ΕΛΛΗΝΙΚΟ ΣΤΑΤΙΣΤΙΚΟ ΙΝΣΤΙΤΟΥΤΟ (Ε.Σ.Ι)

GREEK STATISTICAL INSTITUTE (G.S.I)

31ο Πανελλήνιο Συνέδριο Στατιστικής

«Η Επιστήμη των Δεδομένων στην Πληροφορική και τη Βιοϊατρική»

Πρακτικά του 31ου Πανελληνίου Συνεδρίου Στατιστικής

Proceedings of the 31st Panhellenic Statistics Conference

Λαμία, 4-6 Μαΐου 2018 Lamia, May 4-6, 2018

ΕΛΛΗΝΙΚΟ ΣΤΑΤΙΣΤΙΚΟ ΙΝΣΤΙΤΟΥΤΟ
(Ε.Σ.Ι)
GREEK STATISTICAL INSTITUTE
(G.S.I)

Π Ρ Α Κ Τ Ι Κ Α

31ου Πανελληνίου
Συνεδρίου Στατιστικής

PROCEEDINGS

of the 31st Panhellenic
Statistics Conference

Η επιστήμη των δεδομένων στην Πληροφορική και τη Βιοϊατρική

Data Science in Computing and Biomedicine

Λαμία, 4-6 Μαΐου 2018
Lamia, 4-6 May 2018

ΕΛΛΗΝΙΚΟ ΣΤΑΤΙΣΤΙΚΟ ΙΝΣΤΙΤΟΥΤΟ
(Ε.Σ.Ι)
GREEK STATISTICAL INSTITUTE
(G.S.I.)

Π Ρ Α Κ Τ Ι Κ Α

31ου Πανελληνίου
Συνεδρίου Στατιστικής

Η επιστήμη των δεδομένων στην Πληροφορική και τη Βιοϊατρική

Οργάνωση

ΕΛΛΗΝΙΚΟ ΣΤΑΤΙΣΤΙΚΟ ΙΝΣΤΙΤΟΥΤΟ

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
Τμήμα Πληροφορικής με Εφαρμογές στη Βιοϊατρική

Λαμία, 4-6 Μαΐου 2018

ΕΛΛΗΝΙΚΟ ΣΤΑΤΙΣΤΙΚΟ ΙΝΣΤΙΤΟΥΤΟ
Σολωμού 5, Αθήνα 10683 (Πλατεία Εξαρχείων)
Τηλ. και Fax 210 33.03.909
Email: secretariat@esi-stat.gr
http://www.esi-stat.gr

ISSN: 1792-2461

ΠΕΡΙΕΧΟΜΕΝΑ

ΠΡΟΛΟΓΟΣ …….…….....9
ΧΟΡΗΓΟΙ…….....13
ΠΡΟΓΡΑΜΜΑ ΣΥΝΕΔΡΙΟΥ……..……15
ΕΠΙΤΡΟΠΕΣ ΣΥΝΕΔΡΙΟΥ 25

Εργασίες στα Ελληνικά

ΑΝΑΣΤΑΣΟΠΟΥΛΟΥ Μ., ΡΑΚΙΤΖΗΣ Α.: Αυτοπαλίνδρομα Υποδείγματα Διακριτών Χρονοσειρών με Πεπερασμένο Εύρος Τιμών: Μια Συγκριτική Εφαρμογή σε Δεδομένα Πληθωρισμού των Χωρών-Μελών της Ευρωζώνης.....27

ΓΕΩΡΓΑΚΗΣ Α., ΜΕΛΦΟΣ Ι., ΣΤΑΜΑΤΕΛΛΟΣ Γ: Πίνακες όγκου δασικών συστάδων με την BITTERLICH δειγματοληψία.....42

ΘΕΟΔΟΣΙΑΔΟΥ Ο.,ΤΣΑΚΛΙΔΗΣ Γ.: Επεκτεταμένο φίλτρο Kalman με μη αρνητικούς περιορισμούς και χρηση τετραγωνικής μορφής 53

ΚΑΚΟΥΛΛΟΣ Θ. : Έλεγχοι F Ομοσκεδαστικότητας.....63

ΚΟΥΝΙΑΣ ΣΤΡ. : Η έννοια του τυχαίου και της πιθανότητας 600 π.Χ.-1800 μ.Χ....68

ΚΟΥΤΡΑΣ Μ.Β., ΚΟΚΚΙΝΙΩΤΗ Θ.Κ.: Δημιουργία νέας διδιάστατης κατανομής με γάμμα και βήτα περιθώριες με τη μέθοδο της trivariate reduction.....78

ΚΟΥΤΡΑΣ Μ., ΜΑΥΡΟΥΤΣΟΣ Ν.: Μέθοδοι δημιουργίας νέων οικογενειών κατανομών και μελέτη των ιδιοτήτων τους.....90

ΚΥΡΑΝΑ Ζ., ΓΕΡΟΘΑΝΑΣΗ Κ., ΖΗΜΕΡΑΣ Σ., ΚΥΡΑΝΑΣ Ε.: Εφαρμογή της πολυωνυμικής λογιστικής παλινδρόμησης και τεχνικών ομαδοποίησης στις διατροφικές συνήθειες των Ελλήνων εφήβων.....105

ΜΕΣΣΑΡΙΤΚΗΣ ΣΠ., ΛΕΛΗΣ Α., ΓΕΩΡΓΑΚΗΣ Α.: Σύγκριση μοντέλων όγκου γραμμικής παλινδρόμησης με αντίστοιχα νευρωνικών δικτύων για δασοωιομετρικά δεδομένα του δημοτικού δάσους Νάουσας.....120

ΜΠΑΝΤΟΥΝΑΣ Ι., ΠΙΤΣΕΛΗΣ Γ.: Εκτίμηση αποθεμάτων με χρήση μεθόδων Kalman Filter 132

ΜΠΟΖΙΚΑΣ Α., ΠΙΤΣΕΛΗΣ Γ.: Πρόβλεψη θνησιμότητας με τη χρήση αναλογιστικών υποδειγμάτων αξιοπιστίας.....146

ΜΠΡΑΓΟΥΔΑΚΗΣ Ζ.: Εμπειρική διερεύνηση της ασυμμετρικής προσαρμογής των λιανικών τιμών στην ελληνική αγορά τροφίμων.....161

