Combining the multivariate equations (3)–(5), we obtain the area-level random effect model (6), where y_{dr} is modeled as a linear combination of $\boldsymbol{x}_{dr}$, of a random effect u_{dr} of the specific small area and a sampling error ε_{dr}, for $r = 1, \dots, 3$. Therefore, the area-level random effect model is given by

$$\begin{pmatrix} y_{d1} \\ y_{d2} \\ y_{d3} \end{pmatrix} = \begin{pmatrix} x_{d11} & \cdots & x_{d1p_1} & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & x_{d21} & \cdots & x_{d2p_2} & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & \cdots & 0 & x_{d31} & \cdots & x_{d3p_3} \end{pmatrix}_{3\times p} \begin{pmatrix} b_{11} \\ \vdots \\ b_{1p_1} \\ b_{21} \\ \vdots \\ b_{2p_2} \\ b_{31} \\ \vdots \\ b_{3p_3} \end{pmatrix}_{p\times 1} + \begin{pmatrix} u_{d1} \\ u_{d2} \\ u_{d3} \end{pmatrix} + \begin{pmatrix} \varepsilon_{d1} \\ \varepsilon_{d2} \\ \varepsilon_{d3} \end{pmatrix}. \quad (6)$$

The covariance matrix is provided by

$$\boldsymbol{V}_d = \boldsymbol{Z}_d^T \boldsymbol{V}_{ud} \boldsymbol{Z}_d + \boldsymbol{V}_{\varepsilon d},$$

[Benavent and Morales, (2016)]. For the univariate case FH, both matrices $\boldsymbol{V}_{\varepsilon d}$ and $\boldsymbol{V}_{ud}$ have a diagonal form. At the area-level, equation (6) can be described in matrix form as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\varepsilon}. \quad (7)$$

This form represents the integration of D area-level models. According to model (7), the expected value of the transformed proportions is given by $E[\boldsymbol{y}] = \boldsymbol{X}\boldsymbol{b}$, while the covariance matrix of $\boldsymbol{y}$, can be obtained through

$$\boldsymbol{V} = var(\boldsymbol{y}) = \boldsymbol{Z}^T \boldsymbol{V}_u \boldsymbol{Z} + V_{\varepsilon}. \quad (8)$$

The covariance matrices of sampling errors and random effects, $\boldsymbol{V}_{\varepsilon}$ and $\boldsymbol{V}_u$ respectively, have a block diagonal form, assuming independence between sampling errors and random effects across domains.

To estimate the response variables for the examined domains we utilize the multivariate EBLUP, as an iterative estimation of the covariance matrix $\boldsymbol{V}_u$ is required. The estimation of the 6 elements associated with $\boldsymbol{V}_u$, (3 variance and 3 covariance components) are produced by an iterative residual maximum likelihood (REML) algorithm, resulting in the construction of the Fisher information matrix. The results of this algorithm are later incorporated for the estimated covariance matrix of random effects $\widehat{\boldsymbol{V}}_u$. The estimated covariance matrix participates in the computation of $\widehat{\boldsymbol{V}}$, as $\widehat{\boldsymbol{V}} = \widehat{\boldsymbol{V}}_u + \boldsymbol{V}_{\varepsilon}$. More details about the REML algorithm can be found in [Benavent and Morales, (2016)]. The EBLUP estimate of $\boldsymbol{m} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{Z}\boldsymbol{u}$, is given by

$$\widehat{\boldsymbol{m}}_E = \boldsymbol{X}\widehat{\boldsymbol{b}}_E + \boldsymbol{Z}\widehat{\boldsymbol{u}}_E\,,$$

where

$$\widehat{\boldsymbol{b}}_E = \left(\boldsymbol{X}^T \widehat{\boldsymbol{V}}^{-1} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \widehat{\boldsymbol{V}}^{-1} \boldsymbol{y}\,,$$

and

$$\widehat{\boldsymbol{u}}_E = \widehat{\boldsymbol{V}}_u \boldsymbol{Z}^T \widehat{\boldsymbol{V}}^{-1} \left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{b}}_E\right).$$

The respective estimate of the covariance matrix based on the EBLUP estimators is produced by

$$\widehat{\boldsymbol{V}} = \boldsymbol{Z}^T \widehat{\boldsymbol{V}}_u \boldsymbol{Z} + \boldsymbol{V}_{\varepsilon}.$$

We note that the variances and covariances of the examined response variables are a necessary input of the TriCompFH model, utilized for the construction of the $\boldsymbol{V}_{\varepsilon d}$ matrix. As a result, the transformation of the $\widehat{\boldsymbol{cov}}(\bar{\boldsymbol{z}}_d)$ matrix is requisite. The first derivatives of the *alr* function are

$$Alr_{rr}(\bar{\boldsymbol{z}}_d) = \frac{\partial alr_r(\bar{\boldsymbol{z}}_d)}{\partial \bar{z}_{dr}} = \frac{1}{\bar{z}_{dr}} + \frac{1}{1-\bar{z}_{d1}-\bar{z}_{d2}-\bar{z}_{d3}},$$

$$Alr_{rs}(\bar{\boldsymbol{z}}_d) = \frac{\partial alr_r(\bar{\boldsymbol{z}}_d)}{\partial \bar{z}_{ds}} = \frac{1}{1-\bar{z}_{d1}-\bar{z}_{d2}-\bar{z}_{d3}}, r \neq s.$$

Hence, we generate the covariance matrix $\boldsymbol{Alr}(\bar{\boldsymbol{z}}_d) = \left(Alr_{ij}(\bar{\boldsymbol{z}}_d)\right)_{i,j=1,\dots,3} = diag_{1\leq r\leq 3}\left(\frac{1}{\bar{z}_{dr}}\right) + \left(\frac{1}{1-\bar{z}_{d1}-\bar{z}_{d2}-\bar{z}_{d3}}\right)\mathbf{1}_3\mathbf{1}_3^T$, where $\mathbf{1}_3$ is the 3×1 vector of ones. Thus, for the covariance matrix of random effects, we use

$$\boldsymbol{V}_{\varepsilon d} = \widehat{\boldsymbol{cov}}(\boldsymbol{y}_d) = \boldsymbol{Alr}(\bar{\boldsymbol{z}}_d)\widehat{\boldsymbol{cov}}(\bar{\boldsymbol{z}}_d)\boldsymbol{Alr}(\bar{\boldsymbol{z}}_d)^T.$$

Finally, since the original response variables are proportions, the last step of the compositional FH methodology includes the implementation of the inverse *alr* transform, resulting in $\hat{p}_{dr} = alr^{-1}(\hat{m}_{d1}, \hat{m}_{d2}, \hat{m}_{d3}) = \frac{e^{\hat{m}_{dr}}}{1+e^{\hat{m}_{d1}}+e^{\hat{m}_{d2}}+e^{\hat{m}_{d3}}}$, for $r = 1, 2, 3$. The last proportion can be retrieved by $\hat{p}_{d4} = 1 - \sum_{r=1}^{3}\hat{p}_{dr}$. The estimated proportions derived from the application of the inverse *alr* function can be used for the computation of the mean square errors (MSE) and random effects, for the evaluation of the estimation efficiency of the TriCompFH.

3. RESULTS

This section presents a comprehensive description of the findings derived from the statistical analysis. Initially, the focus is on verifying the assumptions of the multivariate model, including diagnostics for its adequacy, such as assessing the normality of residuals, normality of random effects, homoskedasticity and potential bias in indirect estimates. The results of the small area estimations are subsequently presented. After data processing and handling of outliers, the percentage distributions were found to follow a normal distribution without requiring any transformation. Following the stepwise auxiliary variable selection procedure within the EBLUP-FH framework and minimizing the AIC criterion, the covariates were determined without indication of multicollinearity (Equation 9) [Georgakis et al. (2024b)]. The logarithmic and squared transformation of the auxiliary variables yielded superior model fit in the univariate FH framework, fulfilling the assumptions of normality for both residuals and random effects. Table 3 presents the estimated regression coefficients of the TriCompFH model for the three diameter classes (c1, c3, c4). The calculation of MSEs was performed using 400 bootstrap iterations.

$$\begin{pmatrix} y_1 \\ y_3 \\ y_4 \end{pmatrix} \sim \begin{pmatrix} \log(L4)+h25^2 \\ \log(L4)+h25^2 \\ \text{hLskew}^2 \end{pmatrix} \quad (9).$$

Table 3. *EBLUP coefficient of the multivariate compositional Fay-Herriot model for the three diameter classes c1, c3, c4*

Diameter Class	Variable	β (Estimate)	Standard Error	t-statistic	p-value
c1 [8-12] cm	Intercept (β1)	3.728	0.597	6.245	4.23E-10
c1 [8-12] cm	log(L4)	1.331	0.233	5.704	1.17E-08
c1 [8-12] cm	$h25^2$	-0.006	0.001	-6.639	3.17E-11
c3 [22-34] cm	Intercept (β3)	-1.595	0.352	-4.531	5.87E-06
c3 [22-34] cm	log(L4)	-0.606	0.141	-4.307	1.65E-05
c3 [22-34] cm	$h25^2$	0.003	0.000	6.893	5.47E-12
c4 [>35] cm	Intercept (β4)	-0.341	0.039	-8.691	3.59E-18
c4 [>35] cm	$hLskew^2$	514.988	40.465	12.727	4.20E-37

Before proceeding with the final estimates, statistical assumptions were evaluated to ensure model adequacy. The Shapiro-Wilk test confirmed the normality of residuals and random effects, supporting the model's assumptions. Q-Q plots provided a visual assessment of normality, while the Breusch-Pagan test confirmed homoskedasticity. Figure 1 compares direct and model-based estimates to assess potential estimation bias [Brown et al. (2001)]. The close alignment of the regression line with the bisector ($y = x$) suggests negligible biases in model-based compared to direct sample-based estimates for all classes. To further verify model fit, the Wald test was applied to compare direct and model-based estimates. The results confirmed no statistically significant difference between the two estimate types. For all three DBH classes, the χ^2 test statistics were below the critical value of 24.996 (df = 15, α = 0.05), indicating no statistically significant difference between model-based and direct estimates.

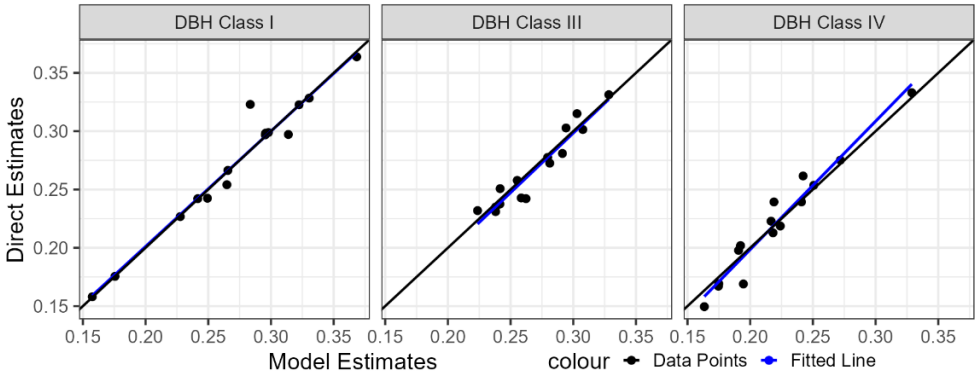


Figure 1. *Comparison of direct and model-based percentage class estimates (c1,c3,c4)*

Figure 2 compares direct and TriCompFH estimates (percentages), showing narrower confidence intervals with increasing sample size, indicating improved precision. The model's intervals consistently include direct estimates.

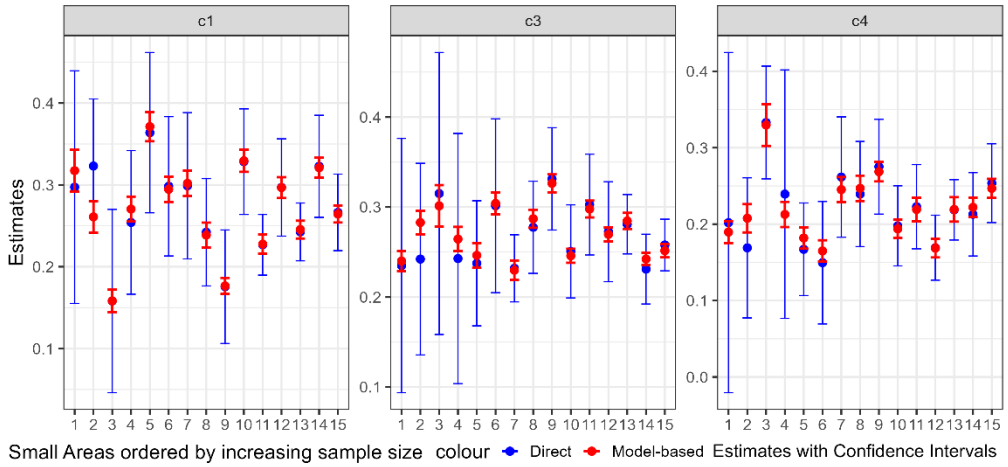


Figure 2. *Estimates of percentages with confidence intervals for all small areas and diameter classes (c1, c3, and c4) comparing direct and model-based estimates. The x-axis displays the sample size per small area.*

To evaluate improvements in small area estimates, the relative efficiency (RE) was calculated as follows: $RE_r = \left(\frac{RMSE\left(\hat{p}_{d,r}^{TriCompFH}\right)}{RMSE\left(\hat{p}_{d,r}^{dir}\right)} \right) \times 100\%$, for $r = 1, 3, 4$. Figure 3 illustrates the RE_r distribution across all small area estimates and classes. RE_r values below 1 indicate that estimates based on TriCompFH are more effective across the three diameter classes (Class I, III, IV), with lower values reflecting more significant improvement over direct (dir) estimates.

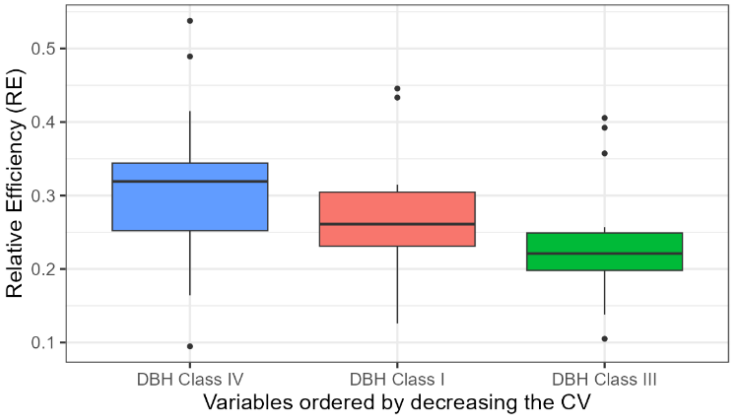


Figure 3. *Relative Efficiency (RE) distribution of small area estimates (model-based vs. direct) for diameter classes (c1, c3, c4)*

Table 4 presents the coefficient of variation (CV) distribution of small area means, assessing the reliability of estimates. The TriCompFH model exhibits substantially lower CV values compared to direct estimates, reinforcing their reliability. The table highlights that all TriCompFH estimates have a CV below 10%, with most below 5%. Direct estimates (Dir) exceed 5% CV, with the majority surpassing 10% CV and over 33% with CVs above 15%, require careful interpretation.

Table 4. *Distribution of CV% of the mean for the three classes (c1, c3, c4), comparing direct (Dir) and trivariate compositional Fay-Herriot model (TFH) estimates*

CV threshold %	CV DIR c1	CV Dir c3	CV Dir c4	CV TFH c1	CV TFH c3	CV TFH c4
0-5	0	0	0	12	14	8
5.1 - 10	4	7	1	3	1	7
10.1 - 15	6	3	8	0	0	0
>15.1	5	5	6	0	0	0

Emphasis is placed on the TriCompFH model's reliability in estimating tree percentage distributions for three diameter classes (c1, c3, c4). The fourth class (c2) is derived by subtracting the sum of the proportions of the three classes from unity for each small area. Since the TriCompFH model's proportion estimates are statistically validated, they are used to estimate the frequency of trees per diameter class. The final tree density estimates (number of trees per hectare) are calculated by multiplying the estimated percentages from the TriCompFH model by the tree density per hectare (obtained from direct estimates based on the sample design). Figure 4 presents these final estimates across all classes, including c2.

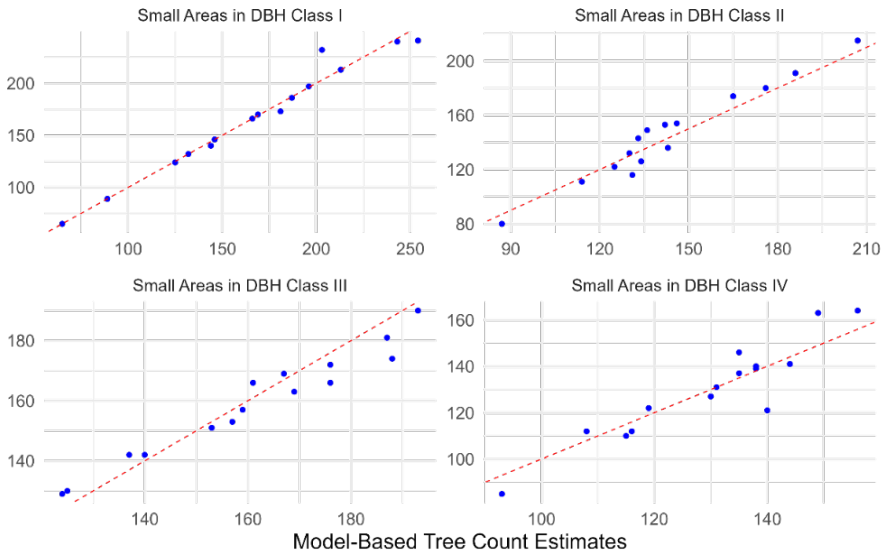


Figure 4. *Direct (y) vs Model-based (x) tree density estimates for four DBH c1-4*

4. DISCUSSION

Accurate SAE of tree diameter distributions is crucial for forest management, yet traditional methods struggle with high uncertainty, especially when sample sizes are limited. While previous studies focus on population-level estimates, this study introduces the TriCompFH model, designed to estimate multiple DBH class distributions across subpopulations like FMUs. Unlike conventional univariate models, TriCompFH preserves compositional constraints and exploits correlations among DBH classes, enhancing predictive reliability.

The model integrates auxiliary satellite-derived canopy height variables, demonstrating remote sensing's potential in forest inventories. Initially applied to estimate tree percentages in three diameter classes (with the fourth as the remainder), these were converted into tree counts per class. For practical forest management, merging the first two classes allows for estimating stem frequency in three broader classes, providing tree density per small area.

TriCompFH utilizes 3D canopy height data from WorldView satellite images and tree census data, with preprocessing ensuring model reliability. Small area estimates for DBH class percentages were converted to stem density. To address areas with 0–3 sample plots, clustering enlarged small areas and optimized correlations between auxiliary and response variables for improved estimates [Georgakis et al. (2023)]. With larger samples, clustering could be omitted, enabling more small areas while maintaining accuracy.

Despite its advantages, TriCompFH faces computational challenges as the number of response variables increase. Extreme values can cause convergence issues, requiring variance smoothing and outlier detection. The bivariate FH model estimated two diameter classes across 18 small areas, but the trivariate model, though successful in 16 areas, required removing an additional group to meet normality assumptions. As multivariate model complexity grows, robust preprocessing becomes essential.

Future research could refine tree density estimates by replacing the design-based with univariate Fay-Herriot estimates and explore more diameter classes (e.g., six classes or 10 cm intervals) in areas with larger sample size for finer stand structure representation. Bayesian hierarchical models may reduce preprocessing needs, while integrating spatial dependencies could enhance applicability where small areas are not pre-clustered [Georgakis et al. (2023)].

5. CONCLUSION

This study presents the Trivariate Compositional Fay-Herriot (TriCompFH) model as a novel approach for SAE of tree diameter distributions. Unlike traditional methods that require large sample sizes to model a detailed distribution, this methodology enables the estimation of multiple diameter distributions within subpopulations, such as FMUs. By integrating compositional (percentages) constraints, (optional) clustering analysis and auxiliary remote sensing data, the model provides reliable small area

estimates when sample based direct estimates are not feasible. By delivering practical estimates of tree density distributions, aiding forest management decisions on stand structure and sustainable harvesting. Although computational challenges arise with increasing model complexity, careful data preprocessing and methodological refinements ensure model robustness. Future research could explore Bayesian hierarchical models, refine diameter class segmentation, and incorporate spatial dependencies to extend the model's applicability. By bridging the gap between traditional field-based inventory methods and advanced statistical modeling, this research offers a scalable and efficient solution for forestry practitioners, enabling data-driven decision-making in sustainable forest management.

ΠΕΡΙΛΗΨΗ

Η ακριβής εκτίμηση για μικρές περιοχές (Small Area Estimation, SAE) των κατανομών διαμέτρου των δέντρων είναι κρίσιμη για τις δασικές απογραφές που έχουν σκοπό τη διαχείριση των δασών, ωστόσο τα παραδοσιακά μοντέλα συχνά απαιτούν μεγάλο μέγεθος δείγματος, και συνήθως δεν εφαρμόζονται σε υποπληθυσμούς. Η παρούσα μελέτη παρουσιάζει το τριμεταβλητό συνθετικό μοντέλο Fay-Herriot (Trivariate Compositional Fay-Herriot, TriCompFH), το οποίο αξιοποιεί συνθετικά δεδομένα και βοηθητικές μεταβλητές από δορυφορικά δεδομένα και δασικές απογραφές, με σκοπό την παραγωγή αξιόπιστων εκτιμήσεων για μικρές περιοχές των κατανομών διαμέτρου των δέντρων. Σε αντίθεση με τα μονομεταβλητά μοντέλα, το TriCompFH αξιοποιεί τις συσχετίσεις μεταξύ των εξαρτημένων μεταβλητών των στηθιαίων διαμέτρων, μειώνοντας την αβεβαιότητα εκτίμησης. Ενσωματώνοντας δεδομένα τηλεπισκόπησης και τεχνικές ομαδοποίησης, ενισχύεται η εκτίμηση σε μικρές περιοχές με ελάχιστα επίγεια δείγματα. Η προτεινόμενη μεθοδολογία υποστηρίζει την τεκμηριωμένη διαχείριση των δασών, παρέχοντας κλιμακούμενες και αξιόπιστες εκτιμήσεις κατανομών διαμέτρου, διευκολύνοντας έτσι τη λήψη αειφορικών αποφάσεων.

Acknowledgements: The authors would like to sincerely thank the anonymous reviewer for the thorough and constructive comments, which led to a significant improvement of this paper and to a more accurate presentation of the mathematical notation.

REFERENCES

Aitchison, J. (1982). The Statistical Analysis of Compositional Data. *Journal of the Royal Statistical Society: Series B (Methodological)*, **44**(2), 139-160.

Anisa, R., Kurnia, A., & Indahwati, I. (2014). Cluster Information of Non-Sampled Area in Small Area Estimation. *IOSR Journal of Mathematics*, **10**, 15-19.

Benavent, R., & Morales, D. (2021). Small area estimation under a temporal bivariate area-level linear mixed model with independent time effects. *Statistical Methods & Applications*, **30**(1), 195-222.

Brown, G., Chambers, R., Heady, P. and Heasman, D. (2001). Evaluation of small area estimation methods—an application to unemployment estimates from the UK LFS. In *Proceedings of Statistics Canada Symposium.* S. Canada (ed.), Statistics Canada, Ottawa, Canada, pp. 1-10.

Coenders, G., & Pawlowsky-Glahn, V. (2020). On interpretations of tests and effect sizes in regression models with a compositional predictor. *SORT-Statistics and Operations Research Transactions*, **201**, 201-220.

Egozcue, J.J., & Pawlowsky-Glahn, V. (2019). Compositional data: The sample space and its structure. *TEST*, **28**(3), 599-638.

Esteban, M.D., Lombardía, M.J., López-Vizcaíno, E., Morales, D., & Pérez, A. (2020). Small area estimation of proportions under area-level compositional mixed models. *TEST*, **29**(3), 793-818.

Franco, C., & Maitra, P. (2023). Combining surveys in small area estimation using area-level models. *WIREs Computational Statistics*, **n/a**(n/a), e1613.

Friedman, H.P., & Rubin, J. (1967). On Some Invariant Criteria for Grouping Data. *Journal of the American Statistical Association*, **62**(320), 1159-1178.

Georgakis, A., Gatziolis, D. and Stamatellos, G. (2023) A Primer on Clustering of Forest Management Units for Reliable Design-Based Direct Estimates and Model-Based Small Area Estimation. *Forests*, **14** (10).

Georgakis, A., Papageorgiou, V.E., Gatziolis, D. and Stamatellos, G. (2024a) Temporal-Like Bivariate Fay-Herriot Model: Leveraging Past Responses and Advanced Preprocessing for Enhanced Small Area Estimation of Growing Stock Volume. *Operations Research Forum*, **5** (1), 9.

Georgakis, A., Papageorgiou, V.E., & Stamatellos, G. (2024b). A new approach to small area estimation: improving forest management unit estimates with advanced preprocessing in a multivariate Fay–Herriot model. *Forestry: An International Journal of Forest Research*, cpae061, 18.

Goerndt, M.E., Monleon, V.J., & Temesgen, H. (2011). A comparison of small-area estimation techniques to estimate selected stand attributes using LiDAR-derived auxiliary variables. *Canadian Journal of Forest Research*, **41**(6), 1189-1201.

Haris, F., & Ubaidillah, **A.** (2020). Mean Square Error of Non-Sampled Area in Small Area Estimation. *Proceedings of the 1st International Conference on Statistics and Analytics (ICSA 2019)*, Bogor, Indonesia, pp. 383-394.

Huber, M.O., Sterba, H., & Bernhard, L. (2014). Site conditions and definition of compositional proportion modify mixture effects in *Picea abies – Abies alba* stands. *Canadian Journal of Forest Research*, **44**(10), 1281-1291.

Kobal, M., Kastelec, D., & Eler, K. (2017). Temporal changes of forest species composition studied by compositional data approach. *iForest - Biogeosciences and Forestry*, **10**(4), 729-738.

Krause, J., Burgard, J.P., & Morales, D. (2022). Robust prediction of domain compositions from uncertain data using isometric logratio transformations in a penalized multivariate Fay–Herriot model. *Statistica Neerlandica*, **76**(1), 65-96.

Marhuenda, Y., Morales, D. and del Carmen Pardo, M. (2014). Information criteria for Fay–Herriot model selection. *Computational Statistics & Data Analysis*, **70**, 268-280.

Poudel, K.P. (2016). Methods for estimating aboveground biomass and its components for Douglas-fir and lodgepole pine trees. *Canadian Journal of Forest Research*, **46**(1), 77-87.

Rao, J.N.K., & Molina, I. (2015). *Small Area Estimation* (2nd ed.). Wiley, Hoboken, NJ.

Rousseeuw, P.J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, **20**, 53-65.

Sholihin, M.R., & Sumarni, C. (2022). *sae.prop: Small Area Estimation using Fay-Herriot Models with Additive Logistic Transformation.*

Silva Pedro, M., Rammer, W., & Seidl, R. (2017). Disentangling the effects of compositional and structural diversity on forest productivity. *Journal of Vegetation Science*, **28**(3), 649-658.

Γεωργάκης, Α. (2024). *Εκτιμήσεις μικρής έκτασης σε ανομήλικα δασικά οικοσυστήματα, με ομαδοποίηση των δασικών διαχειριστικών μονάδων και πολυμεταβλητά Fay-Herriot μοντέλα.* Διδακτορική διατριβή, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης, σ. 192.

Greek Statistical Institute
Proceedings of the 36th Panhellenic & 2nd International Stat Conf (2024), pp.313-332

Skill competitiveness and concentration in the Software Engineering market using prioritization approaches

Konstantinos Georgiou[1], Nikolaos Mittas[2] and Lefteris Angelis[1]
[1]School of Informatics, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece
[2]Department of Chemistry, School of Science, Democritus University of Thrace, 65404, Kavala, Greece
konsgeor@csd.auth.gr, nmittas@chem.duth.gr, lef@csd.auth.gr

ABSTRACT

The labour market is a complex ecosystem where different sectors have different demands and requirements regarding the necessary skills. The emergence of different positions with different responsibilities, in combination with the growing presence of AI in the working landscape has led to a constantly alternating landscape of demanded knowledge and skills. However, instead of adapting to this fluctuating situation and adjusting their job roles in accordance with global developments, companies frequently post job advertisements with an overwhelming and irrelevant number of skills that do not reflect the responsibilities of a job. Hence, prioritizing skills is of great importance for the evolving labour market, across different sectors, while the concept of producing competitive workers with up-to-date skills is crucial. In this study, the Hierarchical Cumulative Voting algorithms is combined with the Herfindahl-Hirschman Index, with the twofold goal of prioritizing skills across Software Engineering (SE) jobs and sectors and assess the competitiveness of examined sectors based on their preferred skills. The results extracted from the analysis showcase that different countries and industries demand different skills that range from soft skills pertaining to communication and collaboration to hard skills that concern technical knowledge and expertise. In addition, the Herfindahl-Hirschman Index unveiled differences among countries and industries, revealing stricter, monopolized or more flexible skill markets. The derived insights indicate that the proposed framework has the potential to pinpoint the skills that a sector requires to remain competitive and relevant in an ever-digitalized world.

Keywords: Hierarchical Cumulative Voting, Herfindahl-Hirschman Index, Skill Competitiveness, Prioritization, Market Concentration

1. INTRODUCTION

The profiling of the labour market, regardless of the examined sector, is a challenging task that requires constant vigilance and observation of global developments and the impact they may have on the required skills for specific positions (Gangl, 2002; Gallie 2017; Ross et al., 2024). Defining the necessary knowledge and competencies that a modern worker should possess in order to be a competitive asset for a company is frequently fuzzy, due to unforeseen circumstances, insufficient training and learning strategies within the corporate environment as well as the increasing gap between the labour market demand (i.e. the skills that companies require from the workers) and the labour supply (i.e. the skills that active and future workers provide to the company) (Cappelli et al., 2015; McGuiness et al., 2016; Rikala et al., 2024).

In recent years, the clear identification of skills that may hold higher priority over others for specific positions becomes increasingly difficult due to the emergence, rise and normalized use of AI systems (Lane et al., 2021; Shioshira et al., 2021). Large Language Models (LLMs) such as ChatGPT, DeepSeek or Bard are continuously being integrated into the pipelines of companies and organizations, facilitating and automating multiple tasks. This inevitably transforms the landscape of work, as several mundane or technical skills may end up being a lower priority for employers, who seek competences beyond the tasks and processes that can be achieved by AI (Tokunova et al., 2023, Zubair et al., 2024;). Hence, monitoring the labour market and tracking the skills that are more highly demanded in specific sectors is crucial for designing upskilling strategies, mitigating the impact of AI in an effort to introduce "human-centric" job positions that also take into account the prowess of LLMs as well as mapping skills that are more generic and thus refer to general knowledge to more specific that may be tailored to each individual (Leng et al., 2022; Huang et al., 2022; Xu et al., 2021).

Furthermore, assigning priorities and detecting highly demanded skills allows for a holistic overview of the labour market and the discovery of sectors where few skills are considered a priority among employers (thus indicating a closed, less flexible skills environment) and sectors where there are multiple skills on demand (thus pointing to a competitive and flexible skills ecosystem which allows different skills to flourish). Combining the process of identifying sector-specific skills that hold higher importance, with the outlining of skill ecosystems that vary in different dimensions (e.g. countries or timeframes) can be a valuable tool for policymakers and stakeholders that could assess different skill

requirements in different projected axes, formulating indicative strategies and policies.

Motivated by these considerations, and in the context of SKILLAB, a Horizon Europe funded project aimed at monitoring the European Union labour market, in this study we introduce a data driven framework that leverages online job vacancies (OJA) data across the EU-27 member states in order to detect primary skills that are present in job descriptions (Kavargyris et al., 2025), while leveraging ongoing research on labour market skill demand from previous years (Georgiou et al., 2021). Leveraging the Hierarchical Cumulative Voting (HCV) algorithm, the introduced mechanism is able to assign priorities in specific skills, detecting skill combinations that characterize the skill demand of a sector. It should be noted that this framework has been designed as in to operate for any given dimension, allowing it to move beyond specific sectors and explore skill priorities for different countries, timeframes and industries. Moreover, the framework leverages the assigned priorities in order to gauge whether skill demand in specific and targeted groups (e.g. skills of a country) is flexible and competitive or whether stricter circumstances are observed, with specific skills dominating. To achieve this, we leverage the Herfindahl-Hirschman Index (Rhoades et al., 1993) for market concentration, transferring its operation to calculate skills diversity in a specific dimension.

Overall, our framework constitutes a collective and data-driven mechanism for profiling skill demand and competitiveness across the EU labour market, providing rich insights for policymakers, industrial stakeholders and educational institutions for the development of skills-centric EU policies, the identification of important skills for in-company training as well as the design of labour market aware curricula that truly capture the demands and intricacies of the working landscape.

In order to showcase the usability and adaptability of the framework, a case study on OJAs pertaining to Software Engineering (SE) is presented, demonstrating the most demanded skills along with projections of skill diversity for different countries, industries and sectors. Hence, the objectives of this study are twofold and are expressed via the following two research questions (RQs), tailored to the case study that will be presented:

RQ_1: What skills are prioritized in different dimensions?

Motivation: In this RQ the primary focus is to identify skills that hold higher importance in SE OJAs across the EU. To support this objective, the HCV algorithm is applied on skills extracted from OJAs, under a structured taxonomy of skills, ranging from more generic to more specific. The motivation behind RQ_1 is the necessity to understand which skills are more important across sectors and across levels of significance, in an effort to properly capture the needs of the labour market and design targeted strategies for upskilling and reskilling.

RQ_2: Are SE "markets" competitive or monopolized?

Motivation: In the second RQ, and by taking into account the skill importance extracted from the previous RQ, the goal is to assess whether the skill ecosystems (meaning a combination of important skills) that characterize a sector (in this study the sector of SE) are competitive or monopolized. In other words, we want to examine whether multiple skills are allowed to thrive and be considered assets for an active or potential worker or if specific skills monopolize a skill ecosystem and hence should be the main focus of individuals that seek to enter the labour market in this sector. To that end, we introduce the term of a **"skill market"** as a concentration of skills in a specific dimension (e.g. a country), while in essence we attempt to assess **"skill competitiveness"**, which expresses whether a skill market has multiple prioritized skills or few.

It should be emphasized that, while for the purposes of interpretation and demonstration, the described methodology is applied on SE OJAs, it can be applied on any dedicated sector to extract insights on skills priorities and skill competitiveness.

2. BACKGROUND INFORMATION

2.1 The ESCO taxonomy

ESCO is the European Union's standardized classification system for skills and competencies. Available in 27 languages, it provides a consistent way to outline the skills, competencies, and qualifications needed in the EU labor market, ensuring comparability across countries. The classification is regularly updated to reflect evolving labor market trends and demands.

The ESCO taxonomy organizes skills into two primary categories, or **pillars**: the **Knowledge pillar**, covering theoretical subject matter within a field, and the **Skills pillar**, which refers to the practical application of knowledge and abilities in work-related tasks. Skills within the ESCO framework are arranged hierarchically, starting with broad, general skills and branching into more specialized ones. This hierarchy is divided into four levels for each pillar, ranging from the most abstract (Level 1, e.g., management or communication skills) to the most detailed (Level 4, e.g., promoting sustainable practices or collecting biological data).

2.2 Hierarchical Cumulative Voting

The HCV algorithm, introduced by Berander et al. (2020) serves as the core method in our study for prioritizing identified skills and assigning them the designated importance. Initially designed for SE requirement prioritization, HCV addresses the shortcomings of conventional voting techniques (e.g., AHP, cumulative voting), which can be inefficient or ill-suited for hierarchical datasets. As mentioned in Section 2.1, skills are organized in a tree-like structure, with higher-level skills (HLSs) and lower-level skills (LLSs). Hence, the tree is organized in levels, with HLSs being more generic and abstract, while LLSs are more refined and specific. In addition, LLSs are direct children of HLSs. The ESCO skill tree is organized into 4 levels, with Level 1 containing 8 core generic skills, and subsequent levels being more diverse and containing a larger number of skills. Each skill is assigned an initial weight (referred to as the *assigned priority*), derived from its *relative frequency*—the count of job postings containing the skill divided by the total job postings analyzed.

The HCV algorithm processes skills level-by-level, computing *intermediate priorities* for each LLS using:

$$p_i LLS_u = C_{HLS_u} * p_a LLS_u * p_a HLS_u \quad (1)$$

Here, C_{HLS_u} (the *compensation factor*) represents the number of HLSs linked to a given LLS, while the p_a terms denote the assigned priorities (relative frequencies) of the respective LLS and its HHS parent. Thus, for each level, intermediate priorities are calculated for *all* LLSs belonging under several HLS parents. For example, if we consider Level 1 containing 8 core skills, and Level

2 containing 70 skills, $Eq\ 1.$ would calculate 560 different intermediate priorities, one for each different skill.

Finally, intermediate priorities are *normalized* to ensure they sum to 1 within each level, using:

$$p_f LLS_u = \frac{p_i LLS_u}{\sum_k p_i LLS_k} \tag{2}$$

In this context, $p_i LLS_u$ expresses the intermediate priority of a LLS, while $\sum_k p_i\ LLS_k$ represents the sum of *all* $p_i LLS_u$ across an entire tree level. Hence, likewise to $Eq\ 1.$, the calculation of the normalized priorities for each LLS ($p_f LLS_u$) is done for each skill in a specific tree level. The final output is, for each tree level, a *normalized priority value,* for each skill in this specific tree level, indicating its priority compared to similar level skills.

In our application of HCV to the ESCO skill hierarchy, assigned priorities are determined empirically from job ad frequencies rather than expert input, facilitating a data-driven approach across all four ESCO levels.

2.3 Herfindahl-Hirschman Index

The HHI is a widely used measure of market concentration, helping to assess the level of competition within an industry. A common example is, considering a specific sector (e.g. Agriculture), to define a market that contains several firms. Each firm has a market share in percentages, indicating the total turnover of a firm in comparison with the total turnover of the market. Hence, market share expresses the penetration of a company in a market. The HHI is then calculated by summing the squares of the market shares of all firms in the market, expressed as:

$$HHI = \sum_{i=1}^{N} s_i^2 \tag{3}$$

where s_i^2 represents the market share of firm i (expressed as a percentage), and N is the total number of firms. The HHI ranges from close to 0 (indicating perfect competition with many small firms) to 10,000 (reflecting a monopoly, where one firm dominates 100% of the market). Regulatory bodies, such as the U.S. Department of Justice and the Federal Trade Commission, use specific HHI

thresholds to evaluate mergers and antitrust concerns: markets with an HHI below 1,500 are considered *unconcentrated*, those between 1,500 and 2,500 are *moderately concentrated*, and those above 2,500 are *highly concentrated*.

The HHI is applied in economics, finance, and industrial organizations to analyze competition dynamics, market stability, and industry trends (Laine, 1995; Krikva 2016; Djolov, 2013; Johan et al., 2022). A rising HHI suggests increasing market concentration, which may reduce competition and lead to higher prices for consumers. Conversely, a declining HHI indicates greater fragmentation and competitive pressures. Policymakers and researchers also use the index to study labor markets, trade, and even skill diversity in workforce analyses, making it a versatile tool for assessing concentration across different domains.

3. METHODOLOGY

The research methodology applied in this study is separated into several distinct stages that facilitate structured data processing along with efficient skill prioritization and assessment of skill competitiveness. Overall, the applied methodology is comprised of *(i) data collection, (ii) data preprocessing, (iii) skills mapping, (iv) skills prioritization* and *(v) skill competitiveness assessment.* In this section, we present each stage and analyze the steps taken towards its completion.

3.1 Data Collection

The data utilised in this study come from the Web Intelligence Hub (WIH) of the EU, which serves as an observatory of the labour market landscape, accumulating OJA data from all EU countries, including the 27 member states. Via the use of the SKILLAB Tracker, a specialized Application Programming Interface (API) developed in the context of the project, 1.3 million OJAs covering the whole of 2024 were retrieved from WIH. The data contained several features pertaining to the profile of an OJA, including *a) the date that the OJA was posted, b) the industry to which it belongs, c) required experience levels, d) the country of origin* and *e) the associated skills, derived from the ESCO taxonomy.* Our focus on this study is to leverage the skills data, in combination with different dimensions of the dataset (e.g. countries) to extract insights relevant to the two posed RQs. Two major points that should be clarified is that skills are assigned to OJAs based on trained machine learning classifiers by the EU and validated by experts to be as relevant as possible. In

addition, the existence of different industries is translated by the demand for software solutions in many different sectors. Finally, the retrieved OJAs belonged to four different occupations: System Analysts, Application Programmers, Web and Multimedia Developers and Software Developers.

3.2 Data Preprocessing

Following their retrieval, OJA data are stored in a database, while duplicate entries are removed. In addition, skills are treated as a single row in the database, with each job covering as many rows as the skills associated with it.

3.3 Skills Mapping

Utilizing the preprocessed dataset of OJAs, each skill was mapped to the ESCO taxonomy to identify its name and level within the taxonomy. Skills were separated in different pillars (Skills and Knowledge) along with their examined levels (Level 1 being associated with generic skills and Level 4 being associated with domain-specific skills that capture technological knowledge).

3.4 Skills Prioritization

Given that each job was associated with a set of skills and following the mapping of these skills to their levels and pillars. The HCV algorithm was applied on the collective dataset of OJAs, for each occupation and pillar. The output of the algorithm is a table which, for each level of the ESCO taxonomy, presents the level-specific skills, with their assigned priorities. For the purposes of this study, in the Results section we will present the most important skills regardless of their level. Also, for simplicity, only the results for the collective dataset and not for separate occupations will be presented. The complete insights can be found in an external repository. The algorithm was applied for all pillars and across different dimensions, to facilitate the extraction of insights with the HHI index in the next section.

3.5 Skill Competitiveness Assessment

The final step of the methodology is directly associated with the HCV algorithm outcome and concerns the application of the HHI index for specific dimensions. Thus, utilizing the priorities extracted from the HCV algorithm, which

corresponded to skills spread across OJAs that concerned countries or industries, we model each skill as a firm and its priority as a market share. Then, the HHI is calculated for the skills contained in a specific dimension. The extracted values were examined based on the defined threshold to assess the competitiveness or monopoly of specific skill markets.

4. RESULTS

In this section, the results and insights from the methodological approach are presented. To facilitate interpretation and ensure clarity, results are presented as per the posed RQs, with RQ_1 focusing on detecting skill priorities across the ESCO levels and RQ_2 leveraging the assigned priorities of the skills to examine the HHI values.

RQ_1: What skills are prioritized in different dimensions?

In the following Tables, the top skills for each level are presented, along with several noteworthy skills for specific countries and industries.

Table 1. *Top skills per ESCO level - Countries*

Skill	Priority	Level
Working with computers	0.37	1
Communication, collaboration and creativity	0.30	1
Information skills	0.16	1
Management skils	0.13	1
Assisting and caring	0.10	1
Accesing & analyzing digital data	0.15	2
Using digital tools for collaboration	0.13	2
Programming systems	0.12	2
Problem solving	0.11	2
Creating artistic/visual content	0.10	2
Accessing & analyzing digital data	0.22	3

Using collaboration/productivity tools	0.20	3
Processing sound/image content	0.12	3
Programming systems	0.09	3
Artistic design/performance	0.08	3
Computer literacy	0.17	4
Using Microsoft Office	0.16	4
Scripting programming	0.10	4
Object-oriented programming	0.09	4
Analyzing software specifications	0.09	4

Table 2. *Highest skills overall - Countries*

Country	Unique Skill	Priority	Level
Estonia	Manage ICT virtualisation environments	0.31	4
Finland	Using digital tools for sound/image processing	0.18	3
Malta	Management skills	0.31	1
Romania	Brainstorm ideas	0.18	4
Sweden	Communicating with colleagues/client	0.24	3
Sweden	Create solutions to problems	0.17	4

The analysis of prioritized skills across EU countries reveals notable patterns in skill demand. The highest-priority skills in Level 1 are *working with computers* and *communication, collaboration, and creativity*, pinpointing the crucial role of digital literacy and interpersonal competencies in the modern workforce. *Management skills* and *assisting/caring* are less emphasized but remain critical for entry-level roles.

Skills on Level 2 such as *accessing & analyzing digital data* and *using digital tools for collaboration* are dominant, indicating a transition towards data-driven

and collaborative work environments. *Programming systems* and *problem-solving* further indicate the growing integration of technical and analytical capabilities.

For Level 3, accessing *& analyzing digital data* reappears with higher priority, alongside *collaboration/productivity tools*, suggesting a high demand for data proficiency and teamwork technologies. *Processing sound/image content* and *artistic performance* highlight niche creative specializations. Finally, *computer literacy* and *using Microsoft Office*, in Level 4, are always relevant in these positions, but technical competencies like *scripting programming* and *object-oriented programming* also emerge, reflecting the industry needs for software development and systems analysis.

Several EU countries also exhibit unique skill priorities, revealing localized labor market demands. For example, Estonia prioritizes *managing ICT virtualization environments*, justifying its position as a digital leader. Finland has a high demand for *digital tools for sound/image processing*, while Malta focuses on *management skills*. Romania prioritizes *brainstorming ideas*, suggesting innovation-centric roles and Sweden demonstrates focus on *communicating with colleagues/clients* and *creating problem solutions* reflecting its collaborative and problem-solving work culture.

Table 3. *Top skills per ESCO level - Industries*

Skill	Priority	Level
Working with computers	0.34	1
Communication, collaboration and creativity	0.30	1
Information skills	0.17	1
Management skils	0.12	1
Assisting and caring	0.06	1
Accesing & analyzing digital data	0.12	2
Solving problems	0.11	2
Creating visual/instructive materials	0.10	2
Using digital collaboration content	0.10	2
Liaising and networking	0.10	2

Accessing & analyzing digital data	0.19	3
Using collaboration/productivity tools	0.18	3
Communicating with clients and colleagues	0.09	3
Managing, gathering & storing digital data	0.07	3
Creating artistic designs/performance	0.07	3
Computer literacy	0.18	4
Using Microsoft Office	0.17	4
Develop creative ideas	0.09	4
Perform data analysis	0.06	4
Object-oriented programming	0.05	4

Table 4. *Highest skills overall - Industries*

Industry	Unique Skill	Priority	Level
Real Estate Activities	Identify customer's needs	0.87	4
Real Estate Activities	Engaging with others to identify needs	0.40	3
Gambling & Betting	Developing solutions	0.37	3
Leather Manufacture	Advocating for individual/community needs	0.11	3
Households as Employers	Documenting & recording information	0.16	2
Mining Support Services	Setting up computer systems	0.33	3
Postal & Courier Services	Administer ICT systems	0.20	4
Manufacture of Machinery	Selling products or services	0.34	3
Wholesale Trade	Make drawings	0.28	4
Services to buildings and Landscape	Use software design patterns	0.18	4

As far as industries are concerned, in Level 1, *Working with computers* and *communication, collaboration, and creativity* dominate, reinforcing the critical

role of digital literacy and interpersonal skills in entry-level roles. *Information skills* and *management skills* follow, while *assisting and caring* appears less prioritized, suggesting these are niche even at foundational levels.

In Level 2, *accessing & analyzing digital data* and *solving problems* emerge as key important skills, alongside *creating visual/instructive materials* and *using digital collaboration tools*. The inclusion of *liaising and networking* indicates the need for building lasting connections and relationships in workplace environments.

In Level 3, data-driven competencies remain prominent, with *accessing & analyzing digital data* and *using collaboration/productivity tools* having leading positions. *Communicating with clients and colleagues* and *managing digital data* reflect the need for both technical and relationship building with clients.

Finally, Level 4 resembles the landscape in countries, with *computer literacy* and *using Microsoft Office* remain, but *developing creative ideas* and *object-oriented programming* indicate different demands, that combine soft skills with technical knowledge.

By inspecting the highest skills in the overall industries, several patterns are identified. More specifically, real estate appears to focus on skills related to customer engagement and needs identification, which can then be addressed via ICT and SE solutions. Moreover, several industries related to manufacturing and trade (Mining Support Services, Postal & Courier Services, Services to buildings and Landscape) report skills relevant with ICT systems and software design.

RQ_2: Are SE "markets" competitive or monopolized?

Examining the HHI values for the countries, most countries keep the HHI in the range of 1500 – 4000. Notable exceptions are Latvia and Slovenia which exhibit high values of HHI across the four ESCO levels. In Latvia, all levels exhibit a value of 10.000, indicating absolute monopoly with specific skills dominating the market. In contrast, Poland exhibits the lowest HHI value, indicating a flexible and competitive skill market that allows for several skills to flourish and become essential for OJAs.

Figure 1. *HHI Values - Countries*

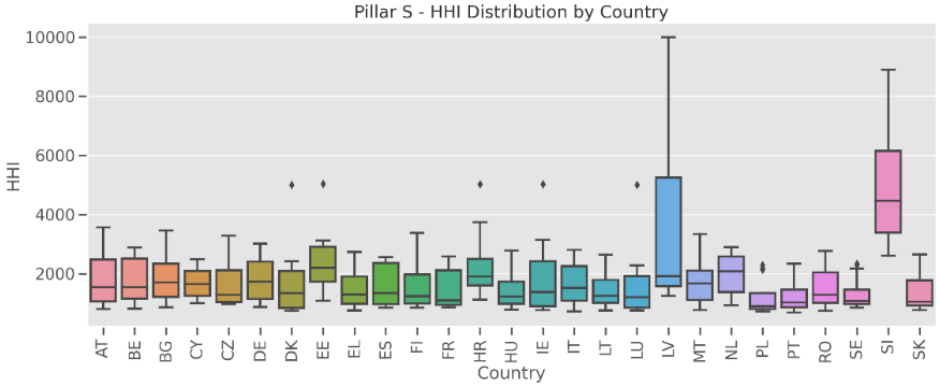


Figure 2. *HHI Values - Industries*

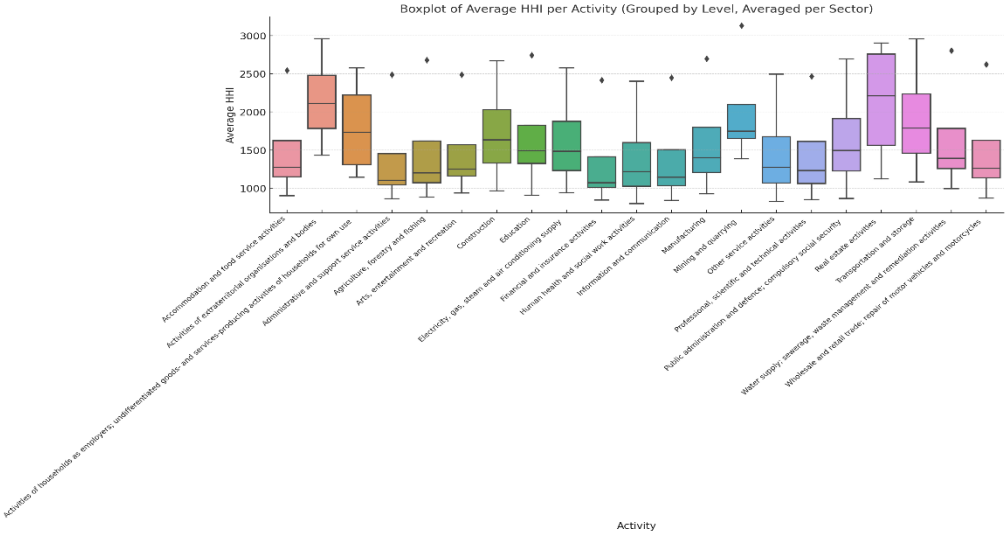


Figure 3. *HHI Values - Industries*

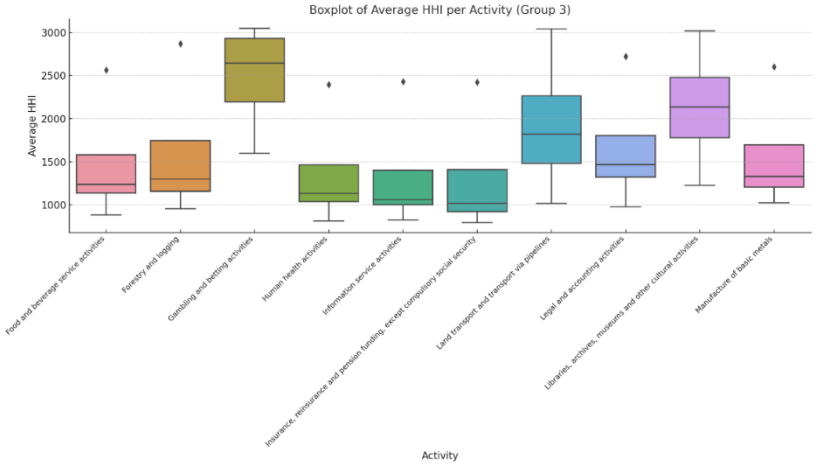


Figure 4. *HHI Values - Industries*

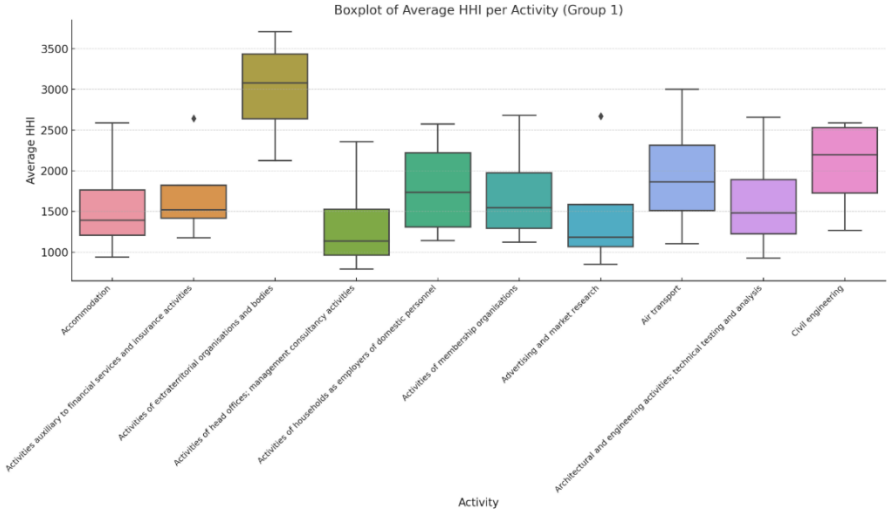
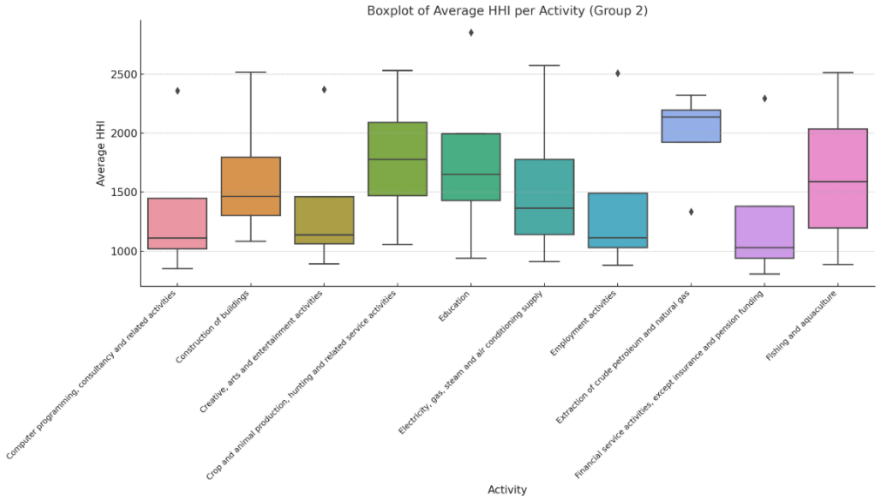


Figure 5. *HHI Values - Industries*



Regarding the examined industries, most HHI values fall into the 1000 – 3000 range, indicating moderately concentrated skill markets. In the following table, some indicative industries are presented.

Table 5. *Indicative HHI values - Industries*

Industry	Average HHI	Max HHI	Min HHI
Postal & Courier	5027	6734	3808
Real Estate	3762	7176	1793
Accommodation	1741	2715	834
Construction	1700	2910	1142
Sports & Recreation	1130	2134	769
Health & Social Work	1223	2295	777

The HHI for specific industries reveals clear differences in the concentration of skill priorities. Postal & Courier (HHI = 5027) and Real Estate (HHI = 3762) showcase the highest concentration, indicating that a small number of core competencies dominate these sectors or that there are sub-sectors that are more diversified or balanced (in the case of Real Estate). In contrast, Accommodation (HHI = 1741) and Construction (HHI = 1700) show a moderate concentration,

indicating a relative balance between technical, managerial, and service-oriented skills. Sports & Recreation (HHI = 1130) and Health & Social Work (HHI = 1223) report the lowest concentration levels, suggesting diversified skill ecosystems with multiple complementary competencies emerging in these sectors. Overall, these findings imply that industries such as Postal & Courier and Real Estate rely on narrower, more specialized skill sets, while sectors like Health, Social Work, and Sports require broader, more interdisciplinary profiles that favor adaptability and transfer of skills across skill ecosystems.

5. DISCUSSION & CONCLUSIONS

In this study, OJAs pertaining to SE were collected from an official EU database with the twofold purpose of not only detecting which skills are primarily demanded from employers, but also to assess whether skills expressed in OJAs from different countries and industries exhibit different levels of competitiveness. To that end, two specialized methodologies, namely the HCV algorithm and the HHI were applied on ESCO skills in order to profile the landscape of EU OJAs in a specific sector.

The results indicate that, based on the level of generality (ranging from generic skills to more refined ones), different countries showcase varied skill demands, with Scandinavian countries illustrating a need for creative thinking and soft skills, while southern countries appear to emphasize on hard skills and IT-oriented knowledge. In addition, as far as industries are concerned, a clear distinction is observed between industries that require technical knowledge of programming and software specifications and socially leaning industries that also require communication, networking and collaboration with customers.

Regarding the HHI insights, most countries and sectors retain a moderate concentration, with few exceptions that showcase strict skill markets or highly flexible ecosystems. Regardless, the application of the HHI in prioritized skills emerges as a prominent method of assessing the competitiveness of different dimensions and exploring localized labour market demands and requirements.

Future directions of this work could involve the application of additional prioritization algorithms, with a comparison of prioritized skills across ESCO levels. Moreover, a variety of market diversity and concentration indicators could be leveraged to further assess skill and skill market competitiveness, compare markets across different industries, while also combining the insights with conclusions and remarks from EU related documents (e.g. policies relevant

to upskilling, reskilling and the green transition) for a holistic overview of the EU labour market.

ΠΕΡΙΛΗΨΗ

Η αγορά εργασίας είναι ένα σύνθετο οικοσύστημα όπου διαφορετικοί κλάδοι και βιομηχανίες έχουν διαφορετικές απαιτήσεις όσον αφορά τις απαραίτητες δεξιότητες. Η δημιουργία σύνθετων θέσεων εργασίας με διαφορετικές αρμοδιότητες, σε συνδυασμό με την αυξανόμενη παρουσία της Τεχνητής Νοημοσύνης στον εργασιακό χώρο, έχει οδηγήσει σε ένα συνεχώς εναλλασσόμενο τοπίο απαιτούμενων γνώσεων και δεξιοτήτων. Ωστόσο, αντί να προσαρμόζονται σε αυτήν την μεταβλητή κατάσταση και να προσαρμόζουν τις αγγελίες εργασίας τους σύμφωνα με τις παγκόσμιες εξελίξεις, οι εταιρείες δημοσιεύουν συχνά αγγελίες εργασίας με μεγάλο και μη σχετικό αριθμό δεξιοτήτων που δεν αντικατοπτρίζουν τις αρμοδιότητες μιας θέσης εργασίας. Ως εκ τούτου, η ιεράρχηση των δεξιοτήτων έχει μεγάλη σημασία για την εξελισσόμενη αγορά εργασίας, σε διαφορετικούς τομείς, ενώ η έννοια της παραγωγής ανταγωνιστικών εργαζομένων με ενημερωμένες δεξιότητες είναι κρίσιμη. Σε αυτή τη μελέτη, ο αλγόριθμος Hierarchical Cumulative Voting συνδυάζεται με τον Δείκτη Herfindahl-Hirschman, με τον διττό στόχο της ιεράρχησης δεξιοτήτων σε όλες τις θέσεις εργασίας στον τομέα της Μηχανικής Λογισμικού και της αξιολόγησης της ανταγωνιστικότητας των εξεταζόμενων τομέων με βάση τις προτιμώμενες δεξιότητές τους. Τα αποτελέσματα που εξήχθησαν από την ανάλυση καταδεικνύουν ότι διαφορετικές χώρες και κλάδοι απαιτούν διαφορετικές δεξιότητες που κυμαίνονται από ήπιες δεξιότητες που σχετίζονται με την επικοινωνία και τη συνεργασία έως τεχνικές δεξιότητες που αφορούν τεχνικές γνώσεις και εξειδίκευση. Επιπλέον, ο Δείκτης Herfindahl-Hirschman αποκάλυψε διαφορές μεταξύ χωρών και κλάδων, αποκαλύπτοντας πιο αυστηρές, μονοπωλιακές ή πιο ευέλικτες αγορές δεξιοτήτων. Τα εξαγόμενα συμπεράσματα δείχνουν ότι το προτεινόμενο πλαίσιο έχει τη δυνατότητα να εντοπίσει τις δεξιότητες που απαιτεί ένας τομέας για να παραμείνει ανταγωνιστικός και σχετικός σε έναν κόσμο που ψηφιοποιείται ολοένα και γρηγορότερα.

Acknowledgments: *The authors acknowledge the funding received from the EU's Horizon Europe Framework Programme SKILLAB under grant agreement No. 101132663.*

REFERENCES

Berander, P., & Jönsson, P. (2006). Hierarchical cumulative voting (hcv)—prioritization of requirements in hierarchies. *International Journal of Software Engineering and Knowledge Engineering*, *16*(06), 819-849.

Cappelli, P. H. (2015). Skill gaps, skill shortages, and skill mismatches: Evidence and arguments for the United States. *ILR review*, *68*(2), 251-290.

Choiri, A. (2025). The Role of Artificial Intelligence (AI) in Economic and Labor Market Transformation. *Digital Marketing, Consumer Behavior, and Economic Trends Journal*, *1*(1), 1-9.

Djolov, G. (2013). The Herfindahl-Hirschman index as a decision guide to business concentration: A statistical exploration. *Journal of Economic and Social Measurement*, *38*(3), 201-227.

Gangl, M. (2002). Changing labour markets and early career outcomes: labour market entry in Europe over the past decade. *Work, employment and society*, *16*(1), 67-90.

Gallie, D. (2017). The quality of work in a changing labour market. *Social Policy & Administration*, *51*(2), 226-243.

Georgiou, K., Mittas, N., Mamalikidis, I., Mitropoulos, A., & Angelis, L. (2021). *Analyzing the roles and competence demand for digitalization in the oil and gas 4.0 era. IEEE Access 9, 151306e151326*.

Huang, S., Wang, B., Li, X., Zheng, P., Mourtzis, D., & Wang, L. (2022). Industry 5.0 and Society 5.0—Comparison, complementation and co-evolution. *Journal of manufacturing systems*, *64*, 424-428.

Johan, S., & Vania, I. (2022, May). The Application of Herfindahl-Hirschman Index in Measuring the Concentration Level of Financial-Technology Industry. In *Tenth International Conference on Entrepreneurship and Business Management 2021 (ICEBM 2021)* (pp. 8-11). Atlantis Press.

Kavargyris, D. C., Georgiou, K., Papaioannou, E., Petrakis, K., Mittas, N., & Angelis, L. (2025). ESCOX: A tool for skill and occupation extraction using LLMs from unstructured text. *Software Impacts*, 100772.

Krivka, A. (2016). On the concept of market concentration, the minimum Herfindahl-Hirschman index, and its practical application. *Panoeconomicus*, *63*(5), 525-540.

Laine, C. R. (1995). The Herfindahl-Hirschman index: a concentration measure taking the consumer's point of view. *The Antitrust Bulletin*, *40*(2), 423-432.

Lane, M., & Saint-Martin, A. (2021). The impact of Artificial Intelligence on the labour market: What do we know so far?. *OECD Social, Employment, and Migration Working Papers*, (256), 0_1-60.

Leng, J., Sha, W., Wang, B., Zheng, P., Zhuang, C., Liu, Q., ... & Wang, L. (2022). Industry 5.0: Prospect and retrospect. *Journal of Manufacturing Systems*, *65*, 279-295.

McGuinness, S., & Ortiz, L. (2016). Skill gaps in the workplace: measurement, determinants and impacts. *Industrial relations journal*, *47*(3), 253-278.

Rhoades, S. A. (1993). The herfindahl-hirschman index. *Fed. Res. Bull.*, *79*, 188.

Rikala, P., Braun, G., Järvinen, M., Stahre, J., & Hämäläinen, R. (2024). Understanding and measuring skill gaps in Industry 4.0—A review. Technological Forecasting and Social Change, 201, 123206.

Ross, A. G., McGregor, P. G., & Swales, J. K. (2024). Labour market dynamics in the era of technological advancements: The system-wide impacts of labour augmenting technological change. *Technology in Society*, *77*, 102539.

Shiohira, K. (2021). Understanding the Impact of Artificial Intelligence on Skills Development. Education 2030. *UNESCO-UNEVOC International Centre for Technical and Vocational Education and Training*.

Tokunova, A., Zvonar, V., Polozhentsev, D., Pavlova, V., & Fedoruk, O. (2024). Economic consequences of artificial intelligence and labor automation: employment recovery, transformation of labor markets, and dynamics of social structure in the context of digital transformation. *Financial Engineering*, *2*, 1-12.

Xu, X., Lu, Y., Vogel-Heuser, B., & Wang, L. (2021). Industry 4.0 and Industry 5.0—Inception, conception and perception. *Journal of manufacturing systems*, *61*, 530-535.

Zubair, S. (2024). AI-Driven Automation: Transforming Workplaces and Labor Markets. *Frontiers in Artificial Intelligence Research*, *1*(3), 373-411.

Greek Statistical Institute
Proceedings of the 37th Panhellenic & 3rd International Stat Conf (2025), pp.333-346

ETHICAL CONSIDERATION FOR THE USE OF DATA IN STATISTICS

Vasileios Ismyrlis
University of the Aegean
vismyrlis@aegean.gr

ABSTRACT

Statistics is a science that needs data to complete its tasks and operations. Nowadays data is collected in many ways and from various sources. Most of the methods utilized collect and use information about human beings. Hence, in all these procedures, certain matters arise about the appropriate use of this data, having ensured the consent and protection of all the participants. Obviously, all the data gathered should be protected and secured, and used only for their initial purpose. In this study, certain instructions, guidelines, and already established measures will be referred to, which cope to ensure that ethical aspects are integrated in the collection, analysis and presentation of data.

Keywords: Statistics, Ethical aspects, Data, Data sources, Data security.

1. INTRODUCTION (THE NEED FOR DATA)

We live in a digital era where data generated from people's interactions is abundant and constantly circulating. Many organizations—both private and public—rely on this data to stay informed about various situations and to develop policies that serve their interests and those of their customers or citizens.

This growing demand for data has led organizations to constantly seek access to the most relevant information.

In addition to direct interactions—such as asking people questions or observing their behavior—researchers now turn increasingly to the Internet and social media, which have become major sources of data. In these cases, data is often collected without direct contact or consent from specific individuals.

These methods result in the accumulation of vast amounts of data, which are then analyzed to produce insights and conclusions. Hence, many questions arise (for the collection, use, sharing, and storage of data):

- Do people know why they are providing the specific information?
- Have people consented to the provision of information?
- Have actions been taken to protect and secure the data collected?
- Are the data presented completely anonymized?
- Is someone responsible for the collected data?
- What can we reveal and what not? E.g., is there any personal information that could not be revealed?
- Are the results presented based on the original data, or are they falsified?

Another major issue that arises is determining who holds ethical responsibility for data. It appears that the key parties involved in this matter include data creators (such as researchers), data providers (suppliers), and data users (Hand, 2018).

However, the primary goal is to establish principles that support the ethical handling of data. Obviously, these principles are intended to guide both current and future uses of data.

2. SOURCES OF DATA

Researchers and statisticians that require data for their research and analysis, should look for data in the following sources:

- Government databases: Hellenic Statistical Organisation (ELSTAT), Bank of Greece (TTE), Institute of SETE-INSETE is a non-profit organisation on the initiative of the Greek Tourism Confederation (SETE), Eurostat
- Academics and research institutions
- Business and financial data: Google Finance, Athens Stock Exchange (XAA)
- Social media and web data: there are numerous
- Data originated from direct interaction with people

Regarding the ways that data are collected, there are two categories of data:

- Primary data is collected firsthand from the sources by researchers using methods such as surveys, interviews, experiments, observations, and focus groups. They are specific to the research purpose.
- Secondary data: already been collected and sometimes published for a different purpose; sources: books, journals, articles, encyclopedias, reports (government, business, news), government records, and online databases.

3. METHODS FOR COLLECTING DATA

There are many methods that researchers use to collect the required data. Some of these methods include communication with people, some of them collect data from digital media and sources, and others are mined through databases. More specifically, some of these methods are:

- Surveys and questionnaires, interviews, observations, experiments, focus groups, case studies
- Web scraping, sensor data collection, social media analytics
- Databases (online or not) and public records mining

The first two methods gather data that is considered primary, so they are collected directly from the researcher for a specific purpose.

The third one concerns secondary data, meaning they have already been collected and they are usually published.

4. HOW TO PROTECT DATA AND PEOPLE

Researchers, statisticians, and organisations need data to analyze and produce results that contribute to making decisions, conducting strategies, etc. (Raman et al, 2023). This indicates that a lot of information is used; therefore, it's critical to take ethical considerations into account when using data.