ΠΑΝΑΓΙΩΤΟΥ Ν., ΤΡΙΑΝΤΑΦΥΛΛΟΥ Ι.: Μελέτη μη παραμετρικών διαγραμμάτων ελέγχου με χρήση Μεθόδων Προσομοίωση.....179

ΠΑΠΑΓΕΩΡΓΙΟΥ Μ., ΦΑΡΜΑΚΗΣ Ν.: Γραφήματα και δειγματοληψία 191

ΣΕΡΑΦΕΤΙΝΙΔΟΥ Ε., ΒΕΡΟΠΟΥΛΟΥ Γ.: Παράγοντες κατά τη διάρκεια της ζωής που καθορίζουν την κατάθλιψη σε άτομα ηλικίας 50 και άνω: διαφοροποιήσεις μεταξύ ανδρών και γυναικών στην Ευρώπη.....204

ΣΤΑΜΑΤΗΣ Α., ΤΑΜΠΟΥΡΗΣ Ε.,ΤΑΡΑΜΠΑΝΗΣ Κ.: Κατηγοριοποίηση των τουριστικών προορισμών της Ελλάδας με βάση τους καλλικρατικούς δήμους: μια εμπειρικη αναλυση.....220

ΤΣΙΜΠΑΝΟΣ Α. , ΑΓΙΑΚΛΟΓΛΟΥ Χ. : Επιλογή υποδειγμάτων με πληροφοριακά κριτήρια σε χωρικές διαδικασίες ….....236

ΧΑΡΑΛΑΜΠΙΔΗΣ Χ.: Γένεση Θεωρίας Πιθανοτήτων CARDANO έναντι PASCAL και FERMAT.....251

ΧΑΣΙΩΤΗΣ Β., ΧΑΤΖΟΠΟΥΛΟΣ Σ,ΚΟΥΝΙΑΣ ΣΤΡ.: Βέλτιστοι 3^m παραγοντικοί σχεδιασμοί εκτιμητικής τάξης iv με n=0 mod 3 παρατηρήσεις.....259

ΧΑΤΖΗΠΑΝΤΕΛΗΣ Θ., ΜΑΡΚΑΚΗ Ε.,ΣΩΤΗΡΟΓΛΟΥ Μ.: Πείραμα συμμετοχικής διαβουλευτικής διαχείρισης θεμάτων: Δήμος Θέρμης….....265

Εργασίες στα Αγγλικά

BOUNTZIS P., PAPADIMITRIOU E., TSAKLIDIS G.: Markovian Arrival Process modeling for the detection of seismicity rate changes for the strong earthquakes in Greece.....278

KALLIGERIS E.-N., KARAGRIGORIOU A., LAMBROU A. , PARPOULA C.: A Comparative Study for the Use of Periodic Regression Models for Detection of Influenza Outbreaks 292

MILIONIS A.E., GALANOPOULOS N.G. Time series with interdependent level and second moment: testing; consequences for modelling applications (part ii).....302

MILIONIS A.E., KARAGIOTA E.V.: systematic risk and market microstructure in a market under conditions of economic crisis. The case of the Athens stock exchange……314

PAPAIOANNOU T.: reproducibility crisis in science: reproducibility and replicability.....329

SIOURIS G.-J., SKILOGIANNI D., KARAGRIGORIOU A.: Low price effect and post model correction of value at risk….....337

SPITIERIS M.: On a new approach for improved multivariate imputations……353

TSAMTSAKIRI P., KARLIS D.: Bayesian estimation for inarch models.....366

VIANA M.: Symmetrically adapted data summaries.....379

ΠΡΟΛΟΓΟΣ

Το 31ο Πανελλήνιο Συνέδριο Στατιστικής, με τίτλο *Η Επιστήμη των Δεδομένων στην Πληροφορική και τη Βιοϊατρική*, διοργανώθηκε από το Ελληνικό Στατιστικό Ινστιτούτο (ΕΣΙ) σε συνεργασία με το Τμήμα Πληροφορικής με Εφαρμογές στη Βιοϊατρική του Πανεπιστημίου Θεσσαλίας.

Διεξήχθη 4-6 Μαΐου 2018 στις εγκαταστάσεις της Σχολής Θετικών Επιστημών του Πανεπιστημίου Θεσσαλίας στη Λαμία και έλαβαν μέρος 180 σύνεδροι, εκ των οποίων μέλη ΔΕΠ Ανωτάτων Ιδρυμάτων της χώρας και του εξωτερικού, αλλά και πολλοί φοιτητές προπτυχιακού και μεταπτυχιακού επιπέδου. Το επιστημονικό πρόγραμμα ήταν ιδιαίτερα πλούσιο περιλαμβάνοντας 23 παράλληλες συνεδρίες με 95 παρουσιάσεις εργασιών και αναρτημένων ανακοινώσεων (posters), καθώς και σεμινάρια για την (α) Στατιστική και ανάλυση δεδομένων με τη χρήση του στατιστικού πακέτου IBM SPSS Statistics 24, (β) Εξόρυξη κειμένων στο R και (γ) Μετα-ανάλυση με το στατιστικό πακέτο Stata 10. Μεταξύ των στόχων του Συνεδρίου ήταν η παρουσίαση σύγχρονων υπολογιστικών ή αλγοριθμικών τεχνικών και μεθόδων Στατιστικής συμπερασματολογίας για την εξαγωγή γνώσης και ανάπτυξης μοντέλων πρόβλεψης. Γενικώς, το Συνέδριο ανέδειξε τον διεπιστημονικό χαρακτήρα της Επιστήμης των Δεδομένων και τον σημαντικό ρόλο της Στατιστικής στην Πληροφορική και τη Βιοϊατρική. Οι χώροι του Πανεπιστημίου Θεσσαλίας ήταν εξαιρετικοί και άρτια εξοπλισμένοι για την υποστήριξη και διοργάνωση του Συνεδρίου.