It is essential to ensure some prerequisites in the collection, use, sharing, and storage of data. Hence, the following guidelines are valid for those who collect, use, analyze, share, and store data. Those responsible for the data should ensure privacy and confidentiality of the personal information, ensure transparency for the ways used to collect the data and for the analysis methods, secure data that is collected, identify the people who are responsible (accountability), provide informed consent when conducting research, establish integrity in the presentation of results and findings (do not try to mislead), avoid potential conflicts and possible bias, consider the social impacts of the project, do not harm, or provoke injustice (social responsibility).

Another important ethical issue is the ease of posting data and generally similar actions that tend to mislead the public (Raman et al., 2023).

Some prompts about the ethical aspects of using datasets from secondary sources are (Raman et al., 2023):

- Do we have permission to use these datasets?
- Did the participants consent to their data being used?
- Who collected the data, and made it publicly available?
- Do the data contain identifiable information?
- Were the variables collected accurately?

- Should we analyze data if we do not know how the data were collected?
- May researchers prioritize certain principles from the Belmont Report over others?

The following prompts also concern secondary data:

- Are the data sufficiently anonymized or old to be free of ethical concerns?
- Should race be used as a variable?
- Are researchers who use certain variables complicit in or perpetuating some undue stigma or discrimination against minority subgroups?

To answer all these questions and worries, we must take into consideration ethical issues regarding the above. Hence, as also referred to above, ethical aspects should be considered regarding the collection, use, sharing, storage, and reporting of the data and the results derived from their analysis.

Many professional associations, such as the American Statistical Association (ASA) and the International Statistical Institute (ISI) have produced ethical guidelines for statisticians. Another source of ethical guidelines is the committees of research in tertiary education (Research Ethics Committees), which provide some guidelines for ethical considerations in research taking place inside the education organizations. Any organization or independent research should also engage stakeholders (employees, suppliers, customers, community, etc.) to improve ethical data usage.

Another issue that complicates issues concerning data, is technology, which enables the collection, and analysis of data. However, artificial intelligence, an aspect of technology, raise many more ethical issues, such as human rights, discrimination, and privacy.

The following paragraphs present the most significant initiatives addressing ethical issues related to data. It covers many aspects concerning these issues. It examines the handling of data in tertiary education's research, it provides recommendations for the professionals of the statistical field (even for statistical education), it presents certain private initiatives, it covers the bioethical field, and registers certain legislative actions.

4.1 Ethics committees

It is usual to create bodies that deal with ethical matters and provide guidelines to follow. An example of them is ethics committees, which examine the following topics (Hand, 2018), concerning mainly primary sources of data:

- identify which ethics body has oversight of the work;
- understand the policies and procedures of the institution;

- understand national laws and regulations (such as those pertaining to discrimination, privacy, consent, and the need to provide an explanation for analysis);
- maintain documentation of data manipulation and modification;
- understand the origin of the data (provenance);
- treat the metadata as rigorously as the data;
- have an explicit data management plan;
- store data securely;
- determine for how long the data must be kept;
- indicate who can access the information;
- guarantee the use of the proper statistical, machine learning, data mining, and other tools;
- have systems in place that allow data to be corrected (and deleted if necessary); and
- be clear about the benefits of the analysis, and who derives the benefits.

4.1.1 Committees for Research in tertiary education

There are specific committees to control ethical issues and in general, the credibility of research projects carried out in universities and research bodies. Below is an example of such a committee in Greek tertiary education, is presented.

It is the Committee for Research Ethics of the University of Macedonia, and the information is taken from its website (Committee for Research Ethics of the University of Macedonia). The purpose and responsibilities of the committee are presented briefly.

Purpose: to provide guarantees at a moral and ethical level concerning the credibility of research projects carried out in universities and research bodies. The value of human beings, the autonomy of the participants, their privacy and personal information, and the natural and cultural environment are all factors that the Committees for Research Ethics regulate when conducting research projects. The committees also monitor compliance with generally accepted principles of research integrity and the criteria of good scientific practice.

Responsibility: to determine whether a specific research project that a university or research body plans to conduct conforms with the relevant laws and, in terms of its content and methodology, with generally recognized ethical standards for research. The Committee for Research Ethics evaluates the research proposal and either approves it or makes recommendations and propositions for its revision in case any ethical obstacles arise.

However, these committees present many dysfunctions due to the complicated legislation and the lack of experience and expertise of the members; hence, certain suggestions are presented in a later paragraph.

4.2 GDPR

A major source of ethical issues and guidelines about data is the General Data Protection Regulation (GDPR) (EU, 2016). It is a regulation applicable as of May 25th, 2018, in all member states of the European Union to harmonize data privacy laws across Europe.

Using quotes from Wolford (2024), some of the GDPR's key features will be discussed.

Data protection and accountability principles are outlined in Articles 5.1-2 of the regulation:

- Lawfulness, fairness and transparency — Processing must be lawful, fair, and transparent to the data subject.
- Purpose limitation — You must process data for the legitimate purposes specified explicitly to the data subject when you collected it.
- Data minimization — You should collect and process only as much data as absolutely necessary for the purposes specified.
- Accuracy — You must keep personal data accurate and up to date.
- Storage limitation — You may only store personally identifying data for as long as necessary for the specified purpose.
- Integrity and confidentiality — Processing must be done in such a way as to ensure appropriate security, integrity, and confidentiality (e.g. by using encryption).
- Accountability — The data controller is responsible for being able to demonstrate GDPR compliance with all of these principles.

For example, some actions to enhance accountability are:

- Assign your team duties related to data protection.
- Keep thorough records of the data you're gathering, its intended use, its storage location, the employee in charge of it, etc.
- Train your staff and implement technical and organizational security measures.
- Have Data Processing Agreement contracts in place with third parties you contract to process data for you.
- Appoint a Data Protection Officer (though not all organizations need one).

Data security

To handle data securely, there are requirements to implement “appropriate technical and organizational measures”. Technical measures mean anything from requiring your employees to use two-factor authentication on accounts where personal data is stored to contracting with cloud providers that use end-to-end encryption.

Data protection by design and by default

Following the implementation of this rule, data protection must be taken into account "by design and by default" in every aspect of an organization. In practice, this means that any new product or activity should be designed with data protection principles in mind. The GDPR covers this principle in Article 25.

When someone is allowed to process data

The situations in which processing personal data is permitted are listed in Article 6. When the data subject gave you specific, unambiguous consent to process the data, when processing is necessary to execute or to prepare to enter into a contract to which the data subject is a party. Other cases include the compliance with a legal obligation and the need to process the data to save somebody's life.

Processing is also allowed when it is necessary to perform a task in the public interest or to carry out some official function. Another case is when there is a legitimate interest in processing someone's personal data. This is the most flexible lawful basis, though the "fundamental rights and freedoms of the data subject" always override your interests, especially if it's a child's data.

A researcher must notify the data subject and document the legal basis for the data processing after he has established it. And if he decides later to change his justification, he needs to have a good reason, document this reason, and notify the data subject.

Consent

Strict new guidelines have been established regarding what constitutes a data subject's consent to have their information processed. First of all, consent must be "freely given, specific, informed and unambiguous." Requests for consent must be "clearly distinguishable from the other matters" and presented in "clear and plain language."

Data subjects have the right to revoke their consent at any time, and you must respect their choice. You can't simply change the legal basis of the processing to one of the other justifications. Children under 13 can only give consent with permission from their parents. Finally, the researcher needs to keep documentary evidence of consent.

Data Protection Officers

Contrary to popular belief, not every data controller or processor needs to appoint a Data Protection Officer (DPO). There are three conditions under which an organization is required to appoint a DPO:

- It is a public authority other than a court acting in a judicial capacity.

- Its core activities require you to monitor people systematically and regularly on a large scale.
- Its core activities are large-scale processing of special categories of data listed under Article 9 of the GDPR or data relating to criminal convictions and offenses mentioned in Article 10.

Some of their basic responsibilities include understanding the GDPR and how it impacts the business, advising staff members about their responsibilities, conducting data protection training, conducting audits and monitoring GDPR compliance, and communicating with regulators.

An organization could also choose to designate a DPO even if it isn't required to. Having someone in this position who handles and protects data has only advantages.

People's privacy rights

In an effort to provide people with greater control over the data they lend to businesses, the GDPR acknowledges a long list of new privacy rights for data subjects. As an organization, it's important to understand these rights to ensure you are GDPR compliant.

Below is a rundown of data subjects' privacy rights: to be informed, to access, to rectification, to erasure, to restrict processing, to data portability, to object. Finally, there are rights in relation to automated decision making and profiling.

4.3 Open Data Institute

Open Data Institute is a private organization dealing with the diffusion of data. It has produced a data ethics canvas with instructions about dealing with data (Open Data Institute, 2021a).

There are four sections of the data ethics canvas:

- Assess the data: By asking you to assess the data sources, their limitations, the rights and legislative context of data collection, use, and sharing that apply to your sector.
- Identify the impact: By asking you to identify your reasons for using data to develop this service/product, the potential positive and negative impact on people and society, and the potential approaches to minimize negative harm.
- Evaluate the engagement: Challenges you to think about the people involved, your approaches to communication and transparency.
- Decide the process: Helps you decide on the next steps: what should you prioritize, when will you iterate, do you need training, etc.

This institute has also produced a framework to develop ethical guidance, called data ethics-data life cycle (Open Data Institute, 2021b). Each stage in the data lifecycle involves different data practices against which ethics should be evaluated. The process of examining the impact is not always clear or straightforward. Understanding where an organization is on the data lifecycle is important in determining the risks associated with its immediate data practices.

4.4 The profession of statistician

4.4.1 International Statistical Institute (ISI)

International Statistical Institute is a non-profit, non-government organization having as a main objective to promote the understanding, development, and good practice of statistics worldwide by providing a welcoming environment for advancing statistical knowledge, learning best practices, sharing state-of-the-art developments, and by creating opportunities to network. It has also produced a declaration on professional ethics with some professional values and ethical principles derived from them. (Declaration on Professional Ethics | ISI, n.d.)

This declaration emphasizes that statisticians, whatever area of expertise they belong to, should integrate ethical principles in their professional competence and should be part of a comprehensive statistical education. The declaration enables the statistician's individual ethical judgments and decisions to be informed by shared values and experience.

When a statistician carries out his responsibilities, his actions must be consistent with the interest of the following groups and not favor any society, employers, clients, funders, colleagues, and subjects.

The declaration describes the shared professional values of statisticians are: respect, professionalism, truthfulness, and integrity.

The ethical principles for statisticians, as outlined in the declaration, are presented below:

Pursuing objectivity, clarifying obligations and roles, assessing alternatives impartially, conflicting interests, avoiding preempted outcomes, guarding privileged information, exhibiting professional competence, maintaining confidence in statistics, exposing and reviewing methods and findings, communicating ethical principles, bearing responsibility for the integrity of the discipline, protecting the interests of subjects.

The above declaration of ISI has also been adopted by the Federation of European National Statistical Societies (FENStats), and it is presented as the Code of Ethics for Statistical Practice in Europe (https://www.fenstats.eu/ethics_code_isi, assessed 22/08/2025).

In the same declaration there are also certain ethical principles derived from professional values. Some of these are pursuing objectivity, clarifying obligation and roles, guarding privileged information, maintaining confidence in statistics, communicating ethical principles, etc.

4.4.2 American Statistical Association (ASA)

ASA is a big community of statisticians, and it has provided some Ethical Guidelines for Statistical Practice (American Statistical Association, 2022), which follow: professional integrity and accountability, integrity of data and methods. They also include responsibilities to stakeholders, to research subjects, to data subjects, or those directly affected by statistical practices, to members of multidisciplinary teams, to fellow statistical practitioners, and the profession. There are also responsibilities of leaders, supervisors, and mentors in statistical practice, and finally, responsibilities regarding potential misconduct.

Finally, as Metcalf (2014) noted, there are numerous ethical challenges concerning the professions dealing with data, and the ways to address them include

- protect vulnerable populations who could be harmed by the profession's activities;
- protect/enhance the good reputation of and trust for the profession;
- establish the profession as a distinct moral community worthy of autonomy from external control and regulation;
- provide a basis for public expectations and evaluation of the profession; serve as a basis for adjudicating disputes among members of the profession and between members and nonmembers;
- create institutions resilient in the face of external pressures; and
  respond to past harms done by the profession

4.5 The Belmont Report

The Belmont Report is a foundational document in research ethics (Ryan et al., 1979), outlining three core principles for protecting human subjects in research (Paxton, 2020): respect for persons, beneficence, and justice.

It guarantees that research involving human subjects is carried out responsibly and ethically and offers ethical guidelines for doing so.

The following requirements are essential to apply the general principles to the conduct of research: informed consent, risk/benefit assessment, and the selection of subjects for research.

4.6 Guidelines for Assessment and Instruction in Statistics Education (GAISE)

The College Gaise Report has been produced by American Statistical Association (ASA) and it is a text with guidelines about statistical problem-solving process. These recommendations are designed to assist the teaching of statistics in classrooms. Its first edition was produced by Aliaga et al. (2005). One of the student goals as presented in the report is that they should demonstrate an awareness of ethical issues associated with sound statistical practice (GAISE College Report ASA Revision Committee, 2016).

The report highlights that instructors should educate their students about the ethical aspects related to the collection of those data (Raman et al., 2023).

Another point that is often brought up is that ethics courses ought to be taken in addition to statistics and/or data analysis courses (De Veaux et al., 2017).

4.7 European Statistics Code of Practice

There are sixteen principles that are presented in the European Statistics Code of Practice, covering the institutional environment, the statistical processes, and the statistical output (https://www.statistics.gr/en/european-statistics-code-of-practice, assessed 21/08/2025).

Some of them are professional independence, commitment to quality, mandate for data collection and access to data, adequacy of resources, cost-effectiveness, accuracy and reliability, relevance, impartiality, and objectivity.

4.8 National Commission for Bioethics and Technoethics

Finally, there is the National Commission for Bioethics and Technoethics, which is an independent advisory body of the State of Greece in the fields of life sciences and new technologies in general.

This commission issued a report in 2024 concerning the committees for research in tertiary education https://bioethics.gr/api/files/download/2411/REPORT_RECs_2024_FINAL_GR.pdf?attachment=false, assessed 01/09/2025).

In this report there are several suggestions for the committees for research, such as seeking collaboration between the committees, ensuring adequate support from the state, providing appropriate training for the members, avoiding certain bad practices, and being wary of situations including conflicting interests. Overall, it is concluded that the committees present serious problems that must be confronted in order to provide quality services.

5. CONCLUSION

Data is a valuable asset across many scientific fields, as it is used to conduct research and generate meaningful results that benefit societies. This continuous need for data must be handled in a way that protects all personal data and information of individuals.

However, many more issues arise, such as the use of technology and its role in modern science. For example, an extension of technology is artificial intelligence, which is an extremely complicated tool that produces many more ethical issues.

Furthermore, biomedical research produces large quantities of data which are not always easy to handle. Hence, many ethical issues are evident and should also be confronted in this field.

All the previously mentioned policies, procedures, guidelines, and comparable actions could contribute to ensure that data is handled and protected in an ethical way. As anyone can realize, certain of them are intended to provide suggestions and recommendations for the professionals of the statistical profession. Therefore, all data professionals should exercise extreme caution and try to put them into practice. Their contribution is very crucial in the effort to protect data of subjects participating in surveys.

REFERENCES

Aliaga, M., Cobb, G., Cuff, C., Garfield, J., Gould, R., Lock, R., Moore,T., Rossman, A., Stephenson, B., Utts, J., Velleman, P., and Wit-mer, J. (2005), "Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report," American Statistical Association [online]. Available at https://www.amstat.org/asa/files/pdfs/GAISE/2005GaiseCollege_Full.pdf, accessed 22.5.2025.

American Statistical Association (2022), "Ethical Guidelines for Statistical Practice," [online]. Available at https://www.amstat.org/your-career/ethical-guidelines-for-statistical-practice, accessed 22.05.2025.

Committee for Research Ethics of the University of Macedonia, https://www.uom.gr/en/ethics, last accessed 22.05.2025.

Declaration on Professional Ethics, ISI. (n.d.), https://isi-web.org/declaration-professional-ethics, accessed 22/05/2025)

De Veaux, R. D., Agarwal, M., Averett, M., Baumer, B. S., Bray, A., Bressoud,T. C., Bryant, L., Cheng, L. Z., Francis, A., Gould, R., Kim, A. Y., Kretch-mar, M., Lu, Q., Moskol, A., Nolan, D., Pelayo, R., Raleigh, S., Sethi,R. J., Sondjaja, M., Tiruviluamala, N., Uhlig, P. X., Washington, T. M.,Wesley, C. L., White, D., and Ye, P. (2017), "Curriculum Guidelines for Undergraduate Programs in Data Science," Annual Review

of Statistics and Its Application, 4, 15–30. https://doi.org/10.1146/annurev-statistics-060116-053930.

EU, (2016), Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), http://data.europa.eu/eli/reg/2016/679/oj, accessed 22/05/2025.

GAISE College Report ASA Revision Committee (2016), "Guidelines for Assessment and Instruction in Statistics Education College Report 2016," [online]. Available at https://www.amstat.org/docs/default-source/ amstat-documents/gaisecollege_full.pdf, accessed 22/05/2025.

Hand, D. J. (2018), "Aspects of Data Ethics in a ChangingWorld:Where are We Now?," Big Data, 6, 176–190. https://doi.org/10.1089/big.2018.0083.

Open Data Institute (2021a), "Data Ethics Canvas," [online]. Available at https://theodi.org/wp-content/uploads/2021/07/Data-Ethics-Canvas-English-Colour.pdf, accessed 22.05.2025.

Open Data Institute (2021b), "Data Ethics Life Cycle," [online]. Available at https://data-ethics-essentials.learndata.info/m2-what-is-data-ethics/#/id/64f04f900667765a0b978c97, accessed 11/05/2025.

Metcalf J. (2014). Ethics codes: History, context, and challenges. Council for Big Data, Ethics, and Society. Available online at https://bdes.datasociety.net/ wp-content/uploads/2016/10/EthicsCodes.pdf, accessed 25/05/2025.

Paxton, A. (2020), "The Belmont Report in the Age of Big Data: Ethics at the Intersection of Psychological Science and Data Science," in Big Data in Psychological Research, eds. S. E. Woo, L. Tay and R. W. Proctor, pp. 347–372, Washington, DC: American Psychological Association.

Raman R., Utts J., Cohen A.I., and Hayat M.J., (2023), Integrating Ethics into the Guidelines for Assessment and Instruction in Statistics Education (GAISE), The American Statistician, 2023, VOL. 77, NO. 3, 323–330: Teacher's Corner. (Integr)

Ryan, K. J., Brady, J. V., Cooke, R. E., Height, D. I., Jonsen, A. R., King,P., Lebacqz, K., Louisell, D. W., Seldin, D. W., Stellar, E., and Turtle,R. H. (1979), "Ethical Principles and Guidelines for the Protection of Human Subjects of Research (The Belmont Report)," Retrieved from US Department of Health & Human Services, Human & Health Ser-vices website: https://www.hhs.gov/ohrp/sites/default/files/the-belmont-report-508c_FINAL.pdf.

Wolford, B. (2024). What is GDPR, the EU's new data protection law? GDPR.eu. https://gdpr.eu/what-is-gdpr/?cn-reloaded=1, accessed 22.05.2025.

Web resources

- https://www.fenstats.eu/ethics_code_isi (assessed 22/08/2025)
- https://www.statistics.gr/en/european-statistics-code-of-practice (assessed 21/08/2025)
- https://bioethics.gr/announcements-26/ethsia-ek8esh-pepragmenwn-2024-3223 (assessed 01/09/2025)

ΠΕΡΙΛΗΨΗ

Η Στατιστική είναι μια επιστήμη που χρειάζεται τα δεδομένα για να ολοκληρώσει τα καθήκοντα και τις λειτουργίες της. Στην εποχή μας τα δεδομένα συγκεντρώνονται με πολλούς τρόπους και από ποικίλες πηγές. Οι περισσότερες μέθοδοι που χρησιμοποιούνται, συλλέγουν δεδομένα και πληροφορίες για ανθρώπινα όντα. Συνεπώς, σε όλες αυτές τις διαδικασίες αρκετά ζητήματα προκύπτουν σχετικά με την κατάλληλη χρήση αυτών των δεδομένων, έχοντας διασφαλίσει τη συναίνεση και την προστασία όλων των συμμετεχόντων. Προφανώς, όλα τα δεδομένα που συγκεντρώνονται πρέπει να προφυλάσσονται, να διασφαλίζονται και να χρησιμοποιούνται μόνο για τον αρχικό σκοπό. Σε αυτή τη μελέτη, συγκεκριμένες οδηγίες, κατευθύνσεις και ήδη θεσμοθετημένα μέτρα θα αναφερθούν, τα οποία έχουν σκοπό να εξασφαλίσουν ότι οι ηθικές διαστάσεις ενσωματώνονται στη συλλογή, ανάλυση και παρουσίαση των δεδομένων.

Λέξεις κλειδιά: Στατιστική, Ηθικές διαστάσεις, Δεδομένα, Πηγές δεδομένων, Ασφάλεια δεδομένων.

Greek Statistical Institute
Proceedings of the 37th Panhellenic & 3rd International Statistics Conference (2025), pp. 347–368

Predicting Skills in Digital Publication Policies: A Linear Discriminant Analysis and Machine Learning Classification Approach

***Dimitrios Christos Kavargyris*[1], *Panagiotis Gkorezis*[2], *Nikolaos Mittas*[3], *Lefteris Angelis*[1]**

[1] School of Informatics, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece
[2] School of Economics, Faculty of Economics and Political Sciences, Aristotle University of Thessaloniki, Greece
[3] Department of Chemistry, School of Science, Democritus University of Thrace, 65404 Kavala, Greece

ABSTRACT

As the digital economy reshapes the global labor landscape, policies must do more than respond—they must anticipate. In an era where skill mismatches are growing faster than solutions, data-driven policy analysis is no longer optional; it is essential. This study develops a supervised learning framework to analyze 981 Digital Publication Policies (DPP) from European and international databases (2019–2024), aiming to identify, classify, and interpret embedded technical and soft skills. Using the European Skills, Competences, Qualifications and Occupations (ESCO) taxonomy for skill extraction and bag-of-words vectorization, each document is encoded as a high-dimensional feature vector comprising skill frequencies, metadata, and contextual attributes. These documents are classified into seven socio-economic categories—such as education, employment, economy, and innovation—based on a schema developed by labor market policy experts. Feature vectors are used to train Random Forest and Gradient Boosting models, with Linear Discriminant Analysis (LDA) applied for dimensionality reduction and enhanced class separability. Unlike prior studies focused solely on prediction, our approach also interprets which skills drive differences between policy areas. Results show that LDA improves model performance, with domains like Training and Digital Economy leading in skill-focused policy alignments. The model also highlights emerging skills with future relevance. Finally, this framework can be integrated into EU-funded projects like SKILLAB to support real-time, skill-informed policy planning.

Keywords: Skill acquisition; Policy Classification; Linear Discriminant Analysis; ESCO

1. INTRODUCTION

As the digital tide rises ever faster (Adewumi, 2022), governments and international organizations increasingly rely on publication policies—strategic documents outlining goals in areas like digital education, employment, and innovation. Issued by bodies such as the European Commission and the ILO (Standing, 2010; European Commission,

2022), these policies function as operational tools for guiding digital transformation, coordinating action, and allocating resources. Despite their importance, they remain under-examined from a quantitative, skill-based perspective, limiting efforts to align education and workforce development with emerging demands.

Digital Publication Policies (DPP)—strategic documents issued by governments and international bodies—aim to guide action in domains such as education, employment, and innovation. A central challenge within these policies is the urgent need to upskill and reskill the European workforce in response to rapid digitalization, automation, and remote work. By 2030, over 70% of EU workers will require digital reskilling (Lang and Triantoro, 2022), with critical shortages already emerging in fields like data literacy and cybersecurity. The COVID-19 pandemic further exposed digital divides across the workforce (Cedefop, 2023). To address these challenges, the EU has launched initiatives such as the SKILLAB project (Aluas et al., 2024; Kavargyris et al., 2025c), which includes the European Skills Policy Tracker—an expert-driven tool to monitor evolving skill needs across policy areas. Recent work by (Kavargyris et al., 2025b) further refines skill extraction from educational policies via network analysis, highlighting methodological innovations and high-demand skills.

Despite growing recognition of the need for reskilling, few empirical tools exist to systematically analyze the skill content of policy documents. Existing methods—often manual or keyword-based—fail to capture the implicit and contextual ways skills are embedded in policy language. Most research focuses on labor market forecasts rather than extracting skills directly from policy texts, leaving decision-makers with limited insight into strategic skill priorities (Decorte et al., 2025). This gap risks misaligned investments and outdated curricula (Clavié and Soulié, 2023). A data-driven framework is urgently needed to classify policies by socio-economic focus and uncover emphasized technical and soft skills.

To address the challenges of fragmented skill policy interpretation and the lack of scalable empirical tools, we propose a supervised learning framework that models the relationship between policy content and skill emphasis using structured, skill-level features. This study is guided by the following Research Questions (RQ):

- RQ_1: How do digital policy documents encode structured relationships between skills, and what thematic skill clusters emerge through association rule analysis?
- RQ_2: How accurately can machine learning classify policies into socio-economic skill categories?
- RQ_3: How can we quantitatively assess the alignment between skills and digital policy areas?

2. RELATED WORK

In today's dynamic labor market, the ability of workers to remain agile and adaptable is increasingly vital (Özdemir and Tan, 2023). As job roles continue to evolve, employees are required to continuously upgrade their skill sets (Sawant et al., 2022). Life-

long learning plays a pivotal role in acquiring and refining competencies essential for navigating technological advancements. In this context, the continual acquisition and renewal of skills are fundamental to preparing the workforce for future challenges. A workforce that prioritizes adaptability and skill development contributes to higher productivity and operational efficiency. Moreover, flexibility enables individuals to respond swiftly to market shifts and innovations, easing transitions and promoting resilience in the face of change.

In response to these labor market shifts, policy frameworks increasingly recognize the centrality of lifelong learning and skills foresight. International bodies such as the (Organisation for Economic Co-operation and Development, 2021) and the (International Labour Organization, 2021) have called for data-driven approaches to workforce planning, advocating for the integration of real-time labor intelligence into national education and training strategies. Public policies now aim not only to respond to existing skill gaps but also to anticipate future ones, using predictive modeling, competency frameworks, and targeted reskilling initiatives to align workforce capabilities with emerging economic demands. This evolving landscape positions public policy as a critical enabler in shaping skill ecosystems that are agile, inclusive, and future-ready.

3. BACKROUND INFORMATION

In this section, we present the theoretical background necessary for this study. Subsection 3.1 introduces the ESCO taxonomy, which provides the standardized framework for skill classification. Subsection 3.2 discusses the background information regarding the skill extraction mechanism. Finally, Subsection 3.3 outlines the structure of the policy documents used in this analysis.

3.1 ESCO taxonomy

IIn this study, we introduce a novel approach by leveraging the European Classification of Skills, Competences, and Occupations (ESCO)[1] taxonomy to extract structured skill data directly from unstructured policy documents. ESCO is a multilingual taxonomy that standardizes the representation of skills, occupational profiles, and qualifications across European labor markets. Based on the International Standard Classification of Occupations (ISCO-08), ESCO is continuously updated (Kavargyris et al., 2025a) to reflect evolving workforce demands. The current version includes over 13,800 competences and approximately 3,000 occupational titles, each accompanied by detailed descriptions, alternative labels, and categorizations into skills or knowledge types. This integration of ESCO into our analytical pipeline enables a scalable and interpretable mechanism for linking policy content to labor market needs (Kavargyris et al., 2025d).

[1] https://esco.ec.europa.eu/

3.2 Skill extraction mechanism

The skill extraction procedure in this study is grounded in the framework proposed by (Djumalieva and Sleeman, 2018), which classifies occupations based on skill requirements identified in job advertisements. The methodology begins with the generation of simplified "surface forms" of skills using natural language processing (NLP) techniques. These forms enhance the detectability of skill mentions in unstructured text (Caratozzolo et al., 2023b). To align extracted terms with standardized classifications, we use the ESCO taxonomy, mapping identified skill mentions directly to recognized concepts. *Skill2vec* embeddings (Van-Duyet et al., 2017) are further utilized to capture contextual similarities among skills, while hierarchical community detection enables the grouping of frequently co-occurring skills into interpretable clusters. This combined approach allows for a structured and scalable identification of both technical and soft skills from DPP, ensuring consistency across a heterogeneous corpus of 981 international documents.

3.3 DPP structure

Before proceeding with the technical implementation, it is essential to define the structure of a DPP as treated in our study. Table 1 illustrates an example of a DPP, along with the selected features both before and after expert-based classification. These features—such as *title*, *targeted audience*, *country*, *policy area*, and *identified skills*—form the foundation for the machine learning pipeline that follows.

From this point forward, the four distinct source databases (EUR-Lex, ILO, Eurofound, and N-Lex) will collectively be referred to as a single, unified resource named the *Policy Database (PD)*. An important and novel feature of our dataset is the classification of each policy document into one of several high-level policy areas, such as *Education*, *Digital Inclusion*, and *Social Inclusion*. This thematic categorization was expert-driven, carried out by the second author of this study, a specialist in human resources and labor policy. The classification process was informed by the skills-matching methodology developed by CEDEFOP, the European Centre for the Development of Vocational Training.

Table 1: *Example structure of a Digital Publication Policy (DPP).*

Field	Example
Title	Bridging Digital Gaps in Rural Education
Country	EU Member States
Targeted Audience	Teachers, Policy Advisors, Curriculum Designers
Date of Document	15/03/2022
Policy Area	Education, Digital Inclusion, Social Inclusion
Skills	Digital literacy, curriculum design, rural ICT access, e-learning development, training needs analysis
Source	https://www.europa.eu/policies/digital-education-initiative
Publication Policy Text	This policy document highlights the urgent need to strengthen digital learning in rural areas by improving internet access, training educators in digital tools, and promoting inclusion through targeted curriculum reforms. Challenges include uneven infrastructure... (cont'd)

Note: The table illustrates the structure and metadata fields of a Digital Publication Policy (DPP) as used in this study.

4. RESEARCH METHODOLOGY

The framework's architecture for identifying and extracting skills from policy documents and classifying them into distinct socio-economic domains is illustrated in Figure 1. As shown, the process begins with the collection of 981 DPPs (the collection of these documents is described in detail in the corresponding Subsection 4.1), sourced from reputable institutional databases, including sourced from reputable institutional databases, including EUR-Lex[2], ILO[3], Eurofound[4], and N-Lex[5]. covering the period from 2019 to 2024 and encompassing a wide range of socio-economic contexts. These documents are publicly available.

In the *skill extraction phase* (Subsection 4.2), documents undergo preprocessing steps such as text cleaning and normalization before ESCO-based extraction is performed using the NESTA toolkit (Caratozzolo et al., 2023a). This step transforms unstructured textual content into structured representations of relevant skills. Each policy is then encoded as a high-dimensional feature vector that combines skill frequencies with contextual metadata, such as the country of origin and the intended user group.

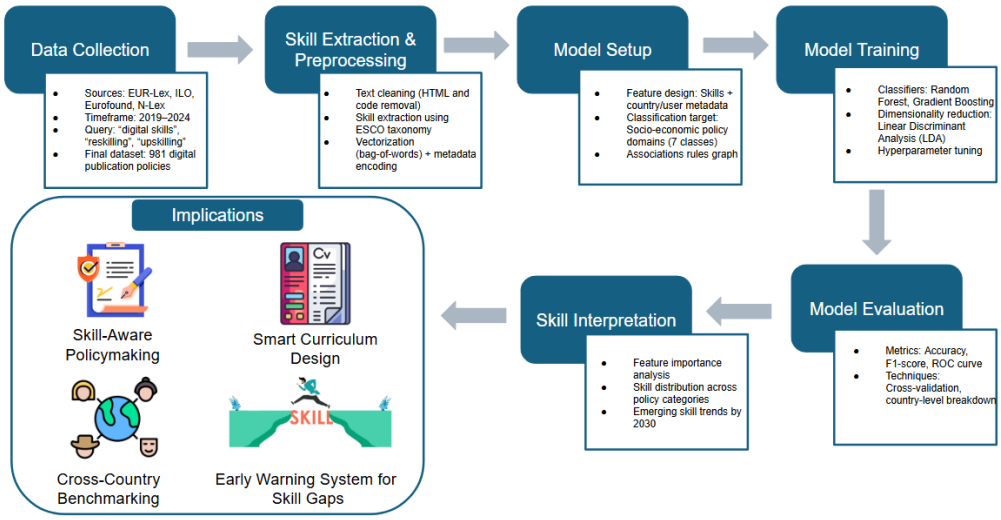
In the *model setup phase* (Subsection 4.3), these feature vectors are used to train

[2] https://eur-lex.europa.eu
[3] https://www.ilo.org/global/publications
[4] https://www.eurofound.europa.eu/publications
[5] https://eur-lex.europa.eu/n-lex

supervised classifiers—specifically Random Forest (RF) and Gradient Boosting (GB) models—while dimensionality reduction is applied using Linear Discriminant Analysis (LDA) to enhance class separability and reduce complexity.

During the *model training and evaluation phase* (Subsection 4.4), performance is assessed using accuracy, F1-score, and the Receiver Operating Characteristic – Area Under the Curve (ROC AUC) metric (Subsection 4.5), supported by grid search hyperparameter tuning and cross-validation. Lastly, the framework incorporates interpretability techniques, including feature importance analysis and association rule graphs, in order to provide insights into the most influential skills and their interdependencies.

Figure 1: *Proposed supervised learning framework*



4.1 Data Collection

To perform our analysis, we first needed to collect a substantial corpus of DPPs reflecting contemporary skill-related discourse. Unlike technical datasets, DPPs are dispersed across diverse institutional repositories with limited standardization in their dissemination formats. Although institutional portals such as EUR-Lex or ILO provide structured access, the heterogeneity in metadata quality, search capabilities, and document formatting renders large-scale collection a time-intensive process.

We selected our DPPs from three primary institutional sources: (i) the International Labour Organization (ILO), a United Nations agency focusing on international labor standards and employment policies; (ii) EUR-Lex, the official legal database of the European Union; and (iii) Eurofound, an EU agency that monitors working and living conditions across member states.

To extract relevant policies, we formulated a comprehensive search query incorporating the terms: "digital skills", "reskilling", "upskilling", "digital transformation",

"digital competence", and "future of work". The search was constrained to the *TITLE-ABS-KEY* fields to ensure semantic relevance. Additionally, we applied a publication date filter spanning from 2019 to 2024, reflecting the most recent and relevant developments in the digital transition and labor policy landscape.

From these three sources, a total of 981 unique DPPs were curated following deduplication and manual relevance checks. These documents form the foundational corpus for subsequent skill extraction, classification, and trend analysis in the context of digital public policy.