Η τελετή έναρξης του Συνεδρίου πραγματοποιήθηκε στο κεντρικό αμφιθέατρο του Πανεπιστημίου Θεσσαλίας στη Λαμία. Την έναρξη του Συνεδρίου κήρυξε ο Πρόεδρος του ΕΣΙ, Ομότιμος Καθηγητής Π. Μωυσιάδης. Εκ μέρους της επιστημονικής και οργανωτικής επιτροπής χαιρετισμούς απεύθυναν ο Αναπληρωτής Καθηγητής Β. Πλαγιανάκος και ο Επίκουρος Καθηγητής Ι. Τριανταφύλλου. Χαιρετισμούς, επίσης απηύθυναν ο Αντιπεριφερειάρχης Στερεάς Ελλάδος Δ. Κυρίτσης και ο Αντιδήμαρχος Λαμιέων Θ. Αρναούτογλου. Ακολούθησαν σε ολομέλεια οι ομιλίες των προσκεκλημένων ομιλητών, του Senior Researcher D. Tasoulis, Winton Capital Management, με τίτλο *Statistics, Computer Science, Finance and other Shenanigans*, και του Adjunct Associate Professor M. Viana, University of Illinois at Chicago, με τίτλο *Symmetrically adapted data summaries*.

Στην πρωινή συνεδρία της Κυριακής, 6ης Μαΐου 2018, παρουσιάστηκαν οι εργασίες των υποψηφίων για το Βραβείο Καλύτερης Εργασίας Νέου Στατιστικού, το οποίο θεσμοθετήθηκε το 2005. Υποψήφιοι ήταν ο κύριος Ζ. Μπραγουδάκης, ο οποίος παρουσίασε την εργασία με τίτλο *Υπάρχει ασυμμετρία προσαρμογής των λιανικών στις βασικές κατηγορίες διατροφής, η*

περίπτωση της ελληνικής αγοράς τροφίμων», η κυρία Ε Σεραφετινίδου, η οποία παρουσίασε την εργασία με τίτλο *Παράγοντες κατά τη διάρκεια της ζωής που καθορίζουν την κατάθλιψη σε άτομα 50 ετών και άνω: Διαφοροποιήσεις μεταξύ ανδρών και γυναικών στην Ευρώπη*», ο κύριος Μ. Σπητιέρης, ο οποίος παρουσίασε την εργασία με τίτλο *Μία νέα προσέγγιση για βελτιωμένες υποκαταστάσεις πολυμεταβλητών ελλιπών δεδομένων*, και ο κύριος Ν. Μπαρούτης, ο οποίος παρουσίασε την εργασία με τίτλο *Adaptive neuro-fuzzy inference systems, artificial neural network and machine learning algorithms on prediction of coronary heart disease*. Η Επιτροπή του Βραβείου Καλύτερης Εργασίας Νέου Στατιστικού κατέθεσε τη βαθμολογία της και στην τελετή λήξης του Συνεδρίου που έλαβε χώραν την ίδια ημέρα, το βραβείο απονεμήθηκε στον κύριο Μ. Σπητιέρη. Κατόπιν εισήγησης της Επιτροπής απονεμήθηκαν και έπαινοι στους υπόλοιπους υποψηφίους. Η Επιτροπή αποτελείτο από τον Καθηγητή Κ. Ζωγράφο, τον Ομ. Καθηγητή Π. Μωυσιάδη, τον Ερευνητή Α΄ Βαθμίδος Π. Προδρομίδη και τον Ομ. Καθηγητή Χ. Χαραλαμπίδη.

Οι κοινωνικές εκδηλώσεις του Συνεδρίου πραγματοποιήθηκαν το Σαββατο, 5η Μαΐου 2018. Το απόγευμα (ώρα 18:00) υπήρξε σύντομη επίσκεψη στο Κέντρο Ιστορικής Ενημέρωσης Θερμοπυλών. Το επίσημο δείπνο του Συνεδρίου έλαβε χώραν στον πολυχώρο «Νύμφες» (7ο χλμ. Λαμίας-Δομοκού, ώρα 21:00). Σε αμφότερες περιπτώσεις η μετάβαση των συνέδρων εξυπηρετήθηκε με δύο λεωφορεία που εξασφαλίσθηκαν από την Οργανωτική Επιτροπή του Συνεδρίου.

Στον τόμο περιλαμβάνονται εργασίες που παρουσιάστηκαν στο Συνέδριο και υποβλήθηκαν για δημοσίευση. Όλες οι εργασίες κρίθηκαν από κριτές με την φροντίδα των υπευθύνων έκδοσης πρακτικών.

Οι παρατηρήσεις και τα σχόλια των κριτών, σύμφωνα με την πάγια πολιτική που ακολουθεί το ΕΣΙ, αφορούν κυρίως στον τρόπο παρουσίασης της εργασίας και στην παρουσία ή απουσία τυπογραφικών και σοβαρών επιστημονικών λαθών. Οι εργασίες πρέπει να έχουν ικανό στατιστικό περιεχόμενο, να αναδεικνύουν το πρόβλημα που μελετούν, να μην περιορίζονται μόνο σε Περιγραφική Στατιστική. Για το σκοπό αυτό υπάρχουν κριτήρια δημοσίευσης εργασιών στα Πρακτικά του ΕΣΙ τα οποία είναι αναρτημένα στην ιστοσελίδα του ΕΣΙ, www.esi-stat.gr. Όλες οι εργασίες, για τις οποίες ζητήθηκε αναθεώρηση, κρίθηκαν εκ νέου από τους κριτές ή από τους υπεύθυνους έκδοσης των πρακτικών.

Συνολικά υποβλήθηκαν 34 εργασίες. Ως κριτές των εργασιών συνεργάστηκαν οι: Ίλια Βόντα, Χαράλαμπος Δαμιανού, Πέτρος Δελαπόρτας, Γεώργιος Δονάτος, Γεώργιος Ηλιόπουλος, Θεόφιλος Κάκουλος Αλεξάνδρα Κοντόλαιμου, Δημήτριος Κουγιουμτζής, Στρατής Κουνιάς, Σταύρος

Κουρούκλης, Βασίλειος Κούτρας, Ιωάννης Κυρίτσης, Γεώργιος Μενεξές, Σταύρος Ντεγιαννάκης, Ιωάννης Παναγόπουλος, Δημήτρης Παπαναστασίου Αθανάσιος Πετραλιάς, Αλέξανδρος Πολυκάρπου, Πρόδρομος Προδρομίδης, Κυριακή Σωτηράκογλου, Γεώργιος Τζαβελάς, Γεώργιος Τσακλίδης, Χαράλαμπος Χαραλαμπίδης, Θεόδωρος Χατζηπαντελής.