To structure skill identification, we employed the *NESTA Skills Taxonomy Toolkit* (Caratozzolo et al., 2023b), which provides methodological guidelines for mapping skills from unstructured text. For model evaluation, performance metrics included Accuracy, F1-score, and the Area Under the Receiver Operating Characteristic Curve (ROC AUC), where ROC AUC measures the ability of a classifier to distinguish between classes across different decision thresholds.

4.2 Skill extraction and Pre-processing

The study employed a standard NLP pipeline tailored to the structure of policy documents. First, tokenization was applied to split text into unigrams, bigrams, and trigrams, serving as candidates for concept extraction. Second, normalization reduced lexical variation by converting terms to their base forms (e.g., singularizing plurals), enabling better clustering of semantically similar terms. Finally, data cleaning removed stopwords, numeric values, acronyms (e.g., NACE), and overly generic terms occurring in more than 100 documents, ensuring relevance to policy-specific terminology.

After classifying 981 DPP into distinct areas, we applied the *NESTA mechanism* to extract relevant skills. Prior to extraction, the texts were preprocessed by removing noise (e.g., stopwords, HTML tags) and applying tokenization, lemmatization, and lowercasing. The resulting skills were then matched to the ESCO taxonomy (Kavargyris et al., 2025d) to ensure standardization and cross-policy comparability.

This procedure enabled the generation of a domain-specific skill list, which served as the foundation for subsequent bibliometric and statistical analyses.

4.2.1 ESCO Mapping Strategy

To extract relevant skills from DPPs, we employed the open-source NESTA skill extraction toolkit, which leverages natural language processing to identify and classify skill phrases in unstructured text. The extraction process follows a two-stage pipeline.

In the first stage, Named Entity Recognition (NER) is used to detect candidate skill expressions—such as "data governance" or "digital pedagogy"—through a combination of spaCy's rule-based tokenizer (Vasiliev, 2020) (which segments text using language-specific rules for whitespace, punctuation, and special cases) and custom pattern-matching rules designed for policy-specific terminology (e.g., acronyms such as NACE or EU

directives). This hybrid approach ensures that multi-word skill phrases are detected consistently in heterogeneous policy documents.

In the second stage, these candidate phrases are mapped to standardized concepts in the ESCO taxonomy using both fuzzy string matching and taxonomy-aware lookup. This includes identifying exact and near matches to ESCO skill labels, aliases, and alternative titles. By combining linguistic and semantic matching techniques, the toolkit ensures accurate and consistent alignment between extracted terms and the ESCO framework, thereby enabling reliable downstream analysis.

4.2.2 Frequency Matrix Construction

Once the skills were extracted and mapped to ESCO concepts using the NESTA framework, we constructed a document–skill frequency matrix $\mathbf{D} \in \mathbf{R}^{n \times m}$, defined as follows:

- $n = 981$: the number of DPPs,
- m: the total number of unique ESCO skills identified across the corpus,
- D_{ij}: the raw frequency of skill j in document i.

To improve signal quality and minimize dimensional noise, several post-processing steps were applied to the document–skill matrix prior to classification. First, *rare skill filtering* was implemented, removing any skill that appeared in fewer than three documents to reduce sparsity. Second, we applied *Term Frequency–Inverse Document Frequency (TF–IDF)* normalization, which rescales raw counts into weighted values that reflect the relative importance of a skill across the corpus. Formally, the TF–IDF weight of skill j in document i is given by:

$$\text{TF–IDF}_{ij} = \text{TF}_{ij} \cdot \log\left(\frac{n}{\text{DF}_j}\right) \tag{1}$$

where TF_{ij} is the term frequency of skill j in document i and DF_j is the number of documents in which skill j appears. Finally, *feature augmentation* was introduced by concatenating contextual metadata—such as country of origin, target audience, and policy domain—with the TF–IDF vectors, thereby enriching the representation and enabling more nuanced classification in downstream machine learning models.

4.3 Model Setup

The classification pipeline is built upon a hybrid feature design that integrates both skill-centric and contextual information. Specifically, each policy document is transformed into a structured vector that combines TF–IDF weighted ESCO skills with auxiliary metadata, including the document's country of origin and its targeted user group (e.g., educators, policymakers, or training providers). This enriched representation ensures that the model captures not only the semantic content of the policy but also its geopolitical and audience-specific framing.

The classification target consists of seven socio-economic policy domains, determined in consultation with labor market policy experts. These domains include *Education*, *Employment*, *Digital Innovation*, *Training*, *Social Inclusion*, and *Governance*. Each document is labeled according to its primary policy orientation, enabling supervised training and interpretability at the domain level.

In addition to predictive modeling, the framework incorporates an association rules graph to support interpretability. By mining frequent co-occurrence patterns among extracted skills within each policy category, the graph visualizes dominant skill bundles and their interrelationships. This layer adds an explanatory dimension to the model's output, offering actionable insights into how skills cluster within strategic policy areas and how these clusters may inform curricular or workforce interventions.

4.4 Model Training

To address the multi-class classification task of assigning DPPs to socio-economic domains, we trained two supervised learning models: RF and GB. These ensemble-based classifiers were selected for their robustness, interpretability, and ability to handle high-dimensional, sparse feature spaces—a natural fit for TF–IDF skill vectors.

Classifiers. RF operates by building an ensemble of decision trees, as described in Eq. 2:

$$\{T_1(x), T_2(x), \ldots, T_k(x)\} \tag{2}$$

, each trained on a bootstrap sample of the dataset. The final class prediction is made via majority voting across the ensemble:

$$\hat{y} = \text{mode}\big(T_1(x), T_2(x), \ldots, T_k(x)\big) \tag{3}$$

where x is the feature vector of a policy document. GB builds an additive model in a forward stage-wise fashion by minimizing a differentiable loss function $L(y, F(x))$ over iterations:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \tag{4}$$

where $h_m(x)$ is the weak learner (typically a decision tree) fitted to the negative gradient of the loss at iteration m, and γ_m is the step size.

Dimensionality reduction. To improve class separability and reduce model complexity, we applied LDA prior to classification. Unlike Principal Component Analysis (PCA), which focuses on variance, LDA maximizes class separability by projecting the data onto a lower-dimensional space that maximizes the between-class scatter S_B and minimizes the within-class scatter S_W. Here, "scatter" refers to the covariance structure of the data within and across classes (Eq. 5). The transformation matrix $\mathbf{W}$ is computed as:

$$\mathbf{W} = \arg\max_{\mathbf{W}} \frac{|\mathbf{W}^T S_B \mathbf{W}|}{|\mathbf{W}^T S_W \mathbf{W}|}. \tag{5}$$

While LDA formally assumes multivariate normality of predictors within each class, real-world data—particularly those derived from text-based representations such as TF-IDF weighted skills and metadata—rarely conform to this distributional requirement. In our case, the feature space is high-dimensional, sparse, and not normally distributed. Nevertheless, prior statistical research has demonstrated that LDA remains robust under such violations. Empirical studies[6] show that LDA continues to deliver reliable class separation and predictive performance even when applied to non-normally distributed predictors (e.g., Gamma or Uniform) (Izenman, 2008; Bate-Eya, 2024). A recent study (Adebiyi et al., 2022) on breast cancer diagnosis applied LDA as a dimensionality reduction technique and demonstrated high predictive accuracy, even though distributional assumptions were not explicitly tested. This robustness arises from the fact that classification accuracy is driven more by differences in group means than by strict adherence to normality. Consistent with these findings, we observed that applying LDA in our framework enhanced class separability and improved downstream model performance, confirming its practical suitability despite deviations from the theoretical assumptions.

4.4.1 Hyperparameter Tuning

We explored variations in the number of decision trees ($n_{\text{estimators}} \in \{100, 200, 300\}$) and the maximum tree depth (max_depth $\in \{10, 20, \text{None}\}$). In the case of GB, the grid included combinations of learning rate ($\eta \in \{0.01, 0.1, 0.2\}$), number of estimators, and maximum tree depth. For the LDA component, the number of extracted components (k) was constrained to $k \leq C - 1$, where $C = 7$ reflects the number of target classes in the dataset. Model configurations were evaluated using the average F1-score and ROC–AUC across folds, with final models selected based on the highest mean performance. This process ensured robust parameter tuning while preserving class distribution throughout training and validation.

4.5 Model Evaluation

To assess the classification performance of our supervised learning framework, we evaluated both the RF and GB classifiers using standard metrics: precision, recall, F1-score, and accuracy. These metrics were computed using the *classification_report* function from *Scikit-learn* (Pedregosa et al., 2011), applied on the test set predictions after

[6]Performance of LDA on non-normally distributed predictors, available at https://www.causeweb.org/usproc/sites/default/files/usresp/2022-2/performance%20of%20lda%20and%20qda%20on%20nonnormally%20distributed%20predictors.pdf

LDA dimensionality reduction. Each model was trained using the LDA-transformed training data and evaluated on a held-out test set comprising 20% of the dataset. The evaluation was performed across all seven socio-economic policy categories, with class-level support taken into account. In addition to these metrics, probability scores were also extracted using the *predict_proba()* function to enable further analysis through ROC curve visualization and AUC computation. This setup ensures a robust, multi-perspective evaluation of model effectiveness, both in terms of overall accuracy and per-class predictive quality.

5. EXPERIMENTAL RESULTS

This section presents the experimental results of the proposed supervised framework. In Subsection 5.1, we provide the exploratory results of the skills that belong to different categories. Subsection 5.2 reports the evaluation of the two classifiers, RF and GB. Finally, in Subsection 5.3, we illustrate how socio-economic skills are differentiated across the various policy areas.

5.1 Descriptive Statistics

The empirical evaluation of our framework offers valuable insights into how DPPs encode skill priorities across socio-economic domains. Through skill extraction, metadata enrichment, and supervised classification, we identify both accurate predictive models and interpretable policy patterns. These results demonstrate the potential of skill-aware machine learning to transform unstructured policy texts into structured, actionable intelligence.

Addressing RQ_1, we begin by analyzing the most frequent full skill expressions. Figure 2 displays the top 15 recurring skills, emphasizing themes such as digital transformation, social inclusion, and education. Prominent entries like “enhance digital learning” and “support digital infrastructure” underscore the emphasis on digital competencies (Pedregosa et al., 2011).

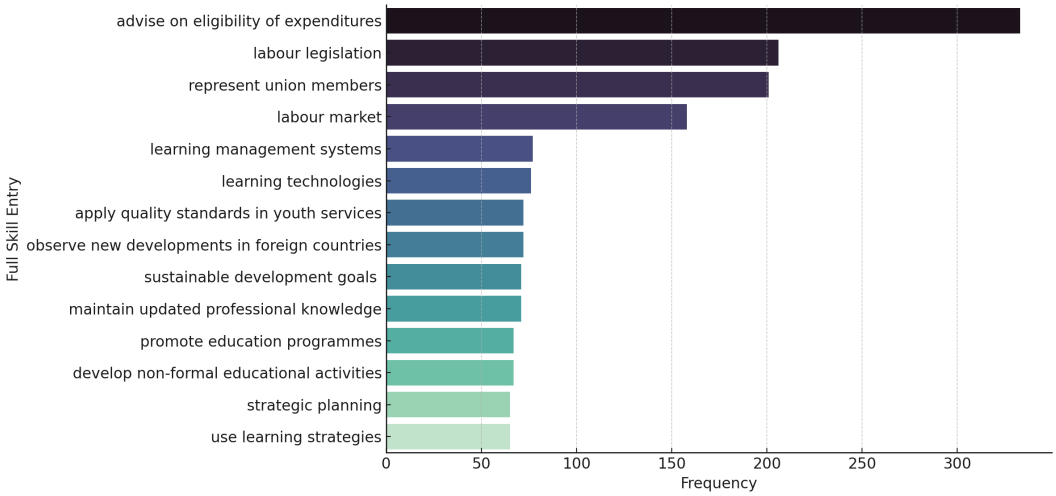


Figure 2: *Top 15 Most Frequent Socioeconomic Skills extracted from DPPs.*

Before delving into classification performance, we first explored the internal structure of the extracted skills using an *association rules graph*. This graph-based representation captures frequently co-occurring skill sets across digital policy publications (DPPs), enabling a relational understanding of how skills cluster in the policy landscape. Unlike frequency-based lists, association rules offer deeper insight into contextual skill bundles—revealing which technical or soft skills are often mentioned together and potentially aligned with shared strategic goals.

Figures 3 and 4 illustrate two such graphs, focusing on "labour market" and "learning management systems" respectively. These ego-centric networks reveal the most frequently co-mentioned skills that form coherent thematic clusters around each central node. The "labour market" graph exposes a wide skill orbit including "promote employment policy", "develop work programmes", and "assess candidates", reflecting a strong orientation toward active labour market measures, employment services, and human capital mobilization.

In contrast, the "learning management systems" graph is heavily connected to pedagogical and digital competences such as "construct individual learning plans", "use learning strategies", and "promote education programmes", capturing the digital transformation of educational delivery. The thickness of the edges further reinforces the intensity of co-occurrence, offering interpretable patterns of skill bundling that can inform curriculum design, training priorities, and policy alignment efforts.

Overall, these associations provide not just visualization but also structural evidence of how digital policies articulate interconnected skill ecosystems.

Figure 3: *Association rules ego-network centered on "Labour Market". The network reveals skill clusters around employment services and human capital, including "promote employment policy" and "assess candidates". Edge thickness indicates the intensity of co-occurrence.*

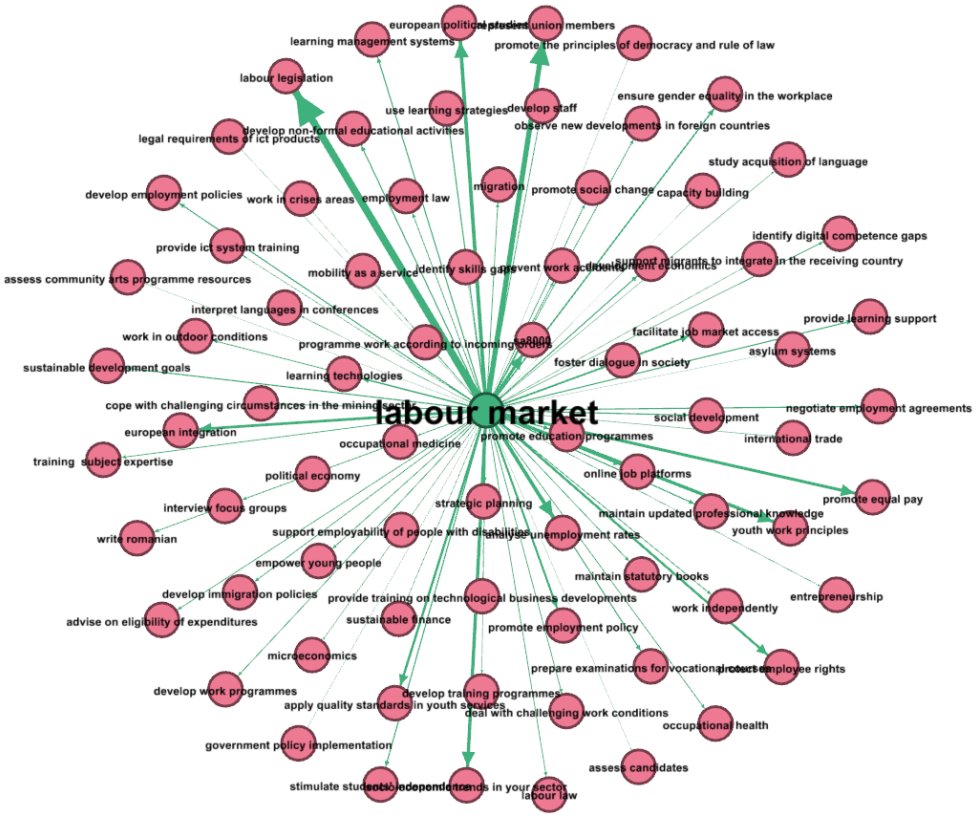
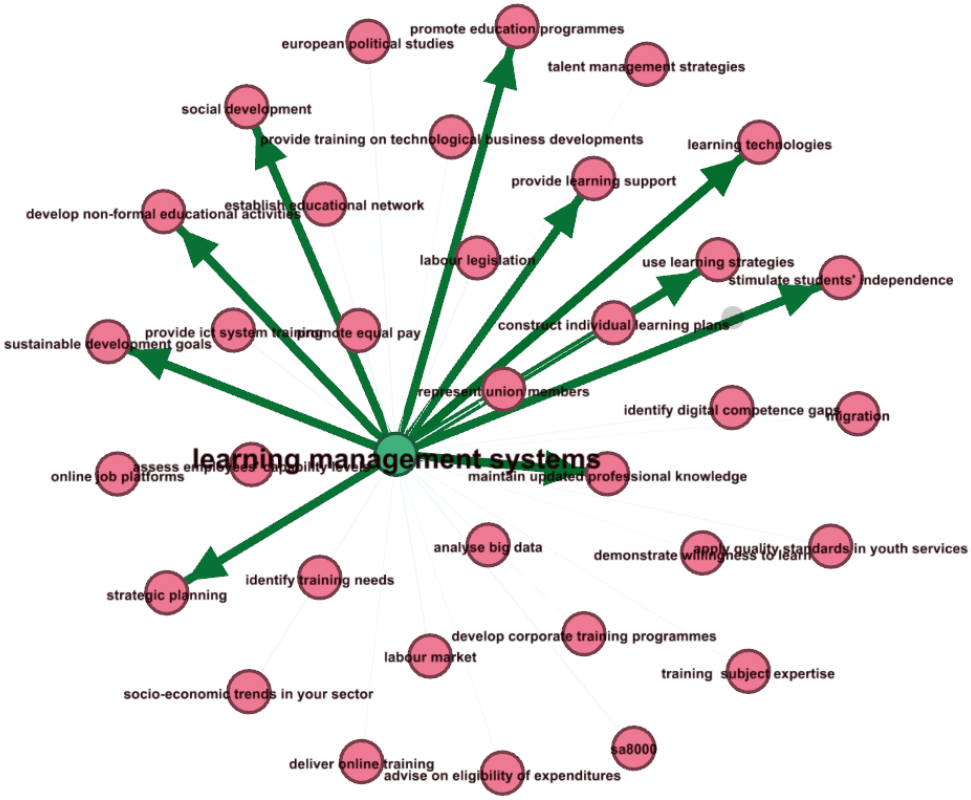


Figure 4: *Association rules ego-network centered on "Learning Management Systems". The graph highlights co-occurring skills, especially pedagogical and digital competences such as "construct individual learning plans" and "promote education programmes". Edge thickness reflects the frequency of co-occurrence.*



5.2 Performance Evaluation

Building on earlier statistical insights, we now assess the predictive value of skill-based representations in classifying DPPs (RQ_2). This section presents the machine learning pipeline, which transforms descriptive patterns into structured outputs using dimensionality reduction and supervised learning. Linear Discriminant Analysis (LDA) was employed to reduce the original 258 features—comprising skills and metadata—into six components that best distinguish policy categories. This not only improves computational efficiency but also enhances interpretability by revealing dominant skill signals.

Furthermore, to validate the LDA results, we tested the statistical significance of feature loadings by implementing a permutation-based null distribution (1,000 iterations). This procedure allowed us to compute empirical p-values for each feature–component association. As shown in Table 2, the interpretation of LDA components is now grounded in statistical evidence, ensuring that dominant features are not merely

descriptive but also supported by significance testing.

Table 2: *Top contributing features for selected LDA components with empirical p-values. Only statistically significant associations ($p < 0.05$) should be emphasized in interpretation.*

Feature	Weight	p-value
LDA Component 1		
public	4.33	0.00
inclusion	4.33	0.00
cooperate	3.04	0.01
colleagues	3.04	0.01
eligibility	2.45	0.00
expenditures	2.45	0.00
advise	2.45	0.00
challenges	2.42	0.00
treatment	2.42	0.00
human	2.42	0.00
LDA Component 3		
migration	5.19	0.000
sustainability	5.09	0.100
tropical	4.75	0.156
asylum	4.29	0.016
change	4.19	0.042
trade	3.48	0.056
finance	3.48	0.062
immigration	3.43	0.366
occupational	3.31	0.060
microeconomics	3.09	0.024

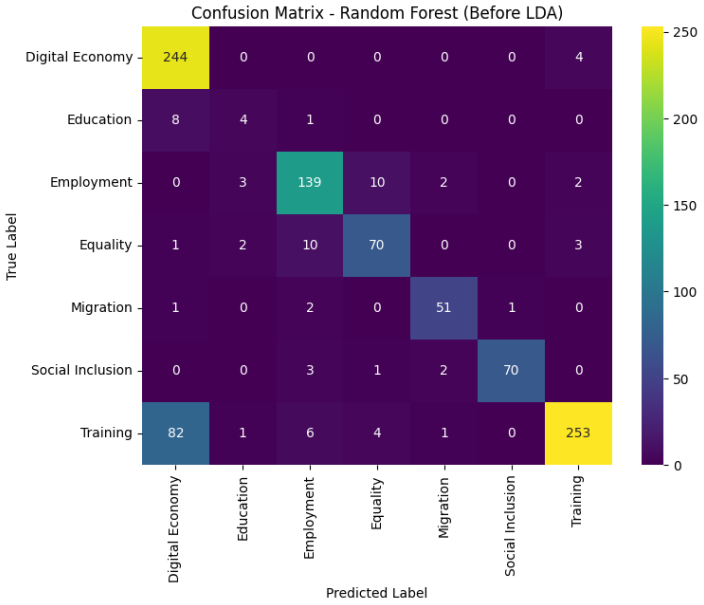
5.2.1 Random Forest Classifier

In a RF classifier, multiple decision trees collaboratively contribute to the final prediction by casting votes, with the majority class determining the output. This ensemble method was applied to the skill–metadata feature space both before and after dimensionality reduction via LDA.

Figure 5 presents the confusion matrix prior to applying LDA. The classification performance is largely comparable, with slight differences—such as 139 correct predictions for "Employment" and 253 for "Training"—suggesting that LDA contributed marginal gains in class separability. The misclassification of "Training" as "Digital Economy" remained consistent across both models, indicating feature overlap between these two domains. These visualizations reinforce that while LDA provides a compact feature space, the classifier maintains robust performance regardless of dimensionality reduction, with marginal improvements in clarity and interpretability.

Similarly, Figure 6 shows the confusion matrix after LDA was applied. The model correctly identified the majority of instances within the "Digital Economy" (244), "Employment" (138), and "Training" (254) classes, although some misclassifications persisted, such as the 82 instances of "Training" wrongly labeled as "Digital Economy."

Figure 5: *Confusion matrix of RF classifier before LDA.*



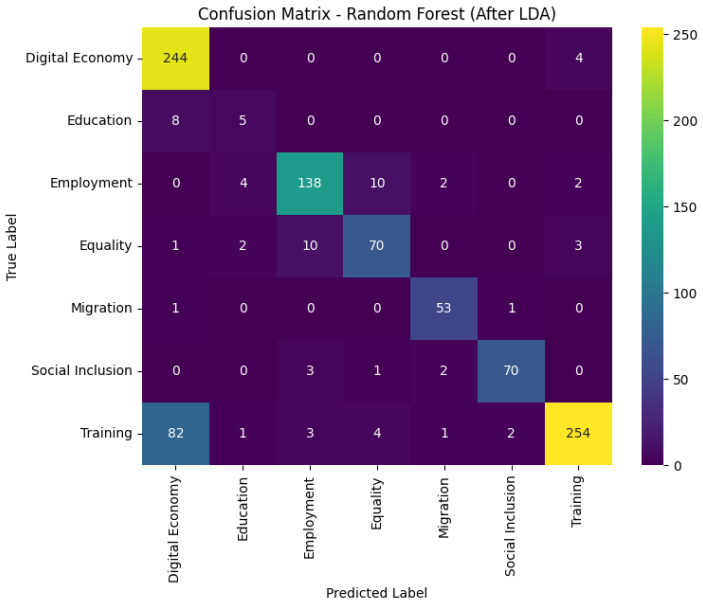
5.2.2 Gradient Boosting Classifier

In the pre-LDA matrix (Figure 7), the performance is comparable, with *Digital Economy* (244), *Employment* (139), and *Training* (253) again showing strong results. However, error dispersion is marginally higher—for example, *Migration* dropped to 51 correct predictions, and *Education* had only 4. The same 82 *Training→Digital Economy* misclassification persisted, confirming overlap between these two domains.

In the post-LDA scenario (Figure 8), the model correctly classified the majority of cases within *Digital Economy* (244), *Employment* (138), and *Training* (254). Some misclassifications remain—for instance, 82 *Training* instances misclassified as *Digital Economy*—but the results suggest slightly sharper separation, especially for *Migration* (53) and *Education* (5).

Overall, the comparison indicates that LDA provides only marginal improvements in class separability. While the overall classification strength of RF remains robust, LDA slightly improves interpretability and reduces error dispersion in mid-sized classes.

Figure 6: *Confusion matrix of RF classifier after LDA.*



5.2.3 Performance metrics

The performance comparison (Table 3) demonstrates that incorporating LDA improves the predictive accuracy of both RF and GB classifiers. The *LDA-only* baseline achieves moderate results, confirming that dimensionality reduction alone provides some discriminative power but is less effective than ensemble methods. The best overall performance is obtained by the LDA+RF model, which yields top scores in accuracy (0.85), precision (0.87), and F1-score (0.85). LDA also enhances GB's performance, particularly in precision and F1-score, suggesting improved pattern recognition. Although RF performs well even without LDA, the dimensionality reduction consistently boosts performance. GB, in particular, benefits more markedly from LDA, compensating for its lower baseline. These results highlight the effectiveness of integrating LDA in skill–policy classification tasks, especially for complex, high-dimensional data.

Figure 7: *Confusion matrix of GB classifier before LDA.*

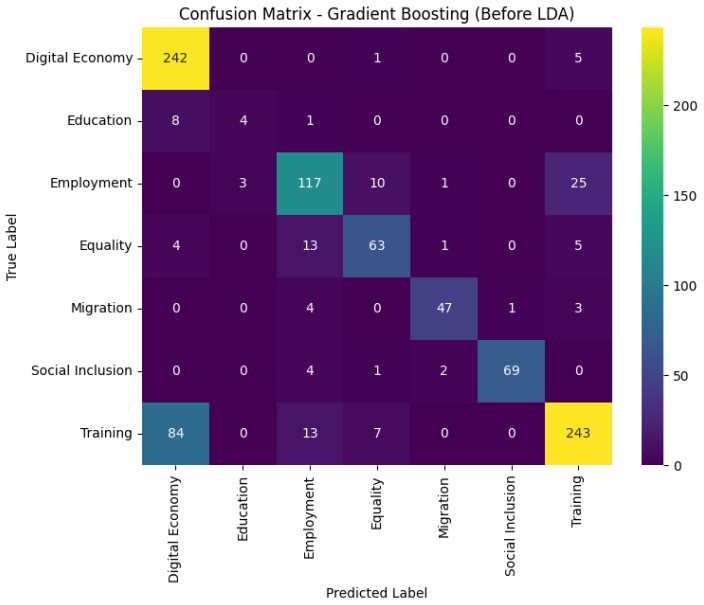


Table 3: *Performance evaluation of the classification models with and without LDA-based feature selection.*

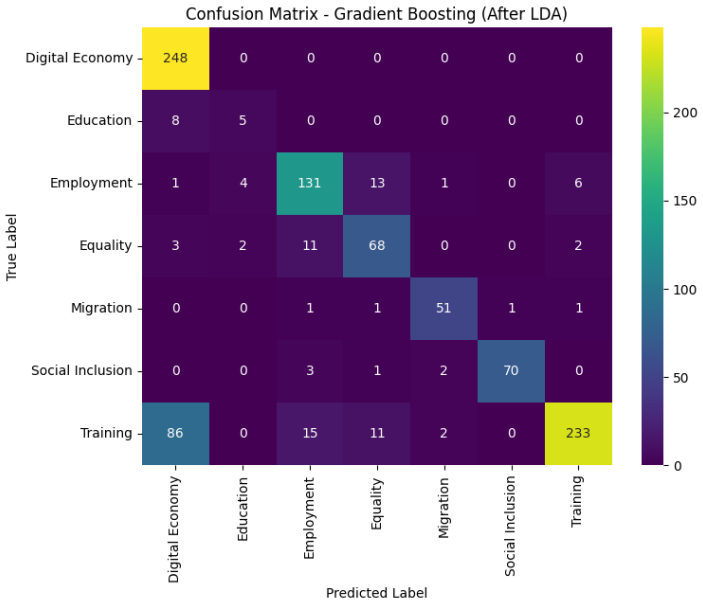
Metric	RF	LDA+RF	GB	LDA+GB	LDA
Accuracy	0.847095	0.850153	0.800204	0.821611	0.738022
Sensitivity	0.847095	0.850153	0.800204	0.821611	0.738022
Specificity	0.972023	0.972599	0.962417	0.967713	0.953801
Precision	0.867381	0.870936	0.812251	0.848037	0.781577
F1 Score	0.846284	0.849859	0.797349	0.819440	0.738406

Note: The additional **LDA** column was included, to provide a direct baseline for dimensionality reduction alone. This highlights that while LDA alone yields moderate performance, its integration with RF and GB consistently improves classification results.

5.3 Predicted Skill–Policy Associations

In response to RQ_3—"How can we quantitatively assess the alignment between skills and digital policy areas?"—this section presents model-derived associations between individual skills and strategic policy domains. Leveraging RF with LDA preprocessing, the model predicts probabilistic links between skills and categories such as Employment, Education, and Digital Economy. The findings in Table 4 indicate strong and interpretable patterns.

Figure 8: *Confusion matrix of GB classifier after LDA.*



For instance, *maintain updated professional knowledge* aligns with *Training* at 100%, while *online job platforms* is most associated with *Employment* (91%). The skill *advise on eligibility of expenditures* correlates with *Social Inclusion* (99.53%), highlighting the model's sensitivity to administrative and social priorities. The dual relevance of *promote equal pay* to both *Digital Economy* (71.49%) and *Training* (25.48%) underscores cross-cutting policy concerns. Overall, the results confirm the model's utility in revealing embedded skill–policy connections, enabling more targeted reskilling interventions and offering a scalable framework for policy analysis.

Table 4: *Predicted Skill–Policy Associations Based on RF Classifier*

Skill	**Digital Economy (%)**	**Education (%)**	**Employment (%)**	**Equality (%)**	**Migration (%)**	**Social Inclusion (%)**	**Training (%)**
represent union members	69.10	23.36	0.00	0.00	0.00	0.00	7.52
maintain updated professional knowledge	0.00	0.00	0.00	0.00	0.00	0.00	100.00
online job platforms	0.00	0.00	91.00	0.00	1.00	0.00	8.00
promote equal pay	71.49	2.00	0.00	1.02	0.00	0.00	25.48
advise on eligibility of expenditures	0.00	0.00	0.00	0.00	0.46	99.53	0.00

Note: Values represent predicted probabilities of association between skills and policy areas, expressed as percentages.

6. DISCUSSION & CONCLUSIONS

In this study, machine learning classifiers—RF and GB—were applied to a curated dataset of 981 DPPs, each enriched with skill annotations. The dataset was pre-processed and partitioned into training and testing subsets. Specifically, 80% of the dataset was used for training and feature modeling, while the remaining 20% was reserved for evaluating the predictive performance of the models.

To enhance classification accuracy, we employed LDA, a dimensionality reduction technique that extracts the most informative features, particularly useful in high dimensional policy text. Following LDA, both classifiers were trained on the transformed dataset. The RF model achieved an accuracy of 85.0%, while GB yielded 82.2%, with comparable results observed across metrics such as precision, recall, and F1-score. These findings suggest that LDA effectively improves the interpretability and performance of skill–policy classifiers, supporting earlier observations in the machine learning literature.

Skills were extracted using the NESTA mechanism and mapped to the ESCO taxonomy for consistency. By combining expert categorization with machine learning, our model assigned high-confidence skill associations to domains such as *Employment*, *Digital Economy*, and *Training*—with some exceeding 99%, particularly in *Training*. This demonstrates the potential for scalable, policy-aware skill forecasting. It also opens pathways for future research using embeddings, deep learning, and multilingual data to align education and labor strategies with shifting policy needs.

Acknowledgments

This work was conducted as part of the PhD research of the first author. The authors acknowledge funding from the European Union's Horizon Europe Framework Programme through the **SKILLAB** project (Grant Agreement No. 101132663). They also gratefully thank the anonymous reviewers for their thorough evaluation and insightful comments, which have significantly improved the clarity and rigor of this study.

References

Adebiyi, M. O., Arowolo, M. O., Mshelia, M. D., and Olugbara, O. O. (2022). A linear discriminant analysis and classification model for breast cancer diagnosis. *Applied Sciences*, 12(22):11455.

Adewumi, A. A. (2022). Protecting intangible cultural heritage in the era of rapid technological advancement. *International Review of Law, Computers & Technology*, 36(1):3–16.

Aluas, M., Angelis, L., Arapakis, I., Arvanitou, E., Georgiou, K., Gogos, A., Voulgarakis, V., et al. (2024). SKILLAB: Skills matter. In *2024 50th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, pages 491–498. IEEE.

Bate-Eya, A. E. (2024). Empirically comparing the performance of lda and qda when

classifying customer sales data with different properties of normality and equality of covariance matrices. *Unpublished manuscript*. Available online.

Caratozzolo, P., Azofeifa, J. D., Mejía-Manzano, L. A., Rueda-Castro, V., Noguez, J., Magana, A. J., and Benes, B. (2023a). A matrix taxonomy of knowledge, skills, and abilities (ksa) shaping 2030 labor market. In *2023 IEEE Frontiers in Education Conference (FIE)*, pages 1–8. IEEE.

Caratozzolo, S., Tolan, S., and Günther, C. W. (2023b). SkillNER: A Named Entity Recognition Approach for Skills Extraction in Online Job Ads. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ACL.

Cedefop (2023). Cedefop report on digital skills and employment – 2023. https://www.cedefop.europa.eu. Accessed: 2025-09-29.

Clavié, B. and Soulié, G. (2023). Large language models as batteries-included zero-shot esco skills matchers. *arXiv preprint arXiv:2307.03539*.

Decorte, J.-J., Van Hautte, J., Develder, C., and Demeester, T. (2025). Efficient text encoders for labor market analysis. *arXiv preprint arXiv:2505.24640*.

Djumalieva, J. and Sleeman, C. (2018). An Open and Data-driven Taxonomy of Skills Extracted from Online Job Adverts. *Nesta Working Paper*, (18/01).

European Commission (2022). European Year of Skills 2022. https://ec.europa.eu/skills/. Accessed: 2025-09-29.

International Labour Organization (2021). World employment and social outlook: Trends 2021.

Izenman, A. J. (2008). *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. Springer Texts in Statistics. Springer, New York.

Kavargyris, D. C., Georgiou, K., Maraki, I., Mittas, N., and Angelis, L. (2025a). Escoplus: A framework for enriching the esco taxonomy with digital skills from stack overflow. In *Euromicro Conference on Software Engineering and Advanced Applications*, pages 32–48, Cham. Springer Nature Switzerland.