Η Επιτροπή Έκδοσης Πρακτικών του ΕΣΙ εκφράζει τις ευχαριστίες της προς όλους τους κριτές για την επιμελημένη και προσεκτική αξιολόγηση των εργασιών.

Η σειρά παρουσίασης των εργασιών στον παρόντα τόμο είναι αλφαβητική με βάση το επώνυμο του πρώτου συγγραφέα. Προηγούνται οι εργασίες στην ελληνική και έπονται οι εργασίες στην αγγλική.

Το Διοικητικό Συμβούλιο (ΔΣ) του ΕΣΙ αισθάνεται την ανάγκη να ευχαριστήσει το Πανεπιστήμιο Θεσσαλίας και την Οργανωτική Επιτροπή για την πολύ καλή οργάνωση και προσφορά τους. (Η σύνθεση της Οργανωτικής Επιτροπή παρατίθεται στη σελίδα 26).

Τέλος, το ΔΣ του ΕΣΙ εκφράζει τις ευχαριστίες του στη Γραμματέα του ΕΣΙ, Αναστασία Αυγερινού για την τεχνική επιμέλεια της έκδοσης των Πρακτικών και των CD.

ΕΚ ΜΕΡΟΥΣ ΤΟΥ ΔΣ ΤΟΥ ΕΣΙ

Οι υπεύθυνοι Έκδοσης Πρακτικών του

31^{ου} Πανελληνίου Συνεδρίου Στατιστικής

Ελευθέριος Αγγελής	Δημήτριος Κωνσταντινίδης
Πολυχρόνιος Μωϋσιάδης	Γεώργιος Παπαδόπουλος
Παναγιώτης Παπαϊωάννου	Πρόδρομος Προδρομίδης

Αθανάσιος Ρακιτζής

ΧΟΡΗΓΟΙ ΣΥΝΕΔΡΙΟΥ

Η Οργανωτική Επιτροπή του 31ου Πανελλήνιου Συνέδριου Στατιστικής ευχαριστεί θερμά τους κάτωθι Χορηγούς. Χωρίς την οικονομική ενίσχυσή τους θα ήταν εξαιρετικά δύσκολο να επιτύχουμε τους στόχους του Συνεδρίου.

ΠΕΡΙΦΕΡΕΙΑ ΣΤΕΡΕΑΣ ΕΛΛΑΔΑΣ

ΔΗΜΟΣ ΛΑΜΙΕΩΝ

ΤΡΑΠΕΖΑ ΤΗΣ ΕΛΛΑΔΟΣ 1928

ONEX
A Partner. Forever.

ΠΡΟΓΡΑΜΜΑ ΣΥΝΕΔΡΙΟΥ

Παρασκευή, 4 -5- 2018

13:00-14:30 *Εγγραφή συνέδρων*

	ΑΙΘΟΥΣΑ Ι **ΠΙΘΑΝΟΤΗΤΕΣ**	*ΑΙΘΟΥΣΑ ΙΙ* **ΜΠΕΫΖΙΑΝΗ ΣΤΑΤΙΣΤΙΚΗ**
	Προεδρεύων: **Χ. Χαραλαμπίδης**	**Προεδρεύων:** **Δ. Καρλής**
14:30	**A. Di Crescenzo, Γ. Ψαρράκος:** Το πιθανοθεωρητικό ανάλογο του θεωρήματος της μέσης τιμής για δεσμευμένες τυχαίες μεταβλητές	**Π. Τσαμτσακίρη, Δ. Καρλής:** Μπεϋζιανή εκτίμηση για το μοντέλο INARCH για διακριτές χρονοσειρές
14:50	**Ν.Δ. Μαχαιράς, Σ.Μ. Τζανίνη**ς: Μία τεχνική αλλαγής μέτρου για μικτές σύνθετες ανανεωτικές διαδικασίες	**P. Economou, K. Kounetas, S. Malefaki:** A Bayesian metafrontier stochastic model
15:10	**Μ. Κούτρας, Δ. Λυμπερόπουλος:** Νέες ανισότητες για N-demimartingales με εφαρμογές στις στατιστικές συναρτήσεις σάρωσης	**K. Bourazas, P. Tsiamyrtzis:** A Bayesian self-starting Shiryaev statistic for phase I data
	ΑΙΘΟΥΣΑ Ι **ΣΤΑΤΙΣΤΙΚΑ ΜΕΤΡΑ ΑΠΟΚΛΙΣΗΣ ΚΑΙ ΕΦΑΡΜΟΓΕΣ**	*ΑΙΘΟΥΣΑ ΙΙ* **ΣΤΑΤΙΣΤΙΚΗ ΣΤΙΣ ΕΠΙΧΕΙΡΗΣΕΙΣ**
	Προεδρεύων: **Τ. Παπαϊωάννου**	**Προεδρεύων:** **Β. Δρακόπουλος**
15:30	**Γ. Αυλογιάρης, Α. Μιχέας, Κ. Ζωγράφος:** Τοπική φ-απόκλιση και εφαρμογές της	**Δ. Σακελλάρης.:** Η Επιστήμη των Δεδομένων στην υπηρεσία του αθλητισμού

15:50	**N. Martin, L. Pardo, K. Ζωγράφος:** Στατιστικά τεστ φ-απόκλισης βασισμένα στη σύνθετη πιθανοφάνεια	**D. Stogiannis, F. Siannis:** Assessing the effectiveness of TV ad spend on sales
16:10	**Α. Μπατσίδης, Π. Οικονόμου**: Επιλογή μεταξύ δύο μη εμφωλευμένων μοντέλων για την περιγραφή απαριθμητών δεδομένων	**V.L. Georgiou, E. Charitidou, E. Lempesis, A. Karamitrou, C. Kotopoulos:** Utilizing R in a big data analytics production system: The case of the IRI product availability prediction implementation
16:30	**Διάλειμμα-καφές**	
	ΑΙΘΟΥΣΑ Ι **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ** **Προεδρεύουσα:** **Φ. Κολυβά-Μαχαίρα**	***ΑΙΘΟΥΣΑ ΙΙ*** **ΚΑΤΑΝΟΜΕΣ ΠΙΘΑΝΟΤΗΤΑΣ** **Προεδρεύων:** **Δ. Κουγιουμτζής**
16:50	**Μ. Μπάρδα, Γ. Παπαδόπουλος, Π. Μπεμπέλη:** Στατιστικά εργαλεία για τη διερεύνηση της αλληλεπίδρασης γονοτύπου-περιβάλλοντος: η περίπτωση του Lupinus mutabilis Sweet	**Μ.Β. Κούτρας, Θ.Κ. Κοκκινιώτη**: Δημιουργία νέας διδιάστατης κατανομής με Γάμμα και Βήτα περιθώριες με τη μέθοδο της trivariate reduction
17:10	**Μ. Χουζούρης, Μ. Μιχαλούδη, Π. Ξένος, Μ. Νεκτάριος, Π. Τήνιος:** Συμπεριφορικές ανωμαλίες και ανορθολογικές συμπεριφορές στην εκτίμηση των γνωστικών ικανοτήτων	**Μ.Β. Κούτρας, Ν. Μαυρούτσος:** Μέθοδοι δημιουργίας νέων οικογενειών κατανομών και μελέτη των ιδιοτήτων τους
17:30	**Θ. Χατζηπαντελής, Ε. Μαρκάκη, Μ. Σωτήρογλου:** Πείραμα συμμετοχικής διαβουλευτικής διαχείρισης θεμάτων	**Β.Ε. Πιπερίγκου:** Κατανομές συναφείς με πολυώνυμα Gegenbauer και εφαρμογές
18:00	***Τελετή ενάρξεως του Συνεδρίου – Χαιρετισμοί***	

15:00-18:00 Σεμινάριο: Μετά-Ανάλυση με τη χρήση του στατιστικού πακέτου STATA (*Π.Γ. Μπάγκος, Π.Ι. Κοντού*)

ΟΜΙΛΙΕΣ ΣΕ ΟΛΟΜΕΛΕΙΑ (Προεδρεύων: Χ. Μωυσιάδης)

18:15 **D. Tasoulis:** Statistics, Computer Science, Finance and other Shenanigans

19:00 **M. Viana:** Symmetrically adapted data summaries

20:00 ***Καλωσόρισμα (Welcome cocktail)***

ΣΑΒΒΑΤΟ 5-5-2018

9:00-12:00 Σεμινάριο: Μετα-Ανάλυση με τη χρήση του στατιστικού πακέτου STATA (*Π.Γ. Μπάγκος, Π.Ι. Κοντού*)

	ΑΙΘΟΥΣΑ Ι **ΜΕΛΕΤΗ ΔΙΑΓΡΑΜΜΑΤΩΝ ΕΛΕΓΧΟΥ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΑΞΙΟΠΙΣΤΙΑΣ** **Προεδρεύων: Ι. Τριανταφύλλου**	*ΑΙΘΟΥΣΑ ΙΙ* **ΣΤΑΤΙΣΤΙΚΗ ΚΑΙ ΜΟΝΤΕΛΟΠΟΙΗΣΗ ΣΤΗ ΣΕΙΣΜΟΛΟΓΙΑ** **Προεδρεύων: Γ. Τσακλίδης**	*ΑΙΘΟΥΣΑ ΙΙΙ* **ΚΑΤΑΝΟΜΕΣ ΠΙΘΑΝΟΤΗΤΑΣ – ΔΕΙΓΜΑΤΟΛΗΨΙΑ** **Προεδρεύουσα: Ι. Βόντα**
9:00	**Ι.Τριανταφύλλου:** Ελεύθερα κατανομής διαγράμματα ελέγχου φάσης ΙΙ με χρήση διατεταγμένων παρατηρήσεων	**P. Bountzis, E. Papadimitriou, G. Tsaklidis:** Detection of seismicity rate changes for the strong earthquakes of Greece by Markovian arrival process modeling	**Β.-Ε. Τσόλκα, Κ. Πετρόπουλος:** Η τριγωνική κατανομή και η κατανομή Beta στην ανάλυση κινδύνου
9:20	**Ν. Παναγιώτου, Ι. Τριανταφύλλου:** Μελέτη μη παραμετρικών διαγραμμάτων ελέγχου με χρήση μεθόδων προσομοίωσης	**O. Mangira, V. Karakostas, G. Vasiliadis:** On the magnitude–frequency distribution for earth-quake sequences in the Central Ionian islands	**C.H. Skiadas, C. Skiadas:** A generalization of the inverse Gaussian with applications to the health state of a population

9:40	**Π. Ρούνη, Ι. Τριανταφύλλου:** Μελέτη μονότονων συστημάτων αξιοπιστίας με χρήση μεθόδων προσομοίωσης	**Ch. Kourouklas, E. Papadimitriou, V. Karakostas, G. Tsaklidis:** Analyzing the correlations of moderate to strong earthquakes interevent times in Greece	**Α. Γεωργάκης, Ι. Μέλφος, Γ. Σταματέλλος:** Πίνακες όγκου δασικών συστάδων με την Bitterlich δειγματοληψία
10:00	**Α.Χ. Ρακιτζής, A. Mukherjee:** Ένα πολυμεταβλητό διάγραμμα ελέγχου για μηδενοδιογκωμένες Poisson διεργασίες	**Δ. Χορόζογλου, Δ. Κουγιουμτζής, Ε. Παπαδημητρίου:** Διερεύνηση της συσχέτισης διαδοχικών σεισμών πριν τον κύριο σεισμό από δεδομένα σεισμικού καταλόγου για τον ελληνικό χώρο	**Ι. Παπατσούμα, Ν. Φαρμάκης:** Συναρτήσεις πυκνότητας πιθανότητας με αρνητικό εκθέτη: Θεωρητική και δειγματοληπτική προσέγγιση
10:20	**A. Apsemidis, S. Psarakis:** The role of support vector machines in statistical process monitoring	**Ρ. Λύκου, Γ. Τσακλίδης, Ε. Παπαδημητρίου:** Διερεύνηση των σημείων αλλαγής σε σεισμικές ακολουθίες των Ιονίων νήσων	**Μ. Παπαγεωργίου, Ν. Φαρμάκης:** Γραφήματα και δειγματοληψία

10:45 **Διάλειμμα-καφές**

11:30-17:30 **Σεμινάριο: Στατιστική Ανάλυση Δεδομένων με τη χρήση του πακέτου IBM *SPSS* Statistics (*Ι. Τριανταφύλλου*)**