Kavargyris, D. C., Georgiou, K., Mittas, N., and Angelis, L. (2025b). Extracting knowledge and highly demanded skills from european union educational policies: a network analysis approach. *Social Network Analysis and Mining*, 15(1):1–41.

Kavargyris, D. C., Georgiou, K., Papaioannou, E., Moysiadis, T., Mittas, N., and Angelis, L. (2025c). Future Skills in the GenAI Era: A Labor Market Classification System Using Kolmogorov–Arnold Networks and Explainable AI. *Algorithms*.

Kavargyris, D. C., Georgiou, K., Papaioannou, E., Petrakis, K., Mittas, N., and Angelis, L. (2025d). ESCOX: A tool for skill and occupation extraction using LLMs from unstructured text. *Software Impacts*, page 100772.

Lang, G. and Triantoro, T. (2022). Upskilling and Reskilling for the Future of Work: A Typology of Digital Skills Initiatives. *Information Systems Education Journal*, 20(4):97–106.

Organisation for Economic Co-operation and Development (2021). *OECD Employment Outlook 2021: Navigating the COVID-19 Crisis and Recovery*. OECD Publishing,

Paris.

Özdemir, A. and Tan, F. Z. (2023). *The impact of digital transformation on organizational agility*. Nobel Yayın Dağıtım, Ankara.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Sawant, R., Thomas, B., and Kadlag, S. (2022). Reskilling and upskilling: To stay relevant in today's industry. *International Review of Business and Economics*, 7(1):Article 4.

Standing, G. (2010). *Work after Globalization: Building Occupational Citizenship*. Edward Elgar, Cheltenham, UK.

Van-Duyet, L., Quan, V. M., and An, D. Q. (2017). Skill2vec: Machine learning approach for determining the relevant skills from job description. *arXiv preprint arXiv:1707.09751*.

Vasiliev, Y. (2020). *Natural Language Processing with Python and spaCy: A Practical Introduction*. No Starch Press, San Francisco, CA.

ΠΕΡΙΛΗΨΗ

Καθώς η ψηφιακή μετάβαση μετασχηματίζει την αγορά εργασίας, αναδεικνύεται η ανάγκη για πολιτικές που βασίζονται σε δεδομένα και προβλέπουν μελλοντικές δεξιότητες. Η μελέτη αυτή αναλύει 981 ψηφιακές πολιτικές δημοσιεύσεις (2019– 2024), αξιοποιώντας μηχανική μάθηση και εξαγωγή δεξιοτήτων βάσει της επίσημης ταξινομίας της ΕΕ. Κάθε πολιτικό κείμενο μετατρέπεται σε διανυσματική αναπαράσταση, περιλαμβάνοντας συχνότητες δεξιοτήτων, μεταδεδομένα και συμφραζόμενα. Η ταξινόμηση γίνεται σε επτά κοινωνικοοικονομικές κατηγορίες. Η Γραμμική Διαχωριστική Ανάλυση εφαρμόζεται για μείωση διαστασιμότητας και ενίσχυση διακριτικής ικανότητας. Τα αποτελέσματα δείχνουν υψηλή ακρίβεια ταξινόμησης και εντοπίζουν δεξιότητες με σημασία σε συγκεκριμένα πολιτικά πεδία. Το πλαίσιο αυτό προσφέρει εργαλείο για τεκμηριωμένη χάραξη πολιτικής, ευθυγραμμισμένη με τις ανάγκες της αγοράς εργασίας.

Greek Statistical Institute
Proceedings of the 37th Panhellenic & 3rd International Stat Conf (2025), pp.369-381

DIMENSIONALITY REDUCTION METHOD FOR FOUR-DIMENSIONAL DATA CUBES BASED ON VARIABILITY DECOMPOSITION

Z. Kyrana[1,*], N. Papafilippou[1], E. Pratsinakis[1], A. M. Michaelidou[2], C. Dordas[1], A. Markos[3], G. Menexes[1]

[1] Laboratory of Agronomy, Faculty of Agriculture, Forestry and Natural Environment, School of Agriculture, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece
{kyranaze, npapafil, epratsina, chdordas, gmenexes}@agro.auth.gr

[2] Laboratory of Dairy Technology, Faculty of Agriculture, Forestry and Natural Environment, School of Agriculture, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece
amichail@agro.auth.gr

[3] Department of Primary Education, Democritus University of Thrace, 68100 Alexandroupoli, Greece
amarkos@eled.duth.gr

[*] Corresponding Author

ABSTRACT

Data cubes are p-dimensional data structures, where $p \geq 3$. In Agriculture, the key point of Statistical-Data Analysis is the study of variability, focusing on its structure and components. The proposed method was based on Principal Components Analysis (PCA) and the idea of decomposing the total variability of a four-dimensional data cube ($p = 4$) into variability between and within cube's "slices" and "hyper-slices". The aim of this study was to propose new insights into the dimensionality reduction of four-dimensional $I \times J \times K \times G$ cubes with independent "slices" and "hyper-slices", and equal number of experimental units per "slices" within each "hyper-slice". As an example, a simulated dataset based on an agricultural experiment was used. The fourth dimension G consisted of three irrigation levels ("hyper-slices"). The third dimension K consisted of five basil varieties ("slices"). The second dimension J consisted of seven quantitative traits-variables, and the first dimension I consisted of different experimental plots (basil plants). The method showed that the structure of the cube's total variability was more closely associated with between-"slices" variability (basil varieties) than with the other sources of variability. Therefore, it is considered wiser to study the structure of the variability in depth, so that significant information can be extracted about any effect of "slices" or/and "hyper-slices" on the overall structure of the cube.

Keywords: multidimensional analysis; "hyper-slices"; principal components analysis

1. INTRODUCTION

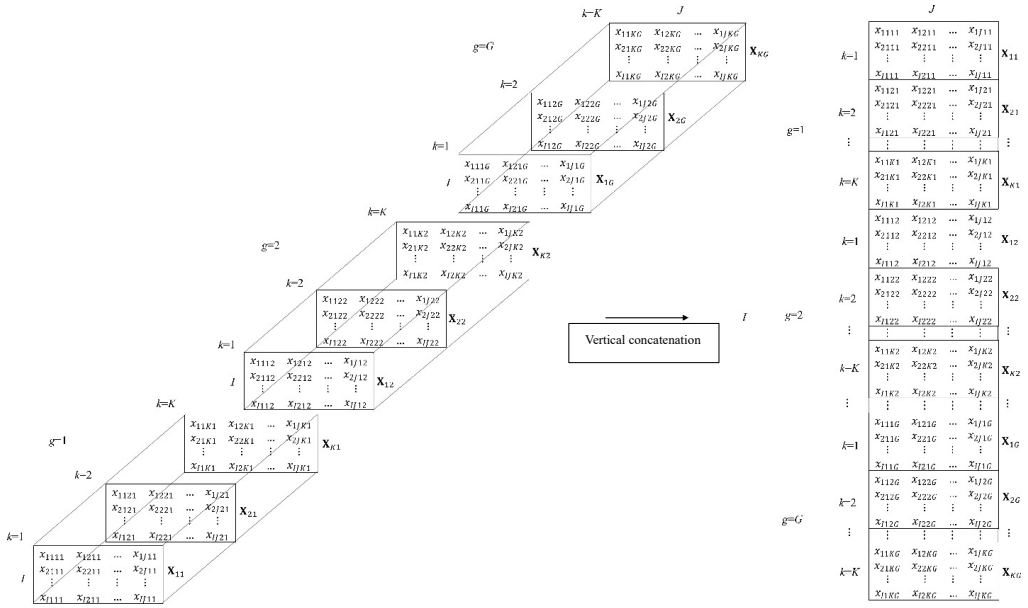
In many scientific fields and experimental procedures, data are represented as complex, multidimensional phenomena in the form of data cubes. Specifically, data cubes are *p*-dimensional data structures, where $p \geq 3$. *p* is determined by the number of independent dimensions of the data. *p* can take any integer value greater than or equal to 3, depending on the structure of the data. For example, in a three-dimensional data cube, the first dimension consists of objects (e.g., different plants), the second dimension consists of variables (e.g., quantitative variables), and the third dimension consists of the variable of "slices" (e.g., different varieties). If an additional fourth dimension ("hyper-slices") was added (e.g., different irrigation levels), the data cube would become four-dimensional.

Generally, data cubes appear under different structural characteristics. In most scientific fields, data cubes consist of dependent "slices" (i.e. repeated measurements of the same objects per "slice"). However, in most agricultural experiments data cubes consist of independent "slices", where the objects are different and equal or unequal in number per "slice" [Kroonenberg (2008)]. This study focused on four-dimensional data cubes with independent "slices", where the objects or else experimental units (e.g. experimental plots, pots or plants) were different and equal in number per "slice". In agricultural experiments it is crucial a design to be "balanced", that is, all experimental treatments are tested on the same number of experimental units (i.e. each treatment are equally replicated).

Regarding the mathematical background of a four-dimensional data cube with independent "slices" and "hyper-slices" and equal number of objects per "slice" and "hyper-slice", let $\underline{\mathbf{X}} = [\mathbf{X}_{11}, \mathbf{X}_{21}, \dots, \mathbf{X}_{K1}, \mathbf{X}_{12}, \mathbf{X}_{22}, \dots, \mathbf{X}_{K2}, \dots, \mathbf{X}_{1G}, \mathbf{X}_{2G}, \dots, \mathbf{X}_{KG}]$ be a four-dimensional $I \times J \times K \times G$ data cube (consequently $p = 4$), with x_{ijkg} elements, where $i = 1, 2, \dots, I$, $j = 1, 2, \dots, J$, $k = 1, 2, \dots, K$ and $g = 1, 2, \dots, G$. The first dimension *I* denotes the rows (experimental units or objects) in the k^{th} "slice" within g^{th} "hyper-slice". The second dimension *J* denotes the columns (traits or variables) of the data cube. The third dimension *K* denotes the independent "slices" of the data cube. The fourth dimension *G* denotes the independent "hyper-slices" of the data cube. In other words, the data cube $\underline{\mathbf{X}}$ can be seen as a collection of either three-dimensional data cubes $\underline{\mathbf{X}}_k$, with dimensions $IK \times J \times G$, or two-dimensional matrices $\mathbf{X}_{kg}$, with dimensions $IKG \times J$, where each matrix represents a "slice" *k* within "hyper-slice" *g* of a data cube [Rotari et. al. (2024)].

The most classic way to manage such a data cube is to transform it into a two-dimensional data matrix, through vertical concatenation [Codesido et. al. (2021)]. Specifically, data cube $\underline{\mathbf{X}}$ transformed into a "stacked" matrix $\mathbf{X}$ with dimensions $IKG \times J$ and x_{ijkg} elements, where $i = 1, 2, \dots, I$, $j = 1, 2, \dots, J$, $k = 1, 2, \dots, K$ and $g = 1, 2, \dots, G$ (see Figure 1).

Figure 1. *Transformation of the four-dimensional data cube* **X**, *with independent "slices" and "hyper-slices", and with different and equal number of objects per "slice" and "hyper-slice", into a "stacked" data matrix* **X** *(through vertical concatenation)*



In international scientific literature, there are several dimensionality reduction methods for data cubes. These methods are applied either to the "stacked" data matrices (e.g. PCA, Factor Analysis-FA, Multiple Correspondence Analysis-MCA, etc.) [Acar and Yener (2009); De Roover et. al. (2013); Pavoine et. al. (2007); Sun and Braatz (2020)], or to the overall structure of the data cube (e.g. Simultaneous Components Analysis-SCA, Multigroup PCA, Common PCA, STATIS, extensions of PARAFAC, etc.) [Acar and Yener (2009); Rotari et. al. (2024); Tenenhaus and Tenenhaus (2014)].

However, these methods are unable to study the variability of a data cube in depth. Also, some researchers applied the Multivariate Analysis of Variance (MANOVA). Specifically, MANOVA examines the differences in mean values between the "slices" with respect to the dependent quantitative variables, as well as their interactions. MANOVA also provides valuable information on how the "slices" differ across variables, focusing primarily on comparing the means of the "slices" and detecting overall multivariate differences [Rencher (2002)]. In contrast, the proposed method focuses on the decomposition and visualization of different variability components through lower-dimensional representations derived from PCA.

Therefore, the aim of this study was to propose new insights into the dimensionality reduction of four-dimensional data cubes by focusing on the variability's structure and components in more depth. Specifically, the proposed method was based on PCA and

the idea of decomposing the total variability into between-"slices" variability and within-"slices" variability.

2. MATERIALS AND METHODS

2.1 Dataset and Methods

A simulated dataset based on an agricultural experiment was used in this study. The agricultural experiment conducted at the Aristotle University of Thessaloniki Farm concerned five varieties of basil, three levels of irrigation and seven traits of 60 basil plants [Kalamartzis et. al. (2023)]. The first step of simulation was the finding of means, standard deviations and correlation matrices among traits per "slice" within each "hyper-slice" of the real "Basil" dataset. The second step was the simulation of a larger normally distributed datasets $\underline{\mathbf{X}}$ (with equal objects per "slice" and "hyper-slice" – balanced structure), with specific descriptive values and correlations of traits, per "slice" within each "hyper-slice" (via Python code).

The simulated dataset $\underline{\mathbf{X}}$ considered as a four-dimensional data cube with dimensions $500 \times 7 \times 5 \times 3$ and with independent "slices" and "hyper-slices". Data cube $\underline{\mathbf{X}}$ transformed into a "stacked" data matrix $\mathbf{X}(7{,}500 \times 7)$, through vertical concatenation. The first dimension consisted of different and equal number of objects (basil plants) per "slice" ($I = 500$). The second dimension consisted of the same seven continuous traits per "slice" ($J = 7$). Specifically, these were the measurements of Height (cm), Weight (gr), Leaf Area Index (LAI), Leaf Temperature (Leaf Temp), Photosynthesis, Water Use Efficiency (WUE), and Water Potential. The third dimension was a categorical variable denoting five "slices" ($K = 5$) corresponding to five basil varieties (Burns, Cinnamon, Sweet, Rend, Thai). The fourth dimension was a categorical variable denoting three "hyper-slices" ($G = 3$) corresponding to three irrigation levels (40%, 70%, 100%). Also, the number of basil plants in the g^{th} "hyper-slice" across all "slices" were $IG = 1{,}500$, while the number of basil plants in the k^{th} "slice" across all "hyper-slices" were $IK = 2{,}500$.

The proposed method was based on the PCA method without rotation [Greenacre et. al. (2022); Jolliffe and Cadima (2016)]. Alternatively, FA or Singular Value Decomposition (SVD) on the appropriate data matrix (i.e. the correlation matrix) could be used [Smilde et. al. (2022)]. In other cases, such as where the data lies on non-Euclidian spaces there are some other methods for dimensionality reduction such as t-SNE [Hinton and Roweis (2002)]. For the purposes of interpretation and visualization and based on the *a priori* criterion [Hair et. al. (2010)], only the first two principal components were selected. The statistical software package used was R 4.3.2.

2.2 Method's mathematical background

The proposed method was based on the idea of decomposing the total variability of the data cube into between-"slices" variability and within-"slices" variability:

$$\mathbf{X_T} = \mathbf{X_B} + \mathbf{X_W}$$

However, in terms of four-dimensional data cubes it's true that:

$$\mathbf{X_T} = \mathbf{X_G} + \mathbf{X_K} + \mathbf{X_{KG}} + \mathbf{X_W}$$

More analytically, assume a "stacked" data matrix $\mathbf{X}$ with dimensions $IKG \times J$ and x_{ijkg} elements, where $i = 1, 2, \ldots, I, j = 1, 2, \ldots, J, k = 1, 2, \ldots, K$ and $g = 1, 2, \ldots, G$.

Let $\mathbf{x}_{ikg} = \begin{bmatrix} x_{i1kg} \\ x_{i2kg} \\ \vdots \\ x_{iJkg} \end{bmatrix}_{J\times 1}$ a J-dimensional vector of variables for object i in "slice" k within "hyper-slice" g. The J-dimensional mean vector for the g^{th} "hyper-slice" is $\bar{\mathbf{x}}_{..g} = \frac{1}{IK}\sum_{k=1}^{K}\sum_{i=1}^{I}\mathbf{x}_{ikg} = \begin{bmatrix} \bar{x}_{.1.g} \\ \bar{x}_{.2.g} \\ \vdots \\ \bar{x}_{.J.g} \end{bmatrix}_{J\times 1}$. The J-dimensional mean vector for the k^{th} "slice" is $\bar{\mathbf{x}}_{.k.} = \frac{1}{IG}\sum_{g=1}^{G}\sum_{i=1}^{I}\mathbf{x}_{ikg} = \begin{bmatrix} \bar{x}_{.1k.} \\ \bar{x}_{.2k.} \\ \vdots \\ \bar{x}_{.Jk.} \end{bmatrix}_{J\times 1}$. The J-dimensional mean vector for the k^{th} "slice" within "hyper-slice" g is $\bar{\mathbf{x}}_{.kg} = \frac{1}{I}\sum_{i=1}^{I}\mathbf{x}_{ikg} = \begin{bmatrix} \bar{x}_{.1kg} \\ \bar{x}_{.2kg} \\ \vdots \\ \bar{x}_{.Jkg} \end{bmatrix}_{J\times 1}$. The J-dimensional mean vector for the entire "stacked" matrix $\mathbf{X}$ is $\bar{\mathbf{x}}_{...} = \frac{1}{IKG}\sum_{g=1}^{G}\sum_{k=1}^{K}\sum_{i=1}^{I}\mathbf{x}_{ikg} = \begin{bmatrix} \bar{x}_{.1..} \\ \bar{x}_{.2..} \\ \vdots \\ \bar{x}_{.J..} \end{bmatrix}_{J\times 1}$.

The Sums of Squares and Cross Products (SSCP) matrix of total variability is defined as a $J \times J$ data matrix $\mathbf{X_t} = \sum_{g=1}^{G}\sum_{k=1}^{K}\sum_{i=1}^{I}(\mathbf{x}_{ikg} - \bar{\mathbf{x}}_{...})(\mathbf{x}_{ikg} - \bar{\mathbf{x}}_{...})^{\mathrm{T}}$. The SSCP matrix of between-"slices" variability is defined as a $J \times J$ data matrix $\mathbf{X_k} = \sum_{k=1}^{K} IG(\bar{\mathbf{x}}_{.k.} - \bar{\mathbf{x}}_{...})(\bar{\mathbf{x}}_{.k.} - \bar{\mathbf{x}}_{...})^{\mathrm{T}}$. The SSCP matrix of between-"hyper-slices" variability is defined as a $J \times J$ data matrix $\mathbf{X_g} = \sum_{g=1}^{G} IK(\bar{\mathbf{x}}_{..g} - \bar{\mathbf{x}}_{...})(\bar{\mathbf{x}}_{..g} - \bar{\mathbf{x}}_{...})^{\mathrm{T}}$. The SSCP matrix of interaction between "slices" and "hyper-slices" variability is defined as a $J \times J$ data matrix $\mathbf{X_{kg}} = \sum_{g=1}^{G}\sum_{k=1}^{K} I\left(\bar{\mathbf{x}}_{.kg} - \bar{\mathbf{x}}_{..g} - \bar{\mathbf{x}}_{.k.} + \bar{\mathbf{x}}_{...}\right)(\bar{\mathbf{x}}_{.kg} -$

$\bar{\mathbf{x}}_{..g} - \bar{\mathbf{x}}_{.k.} + \bar{\mathbf{x}}_{...})^{\mathrm{T}}$. The SSCP matrix of within "slices" and "hyper-slices" variability is defined as a $J \times J$ data matrix $\mathbf{X_w} = \sum_{g=1}^{G}\sum_{k=1}^{K}\sum_{i=1}^{I}(\mathbf{x}_{ikg} - \bar{\mathbf{x}}_{.kg})(\mathbf{x}_{ikg} - \bar{\mathbf{x}}_{.kg})^{\mathrm{T}}$.

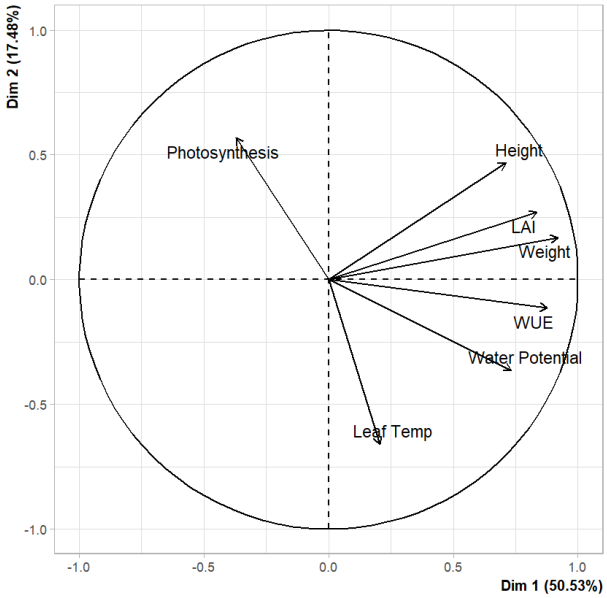
The corresponding $IKG \times J$ data matrix of total variability is defined as $\mathbf{X_T} = \mathbf{X}$. The $IKG \times J$ data matrix of between-"slices" variability is defined as $\mathbf{X_K} = repmat\left(\begin{bmatrix}\mathbf{1}_I\bar{\mathbf{x}}_{.1.}^{\mathrm{T}}\\ \mathbf{1}_I\bar{\mathbf{x}}_{.2.}^{\mathrm{T}}\\ \vdots \\ \mathbf{1}_I\bar{\mathbf{x}}_{.K.}^{\mathrm{T}}\end{bmatrix}, G\right)$, where $repmat(\mathbf{X_K}, G)$ replicates vertically the matrix $\mathbf{X_K}$ G times. The $IKG \times J$ data matrix of between-"hyper-slices" variability is defined as $\mathbf{X_G} = \begin{bmatrix}\mathbf{1}_{IK}\bar{\mathbf{x}}_{..1}^{\mathrm{T}}\\ \mathbf{1}_{IK}\bar{\mathbf{x}}_{..2}^{\mathrm{T}}\\ \vdots \\ \mathbf{1}_{IK}\bar{\mathbf{x}}_{..G}^{\mathrm{T}}\end{bmatrix}$. The $IKG \times J$ data matrix of the interaction between "hyper-slices" and "slices" variability is defined as $\mathbf{X_{KG}} = \begin{bmatrix}\mathbf{1}_I\bar{\mathbf{x}}_{.11}^{\mathrm{T}}\\ \mathbf{1}_I\bar{\mathbf{x}}_{.21}^{\mathrm{T}}\\ \vdots \\ \mathbf{1}_I\bar{\mathbf{x}}_{.KG}^{\mathrm{T}}\end{bmatrix} - \mathbf{X_G} - \mathbf{X_K}$. The $IKG \times J$ data matrix of within "hyper-slices" and "slices" variability is defined as $\mathbf{X_W} = \mathbf{X} - \begin{bmatrix}\mathbf{1}_I\bar{\mathbf{x}}_{.11}^{\mathrm{T}}\\ \mathbf{1}_I\bar{\mathbf{x}}_{.21}^{\mathrm{T}}\\ \vdots \\ \mathbf{1}_I\bar{\mathbf{x}}_{.KG}^{\mathrm{T}}\end{bmatrix}$.

The next step is the conversion of SSCP matrices into correlation matrices **R**. The final step is the application of SVD on each correlation matrix or the application of PCA on each variability data matrix.

3. RESULTS

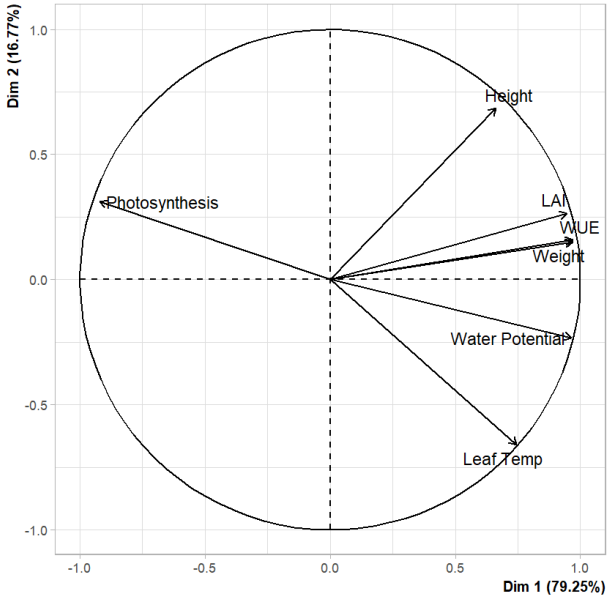
The results of the proposed method on $\underline{\mathbf{X}}$ are presented below. Specifically, the PCA loading diagram for the total variability $\mathbf{X_T}$ is shown in Figure 2. Height, Weight, LAI, WUE, Water Potential contributed positively to the first principal component (PC1), while Photosynthesis contributed positively and Leaf Temp negatively to the second principal component (PC2). PC1 explained 50.5% of the variance, PC2 explained 17.5% of the variance, while both principal components explained 68% of the total variance.

Figure 2. *PCA loading plot for total variability. The axes represent the first two principal components (principal component 1, 2 - Dim 1, 2) with the corresponding percentages of the variance explained by each principal component, and the vectors (arrows) represent the loadings of each variable on the principal components (analysis and visualization in R)*



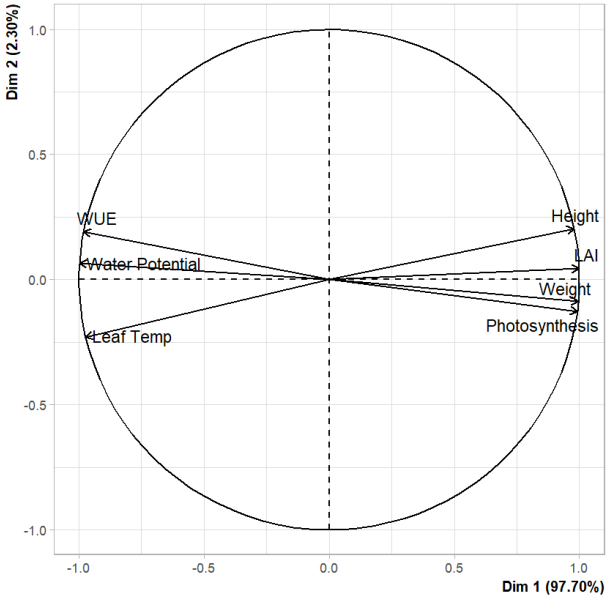
The PCA loading diagram for the between-"slices" variability $\mathbf{X_K}$ is shown in Figure 3. Weight, LAI, Leaf Temp, WUE, Water Potential contributed positively, and Photosynthesis contributed negatively to PC1, while Height contributed positively to PC2. PC1 explained 79.3% of the variance, PC2 explained 16.8% of the variance, while both principal components explained 96.1% of the total variance.

Figure 3. *PCA loading plot for between-"slices" variability. The axes represent the first two principal components (principal component 1, 2 - Dim 1, 2) with the corresponding percentages of the variance explained by each principal component, and the vectors (arrows) represent the loadings of each variable on the principal components (analysis and visualization in R)*



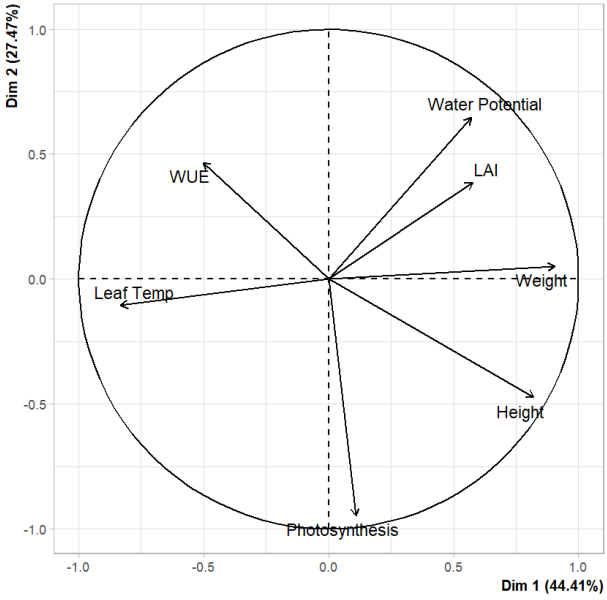
The PCA loading diagram for the between-"hyper-slices" variability $\mathbf{X_G}$ is shown in Figure 4. Height, Weight, LAI and Photosynthesis contributed positively to PC1, while Leaf Temp, WUE and Water Potential contributed negatively to PC1. PC1 explained 97.7% of the variance, PC2 explained 2.3% of the variance, while both principal components explained 100% of the total variance.

Figure 4. *PCA loading plot for between-"hyper-slices" variability. The axes represent the first two principal components (principal component 1, 2 - Dim 1, 2) with the corresponding percentages of the variance explained by each principal component, and the vectors (arrows) represent the loadings of each variable on the principal components (analysis and visualization in R)*



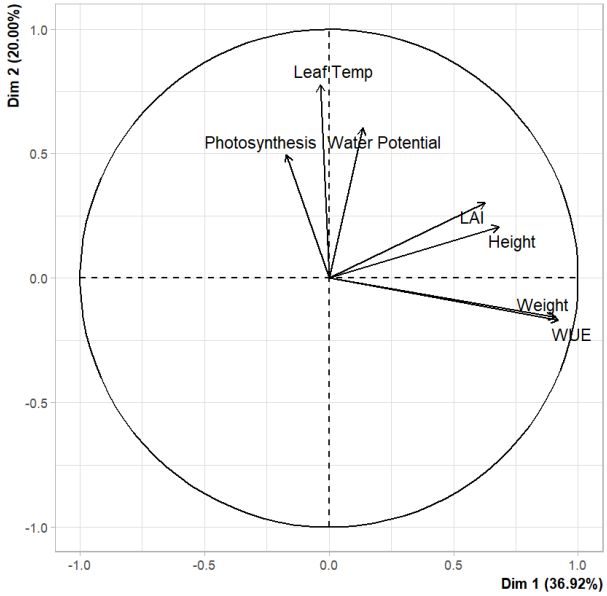
The PCA loading diagram for the interaction between the "slices" and "hyper-slices" variability $\mathbf{X_{KG}}$ is shown in Figure 5. Height, Weight and LAI contributed positively and Leaf Temp and WUE negatively to PC1, while Water Potential contributed positively and Photosynthesis negatively to PC2. PC1 explained 44.4% of the variance, PC2 explained 27.5% of the variance, while both principal components explained 71.9% of the total variance.

Figure 5. *PCA loading plot for the interaction between "slices" and "hyper-slices" variability. The axes represent the first two principal components (principal component 1, 2 - Dim 1, 2) with the corresponding percentages of the variance explained by each principal component, and the vectors (arrows) represent the loadings of each variable on the principal components (analysis and visualization in R)*



The PCA loading diagram for the within "slices" and "hyper-slices" variability $\mathbf{X_W}$ is shown in Figure 6. Height, Weight, LAI and WUE contributed positively to PC1, while Leaf Temp, Photosynthesis and Water Potential contributed positively to PC2. PC1 explained 36.9% of the variance, PC2 explained 20.0% of the variance, while both principal components explained 56.9% of the total variance.

Figure 6. *PCA loading plot for the within "slices" and "hyper-slices" variability. The axes represent the first two principal components (principal component 1, 2 - Dim 1, 2) with the corresponding percentages of the variance explained by each principal component, and the vectors (arrows) represent the loadings of each variable on the principal components (analysis and visualization in R)*



3. CONCLUSION

This study showed an alternative dimensionality reduction method for four-dimensional data cubes, based on PCA and the study of variability. This method provided an opportunity to examine in detail the influence of the cube "slices" and "hyper-slices" on the overall data cube structure.

The results of the total variability showed an almost common structure of variability with between-"slices" variability. In contrast, the results of the total variability showed a different structure from between-"hyper-slices" variability, interaction between "slices" and "hyper-slices" variability and within "slices" and "hyper-slices" variability.

Based on the above, the proposed method provided interesting insights into the structure and components of variability, through visualizations in lower dimensions. It is worth nothing that the choice of the SSCP matrix or matrices depends on the aims of the study. Also, the proposed method could be used complementary to the MANOVA method by providing deeper insights into data cubes' variability.

Therefore, studying the variability structure in depth may provide useful information regarding the relative contribution of “slices” and “hyper-slices” to the overall data cube structure.

ACKNOWLEDGEMENTS

We would like to thank Dr Iakovos Kalamartzis and the Laboratory of Agronomy, School of Agriculture, Aristotle University of Thessaloniki, for granting us access to the experiment's primary data.

ΠΕΡΙΛΗΨΗ

Οι κύβοι δεδομένων αποτελούν p-διάστατες δομές, όπου $p \geq 3$. Στις Γεωπονικές Επιστήμες, το σημαντικότερο παραδοτέο της Στατιστικής Ανάλυσης Δεδομένων είναι η μελέτη της μεταβλητότητας (παραλλακτικότητας, διακύμανσης, διασποράς, αδράνειας), εστιάζοντας στη δομή και στις συνιστώσες της. Η προτεινόμενη μέθοδος βασίστηκε στην Ανάλυση σε Κύριες Συνιστώσες και στην ιδέα της διάσπασης της συνολικής μεταβλητότητας ενός τετραδιάστατου κύβου δεδομένων ($p = 4$) σε μεταβλητότητα μεταξύ και σε μεταβλητότητα εντός των “φετών” και των “υπερ-φετών” του κύβου. Σκοπός της μελέτης ήταν η πρόταση νέων ιδεών σχετικά με τη μείωση διαστάσεων τετραδιάστατων κύβων δεδομένων $I \times J \times K \times G$, με ανεξάρτητες “φέτες” και “υπερ-φέτες”, και ίσο αριθμό πειραματικών μονάδων ανά “φέτα” εντός κάθε “υπερ-φέτας”. Ως παράδειγμα χρησιμοποιήθηκε ένα προσομοιωμένο σύνολο δεδομένων βασισμένο σε γεωπονικό πείραμα. Η τέταρτη διάσταση G αποτελούνταν από τρία επίπεδα άρδευσης (“υπερ-φέτες”). Η τρίτη διάσταση K αποτελούνταν από πέντε ποικιλίες βασιλικού (“φέτες”). Η δεύτερη διάσταση J αποτελούνταν από επτά ποσοτικά χαρακτηριστικά-μεταβλητές και η πρώτη διάσταση I αποτελούνταν από διαφορετικά πειραματικά τεμάχια (φυτά βασιλικού). Η μέθοδος έδειξε ότι η δομή της συνολικής μεταβλητότητας του κύβου συνδέεται περισσότερο με τη μεταβλητότητα μεταξύ των “φετών” (ποικιλίες βασιλικού) σε σχέση με τις υπόλοιπες πηγές μεταβλητότητας. Ως εκ τούτου, κρίνεται σοφότερο να πραγματοποιείται η μελέτης της δομής της μεταβλητότητας εις βάθος, έτσι ώστε να εξαχθεί σημαντική πληροφορία για την τυχόν επίδραση των “φετών” ή/και των “υπερ-φετών” στη δομή της ολικής μεταβλητότητας του κύβου.