	ΑΙΘΟΥΣΑ Ι **ΜΑΘΗΜΑΤΙΚΗ ΣΤΑΤΙΣΤΙΚΗ** **Προεδρεύων:** **Μ. Κούτρας**	*ΑΙΘΟΥΣΑ ΙΙ* **ΑΝΑΛΥΣΗ ΧΡΟΝΟΣΕΙΡΩΝ** **Προεδρεύων:** **Α. Καραγρηγορίου**
11:10	**Γ.Ηλιόπουλος:** Συμπερασματολογία για διάταξη κατανομών με λανθασμένα μοντέλα λόγου πυκνοτήτων	**Χ. Παρπούλα, Α. Καραγρηγορίου, Α. Λάμπρου:** Ανάλυση επιδημιολογικών χρονοσειρών με μοντέλα περιοδικής παλινδρόμησης τύπου Serfling

11:30	**Κ. Πετρόπουλος, L.K. Patra, S. Kumar:** Βελτιωμένοι εκτιμητές για την εντροπία στο μοντέλο μίξης εκθετικών κατανομών	**Ε.-Ν. Καλλιγέρης, Α. Καραγρηγορίου, Χ. Παρπούλα:** Μικτά περιοδικά ARMA μοντέλα για την μοντελοποίηση της επιδημικής δραστηριότητας σε συστήματα βιοεπιτήρησης
11:50	**Π. Μπομποτάς:** Εκτίμηση της μικρότερης παραμέτρου κλίμακας διπαραμετρικών εκθετικών κατανομών	**Π. Κόλιας:** Κατηγοριοποίηση σήματος οπτοκινητικού νυσταγμού με τη μέθοδο εξαγωγής χαρακτηριστικών
12:30	**T. Cacoullos:** F-tests of homoscedasticity under multinormality	**A. Milionis, N. Galanopoulos:** Time series with interdependent level and second moment: testing, consequences for modelling, applications (Part II)
13:00	**ΜΕΣΗΜΒΡΙΝΗ ΔΙΑΚΟΠΗ**	
	ΑΙΘΟΥΣΑ Ι **ΧΡΗΜΑΤΟΟΙΚΟΝΟ-ΜΙΚΑ –ΑΝΑΛΟΓΙΣΜΟΣ** **Προεδρεύων:** **Δ. Κωνσταντινίδης**	***ΑΙΘΟΥΣΑ ΙΙ*** **ΣΥΝΕΔΡΙΑ ΝΕΩΝ ΣΤΑΤΙΣΤΙΚΩΝ** **Προεδρεύων:** **Α. Ρακιτζής**
15:15	**Δ. Σκυλογιάννη, Α. Καραγρηγορίου, Γ.-Ι. Σιούρης:** Η επίδραση της χαμηλής τιμής (low price effect) σε παράγωγα και εκ των υστέρων διόρθωση του επιλεγόμενου μοντέλου πρόβλεψης σύμφωνα με το Percentage Value at Risk (PVaR)	**Σ. Ελευθεράκη:** Εισαγωγή στην Αστροστατιστική: Στατιστικές μέθοδοι και τεχνικές για την ανάλυση πολυμεταβλητών δεδομένων σε σφαιρωτά σμήνη
15:35	**G.-J. Siouris, A. Karagrigoriou, D. Skilogianni:** Conditional mean over discretized tail density function	**N. Savvopoulos, D. Aretopoulou, S. Katsos:** Estimating the paid work participation, wage and labor supply functions of men and women in Greece
15:55	**Ο. Θεοδοσιάδου, Γ. Τσακλίδης:** Επεκταμένο φίλτρο Kalman με μη αρνητικό διάνυσμα καταστάσεων	**Ι.Α. Θάνος:** Μεικτές σύνθετες κατανομές Poisson

16:15	**Ι. Μπαντούνας, Γ. Πιτσέλης:** Εκτίμηση αποθεμάτων με χρήση μεθόδων Kalman filter	**Σ. Ευθυμιόπουλος, Σ.-Γ. Σταυρακάκης:** Επιληπτική Χρονοβιολογία	
	ΑΙΘΟΥΣΑ Ι **ΙΣΤΟΡΙΑ ΠΙΘΑΝΟΤΗΤΩΝ ΚΑΙ ΣΤΑΤΙΣΤΙΚΗΣ** **Προεδρεύων: Θ. Κάκουλλος**	***ΑΙΘΟΥΣΑ ΙΙ*** **ΒΙΟΣΤΑΤΙΣΤΙΚΗ** **Προεδρεύων: Δ. Παναγιωτάκος**	***ΑΙΘΟΥΣΑ ΙΙΙ*** **ΕΠΙΧΕΙΡΗΣΙΑΚΗ ΕΡΕΥΝΑ ΚΑΙ ΟΙΚΟΝΟΜΕΤΡΙΑ** **Προεδρεύων: Ν. Φαρμάκης**
16:40	**Χ. Χαραλαμπίδης:** Γένεση θεωρίας Πιθανοτήτων: Cardano εναντίον Pascal και Fermat	**N.D. Mparoutis:** Data preprocessing on UCI dataset for the coronary heart disease	**Κ.Π. Κατσάνου, Ι. Δημητρίου:** Αποτίμηση απόδοσης ενός δυναμικού πρωτοκόλλου πρόσβασης ασύρματου συνεργατικού δικτύου
17:00	**Σ. Κουνιάς:** Η έννοια του τυχαίου και της πιθανότητας 1600 π.Χ-1800 μ.Χ.	**Σ. Αναστασιάδου:** Η σχέση του γονιδί-ου apolipoprotein E genotype με το Alzheimer disease: Μια meta-analysis	**A.E. Milionis, E. Karagiota:** Systematic risk and market microstructure in a stock market under conditions of economic crisis. The case of the Athens Stock Exchange
17:20	**T. Papaioannou:** Reproducibility crisis in science	**K. Pateras, S. Nikolakopoulos S., K.CB. Roes:** The enlightening journey of three data generating models	**A.E. Milionis:** Statistical analysis of election results. An alternative approach and interpretation
18:00	**ΕΚΔΡΟΜΗ**		
21:00	**ΔΕΙΠΝΟ**		

ΚΥΡΙΑΚΗ 6-5-2018

9:00-13:00 Σεμινάριο: Εξόρυξη Γνώσης από κείμενα με χρήση της γλώσσας προγραμματισμού *R (Σ. Τασουλής)*

	ΑΙΘΟΥΣΑ Ι ΕΙΔΙΚΗ ΣΥΝΕΔΡΙΑ: ΒΕΛΤΙΩΣΗ ΤΗΣ ΠΟΙΟΤΗΤΑΣ ΤΩΝ ΠΑΡΕΧΟΜΕΝΩΝ ΥΠΗΡΕΣΙΩΝ ΥΓΕΙΑΣ **Προεδρεύων: Γ. Παπαδόπουλος**	***ΑΙΘΟΥΣΑ ΙΙ*** ΠΕΙΡΑΜΑΤΙΚΟΙ ΣΧΕΔΙΑΣΜΟΙ - ΑΝΑΛΥΣΗ ΔΙΑΚΡΙΤΩΝ ΧΡΟΝΟΣΕΙΡΩΝ **Προεδρεύων: Σ. Κουνιάς**
9:00	**Θ. Πασχάλης, Χ. Γεωργακοπούλου, Σ. Γούλας, Σ. Μπερσίμης:** Σύνθετοι δείκτες ποιότητας για την αξιολόγηση της παροχής υπηρεσιών υγείας	**Β. Χασιώτης, Στ.Α. Χατζόπουλος, Σ. Κουνιάς:** Βέλτιστοι 3^m παραγοντικοί σχεδιασμοί εκτιμητικής τάξης IV με $N \equiv 0 \bmod 3$ παρατηρήσεις
9:20	**Χ. Γεωργακοπούλου, Θ. Πασχάλης, Σ. Γούλας, Σ. Μπερσίμης, Α. Σαχλάς, Β. Πλαγιανάκος:** Αξιολόγηση της ποιότητας των υπηρεσιών υγείας με χρήση σύνθετων δεικτών	**Χ. Ευαγγελάρας:** Κατασκευή βέλτιστων υπερκορεσμένων σχεδιασμών δύο επιπέδων από κορεσμένους και σχεδόν κορεσμένους ορθογώνιους σχηματισμούς
9:40	**Σ. Γούλας, Θ. Πασχάλης, Χ. Γεωργακοπούλου, Σ. Μπερσίμης, Α. Σαχλάς, Β. Πλαγιανάκος:** Εφαρμογές των σύνθετων δεικτών για την διαχρονική αξιολόγηση της ποιότητας των υπηρεσιών υγείας	**D. Karlis:** Modelling longitudinal discrete data with a BINAR regression model
10:00	**Σ. Τασουλής, Θ. Πασχάλης, Χ. Γεωργακοπούλου, Σ. Γούλας, Σ. Μπερσίμης, Α. Σαχλάς, Β. Πλαγιανάκος:** Τεχνικές μηχανικής μάθησης για την ανάλυση	**Μ. Αναστασοπούλου, Α. Ρακιτζής:** Αυτοπαλίνδρομα υποδείγματα διακριτών χρονοσειρών με πεπερασμένο εύρος τιμών: Μια εφαρμογή σε δεδομένα πληθωρισμού των χωρών-μελών της Ευρωζώνης
10:20	**Διάλειμμα-καφές**	

	ΑΙΘΟΥΣΑ Ι	*ΑΙΘΟΥΣΑ ΙΙ*
	Προεδρεύων: Σ. Μπερσίμης	**Προεδρεύοντες: Κ. Ζωγράφος, Π. Μωυσιάδης, Π. Προδρομίδης, Χ. Χαραλαμπίδης**
10:45	**Σ. Μπερσίμης, Α. Σαχλάς, Π. Οικονόμου:** Παρακολούθηση του αριθμού και της χωρικής κατανομής των νέων κρουσμάτων χρησιμοποιώντας κατάλληλα διαγράμματα ελέγχου και κυρτά πολυγωνικά κέλυφα	**Ζ. Μπραγουδάκης:** Υπάρχει ασυμμετρία προσαρμογής των λιανικών τιμών στις βασικές κατηγορίες διατροφής; Η περίπτωση της ελληνικής αγοράς τροφίμων
11:05	**Σ. Μπερσίμης, Α. Σαχλάς, Π. Οικονόμου:** Μια νέα τεχνική για την παρακολούθηση της Δημόσιας Υγείας με χρήση διαγραμμάτων ελέγχου και συναρτήσεων σάρωσης	**Ε. Σεραφετινίδου, Γ. Βερροπούλου:** Παράγοντες κατά τη διάρκεια της ζωής που καθορίζουν την κατάθλιψη σε άτομα ηλικίας 50 & άνω: Διαφοροποιήσεις μεταξύ ανδρών και γυναικών στην Ευρώπη
11:25	**Σ. Ψαράκης, Α. Σαχλάς, Σ. Μπερσίμης:** Risk-adjusted διαγράμματα ελέγχου	**Σπητιέρης Μ.:** Μια νέα προσέγγιση για βελτιωμένες υποκαταστάσεις πολυμεταβλητών ελλιπών δεδομένων
11:45	**G. Bartzis, W. Ligterink, F.v. Eeuwijik:** Estimating metabolite networks using a multi-step network approach integrating information from lower levelled omic graphical structures	**N.D. Mparoutis, E. Sergaki:** Adaptive neuro-fuzzy inference systems, artificial neural network and machine learning algorithms on prediction of coronary heart disease

12:10-14:00 **<u>Παρουσιάσεις Αναρτημένων Ανακοινώσεων</u>**

Σ. Ζήμερας, Ζ. Κυρανά: Το φαινόμενο της βίας στα ελληνικά ποδοσφαιρικά γήπεδα

Κ. Μακρής, Ι. Βόντα: Στατιστική ανάλυση επιδημιολογικών χρονοσειρών

Κ.Γ. Πανταβού, Π.Γ. Μπάγκος: Η εποχή γέννησης και ο κίνδυνος εμφάνισης σκλήρυνσης κατά πλάκας: Μια πολυμεταβλητή μεταανάλυση

Κ. **.Γ. Πανταβού, Π.Γ. Μπάγκος:** Η διακύμανση της συχνότητας αλληλόμορφων και η επίδρασή της στη μετα-ανάλυση

Κ.Ε. Βέννου, Π.Ι. Κοντού, Π.Γ. Μπάγκος: Μετα-ανάλυση πολλαπλών εκβάσεων σε δεδομένα γονιδιακής έκφρασης