REFERENCES

Acar, E. and Yener, B. (2009). Unsupervised Multiway Data Analysis: A Literature Survey. *IEEE Transactions on Knowledge & Data Engineering*, **21**(1), 6-20.

Codesido, S., Hanafi, M., Gagnebin, Y., González-Ruiz, V., Rudaz, S. and Boccard, J. (2021). Network principal component analysis: a versatile tool for the investigation of multigroup and multiblock datasets. *Bioinformatics*, **37**(9), 1297-1303.

De Roover, K., Ceulemans, E., Timmerman, M. E., Nezlek, J. B. and Onghena, P. (2013). Modeling differences in the dimensionality of multiblock data by means of clusterwise simultaneous component analysis. *Psychometrika*, **78**(4), 648–668.

Greenacre, M., Groenen, P. J. F., Hastie, T., Iodice D'Enza A., Markos, A. and Tuzhilina, E. (2022). Principal component analysis. *Nat Rev Methods Primers*, **2**, 100.

Hair, J. F., Black, W. C., Babin, B. J. and Anderson, R. E. (2010). *Multivariate Data Analysis: A Global Perspective* (7th Ed.). New York: Pearson Education Inc.

Hinton, G. E. and Roweis, S. (2002). Stochastic neighbor embedding. *Advances in neural information processing systems*, **15**.

Jolliffe, I. T. and Cadima, J. (2016). Principal component analysis: a review and recent developments. *Phil Trans R Soc A*, **374**(2065), 20150202.

Kalamartzis, I., Papakaloudis, P. and Dordas, C. (2023). Basil (Ocimum basilicum) Landraces Can Be Used in a Water-Limited Environment. *Plants*, **12**(13), 2425.

Kroonenberg, P. M. (2008). *Applied Multiway Data Analysis*. New Jersey: Wiley Series in Probability and Statistics.

Pavoine, S., Blondel, J., Baguette, M. and Chessel, D. (2007). A new technique for ordering asymmetrical three-dimensional data sets in ecology. *Ecology*, **88**(2), 512-523.

Rencher, A. C. (2002). *Methods of multivariate analysis*. New Jersey: Wiley Series in Probability and Statistics.

Rotari, M., Diaz, V. F., De Ketelaere, B. and Kulahci, M. (2024). An extension of PARAFAC to analyze multi-group three-way data. *Chemometrics and Intelligent Laboratory Systems*, **246**, 105089.

Smilde, B. J., Botman, E., de Ruiter, R. D., Smit, J. M., Teunissen, B. P., Lubbers, W. D., Schwarte, L. A., Schober, P. and Eekhoff, E. M. W. (2022). Monitoring and Management of Fibrodysplasia Ossificans Progressiva: Current Perspectives. *Orthopedic research and reviews*, **14**, 113–120.

Sun, W. and Braatz, R. D. (2020). Opportunities in tensorial data analytics for chemical and biological manufacturing processes. *Comput Chem Eng*, **143**, 107099.

Tenenhaus, A. and Tenenhaus, M. (2014). Regularized generalized canonical correlation analysis for multiblock or multigroup data analysis. *Eur J Oper Res*, **238**(2), 391-403.

Greek Statistical Institute
Proceedings of the 37th Panhellenic & 3rd International Stat Conf (2025), pp.383-399

Skills in Policies: Understanding Digital Competencies in European Policies through Causal Discovery

Evangelia Liotiri[1], Maria Ganopoulou[1], Dimitrios Christos Kavargyris[1] and Lefteris Angelis[1]
[1]Department of Informatics, Aristotle University of Thessaloniki

evaliotiri@csd.auth.gr , marigano@csd.auth.gr , dkavargy@csd.auth.gr and lef@csd.auth.gr

ABSTRACT

Green policies have become a cornerstone of the European Union's strategic agenda, emphasizing sustainability, climate resilience, and digital transformation across sectors. Yet, translating these ambitious objectives into actionable skill frameworks remains a significant challenge for policymakers and institutions. To address this gap, this study applies causal discovery techniques to analyze how digital and green competencies are embedded within EU policy discourse. By systematically examining 1,201 policy documents retrieved from EUR-Lex (2019-2024) and extracting ESCO skills across six hierarchical layers, we construct Directed Acyclic Graphs, to reveal the underlying causal structures. Using the Hill-Climb algorithm optimized with the Bayesian Information Criterion, we identify key causal pathways linking skills and policy descriptors. Additionally, we leverage the Markov Blanket method to isolate the minimal, policy-relevant sets of skills, offering actionable insights for strategic workforce development and evidence-based policy design in the context of Europe's twin transitions.

Keywords: Digital Competencies, ESCO Skills, Causal Discovery, European Policies, Hill-Climb Algorithm, Markov blanket

1. INTRODUCTION

The European Union has set forth a bold agenda to drive digital transformation and achieve climate neutrality under the European Green Deal. Central to this vision is a workforce equipped with digital and sustainability competencies capable of navigating rapid technological advances and environmental challenges. Yet, a significant skills mismatch persists: Eurostat reports that over 44 % of EU citizens lack basic digital competencies, impeding both individual employability and broader economic resilience. Concurrently, the Digital Skills and Jobs Coalition highlights chronic shortages of ICT specialists across Member States (European Commission, 2021a), while the Green Deal's target of 1 million new green jobs by 2030 (and nearly 2 million by 2050) underscores the urgency of effective reskilling and upskilling programs (European Commission, 2021b).

To standardize skill and policy vocabularies, the EU relies on two expert-curated frameworks: ESCO (European Skills/Competences, Qualifications and Occupations) for labor-market skills, and EuroVoc for policy descriptors (ESCO, 2024; EuroVoc, 2024). The ESCO taxonomy contains approximately 14,000 different skills and remains updated to this day, while a study by Kavargyris et al. (2025d) attempted to extend this taxonomy with digital skills. However, these taxonomies alone cannot reveal the directional (causal) relationships that link specific competences to policy objectives.

In response, recent research integrates causal inference and causal discovery methods to untangle complex interdependencies in policy contexts. Quasi-experimental designs such as difference-in-differences and propensity score matching, have been used to assess the impacts of education and labor reforms (Cordero et al. 2017). Meanwhile, structure-learning algorithms for Directed Acyclic Graphs (DAGs) provide data-driven maps of causal pathways; for example, Bayesian network models have been applied to support decision-making in diverse policy domains (Marchant et al., 2025).

Building on these advances, this study applies causal discovery to a corpus of 1,201 EU policy documents (2019–2024), extending the work of Kavargyris et al. (2025a), which quantified similar analyses within the domain of education policies. The methodology involves the extraction of ESCO-aligned skills across six hierarchical layers to capture varying levels of policy granularity. Subsequently, Directed Acyclic Graphs (DAGs) are learned using the Hill-Climb algorithm optimized with the Bayesian Information Criterion (BIC) to reveal causal dependencies between skills. For each descriptor, the Markov blanket, namely, the minimal set of parent, child and co-parent skills, isolates the competencies most critical to both policy formation and implementation.

Two key research questions (RQ) guide this analysis:

- **RQ_1:** How do ESCO skills associated with EuroVoc descriptors evolve across hierarchical ESCO layers?
- **RQ_2:** How can causal models distinguish between skills that drive policy formation and those that emerge as consequences of policy implementation?

Our findings offer actionable insights for aligning curricula and training programs with the EU's digital and green transition goals, supporting evidence-based policy design and workforce development.

The paper is organized as follows. Section 1 provides an introduction to the EU's green-digital agenda, outlines the skills mismatch challenge, and reviews relevant causal-discovery literature. Section 2 details the data and methodological framework: policy document retrieval and annotation, ESCO skill extraction, co-occurrence analysis, mutual-information-based feature selection, Hill-Climb structure learning with BIC scoring, and Markov-blanket extraction. Section 3 presents the results, including layer-

specific DAGs for key EuroVoc descriptors and their associated Markov blankets. Section 4 discusses avenues for future work such as empirical validation against labor-market data, incorporation of expert priors, and embedding-based feature enrichment.

2. DATA AND METHOD

2.1 Data Collection and Preprocessing

A total of 1,201 policy documents covering January 2019-December 2024 were retrieved from EUR-Lex (the EU's official law database) using the focused search queries "digital skills," "reskilling initiatives," "climate action," "renewable energy," and "sustainability." Legislative and strategic instruments (Regulations, Directives, Decisions, Communications) were then filtered for relevance to green and digital transition themes, after which each document's full HTML source was downloaded. Metadata fields (see Table 1) were extracted automatically, and embedded tables were transcribed manually to preserve their semantic structure.

Table 1. *Metadata fields for search queries*

Field	Value
Title	Regulation (EU) 2023/1115 of the European Parliament and of the Council of 31 May 2023 establishing a framework for sustainable food systems
Subtitle	PE/13/2023/REV/2
CELEX Number	32023R1115
Form	Regulation
Author	European Parliament, Council of the European Union
Procedure Number	2022/0301/COD
Date of Document	31/05/2023
EuroVoc Descriptor	food security, sustainable agriculture, ecological transition, biodiversity protection, reduction of greenhouse gases, EU environmental policy
ELI	http://data.europa.eu/eli/reg/2023/1115/oj
Policy Text	REGULATION (EU) 2023/1115 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 31 May 2023 on the sustainable transformation of EU food systems

Every document in our corpus is annotated with expert-assigned EuroVoc descriptors (EuroVoc, 2024). EuroVoc provides a controlled multilingual thesaurus for EU policy content, ensuring consistent thematic tagging across languages and institutions. We extracted each document's EuroVoc labels (e.g., "green economy," "energy policy") to serve as target variables in subsequent causal modeling.

To identify workforce competencies referenced in policy texts, a Natural Language Processing pipeline was applied based on NESTA's methodology (Nesta, 2023). This pipeline includes tokenization, lemmatization, and fuzzy matching against the ESCO taxonomy (Kavargyris et al., 2025b), Europe's structured classification of over 13,800 skills and competences, including approximately 570 green-tagged entries (European Commission, 2024). Each ESCO skill mention was mapped to its canonical label, yielding a binary document-skill matrix.

For analytical clarity, the extracted ESCO skills were grouped into six hierarchical layers:

1. Vision & Leadership: Strategic, abstract competencies (e.g., leadership, infrastructure design)
2. Enabling Digital & Social Skills: Digital coordination, stakeholder networking, resource planning
3. Technical Expertise: Hands-on domain skills (e.g., compliance with environmental laws, precision manufacturing)
4. Decision & Planning: Analytical and organizational tasks (e.g., thematic mapping, research design)
5. Monitoring & Evaluation: Evaluative functions (e.g., environmental auditing, performance reporting)
6. Support & Execution: Logistical and operational tasks (e.g., secure packaging, facility management)

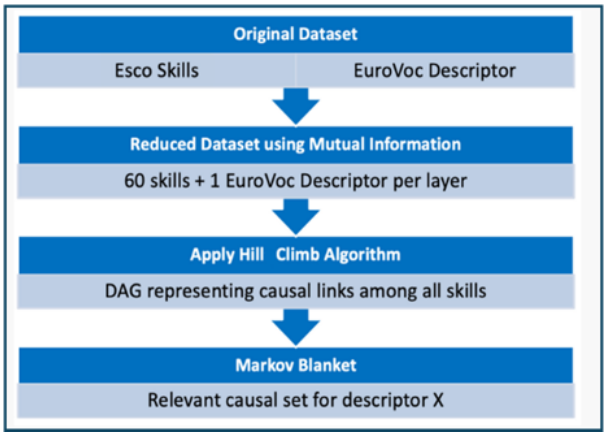
This six-layer schema informed our downstream feature selection and causal-discovery analyses.

An initial exploratory analysis was performed by computing the co-occurrence frequencies of EuroVoc descriptors and ESCO skills (Kavargyris et al., 2025c) across the corpus, constructing a binary co-occurrence network. This step identified overlaps and gaps between policy themes and green-related skills, and guided the selection of descriptors for causal modeling.

2.2 Methodology

The overall analytical pipeline is summarized in Figure 1. As shown, the methodological workflow comprises four key stages. First, we constructed the primary dataset by encoding each of 1,201 EU policy documents (2019–2024) from the EUR-Lex portal as a high-dimensional binary vector indicating the presence or absence of specific ESCO skills (ESCO, 2024) and expert-assigned EuroVoc descriptors. Second, to mitigate sparsity and focus on the most informative features, we applied Mutual Information (MI) selecting the top 60 ESCO skills per layer based on their MI with each descriptor (Kraskov et al., 2004). Third, we used a score-based Hill-Climb algorithm guided by the Bayesian Information Criterion to learn a DAG capturing the most plausible causal relations among the filtered skills and the target descriptor. Finally, for each descriptor node, we extracted its Markov Blanket, the union of its parents, children, and co-parents, as defined by Pearl (2009), thereby isolating the minimal skill subset directly responsible for or influenced by that policy concept. This pipeline delivers a transparent, data-driven framework for interpreting causal interdependencies between ESCO-defined competences and EU policy themes.

Figure 1. *Overview of the analytical pipeline*



2.3 Mutual Information and Feature Selection

Our initial dataset yields an extremely sparse binary matrix, as each policy document references only a small subset of the full ESCO skill taxonomy. To isolate the most informative skill–descriptor relationships, we compute the mutual information (MI) between each binary ESCO skill indicator and each binary EuroVoc descriptor indicator (Kraskov et al., 2004; Vergara & Estévez, 2014).

Mutual information was computed for every ESCO skill against each EuroVoc descriptor. Skills ranking below the 60th position exhibited near-zero MI with the selected descriptors, indicating negligible statistical dependence. Retaining these low-MI features would therefore expand the search space without improving explanatory power, while markedly increasing computational cost. Consequently, the 60 highest-

MI skills per layer were selected. This threshold reduces dimensionality from several thousand variables to 60 per layer, preserves predictive fidelity by focusing on the most informative features, and mitigates overfitting through the exclusion of low-signal noise. The resulting skill subsets serve as the input for the causal-discovery pipeline (Section 2.4).

2.4 Causal Discovery using Hill-Climb Algorithm

Hill-Climb is a greedy, score-based search algorithm that builds a causal graph by iteratively proposing local modifications such as adding, deleting, or reversing a single edge and accepting the change that most improves a chosen score. Commonly used scoring functions include the Akaike Information Criterion (AIC), the Bayesian Dirichlet equivalent uniform (BDeu) score, and the BIC.

In this study, BIC was selected for its strong, data-driven penalty on model complexity. BIC's penalty grows with the logarithm of the sample size, which more rigorously discourages the inclusion of weakly supported edges as the amount of data increases. This property is especially valuable in high-dimensional, sparsely populated settings, where the number of potential edges can explode, and overfitting is a serious risk. By maximizing the BIC score, the hill-climb search yields sparser, more interpretable graphs that retain only those causal links backed by sufficient empirical evidence (Schwarz, 1978; Koller & Friedman, 2009).

The search procedure is the following:

1. Start from an empty DAG (no edges) over all MI-filtered skills and the descriptor.
2. At each iteration, consider three valid, acyclic modifications:
    - Add an edge $Xj \rightarrow Xk$ if Xk has fewer than K parents and no cycle is introduced.
    - Delete an existing edge $Xj \rightarrow Xk$.
    - Reverse an existing edge $Xj \rightarrow Xk$ to $Xk \rightarrow Xj$, provided the reversal remains acyclic.
3. Leveraging BIC's decomposability, update only the affected nodes' contributions for each potential move.
4. Apply the move that yields the largest increase in BIC. Repeat until no further improvement is possible, indicating convergence to a local optimum (Scutari et al., 2018).

By iteratively applying these locally optimal edge modifications under the BIC the hill-climb procedure yields a causal graph that balances model fit and complexity, setting the stage for focused Markov-blanket extraction in Section 2.5.

2.5 Markov Blanket Application

After generating a candidate DAG via the Hill-Climb algorithm, we extract the Markov Blanket (MB) for each target EuroVoc descriptor to isolate the minimal set of skills most directly relevant for causal interpretation. The Markov Blanket of a node Y in a DAG G over variables V is defined as the union of:

- Parents $Pa(Y)$: skills with direct edges into Y, representing causal inputs.
- Children $Ch(Y)$: skills with direct edges from Y, representing likely effects.
- Co-parents: for each $X \in Ch(Y)$, the other parents $\mathrm{Pa(X)} \setminus \mathrm{Y}$, representing potential confounders.

Formally,

$$MB(Y) = \underbrace{Pa(Y)}_{\text{Parents}} \cup \underbrace{Ch(Y)}_{\text{Children}} \cup \underbrace{\bigcup_{X \in Ch(Y)} (Pa(X) \setminus \{Y\})}_{\text{Co-parents of Y's children}} \qquad (1)$$

By construction, conditioning on $MB(Y)$ renders Y independent of all other variables (Pearl, 2009).

Because our variables are binary ESCO skill indicators and binary descriptor indicators, each conditional probability table entry corresponds to a Bernoulli parameter, and the MB represents the smallest local subset of skills that captures all direct influences and effects. Focusing subsequent analysis on MB skills both reduces dimensionality further and highlights the competencies that are most critical for policy formation (as causal inputs) and implementation (as resulting effects) (Tsamardinos et al., 2009).

2.6 Technical implementation in Python

To operationalize the causal-discovery pipeline, a Python-based workflow was implemented using established libraries for data manipulation, causal inference, and network visualization. All analyses were conducted in Python 3.10. The pandas library handled data loading and preprocessing (McKinney, 2010), pgmpy facilitated Bayesian-network structure learning (Python Library Contributors, 2024), networkx supported graph construction and layout (Hagberg, Swart, & S Chult, 2008), and matplotlib generated plots (Hunter, 2007). Causal-discovery routines were executed via pgmpy's HillClimbSearch with BIC scoring.

First, the pre-filtered dataset was loaded from a CSV file into a pandas DataFrame, and the relevant ESCO skill columns for the selected layer were extracted. The binary EuroVoc descriptor column ("energy policy") was appended as the target variable, and all categorical fields were converted to integer type, yielding a binary feature matrix appropriate for DAG learning.

Next, structure learning was performed with pgmpy.models.HillClimbSearch, which greedily adds, deletes, or reverses edges to maximize the BIC score under a specified in-degree constraint. A pgmpy.models.BayesianNetwork was then instantiated from the learned edge list, and the descriptor's MB, its parents, children, and co-parents, was

extracted to isolate the core causal neighborhood (Pearl, 2009; Tsamardinos et al., 2006).

Finally, the full DAG was rendered using networkx and matplotlib, applying a color scheme that denotes the target descriptor (red), its Markov blanket (green), and other skills (blue). This visualization facilitates intuitive exploration of the resulting causal structure and highlights the policy-relevant skill interdependencies.

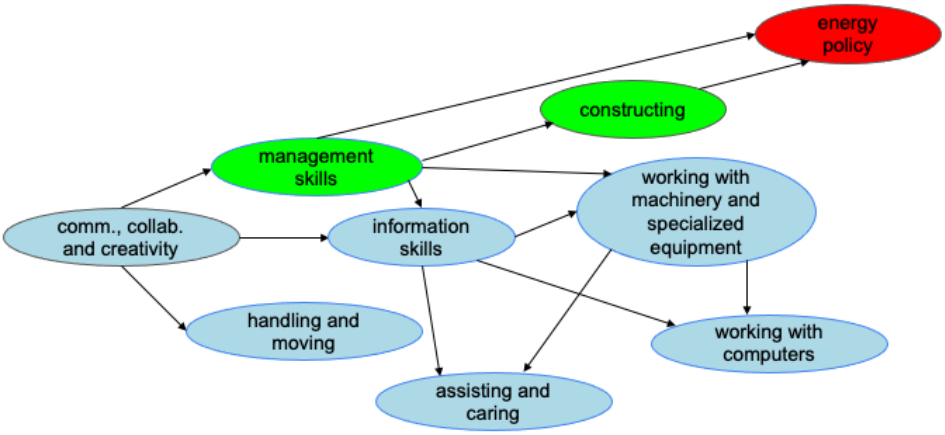
3. RESULTS

The causal-discovery pipeline was applied to five EuroVoc descriptors, Green Economy, Energy Policy, Circular Economy, Self-Sufficiency in Energy, and Environmental Degradation, to generate layer-specific DAGs from MI-filtered ESCO skills and extract each descriptor's Markov blanket. Although DAGs and Markov blankets were obtained for all descriptors, this section examines the "Energy Policy" descriptor in detail. Due to space constraints, we have presented only the detailed layer-by-layer results for the "Energy Policy" descriptor. For completeness, Table 2 (at the end of this section) provides a consolidated overview of the Markov blankets for all five EuroVoc descriptors, listing each descriptor's parent and child skills across Layers 1–6. This summary table enables cross-descriptor comparison and highlights the full spectrum of causal skill–policy relationships uncovered in our analysis.

3.1 DAGs extracted from Hill-Climb Algorithm (RQ_1)

In this section, and to address RQ_1, the results of the DAGs inferred via the Hill-Climb algorithm are examined. Figure 2 displays the DAG generated by algorithm in Layer 1, with the "Energy Policy" descriptor highlighted in red and its Markov-blanket skills in green. Even at this first, relatively simple level, multiple interconnections among the few nodes already complicate direct reading of causal paths. As layer depth increases, especially in Layer 4's 60-node DAG, dense interconnections render visual inspection infeasible. Isolating each descriptor's Markov blanket overcomes this complexity by extracting only the essential parents, children, and co-parents enabling precise causal analysis of the EuroVoc descriptor.

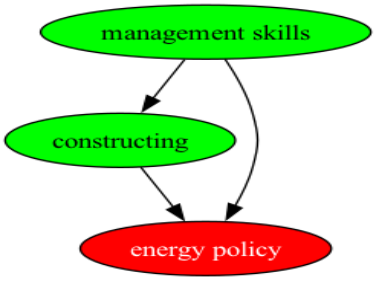
Figure 2. *DAG extracted from Hill-Climb Algorithm. This is from ESCO Layer 1*



3.2 Markov blankets for ESCO Layers 1-6

For each EuroVoc descriptor, an induced subgraph is extracted containing only those MI-filtered ESCO skills from Layers 1–6 that share a direct causal edge (as parent or child) with the descriptor in the learned DAG. Isolating these nodes highlights the competencies that most immediately define or result from each policy concept. This targeted subgraph approach enhances interpretability by focusing on core causal relationships, avoiding the complexity of complete, high-dimensional DAGs.

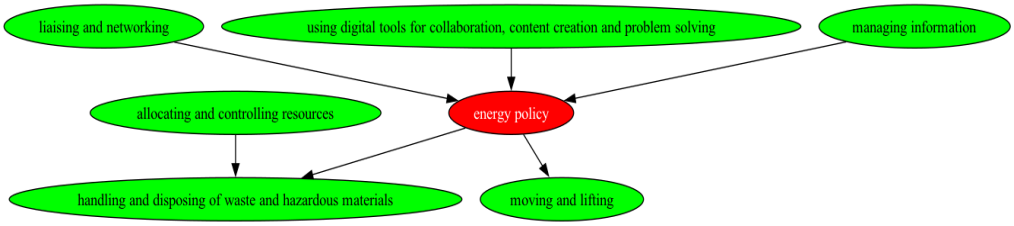
Figure 3. *MB for Energy Policy descriptor from ESCO Layer 1*



In Layer 1 (see Figure 3), the descriptor "Energy Policy" exhibits direct incoming edges from the strategic competencies "Leadership & Planning" and "Infrastructure Design & Execution". These two skills carry the strongest predictive information for the "Energy Policy" descriptor, establishing them as necessary prerequisites in the causal structure. This supports the interpretation that high-level decision-making and

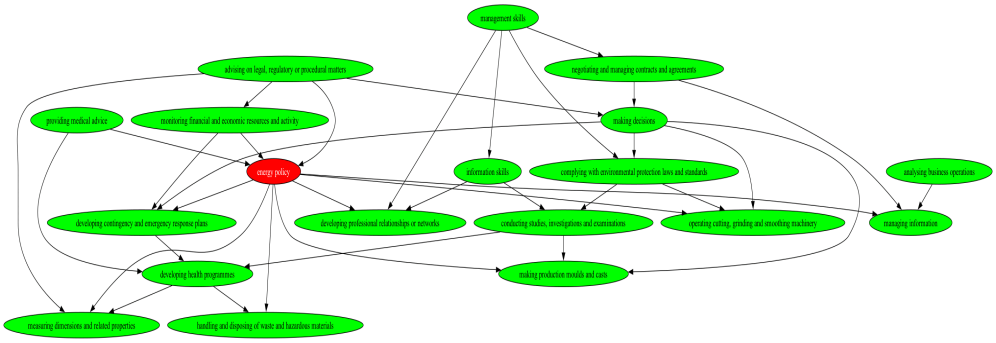
robust infrastructure planning are foundational inputs in the emergence of energy policy documents.

Figure 4. *MB for Energy Policy descriptor from ESCO Layer 2*



In Layer 2 (see Figure 4), Using digital tools for collaboration, content creation and problem solving, Managing Information, Liaising and networking, and Allocating and controlling resources emerge as direct parents of the "Energy Policy" descriptor. These skills exhibit the highest mutual information with the descriptor and appear as parents in the generated DAG, indicating their role as key prerequisites in the policy-formation process. In contrast, Handling and disposing of waste and hazardous materials and Moving and lifting emerge as children of "Energy Policy," appearing only after the descriptor is present. This pattern reflects that, once an energy policy framework is established, managing environmental by-products and coordinating large-scale transport become immediate operational imperatives. Such layer-specific parent-child relationships underscore the dual role of Layer 2 skills in both defining and implementing energy policy directives.

Figure 5. *MB for Energy Policy descriptor from ESCO Layer 3*

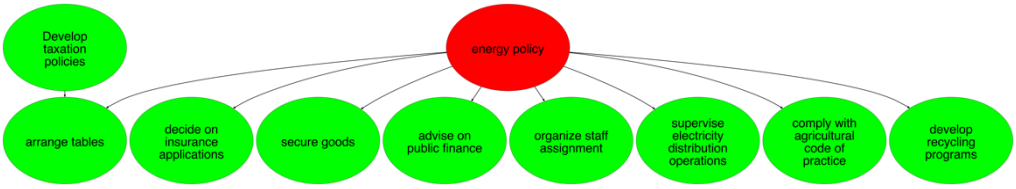


Mid-tier technical expertise serves both as a foundational input and an operational output within the "Energy Policy" causal network (see Figure 5). As formative drivers, competencies in Management and Contract Negotiation ensure effective project governance and legal robustness during policy drafting. Skills in Information Management, Regulatory Compliance, and Financial Monitoring further demonstrate

that rigorous data handling and economic alignment are indispensable to sound policy design. The abilities to Develop Contingency Plan and Advise on Legal or Procedural Matters highlight the necessity of crisis anticipation and navigation of complex regulatory environments, while Professional Networking emphasizes the role of interdisciplinary collaboration among engineers, health practitioners, and environmental specialists.

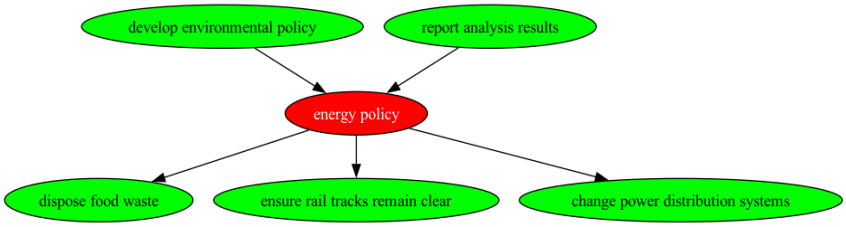
Conversely, once the "Energy Policy" framework is activated, a distinct suite of technical tasks emerges. Skills in Operational Decision Making, Machinery Operation, and Component Fabrication reflect the immediate on-site engineering activities. Subsequent competencies such as Hazardous-Waste Management and Business Process Analysis indicate that technical experts quickly address environmental safeguards and operational efficiency. The activation of Medical Advising and Precision Measurement underscores further downstream focus on health monitoring and quality control. Together, these dual roles illustrate how Layer 3 skills both shape policy formulation and execute its mandates in practice.

Figure 6. *MB for Energy Policy descriptor from ESCO Layer 4*



In Layer 4 (see Figure 6), several direct child nodes of "Energy Policy" were omitted to avoid excessive complexity in the graph. The excluded child nodes are "promote inclusion in organizations", "treat contaminated water", "manage medical supply chains", "organize storage facilities", "integrate measures in architectural designs", and "calculate debt costs". In Layer 4, the DAG identifies Research New Ideas and Create Thematic Maps as direct causal inputs to "Energy Policy," demonstrating that analytical research and spatial planning are critical in formulating energy directives. Downstream, the policy node feeds into a diverse set of operational competencies, including fiscal functions, logistical oversight, environmental services, and organizational activities. These directed edges quantitatively confirm that "Energy Policy" both depends on specialized analytical skills to set its targets and generates a multifaceted implementation footprint that spans budget design, infrastructure management, environmental remediation, and stakeholder coordination. Such complex, layer-specific causal pathways underscore the value of structure learning in revealing how high-level policy objectives translate into concrete operational tasks (Scutari et al., 2018).

Figure 7. *MB for Energy Policy descriptor from ESCO Layer 5*



In Layer 5 (see Figure 7), "Energy Policy" is directly preceded by Develop Environmental Policy and Report Analysis Results, highlighting the pivotal role of regulatory drafting and data-driven evaluation in shaping policy content. These parent skills underscore that rigorous environmental rule-making and interpretive reporting are foundational inputs at the monitoring and evaluation tier. When the policy comes into effect, directed edges lead to specialized implementation competencies such as, Dispose Food Waste, Ensure Rail Tracks Remain Clear, and Change Power Distribution Systems. This pattern quantitatively confirms that evaluative inputs translate into targeted monitoring and maintenance tasks, illustrating the causal transition from policy design to infrastructure upkeep and environmental management.

Figure 8. *MB for Energy Policy descriptor from ESCO Layer 6*



In Layer 6 (see Figure 8), the sole parent skill is Match Goods with Appropriate Packaging According to Security Procedures, reflecting the critical role of secure logistics in the formulation of energy policy, particularly for the safe transport of sensitive infrastructure components and hazardous materials. The absence of any child nodes indicates that this deep-logistics competency functions exclusively as a prerequisite, rather than an outcome of policy enactment. This suggests that, at the most specialized logistical tier, packaging and transport security considerations inform the design of "Energy Policy" but do not themselves emerge from its implementation.

3.3 Overview Markov blankets for ESCO Layers 1-6 (RQ_2)

Across all five EuroVoc descriptors, our Markov-blanket analysis reveals a systematic "Causal → Policy → Effect" pattern. The parent skills represent the competencies that must be present for a descriptor to emerge, while the child skills materialize immediately upon policy enactment. For example, "Energy Policy" is driven by strategic management, infrastructure planning, digital coordination, resource allocation, and technical foresight, which are competencies that collectively enable the

formulation of robust energy directives. Once established, the policy precipitates tasks such as hazardous-waste disposal, logistics coordination, and power-grid upgrades. Similarly, “Green Economy” depends on stakeholder networking and digital liaison to shape sustainable economic strategies, which in turn stimulate SME eco-loan facilitation, urban retrofit services, and green-business incubators. “Circular Economy” is contingent on precision manufacturing and reverse-logistics planning, yielding immediate recycling operations and product eco-redesign cycles. “Self-Sufficiency in Energy” arises from community consultation and local resource analytics, subsequently driving behavior-change hotlines, microgrid maintenance, and resilience monitoring. Finally, “Environmental Degradation” is conditioned by crisis-response planning and ecological risk assessment, producing species-care programs, cooling-infrastructure operations, and environmental system modelling. This consistent layered structure, quantified via causal discovery algorithm and Markov-blanket extraction, provides a data-driven framework for understanding how foundational skills define policy formation and how policies trigger specific implementation activities.

4. DISCUSSION AND FUTURE WORK

This study demonstrates how causal structure learning can reveal the hidden architecture of policy formulation and implementation across EU strategic domains. The multilayered design uncovers not only the critical skills feeding into policy descriptors but also the operational landscape each policy activates. Across all five EuroVoc descriptors, our analysis reveals a robust “Causal → Policy → Effect” cascade. For each descriptor, a specific constellation of upstream competencies, ranging from strategic leadership and digital coordination to domain-specific technical expertise, emerges as necessary precursors in the generated DAG. Upon activation of a descriptor node, a distinct set of downstream operational tasks is triggered (e.g., hazardous-waste protocols in Energy Policy, SME eco-loan facilitation in Green Economy, continuous recycling operations in Circular Economy).

To support interpretability and reproducibility, comprehensive table was included for cataloguing (see Table 2) all causal inputs and outputs for each EuroVoc descriptor. This tabular summary offers a practical reference for researchers and policymakers by clearly mapping the “Causal → Policy → Effect” structure. In causal-graph terms, the Markov blanket of each descriptor cleanly partitions into “incoming” skills (its parents and co-parents), which drive policy formulation, and “outgoing” skills (its children), which represent implementation. This layered pattern reinforces that EU policy development systematically leverages digital, strategic, and technical inputs to define policy aims, which in turn generate concrete, implementation-level responsibilities, evidenced by the directed edges in the DAG.

***Table 2**. Causal skills and effects for all the EuroVoc descriptors*

Causal Skills	EuroVoc Descriptor	Effect Skills
Strategic management, Infrastructure/ Constructing foresight, Digital Coordination, Resource Planning	Energy Policy	Hazardous-waste disposal, Heavy logistics & rail clearing, Power distribution/grid upgrades
Advisory & stakeholder networking, Digital Liaison, Sustainability communication	Green Economy	SME eco-loan facilitation, Urban retrofit services, Green-business incubation
Precision manufacturing & assembly, Resource-efficient product design, Reverse-logistics planning	Circular Economy	Continuous recycling operations, Product eco-redesign cycles, High-rate material recapture
Community consultation & audits, Local resource allocation analytics, Budget mapping	Self-Sufficiency in Energy	Behaviour-change hotlines, Micro-grid maintenance, Resilience monitoring services
Crisis-response planning, Search-and-rescue logistics, Ecological risk assessment	Environmental Degradation	Species-care programmes, Cooling-infrastructure operations, Enviromental system modeling

4.1 Generalizability and Limitations

A potential limitation of this study concerns biases related to data selection and classification. The dataset, derived exclusively from the EUR-Lex repository, may overrepresent formal legislative and strategic documents while underrepresenting local or operational policy initiatives. Furthermore, the keyword-based retrieval strategy emphasizes topics such as digital and green transitions, potentially overlooking other domains like social or cultural policies. The reliance on expert-curated taxonomies-namely EuroVoc and ESCO-introduces an additional layer of bias, as these frameworks may not fully capture emerging or cross-disciplinary competencies. Despite these limitations, their use ensures methodological consistency, comparability, and alignment with official European policy frameworks, thereby supporting the study's reproducibility and interpretability.

4.2 Future Research Directions

Empirical validation against labor-market data will determine whether the Markov-blanket competencies, such as Digital Collaboration and Information Management, are mirrored in actual demand. Mining EU job postings and training curricula for the presence of these skills in energy-sector recruitment will establish the extent to which the inferred "Causal → Policy → Effect" pathways correspond to industry requirements (Cedefop, 2022).

Incorporating domain-expert constraints into the causal-discovery process can further enhance model plausibility. By enforcing that regulatory-compliance skills precede the Energy Policy descriptor and forbidding reverse edges from implementation tasks to strategic inputs, structure learning is guided toward more credible graphs. Prior studies have shown that combining expert priors with score-based search reduces spurious connections and strengthens interpretability (Chickering, 1996; Scutari et al.; Pearl, 2009).

Finally, enriching feature selection with semantic embeddings promises to capture subtler skill-descriptor relationships. Leveraging pre-trained language models such as BERT to compute vector-space similarities alongside mutual-information scores will broaden the set of relevant competencies, particularly for descriptors with limited coverage, and improve the robustness of subsequent causal analyses (Devlin et al., 2019; Vergara & Estévez, 2014).

Collectively, these refinements are expected to deepen causal understanding and enhance the practical utility of the models for policy design and skills development. The findings also translate into actionable insights for stakeholders, offering a foundation for aligning educational programs, workforce training, and policy initiatives. By connecting analytical evidence with strategic decision making, this research supports a more integrated and evidence-based approach to developing the digital and green skills needed for Europe's twin transitions. Practical applications of this framework include integrating digital dashboards to monitor skill gaps, embedding Markov blanket insights into national reskilling programs, and aligning university curricula with foundational skills such as digital coordination, data literacy, and sustainability analysis. Policymakers can also apply this approach to regional development planning by linking renewable energy projects with local training initiatives or by prioritizing funding for projects that clearly demonstrate causal links between targeted skills and measurable outcomes.

Acknowledgments: The research in this paper is part of the PhD dissertation of the first author. The authors also acknowledge the funding received from the European Union's Horizon Europe Framework Programme SKILLAB under grant agreement No. 101132663.

ΠΕΡΙΛΗΨΗ

Οι πολιτικές βιωσιμότητας και η ψηφιακή μετάβαση αποτελούν ακρογωνιαίο λίθο της στρατηγικής ατζέντας της Ευρωπαϊκής Ένωσης. Ωστόσο, η μεταφορά των υψηλών αυτών στόχων σε συγκεκριμένα προγράμματα δεξιοτήτων εξακολουθεί να αποτελεί μείζον ζήτημα για τους φορείς χάραξης πολιτικής. Στο παρόν άρθρο εφαρμόζονται τεχνικές ανακάλυψης αιτιαδών σχέσεων (causal discovery) για να μελετηθεί ο τρόπος ενσωμάτωσης των ψηφιακών και «πράσινων» δεξιοτήτων στην Ευρώπη. Η μελέτη βασίζεται στη συστηματική ανάλυση 1.201 επίσημων εγγράφων πολιτικής από τη βάση EUR-Lex (2019–2024), από τα οποία εξήχθησαν δεξιότητες με βάση το πλαίσιο ESCO, οργανωμένες σε έξι ιεραρχικά επίπεδα. Με τη χρήση κατευθυνόμενων ακυκλικών γραφημάτων (DAGs) μέσω του αλγορίθμου Hill-Climb και τη βελτιστοποίησή τους με το Bayesian Information Criterion, εντοπίστηκαν οι κύριες αιτιώδεις σχέσεις μεταξύ δεξιοτήτων και δεικτών EuroVoc. Επιπλέον, μέσω της εξαγωγής των Markov blankets απομονώνεται το ελάχιστο υποσύνολο δεξιοτήτων που καθορίζει και επεξηγεί κάθε δείκτη EuroVoc. Τα ευρήματα παρέχουν πολύτιμες πληροφορίες για τον στοχευμένο σχεδιασμό εκπαιδευτικών προγραμμάτων και τη διαμόρφωση πολιτικών, ενισχύοντας την τεκμηρίωση της «πράσινης» και ψηφιακής μετάβασης στην Ευρώπη.

REFERENCES

Cedefop. (2022). *Skills OVATE – Real-Time Spotlight on Skill Demand*. https://www.cedefop.europa.eu/en/tools/skills-ovate

Chickering, D. M. (1996). Learning Bayesian networks is NP-complete. In *Learning from Data: Artificial Intelligence and Statistics V* (pp. 121–130).

Cordero, J. M., Cristóbal, V., & Santín, D. (2018). Causal inference on education policies: A survey of empirical studies using PISA, TIMSS and PIRLS. *Journal of Economic Surveys, 32*(3), 879–915. https://doi.org/10.1111/joes.12217

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019* (pp. 4171–4186). https://doi.org/10.18653/v1/N19-1423

European Commission. (2021a). *Digital Skills and Jobs Coalition*. https://digital-strategy.ec.europa.eu/en/policies/digital-skills-coalition

European Commission. (2021b). *Questions & Answers: European Green Deal – Boosting Green Jobs (Q&A_21_6823)*. https://ec.europa.eu/commission/presscorner/detail/en/qanda_21_6823

ESCO. (2024). *European Skills, Competences, Qualifications and Occupations (ESCO)*. https://esco.ec.europa.eu/en

EUR-Lex. (2024). *Access to European Union law*. https://eur-lex.europa.eu

Hagberg, A., Swart, P., & Chult, D. S. (2008). Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference (SciPy 2008)*.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering, 9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

Kavargyris, D. C., Georgiou, K., Maraki, I., Mittas, N., & Angelis, L. (2025d, September). ESCOPlus: A framework for enriching the ESCO taxonomy with fdigital skills from Stack Overflow. In *Euromicro Conference on Software Engineering and Advanced Applications (SEAA)* (pp. 32–48). Cham: Springer Nature Switzerland.

Kavargyris, D. C., Georgiou, K., Mittas, N., & Angelis, L. (2025a). Extracting knowledge and highly demanded skills from European Union educational policies: A network analysis approach. *Social Network Analysis and Mining, 15*, 1–41.

Kavargyris, D. C., Georgiou, K., Papaioannou, E., Moysiadis, T., Mittas, N., & Angelis, L. (2025b). Future skills in the GenAI era: A labor market classification system using Kolmogorov–Arnold networks and explainable AI. *Algorithms, 18*(9), 554.

Kavargyris, D. C., Georgiou, K., Papaioannou, E., Petrakis, K., Mittas, N., & Angelis, L. (2025c). ESCOX: A tool for skill and occupation extraction using LLMs from unstructured text. *Software Impacts*, 100772.

Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: Principles and techniques*. MIT Press.

Kraskov, A., Stögbauer, H., & Grassberger, P. (2004). Estimating mutual information. *Physical Review E, 69*(6), 066138. https://doi.org/10.1103/PhysRevE.69.066138

Marchant, S., Moreira, A., & Tennant, R. (2025). Bayesian causal discovery for policy decision-making. *Data & Policy, 7*, E6.

McKinney, W. (2010). Data structures for statistical computing in Python. In *Proceedings of the 9th Python in Science Conference* (pp. 51–56).

Nesta. (2023). *NLP methods for ESCO skill extraction* [Technical report]. Nesta.

Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge University Press.

Publications Office of the European Union. (2024). *EuroVoc thesaurus*. https://eur-lex.europa.eu/browse/eurovoc.html

Python Library Contributors. (2024). *pgmpy documentation*. https://pgmpy.org

Scutari, M., Brown, L. E., & Clark, T. (2018). Learning Bayesian networks from big data with greedy search. *Journal of Machine Learning Research, 18*, 1–36.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Tsamardinos, I., Brown, L. E., & Aliferis, C. F. (2006). The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning, 65*(1), 31–78. https://doi.org/10.1007/s10994-006-6889-5

Vergara, J. R., & Estévez, P. A. (2014). A review of feature selection methods based on mutual information. *Neural Computing and Applications, 24*(1), 175–186. https://doi.org/10.1007/s00521-013-1368-0

Greek Statistical Institute
Proceedings of the 37th Panhellenic & 3rd International Stat Conf (2025), pp.401-422

COMPARISON OF REGRESSION AND SPARSE REGRESSION MODELS ON MULTIVARIATE TIME SERIES OF STOCHASTIC AND DETERMINISTIC SYSTEMS

***Petros Petridis*[1], *Dimitris Kugiumtzis*[1]**
[1] School of Electrical and Computer Engineering, Aristotle University of Thessaloniki, University Campus, 54124, Thessaloniki
ppetrida@ece.auth.gr, dkugiu@auth.gr

ABSTRACT
During the process of modeling and forecasting multivariate time series, the primary focus is often on optimizing the performance of forecasting models. However, many widely utilized machine learning algorithms do not offer insights into the underlying structure of the equations governing the components of a multivariate time series. This paper explores methods that elucidate the structural form of the equations representing the components of the time series and conducts a comparative analysis of their forecasting errors. Specifically, the study evaluates the performance of regression models bearing analytic form on multivariate time series of stochastic and deterministic systems. These models are simple polynomial regression models, the Sparse Identification of Nonlinear Dynamics (SINDy) algorithm and symbolic regression models. Additionally, the Partial Mutual Information from Mixed Embedding (PMIME) algorithm is employed for lag variable selection and its forecasting capability is assessed when integrated with a polynomial regression model.

Keywords: Multivariate Time Series, Forecasting, Polynomial Regression, Symbolic Regression, Variable Selection.

1. INTRODUCTION

A primary objective in time series analysis is the forecasting of future values based on historical observations. Numerous forecasting algorithms exist, each employing different methodologies and relying on distinct assumptions, with their forecasting models presenting various strengths and limitations. Beyond the choice of the forecasting model, the forecasting of future values of a time series depends on its underlying structure and the presence of noise in its observations.

In particular, in multivariate time series, where the observed variables (representing components of the underlying system) interact with each other, classical linear models for univariate time series forecasting -known as Autoregressive (AR) models- are generalized to Vector Autoregressive (VAR) models. These models are actually regression models on the observed variables (one time series per variable) but at different lags, referred to as lag variables. While VAR models cannot capture nonlinear relationships, they provide insights into the structural lag dependencies in the underlying system [Lütkepohl (2006)]. Additionally, Structural Equation Models (SEM) are frequently used to estimate causal relationships among both observed and latent variables, also offering structural information about the underlying system. Also, various SEM extensions are capable of detecting nonlinear relationships among system components [Kline (2023)]. Nonlinear interactions among the components observed by a multivariate time series can be more effectively captured using complex deep learning models such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks. However, these models do not provide information about the system's structure, effectively functioning as black boxes and exhibiting a tendency toward overfitting [Gharehbaghi (2023)].

One approach to estimating nonlinear relationships is through polynomial regression, wherein higher-order lag variables and interaction terms among the lag variables are employed as independent variables. Nevertheless, in high-dimensional time series or in high-order polynomial regressions under noisy conditions, models tend to overfit, resulting in poor forecasting performance. Consequently, dimensionality reduction techniques are essential. Such techniques include variable extraction methods, such as Principal Component Regression (PCR) or Partial Least Squares (PLS), which produce a new set of fewer variables derived as linear functions of the lag variables. Alternatively, variable selection techniques are used to identify a subset of lag variables (up to a certain maximum lag) that serve as independent predictors for a specific component of the multivariate time series [Wainwright (2019)].

The present paper evaluates the forecasting accuracy of four different regression models on multivariate time series from various stochastic and deterministic systems. These models not only aim to forecast future values but also extract structural information from the time series and estimate nonlinear relationships using different sparsity and feature selection strategies. In each model, each observed variable at one time step ahead is treated as the dependent variable (response variable). The first model is a basic second-order polynomial regression, using as independent variables all lag variables up to a maximum lag. The second model is also a second-order polynomial regression, but formed on a subset of the lag variables selected using the Partial Mutual Information from Mixed Embedding (PMIME) algorithm [Kugiumtzis (2013), Fotiadis, Vlachos & Kugiumtzis (2024)]. This approach results in sparse regression models via variable selection from the pool of lag variables of all components up to a maximum lag. The third forecasting model is the Sparse Identification of Nonlinear Dynamics (SINDy), which generates sparse forecasting models via Sequential Threshold Least Squares method [Silva et al. (2020)]. Finally, the fourth regression

model is symbolic regression through Genetic Programming, implemented using GPTIPS2 (GPTIPS version 2.0), which theoretically evolves sparse models through stochastic processes aimed at discovering optimal regression representations [Searson (2015)].

The structure of the paper is as follows: Section 2 presents the theoretical background and methodology, Section 3 discusses the simulation study and Section 4 outlines the main conclusions derived from the simulation study.

2. METHODOLOGY

The paper begins by presenting the issue of high dimensionality in the forecasting of time series along with the necessity for dimensionality reduction. Subsequently, it briefly introduces the information-theoretic measures employed by the PMIME algorithm for the selection of lag variables (independent variables) that drive or explain a target component (response). The PMIME algorithm is then presented in detail. Finally, the four forecasting models are briefly presented.

2.1 Problem Statement

Consider a multivariate time series comprising K continuous-valued variables $X_1, X_2, \ldots, X_K$, each of length n, $\{x_{1,t}, x_{2,t}, \ldots, x_{K,t}\}_{t=1}^{n}$. The primary objective is to estimate the value of a particular variable, say X_1, at a future time $t+1$, $x_{1,t+1}$, based on the past observations of all variables $x_{i,t}$, $\forall i \in \{1, \ldots, K\}$. By defining a maximum lag L, one can construct a set of lag variables

$$W = \{x_{1,t}, x_{1,t-1}, \ldots, x_{1,t-L+1}, x_{2,t}, \ldots, x_{2,t-L+1}, \ldots, x_{K,t}, \ldots, x_{K,t-L+1}\},$$

which serves as the set of candidate predictors of $x_{1,t+1}$. Nevertheless, the existence of nonlinear dependencies between the response variable and the lag variables necessitates the development of models capable of capturing such nonlinearities.

Linear models, such as Autoregressive (AR) and Vector Autoregressive (VAR) models, are limited in their capacity to capture nonlinear relationships between predictors and responses. Although polynomial regression models can approximate certain nonlinear dependencies similarly to Taylor expansions, the incorporation of all lag variables in noisy, high-dimensional settings often leads to overfitting due to the proliferation of model parameters. For instance, in a multivariate time series with K variables and maximum lag L, the total number of lag variables equals $K \cdot L$. A second-order polynomial regression model in this context comprises $K \cdot L + \frac{K \cdot L(K \cdot L+1)}{2} + 1$ terms: the intercept term, $K \cdot L$ linear (first-order) terms and $\frac{K \cdot L(K \cdot L+1)}{2}$ nonlinear (second-order) terms, i.e., $K \cdot L$ squared terms and $\frac{K \cdot L(K \cdot L-1)}{2}$ interaction terms.

Hence, it is imperative to pursue forecasting modeling approaches that both accommodate potential nonlinear dependencies between the response and lag

predictors and maintain sparsity to prevent overfitting, particularly when noise contaminates the observed time series.

2.2 Information Measures

To quantify the degree of dependence, encompassing both linear and nonlinear relationships between two variables, Mutual Information (MI) and Conditional Mutual Information (CMI) are employed. The latter represents the MI between two variables conditioned on the presence (knowledge) of one or more additional variables [Cover & Thomas (2005), Mesner & Shalizi (2019)]. Both of these information measures are defined in terms of entropy. Let $f_X(x)$ be the probability density function of a continuous variable X, then the differential entropy is defined as:

$$H(X) = -\int_{-\infty}^{+\infty} f_X(x) \cdot \log\big(f_X(x)\big)\, dx = E\big[-\log\big(f_X(x)\big)\big].$$

The Mutual Information between two continuous variables X and Y is then defined as:

$$I(X;Y) = \int_{-\infty}^{+\infty} \int_{-\infty}^{\infty} f_{XY}(x,y) \cdot \log\left(\frac{f_{XY}(x,y)}{f_X(x) f_Y(y)}\right) dxdy = \tag{1}$$

$$H(X) + H(Y) - H(X,Y).$$

Similarly, the Conditional Mutual Information between two continuous variables X and Y, conditioned on a third variable Z, is defined as:

$$I(X;Y|Z) = H(X,Y) + H(Y,Z) - H(Z) - H(X,Y,Z). \tag{2}$$

It should be noted that in this study, the estimation of the entropies defining MI and CMI and consequently the estimation of MI and CMI themselves, is performed using the k-Nearest Neighbors (kNN) method [Kraskov, Stögbauer & Grassberger (2004)]. In the present study, the kNN estimator was implemented with $k = 5$, corresponding to the default value of the PMIME algorithm.

2.3 PMIME Algorithm

The PMIME algorithm is a variable selection method that uses information-theoretic measures to select lag variables that drive or explain a response variable. More specifically, PMIME employs an iterative procedure utilizing Mutual Information (MI) and Conditional Mutual Information (CMI), as defined in equations (1) and (2), to identify, for each response variable, a subset of lag variables $\boldsymbol{w}$, $\boldsymbol{w} \subseteq W$ that directly explain (are correlated with or interact with) the response variable. For example, considering the response variable $x_{1,t+1}$, PMIME initially defines the set $\boldsymbol{w}$ as empty. In the first step of the iterative process, the lag variable $w_1 \in W$ that maximizes the mutual information with the response $x_{1,t+1}$ is selected, i.e.,

$$w_1 = \arg\max_{w \in W} I(x_{1,t+1}; w).$$

Next, it is tested whether the MI between the response $x_{1,t+1}$ and the selected lag variable w_1 is statistically significant, under the null hypothesis $H_0: I(x_{1,t+1}; w_1)=0$. Due to the lack of a parametric asymptotic distribution for MI (and CMI) for continuous variables, the statistical test of H_0 is performed by resampling, following the random shuffling technique applied to one of the two time series, for example the one for the lag variable w_1 (for a more rigorous but time consuming test, see [Fotiadis et al, 2024]). One realization of this resampling generates a randomized time series

$$w_1^{*1} = \{w_{1,t_1}, w_{1,t_2}, \dots, w_{1,t_n},\},$$

where $t_i \in \{1, \dots, n\}$, $t_i \neq t_j$ for $i \neq j$, with $n' = n - L$. Thus, the new variable pair $(x_{1,t+1}; w_1^{*1})$, is not expected to have any linear or nonlinear relationship since the paired value correspond to different time points, thereby satisfying the null hypothesis and resulting in an estimated mutual information $I(x_{1,t+1}; w_1^{*1})$ close to zero. Repeating the above procedure M times, the respective M mutual information values $I(x_{1,t+1}; w_1^{*i})$, $i \in \{1, \dots, M\}$, form the empirical null distribution of the MI statistic for the significance test. Considering a significance level $\alpha = 0.05$, the selected lag variable of the first iteration of PMIME (w_1) is added to the set $\boldsymbol{w}$ ($\boldsymbol{w} = \{w_1\}$) if $I(x_{1,t+1}; w_1)$ exceeds the 95th percentile of the empirical MI distribution. Otherwise, $\boldsymbol{w}$ remains empty, the procedure terminates and it is concluded that the response variable $x_{1,t+1}$ is not explained by any lag variable in the set W. In the present study, the number of random shuffles used in the significance tests was set to $M = 100$, corresponding to the default value of the PMIME algorithm.

In subsequent iterations of the PMIME algorithm, a similar procedure is followed, but instead of MI, CMI conditioned on the already selected set $\boldsymbol{w}$ is used. More precisely, at iteration k, the lag variable $w_k \in W$ with the maximum CMI with the response $x_{1,t+1}$ conditioned on $\boldsymbol{w}$ is sought, i.e.,

$$w_k = \arg\max_{w \in W} I(x_{1,t+1}; w|\boldsymbol{w}).$$

Similarly, the statistical significance of $I(x_{1,t+1}; w_k|\boldsymbol{w})$, for the selected lag variable w_k is tested via resampling, analogously to the MI test in the first iteration, with random shuffling applied again to the time series of the lag variable w_k. If the null hypothesis $H_0: I(x_{1,t+1}; w_k|\boldsymbol{w}) = 0$, is not rejected, the procedure terminates, otherwise, w_k is added to $\boldsymbol{w}$ and the algorithm proceeds to the next iteration $k + 1$. The output of the PMIME algorithm is the selected subset of lag variables $\boldsymbol{w}$ that explain the response variable $x_{1,t+1}$. This is repeated for each response variable $x_{i,t+1}$, $i \in \{2, \dots, K\}$.

2.4 Forecasting Models

Model 1: Polynomial model

The first forecasting model considered in this study is a polynomial regression model of order M, denoted as pol(M), which will be used as a baseline model. For the same setting as above, pol(M), utilizes as predictors of $x_{1,t+1}$ all polynomial terms up to order M (including simple and interaction terms) derived from the set W.

Thus, the independent variables of the regression problem in vector form are

$$\boldsymbol{x_t} = \left[1, x_{1,t}, x_{1,t-1}, \dots, x_{1,t-L+1}, x_{1,t}^2, x_{1,t} \cdot x_{1,t-1}, \dots, x_{1,t} \cdot x_{K,t-L+1}, \dots, x_{K,t-L+1}^2, x_{1,t}^3, \dots, x_{K,t-L+1}^M\right]$$

where 1 stands for the intercept. The regression problem can be written as

$$x_{1,t+1} = \boldsymbol{\beta}^{\mathrm{T}} \cdot \boldsymbol{x_t} + \varepsilon_{1,t+1}, \tag{3}$$

where $\boldsymbol{\beta}$ is the vector of regression coefficients, T denotes transpose and $\varepsilon_{1,t}$ is white noise input. The forecasted value $x_{1,t+1}$ is given as

$$\hat{x}_{1,t+1} = \boldsymbol{b}^{\mathrm{T}} \cdot \boldsymbol{x_t}, \tag{4}$$

where $\boldsymbol{b}$ is the vector of estimated regression coefficients, obtained via the Ordinary Least Squares method.

It should be noted that the number of equations in (3), formed from n observations of the multivariate time series, is $n' = n - L$. Furthermore, according to this formulation, K independent models of the same type are assumed, one for each response $x_{i,t+1}$, $\forall i \in \{1, \dots, K\}$. Consequently, the noise terms of different models at the same time point are uncorrelated, i.e., $\varepsilon_{i,t}$ is uncorrelated with $\varepsilon_{j,t}$ for $i \neq j$. Nevertheless, although the polynomial forecasting model of order M can capture nonlinear relationships (Taylor series approximation), the high dimension K of the multivariate time series, combined with the high order of the polynomial model M and the presence of noise in (3), will lead to overfitting and biased estimation of the coefficients $\boldsymbol{b}$. Therefore, there is a need for dimensionality reduction of the independent variables $\boldsymbol{x_t}$.

Model 2: PMIME and Polynomial Model

The second model suggests, prior to constructing the polynomial model, to select the lag variables using the PMIME. Subsequently, using only the selected lag variables in $\boldsymbol{w}$ and not all the lag variables in W, a polynomial regression model of order M is fitted. This forecasting model is denoted as polPMIME(M). For example, considering the response variable $x_{1,t+1}$, the PMIME algorithm is first applied, which returns the lag variables explaining it, say $\boldsymbol{w} = \left[x_{1,t}, x_{2,t-1},\right]$. Then, the vector of independent variables of polPMIME(2) is

$$\boldsymbol{x_t} = \left[1, x_{1,t}, x_{2,t-1}, x_{1,t}^2, x_{1,t} \cdot x_{2,t-1}, x_{2,t-1}^2\right],$$

based on which the future values of X_1 are estimated via equation (4). Thus, the present forecasting model can capture nonlinear relationships between the response variable and the selected lag variables, while, at the same time, if the order M is not excessively high, it remains sparse, thereby avoiding the phenomenon of overfitting.

Model 3: Sparse Identification of Nonlinear Dynamics Algorithm

The third forecasting algorithm to be employed is the Sparse Identification of Nonlinear Dynamics (SINDy), which constructs sparse forecasting models through the Sequential Threshold Least Squares estimation method. Specifically, the algorithm utilizes a function library, denoted as $Lib = [mul_1, mul_2, mul_3, sqrt, cos, sin, exp, abs, \dots]$, where each function accepts a matrix of variables as input, applies the corresponding operation to its column vectors and returns the transformed matrix as output. The function mul_i returns all terms up to the i-th order, including simple and interaction terms derived from the input matrix. These functions are applied to the set of lag variables W, over all available observations of the time series, thereby constructing the initial matrix of independent variables:

$$\boldsymbol{\Theta_1} = \begin{bmatrix} mul_1(x_{1,1}), \dots mul_1(x_{K,1}) & \cdots & abs(x_{1,1}), \dots, abs(x_{K,1}) & \cdots \\ \vdots & \ddots & \vdots & \ddots \\ mul_1(x_{1,n-L}), \dots mul_1(x_{K,n-L}) & \cdots & abs(x_{1,n-L}), \dots, abs(x_{K,n-L}) \cdots \end{bmatrix}.$$

The goal is to estimate a specific response, say $x_{1,t+1}$, which in column vector form over all observations can be written as:

$$\boldsymbol{x}_{1,t+1} = \begin{bmatrix} x_{1,L+1} \\ \vdots \\ x_{1,n} \end{bmatrix}.$$

The Sequential Threshold Least Squares method is essentially an iterative Ridge Regression. In the first step ($i = 1$), the algorithm estimates the regression coefficients by minimizing the sum:

$$||\boldsymbol{x}_{1,t+1} - \boldsymbol{\Theta_i} \cdot \boldsymbol{b}||^2 + \alpha \cdot ||\boldsymbol{b}||^2, \tag{5}$$

where $\boldsymbol{b}$ is the vector of estimated regression coefficients for a given regularization parameter α (as in Ridge Regression). Next, the algorithm eliminates the columns of $\boldsymbol{\Theta_1}$, whose corresponding coefficients have magnitudes smaller than a predefined threshold ε. The remaining columns form a reduced matrix $\boldsymbol{\Theta_2}$ which is then used in the next iteration. At each iteration i, the regression coefficients are re-estimated via Equation (5), progressively removing independent variables that do not significantly contribute to the model (i.e., whose coefficient magnitudes fall below the threshold ε). This iterative process continues until some final iteration i^*, in which the dimension of Θ_{i^*+1} equals that of Θ_{i^*}, meaning no further variables are removed. Ordinary Least Squares is applied on the selected subset of independent variables Θ_{i^*} to obtain the final (unbiased) model coefficients $\boldsymbol{b}$. It should be noted that model estimation must be performed using Cross-Validation, across various combinations of α, ε and Lib.

Model 4: Genetic Programming Symbolic Regression

Finally, the last forecasting algorithm employed is Genetic Programming Symbolic Regression (GPSR), which constructs forecasting models through evolutionary stochastic processes. Initially, the algorithm generates N_p distinct forecasting models, each expressed in the form:

$$Model_i = b_{\mathrm{i},0} + b_{\mathrm{i},1} \cdot Tree_{\mathrm{i},1} + \cdots + b_{i,m} \cdot Tree_{i,m},$$

$i \in \{1, \dots, N_p\}$ and each $Tree_{i,j}$, $j \in \{1, \dots, \mathrm{m}\}$, where $m \in \{1, \dots, \mathrm{m_g}\}$ denoting the maximum number of genes (trees) is constructed by applying randomly selected functions from a predefined function library, e.g., $Lib = [mul, sqrt, cos, sin, exp, abs, add \dots]$, where the coefficients $b_0, b_1, \dots, b_m$ are computed via Ordinary Least Squares. For instance, a tree $Tree_j$ may take the form:

$$\mathrm{Tree_j} = add\left(mul\left(5, x_{1,t-1}\right), sqrt\left(add\left(x_{1,t-2}, x_{K,t-L}\right)\right)\right),$$

which can be written algebraically as:

$$Tree_j = 5 \cdot x_{1,t-1} + \sqrt{x_{1,t-2} + x_{K,t-L}}.$$

Subsequently, through an iterative evolutionary process over N_g generations, the algorithm refines these models using three core genetic operations: mutation, crossover and direct reproduction. Mutation introduces small, random modifications to the model trees by replacing certain subtrees with newly generated random structures, crossover combines two parent models by exchanging randomly selected subtrees between them, producing new offspring that inherit traits from both parents and direct reproduction copies certain high-performing models unchanged into the next generation, ensuring that the most effective solutions are retained. Parent selection is performed using a tournament selection mechanism: a random subset of models from the previous generation is chosen and the best-performing individual within this subset is selected for reproduction based on forecasting accuracy.

The final output of the algorithm is the best-performing model, selected based on its forecasting accuracy on out-of-sample data (i.e., data not used during training). In the current study

3. SIMULATION

To evaluate the performance of the forecasting models, each multivariate time series of length n, is divided into two sets: a test set comprising the last $n_2 = 200$ observations, on which the forecasting error is computed and a training set consisting of the remaining $n_1 = n - n_2$ observations. The training set is used both for the selection of lag variables, applicable for the PMIME method and for the estimation of the forecasting models, pol(M), polPMIME(M), SINDy and GPSR. Multi-step-ahead forecasts are performed iteratively: the one-step-ahead forecast is fed as input into the model to generate the two-step-ahead forecast and so on. In the multi-step forecasting

procedure, predictions from previous steps were used as inputs for subsequent steps for all variables in the multivariate system, without using observed intermediate values from the test set. The forecasting error at each forecast horizon is quantified using the Normalized Root Mean Square Error (NRMSE), defined as the square root of the mean squared error normalized by the standard deviation of the target variable within the test set.

It is noted that if, for the one-step-ahead forecast, the computed NRMSE exceeds 3 (three times worse than the forecast of the sample mean) in more than 20% of the realizations (i.e., in over 20 out of 100 trials), then the NRMSE values for that horizon and all subsequent horizons are not displayed in the following figures. These thresholds were chosen as pragmatic visualization criteria because an NRMSE greater than 3 indicates performance substantially worse than the sample-mean benchmark, while requiring this to occur in more than 20% of realizations ensures that curves are hidden only when poor performance is systematic, rather than driven by a few extreme cases.

For the first baseline model, a degree $M = 2$ is applied for all lag variables (pol(2)). The second model also employs $M = 2$ (polPMIME(2)), but only on the subset of lag variables selected via PMIME. In the SINDy model, a 5-fold cross-validation is performed on the training data of length n_1, across various function libraries Lib, threshold ε and regularization factor α. The tested α values are 0 (corresponding to Ordinary Least Squares) and 0.3 (corresponding to Ridge Regression), the ε values range from 0.0001 to 0.4 in steps of 0.04 and the function libraries tested are the following $i) mul_1$, ii) mul_2, iii) mul_3, iv) $[mul_1, cos, sin, exp]$, v) $[mul_2, cos, sin, exp]$ and vi) $[mul_3, cos, sin, exp]$. The optimal combination of function library Lib, α and ε is selected based on the Mean Squared Error (MSE). All other model parameters are kept at their default settings. In the Genetic Programming Symbolic Regression (GPSR) algorithm, a validation set comprising 20% of the training data (with length $0.2 \cdot n_1$), is used to tune the parameters N_p and N_g, by minimizing the Root Mean Squared Error (RMSE). The values tested are $N_p \in \{50,100,150,200,300,400\}$ and $N_g \in \{10,25,50,100\}$. Additional settings include a maximum of 5 trees per model ($m_g = 5$), maximum tree depth of 5 and a function library defined as $Lib = [times, minus, plus, rdivide, square, tanh, exp, log, mult3, add3, sqrt, cube, negexp, neg, abs]$, where it includes arithmetic operators ($plus$, $minus$, $times$, $rdivide$), common functions ($square$, $cube$, $sqrt$, exp, log, $tanh$, abs), special terms like $negexp$ ($exp(-x)$) and neg ($-x$) and multivariate functions such as $mult3(x, y, z)$ and $add3(x, y, z)$ allow three-input multiplication and addition. All other model parameters were kept at their default values. The maximum lag used for generating the lag variables across all forecasting models is set to $L = 5$.

For system examined, 100 realizations (i.e., 100 multivariate time series) are generated. These are used to produce statistical results concerning both the accuracy of

identifying the true lag variables for each response variable using PMIME and the forecasting accuracy of each model, pol(2), polPMIME(2), SINDy and GPSR.

3.1 System 1

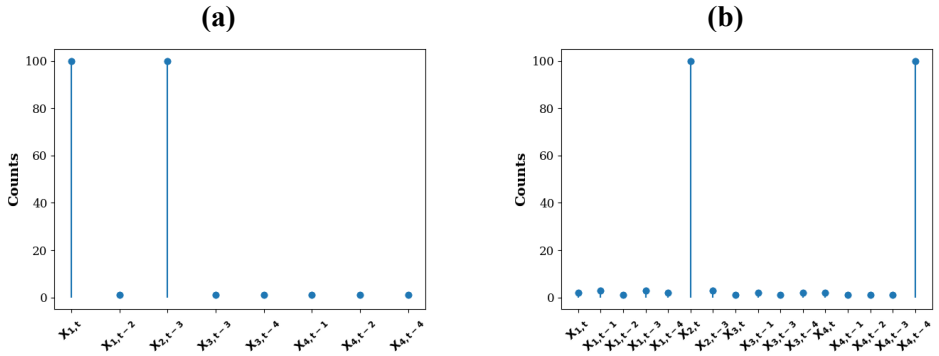
The first system under consideration is a linear stochastic system of order 5 with 4 variables, denoted as VAR(5) [Winterhalder et al. (2005)]:

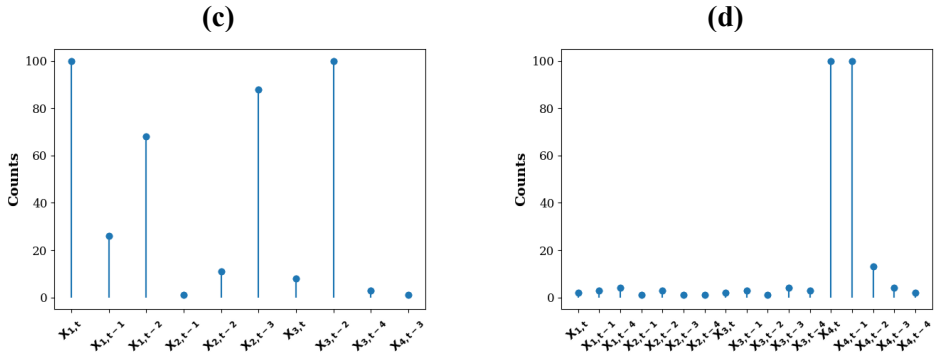
$$x_{1,t} = 0.8x_{1,t-1} + 0.65x_{2,t-4} + \varepsilon_{1,t}$$
$$x_{2,t} = 0.6x_{2,t-1} + 0.6x_{4,t-5} + \varepsilon_{2,t}$$
$$x_{3,t} = 0.5x_{3,t-3} - 0.6x_{1,t-1} + 0.4x_{2,t-4} + \varepsilon_{3,t}$$
$$x_{4,t} = 1.2x_{4,t-1} - 0.7x_{4,t-2} + \varepsilon_{4,t}$$

The noise components $\varepsilon_{i,t}, \forall i \in \{1, \dots, 4\}$, are mutually independent and follow a normal distribution with mean 0 and standard deviation 1. The total length of the original time series is $n = 712$ which is split into a training series of length $n_1 = 512$ and a test series of length $n_2 = 200$.

Figure 1 presents the lag variables selected by the PMIME algorithm in System 1 for each response variable. It is observed that the PMIME algorithm successfully identifies the correct lag variables for each response with approximately 100% accuracy. The frequencies of incorrectly selected lag variables are generally quite low, except for the response variable $x_{3,t+1}$, where the lags $x_{1,t-1}$ and $x_{1,t-2}$, are frequently selected despite not appearing in the corresponding equation. Similar occurrences of spurious lag variables are observed in other responses as well, though with much lower frequency.

Figure 1. *The selected lag variables by the PMIME algorithm on System 1 ($L = 5$, $n_1 = 512$) for response variables a) $x_{1,t+1}$, b) $x_{2,t+1}$, c) $x_{3,t+1}$ and d) $x_{4,t+1}$.*



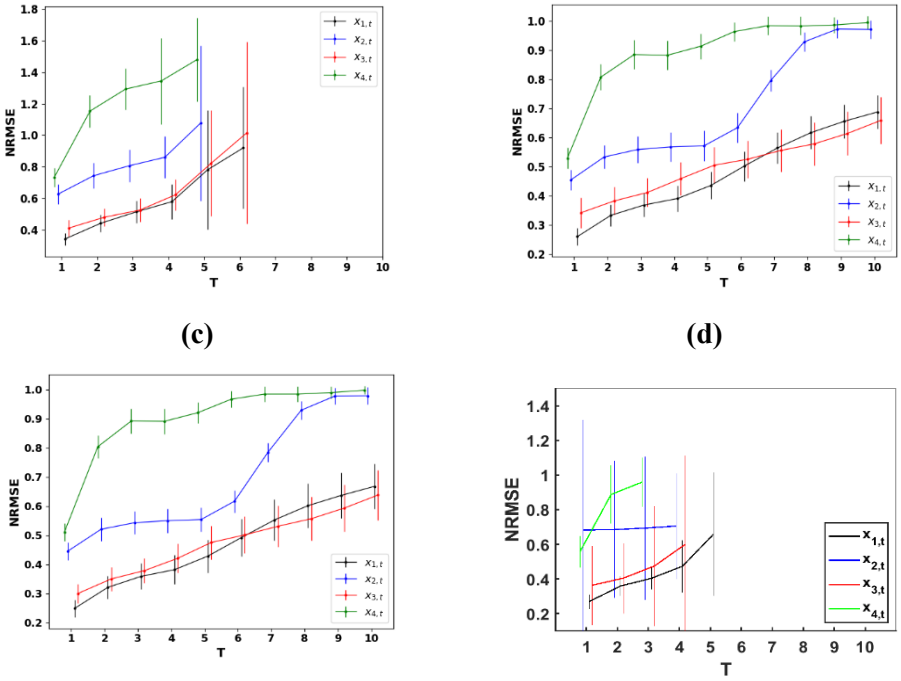


Subsequently, the performance of the forecasting models pol(2), polPMIME(2), SINDy and GPSR is evaluated across the 100 realizations. Figure 2 presents the NRMSE for each response variable and forecasting horizons from 1 to 10 steps ahead across the four models. The pol(2) model exhibits the highest NRMSE for the first forecast steps and after a certain horizon (around step 6 or 7), the NRMSE increases dramatically. Both the polPMIME(2) and SINDy models produce similar forecasts, which are near-optimal given the stochastic nature of the system and show low NRMSE variance. Finally, the GPSR model exhibits slightly worse performance in the early forecast steps, with higher variance in NRMSE and its forecasts explode for larger forecast steps as for pol(2).

The pol(2) model suffers from overfitting due to the inclusion of many irrelevant independent variables, which assign to noise and degrade forecasting performance. In contrast, polPMIME(2) only loses forecasting accuracy when a relevant explanatory variable is not selected by PMIME, but it remains sparse and thus more robust to noise. Similarly, the SINDy algorithm loses accuracy either when a non-informative term is not removed by the Sequential Threshold Least Squares method, due to biased regression coefficients in the presence of noise, or when a term from a relevant variable is erroneously excluded. However, such cases are rare and the sparse model remains accurate. Finally, although the GPSR algorithm achieves satisfactory accuracy in early forecast steps, it tends to overfit the training data, resulting in forecasts that diverge in longer horizons. These results highlight the critical importance of dimensionality reduction when constructing regression models for stochastic systems, even when the underlying dynamics are linear.

Figure 2. *NRMSE of the response variables in System 1 ($n_1 = 512$, $n_2 = 200$) for the forecasting models: a) pol(2), b) polPMIME(2), c) SINDy and d) GPSR. Note the difference in the scale of the vertical axis in the four panels.*

(a) **(b)**



3.2 System 2

The second system [Kugiumtzis (2013)], is a deterministic, nonlinear, chaotic system, consisting of a coupling of K Hénon map subsystems (each represented by one variable). The coupling strength is governed by the parameter C and the coupling structure follows an open ring topology, where each variable is influenced by its two neighboring variables (with a lag of 1), except for the two edge subsystems which influence only their single adjacent neighbor. System 2 is defined as:

$$x_{i,t} = 1.4 - x_{i,t-1}^2 + 0.3x_{i,t-2}, i \in \{1, K\}$$

$$x_{i,t} = 1.4 - \left(0.5C\left(x_{i-1,t-1} + x_{i+1,t-1}\right) + (1-C)x_{i,t-1}\right)^2 + 0.3x_{i,t-2}, \\ i \in \{2, \dots, K-1\}$$

Figure 3 illustrates the lag variables selected by the PMIME algorithm for System 2, for the response variables $x_{1,t+1}$ and $x_{2,t+1}$, using $L = 5$ and a training time series length of $n_1 = 512$, with number of components $K = 3$ and coupling strength $C = 0.2$, which corresponds to weak coupling-a regime where the subsystems interact but do not synchronize. The PMIME algorithm identifies with 100% accuracy the correct lag variables for $x_{1,t+1}$, which are $x_{1,t}$ and $x_{1,t-1}$. In the case of $x_{2,t+1}$, PMIME correctly selects $x_{2,t}$ and $x_{2,t-1}$, with 100% accuracy and also identifies the driving variable X_1, but at lag 2 instead of lag 1 ($x_{1,t-1}$ instead of the correct $x_{1,t}$). Moreover, the driver variable X_3 is also selected more frequently with lag 2 rather than lag 1. It should be noted that for all values $K \geq 3$, the selection frequency of lag variables for

$x_{K,t+1}$ is similar to that of $x_{1,t+1}$, while for response variables from $x_{3,t+1}$ to $x_{\mathrm{K}-1,t+1}$, the results resemble those of $x_{2,t+1}$. Furthermore, in the case where $C = 0$, i.e., no coupling, the results for all response variables match those of $x_{1,t+1}$, as PMIME achieves 100% accuracy under such conditions.

Figure 3. *The selected lag variables by the PMIME algorithm on System 2 ($C = 0.2$, $K = 3$, $L = 5$, $n_1 = 512$) for responses a) $x_{1,t+1}$ and b) $x_{2,t+1}$.*

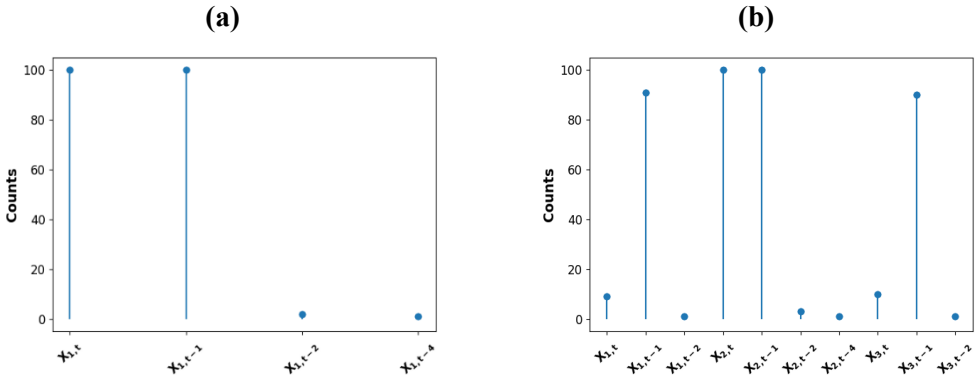
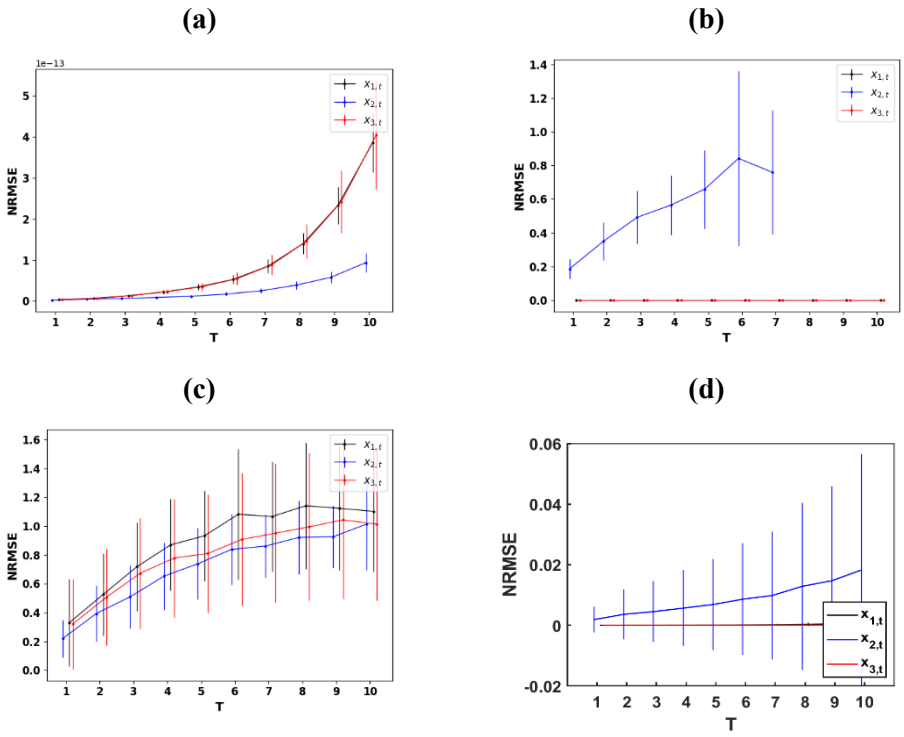


Figure 4 presents the NRMSE values of each response variable for 1 to 10 step-ahead forecasts produced by the four forecasting models for System 2 (with $K = 3$, $C = 0.2$). It is observed that the pol(2) model achieves the highest accuracy (NRMSE approaches zero) across all forecasting horizons, as the system contains no noise and the second-degree polynomial used is sufficient to capture the inherent second-order nonlinear relationships. In the case of polPMIME(2), the forecasting performance for $x_{1,t+1}$ and $x_{K,t+1}$ is identical to that of pol(2), since the PMIME algorithm selects the correct lag variables with 100% accuracy and the second-degree polynomial is adequate to represent the system's nonlinearity. For $x_{2,t+1}$, however, the forecasting accuracy is slightly reduced. This is due to PMIME selecting the correct driver variables (X_1 and X_3) but with incorrect lag (lag 2 instead of lag 1). The SINDy algorithm, although it achieves about the same accuracy as polPMIME(2) for $x_{2,t+1}$, this does not improve for $x_{1,t+1}$ and $x_{K,t+1}$ as it does for pol(2) and polPMIME(2). This is attributed to the small coupling constant $C = 0.2$, which when squared becomes $C^2 = 0.04$, resulting in the elimination of the corresponding terms from the $\boldsymbol{\Theta}_i$ matrices in some iterative step i. This reduction in model complexity leads to decreased forecasting accuracy, highlighting a limitation of SINDy in systems with weak coupling. Lastly, the second-best forecasts are obtained using the GPSR algorithm, which is capable of accurately estimating the nonlinear terms present in System 2.

Figure 4. *NRMSEs of the response variables in System 2 ($C = 0.2$, $K = 3$, $n_1 = 512$, $n_2 = 200$) for the forecasting models a) pol(2), b) polPMIME(2), c) SINDy and d) GPSR. Note the difference in the scale of the vertical axis in the four panels.*



3.3 System 3

System 3 is similar to System 2 but with a different coupling structure, where each subsystem influences the next in a unidirectional manner [Papana, Kugiumtzis & Larsson (2012)]. The system is defined as:

$$x_{i,t} = 1.4 - x_{i,t-1}^2 + 0.3x_{i,t-2},\ i = 1$$

$$x_{i,t} = 1.4 - C\left(x_{i-1,t-1}\, x_{i,t-1}\right) + (1 - C)x^2{}_{i,t-1} + 0.3x_{i,t-2},\ i \in \{2, \ldots, K\}$$

Figure 5 presents the lag variables selected by the PMIME algorithm for System 3, focusing on the response variables $x_{1,t+1}$ and $x_{2,t+1}$ ($L = 5, n_1 = 512, K = 3, C = 0.3$), where $C = 0.3$ corresponds to weak coupling. The PMIME selects the correct lag variables for $x_{1,t+1}$ with 100% accuracy, namely $x_{1,t}$ and $x_{1,t-1}$. In the case of $x_{2,t+1}$, the lags $x_{2,t}$ and $x_{2,t-1}$ are also correctly selected with 100% accuracy. Furthermore, the driver variable X_1 is identified, but with lag 2 instead of 1, i.e., $x_{1,t-1}$ instead of $x_{1,t}$. It is also noted that for any $K \geq 3$, the selection pattern for response variables from $x_{3,t+1}$ to $x_{\mathrm{K},t+1}$ resembles that of $x_{2,t+1}$.

Figure 5. *The selected lag variables by the PMIME algorithm on System 3 ($C = 0.3$, $K = 3$, $L = 5$, $n_1 = 512$) for responses a) $x_{1,t+1}$ and b) $x_{2,t+1}$.*

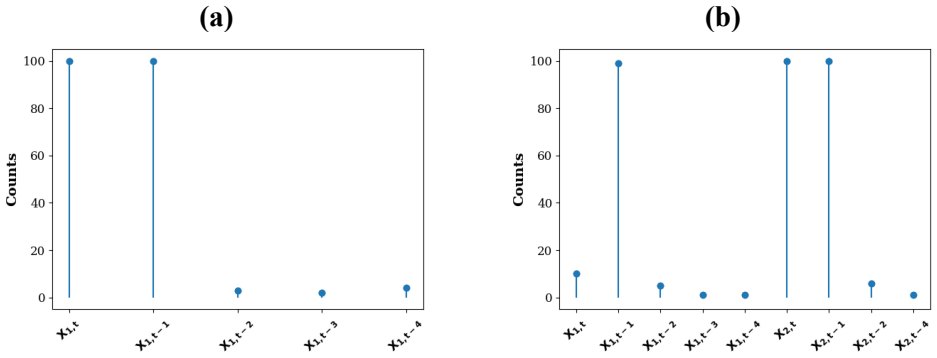
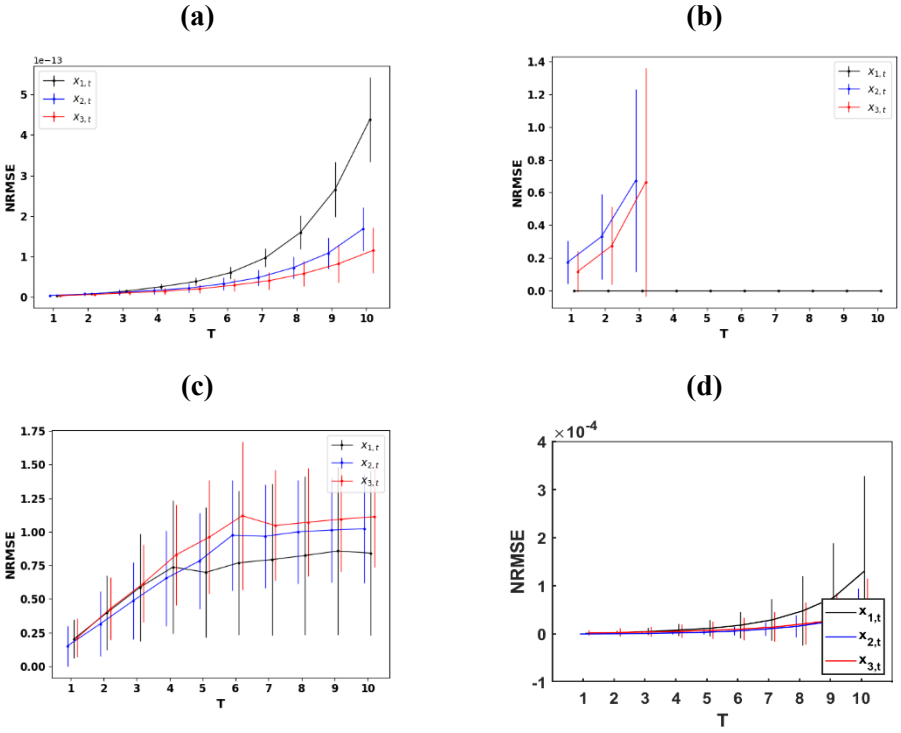


Figure 6 presents the NRMSE of each response variable for 1 to 10 step-ahead forecasts using the four forecasting models on System 3 ($K = 3$, $C = 0.3$). It is observed that the pol(2) model achieves the highest accuracy for all forecast horizons, similarly to what was observed in System 2. Likewise, the polPMIME(2) model achieves similar forecasting accuracy to pol(2) for the response variable $x_{1,t+1}$ but for $x_{2,t+1}$ and $x_{3,t+1}$ the accuracy is reduced. This is due to the PMIME algorithm selecting the correct driver variables, but with an incorrect lag (lag 2 instead of lag 1) in both cases. It is also noted that the NRMSE increases faster with the forecasting step compared to System 2. This is attributed to the higher coupling strength, which leads to a stronger influence of $x_{1,t}$ on $x_{2,t+1}$ that is not accurately captured by the selected lag variables. Similarly to System 2, the SINDy algorithm, while reaching satisfactory accuracy, does not achieve the optimal performance of pol(2). Finally, the second-best forecasts are produced by the GPSR algorithm, which is able to accurately estimate the nonlinear terms of System 3, thus achieving a high level of forecasting accuracy.

Figure 6. *NRMSEs of the response variables in System 3 ($C = 0.3$, $K = 3$, $n_1 = 512$, $n_2 = 200$) for the forecasting models a) pol(2), b) polPMIME(2), c) SINDy and d) GPSR. Note the difference in the scale of the vertical axis in the four panels.*



3.4 System 4

The fourth system [Gourévitch et al. (2006)] is a stochastic, nonlinear system composed of three variables:

$$x_{1,t} = 3.4x_{1,t-1}\left(1 - x_{1,t-1}^2\right)e^{-x_{1,t-1}^2} + \sigma \cdot \varepsilon_{1,t}$$

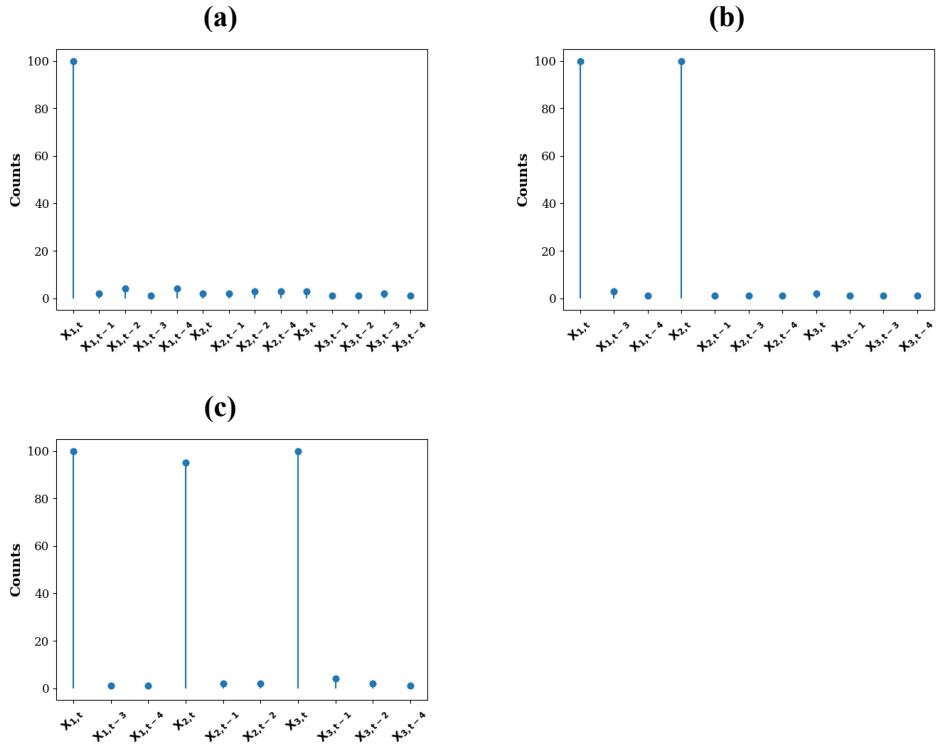
$$x_{2,t} = 3.4x_{2,t-1}\left(1 - x_{2,t-1}^2\right)e^{-x_{2,t-1}^2} + 0.5x_{1,t-1}x_{2,t-1} + \sigma \cdot \varepsilon_{2,t}$$

$$x_{3,t} = 3.4x_{3,t-1}\left(1 - x_{3,t-1}^2\right)e^{-x_{3,t-1}^2} + 0.3x_{2,t-1} + 0.5x_{1,t-1}^2 + \sigma \cdot \varepsilon_{3,t}$$

The noise terms $\varepsilon_{1,t}$, $\varepsilon_{2,t}$ and $\varepsilon_{3,t}$ are mutually independent and follow a normal distribution with mean zero and standard deviation equal to one.

As shown in Figure 7, the PMIME algorithm selects the correct lag variables for each response variable with nearly 100% accuracy, occasionally including a small number of additional lag variables with low frequency.

Figure 7. *The selected lag variables by the PMIME algorithm on System 4 ($\sigma = 0.4$, $L = 5$, $n_1 = 512$) for responses a) $x_{1,t+1}$, b) $x_{2,t+1}$ and c) $x_{3,t+1}$.*



Regarding forecasting (see Figure 8), the pol(2) model yields the lowest accuracy for one-step-ahead forecast, maintaining worse performance than the mean predictor ($NRMSE = 1$) across all horizons. This is due to the fact that a polynomial degree $M = 2$ cannot capture the complex nonlinear relationships in System 4 and also the inclusion of redundant terms leads to overfitting. The polPMIME(2) model performs slightly better in one-step-ahead forecasts, as the sparsity of the model mitigates overfitting but the insufficiency of polynomial degree of order 2 persists. Clearly better performance is achieved by the SINDy and GPSR algorithms for one-step-ahead forecasts, as they can better capture the system's nonlinearity. Nevertheless, for forecasts beyond one step ahead, none of the models perform satisfactorily. This is attributed to the stochastic nature of the system combined with its highly nonlinear structure, which renders it unpredictable even in the short term.

Figure 8. *NRMSEs of the response variables in System 4 ($\sigma = 0.4$, $n_1 = 512$, $n_2 = 200$) for the forecasting models: a) pol(2), b) polPMIME(2), c) SINDy and d) GPSR. Note the difference in the scale of the vertical axis in the four panels.*

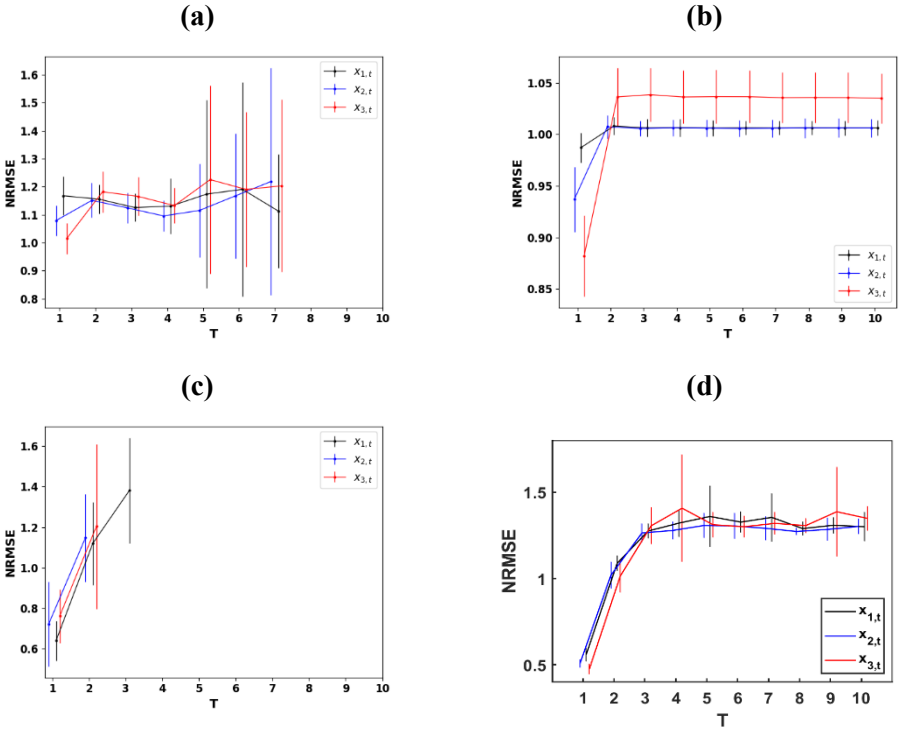


Table 1 reports the mean computational time over 10 iterations for each forecasting algorithm and system, measured on a machine equipped with a 12th Gen Intel Core i5-12450HX processor (2.40 GHz) and 16 GB of RAM, without any parallelization. The reported time includes both model fitting (i.e., selection of appropriate parameters) and multi-step forecasting (from 1 to 10 steps ahead) for every response variable. It is observed that the pol(2) model exhibits the lowest computational cost, as it only performs coefficient estimation via Ordinary Least Squares. Slightly higher execution time is observed for polPMIME(2), due to the additional computational cost of running the PMIME algorithm to select appropriate lag variables for each response variable and so, it depends on the dimensionality of the time series (K) and on the number of lag variables actually driving each response. Significantly higher execution times are recorded for the GPSR and SINDy algorithms. More specifically, the execution time of the SINDy algorithm depends on the dimensionality of the Θ_i matrices, as larger matrices (e.g., System 1 with higher dimensionality (K) results to higher number of initial independent variables/terms) likely require more iterations of the algorithm to eliminate non-predictive terms. Finally, GPSR exhibits the highest computational cost among all models, due to the fitting process.

Note that for both SINDy and GPSR algorithms, the reported computational times refer to cross-validation and validation set procedures and are also attributable to the extent of the hyperparameter search space. Therefore, a smaller hyperparameter grid (if the optimal values are somehow identified) results in reduced execution times.

Table 1. *Mean computational time (in seconds) over 10 iterations required for fitting and multi-step forecasting (1 to 10 steps ahead) of each forecasting algorithm on each system for each response variable.*

	pol(2)	polPMIME(2)	SINDy	GPSR
System 1	0.22	12.83	672.43	1613.80
System 2	0.11	8.21	670.67	930.82
System 3	0.12	9.65	668.24	929.77
System 4	0.15	6.56	666.59	1255.96

4. CONCLUSIONS

In this paper, the performance of the Partial Mutual Information from Mixed Embedding (PMIME) algorithm was presented and three forecasting models that perform feature selection in some sense, were compared against the standard second-order polynomial forecasting model, using both multivariate stochastic and deterministic systems.

The simulation study revealed that in stochastic systems, the second-order polynomial model tends to overfit the data due to the large number of terms (lag variables) included in the regression. Furthermore, in the case of nonlinear systems, where capturing nonlinear dependencies between response variables and lag variables is crucial, increasing the polynomial degree further exacerbates the overfitting issue. Significantly better results were observed with the second-order polynomial model whose terms are selected using the PMIME algorithm. This model remains sparse, making it more robust to noise, while also allowing for an increased polynomial degree to better capture nonlinearity. However, its forecasting accuracy may be compromised either when PMIME fails to correctly identify the relevant lag predictors, or when the polynomial degree is insufficient to capture the underlying nonlinear relationships. The computational cost of this method is a little higher than the second-order polynomial model and it increases with the number of observed variables and the number of lag variables that drive each response (for fixed algorithm's other parameters i.e. M, L). The SINDy algorithm also demonstrated strong capability in accurately capturing nonlinear dynamics and generating sparse, noise-robust models. However, it occasionally failed to include predictors with small coefficient magnitudes, as these were mistakenly removed during the thresholding process, which negatively impacted forecasting accuracy. The computational cost of SINDy is higher than pol(2) and polPMIME(2) algorithms and increases with the number of observed variables (for fixed algorithm's other parameters). Finally, the GPSR algorithm showed potential in

capturing nonlinear relationships. Nonetheless, instances of overfitting were observed, despite setting constraints on the maximum number of genes and maximum depth of trees used in the forecasting models. This issue could be addressed by further reducing these parameters or by introducing a penalty function to discourage model complexity, albeit possibly at the expense of forecasting accuracy. The computational cost of GPSR is the highest among the four forecasting algorithms, primarily due to the fitting process and the parameter validation search.

Nevertheless, it should be noted that both pol(M) and SINDy exhibit high memory demands, as longer time series, more observed variables and higher maximum lags (L) lead to larger design matrices. Also increasing the polynomial degree M or expanding SINDy's function libraries further amplifies this. In contrast, polPMIME(M) (for moderate M) and GPSR avoid such limitations, as they do not construct full lag-variable matrices.

Future work should focus on improving variable selection methods through enhanced or hybrid techniques to increase forecasting accuracy and robustness to noise, especially in complex systems. Additionally, adopting adaptive polynomial regression degree selection, guided by criteria such as cross-validation or information criteria (AIC/BIC), could improve the estimation of nonlinearity while avoiding overfitting. Finally, exploring ensemble or combined models that integrate the strengths of individual approaches (e.g., PMIME, SINDy, GPSR) may lead to more stable and accurate forecasts across a wider range of systems.

ΠΕΡΙΛΗΨΗ

Κατά τη διάρκεια της διαδικασίας μοντελοποίησης και πρόβλεψης πολυμεταβλητών χρονοσειρών, η κύρια έμφαση δίνεται συχνά στη βελτιστοποίηση της απόδοσης των μοντέλων πρόβλεψης. Ωστόσο, πολλοί ευρέως χρησιμοποιούμενοι αλγόριθμοι μηχανικής μάθησης δεν παρέχουν πληροφορίες σχετικά με τη δομική μορφή των εξισώσεων που διέπουν τα επιμέρους στοιχεία του συστήματος που παράγει μια πολυμεταβλητή χρονοσειρά. Η παρούσα εργασία εξετάζει μεθόδους που διερευνούν τη δομική μορφή των εξισώσεων του συστήματος και πραγματοποιεί συγκριτική ανάλυση των σφαλμάτων πρόβλεψης. Συγκεκριμένα, αξιολογείται η απόδοση -σε στοχαστικά και αιτιοκρατικά πολυμεταβλητά συστήματα χρονοσειρών- της απλής πολυωνυμικής παλινδρόμησης, του αλγορίθμου Sparse Identification of Nonlinear Dynamics (SINDy) και της συμβολικής παλινδρόμησης. Επιπλέον, παρουσιάζεται ο αλγόριθμος Partial Mutual Information from Mixed Embedding (PMIME) για την επιλογή μεταβλητών χρονικής υστέρησης και αξιολογείται η ικανότητά του στην πρόβλεψη όταν συνδυάζεται με το μοντέλο πολυωνυμικής παλινδρόμησης.

Λέξεις-κλειδιά: Πολυμεταβλητές Χρονοσειρές, Πρόβλεψη, Πολυωνυμική Παλινδρόμηση, Συμβολική Παλινδρόμηση, Επιλογή Χαρακτηριστικών.

REFERENCES

Cover, T. M., & Thomas, J. A. (2005). *Elements of Information Theory*. Wiley. https://doi.org/10.1002/047174882X.

Fotiadis, A., Vlachos, I., & Kugiumtzis, D. (2024). The causality measure of partial mutual information from mixed embedding (PMIME) revisited. *Chaos, 34*, 033113. https://doi.org/10.1063/5.0189056.

Gharehbaghi, A. (2023). *Deep Learning in Time Series Analysis*. CRC Press. https://doi.org/10.1201/9780429321252.

Gourévitch, B., Bouquin-Jeannès, R. L., & Faucon, G. (2006). Linear and nonlinear causality between signals: Methods, examples and neurophysiological applications. *Biological Cybernetics, 95*(4), 349–369. https://doi.org/10.1007/s00422-006-0098-0.

Kline, R. B. (2023). *Principles and Practice of Structural Equation Modeling* (5th ed.). Guilford Press.

Kraskov, A., Stögbauer, H., & Grassberger, P. (2004). Estimating mutual information. *Physical Review E, 69*, 066138. https://doi.org/10.1103/PhysRevE.69.066138.

Kugiumtzis, D. (2013). Direct coupling information measure from non-uniform embedding. *Physical Review E, 87*, 062918. https://doi.org/10.1103/PhysRevE.87.062918.

Lütkepohl, H. (2006). *New Introduction to Multiple Time Series Analysis*. Springer. https://doi.org/10.1007/978-3-540-27752-1.

Mesner, O. C., & Shalizi, C. R. (2021). Conditional mutual information estimation for mixed discrete and continuous data. *IEEE Transactions on Information Theory, 67*(1), 464–484. https://doi.org/10.1109/TIT.2020.3024886.

Papana, A., Kugiumtzis, D., & Larsson, P. (2012). Detection of direct causal effects and application to epileptic electroencephalogram analysis. *International Journal of Bifurcation and Chaos, 22*, 1250222. https://doi.org/10.1142/S0218127412502227.

Searson, D.P. (2015). GPTIPS 2: An Open-Source Software Platform for Symbolic Data Mining. In: Gandomi, A., Alavi, A., Ryan, C. (eds) *Handbook of Genetic Programming Applications. Springer*, Cham. https://doi.org/10.1007/978-3-319-20883-1_22

Silva, B., Champion, K., Quade, M., Loiseau, J.-C., Kutz, J., & Brunton, S. (2020). PySINDy: A Python package for the Sparse Identification of Nonlinear Dynamics from Data. *arXiv*. https://doi.org/10.48550/arXiv.2004.08424

Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press. https://doi.org/10.1017/9781108627771.

Winterhalder, M., Schelter, B., Hesse, W., Schwab, K., Leistritz, L., Klan, D., Bauer,

R., Timmer, J., & Witte, H. (2005). Comparison of linear signal processing techniques to infer directed interactions in multivariate neural systems. *Signal Processing, 85*(11), 2137–2160. https://doi.org/10.1016/j.sigpro.2005.07.011


ΕΛΛΗΝΙΚΟ ΣΤΑΤΙΣΤΙΚΟ ΙΝΣΤΙΤΟΥΤΟ
ΗΡΑΚΛΕΙΤΟΣ
Ε. Σ. Ι.