Δ. Πανάρετος, Μ. Βαμβακάρη, Γ. Τζαβελάς, Δ. Παναγιωτάκος: Διερεύνηση του ρόλου των ορθογωνίων και μη, μετασχηματισμών των αξόνων στην παραγοντική ανάλυση, σε σχέση με την επαναληψιμότητα των εξαγόμενων προτύπων κάτω από διάφορα σενάρια τυχαίου σφάλματος σε αρχικά δεδομένα

Σ. Μεσσαριτάκης, Α. Λέλης, Α. Γεωργάκης: Σύγκριση μοντέλων όγκου απλής και πολλαπλής γραμμικής παλινδρόμησης με αντίστοιχα νευρωνικών δικτύων σε δασοβιομετρικά δεδομένα του Δημοτικού Δάσους Νάουσας

Β. Τζιούφας, Στ.Α. Χατζόπουλος, Φ. Κολυβά-Μαχαίρα: Κατασκευή D-βέλτιστων κορεσμένων παραγοντικών σχεδιασμών με χρήση βάσεων Graver και Groebner

Ι.Α. Ταμπόσης, Κ.Δ. Τσιρίγος, Μ.Κ. Θεοδωροπούλου, Π.Ι. Κοντού, Π.Γ. Μπάγκος: Αλγόριθμοι εκπαίδευσης για hidden Markov models βασισμένοι στην ημιεπιβλεπόμενη μάθηση

14:00 ΜΕΣΗΜΒΡΙΝΗ ΔΙΑΚΟΠΗ

16:00-18:00 Σεμινάριο: Εξόρυξη Γνώσης από κείμενα με χρήση της γλώσσας προγραμματισμού *R (Σ. Τασουλής)*

	ΑΙΘΟΥΣΑ Ι **ΑΝΑΛΥΣΗ ΘΝΗΣΙΜΟΤΗΤΑΣ** **Προεδρεύων: Π. Μπάγκος**	*ΑΙΘΟΥΣΑ ΙΙ* **ΕΦΑΡΜΟΣΜΕΝΗ ΠΟΛΥΜΕΤΑΒΛΗΤΗ ΑΝΑΛΥΣΗ** **Προεδρεύων: Κ. Ζωγράφος**	*ΑΙΘΟΥΣΑ ΙΙΙ* **ΟΙΚΟΝΟΜΙΚΗ ΣΤΑΤΙΣΤΙΚΗ - ΟΙΚΟΝΟΜΕΤΡΙΑ** **Προεδρεύων: Π. Προδρομίδης**
16:30	**Y. Dimotikalis, C.H. Skiadas:** Application of max entropy principle to human mortality life table data	**Α. Σαρικάκη, Τ. Δάρας:** Ανάλυση κύριων συνιστωσών	**Ζ. Γεωργαντά, Ν. Λογοθέτης:** Δημόσιο χρέος και έλλειμμα 2009: Greece's zero milestone

16:50	**Α. Μποζίκας, Γ. Πιτσέλης:** Πρόβλεψη θνησιμότητας με τη χρήση αναλογιστικών μοντέλων αξιοπιστίας	**Α. Σταμάτης, Ε. Ταμπούρης, Κ. Ταραμπάνης:** Οδηγούμενη από δεδομένα ανάλυση των τουριστικών προορισμών της Ελλάδας	**Α. Τσιμπάνος, Χ. Αγιακλόγλου:** Επιλογή υποδειγμάτων με πληροφοριακά κριτήρια σε χωρικές διαδικασίες
17:10	**Κ.Ν. Ζαφείρης, Χ. Σκιαδάς, Α. Τσώνη:** Η υγεία στην Ευρωζώνη σήμερα	**Σ. Ζήμερας, Ε. Κυρανάς, Κ. Γεροθανάση, Ζ. Κυρανά:** Τεχνικές ομαδοποίησης των διατροφικών συνηθειών των Ελλήνων εφήβων	**Π. Προδρομίδης:** Η εξέλιξη του αριθμού των απασχολουμένων ανά κλάδο, των ανέργων και των μη-μετεχόντων στην αγορά εργασίας ανά περιφέρεια στην διάρκεια της ύφεσης (2008-2017)

18:00 **ΟΛΟΜΕΛΕΙΑ: ΛΗΞΗ ΣΥΝΕΔΡΙΟΥ**

<u>Επιστημονική Επιτροπή</u>
Βασδέκης Βασίλειος, Οικονομικό Πανεπιστήμιο Αθηνών
Βόντα Φίλια, Μετσόβιο Πολυτεχνείο
Δαμιανού Χαράλαμπος, Πανεπιστήμιο Αθηνών
Δρακόπουλος Βασίλειος, Πανεπιστήμιο Θεσσαλίας
Καραγρηγορίου Αλέξανδρος, Πανεπιστήμιο Αιγαίου
Κολυβά-Μαχαίρα Φωτεινή, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
Μαλεφάκη Σόνια, Πανεπιστήμιο Πατρών
Μπάγκος Παντελής, Πανεπιστήμιο Θεσσαλίας
Μπατσίδης Απόστολος, Πανεπιστήμιο Ιωαννίνων
Μπουρνέτας Απόστολος, Πανεπιστήμιο Αθηνών
Μωυσιάδης Πολυχρόνης, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
Παναγιωτάκος Δημοσθένης, Χαροκόπειο Πανεπιστήμιο
Παπαϊωάννου Τάκης, Πανεπιστήμια Πειραιώς και Ιωαννίνων
Πλαγιανάκος Βασίλειος, Πανεπιστήμιο Θεσσαλίας
Ρακιτζής Αθανάσιος, Πανεπιστήμιο Αιγαίου
Τριανταφύλλου Ιωάννης, Πανεπιστήμιο Θεσσαλίας
Τσακλίδης Γεώργιος, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
Χριστοφίδης Τάσος, Πανεπιστήμιο Κύπρου

<u>Οργανωτική Επιτροπή</u>
Αδάμ Μαρία, Πανεπιστήμιο Θεσσαλίας
Δρακόπουλος Βασίλειος, Πανεπιστήμιο Θεσσαλίας
Μπάγκος Παντελής, Πανεπιστήμιο Θεσσαλίας
Μπερσίμης Σωτήριος, Πανεπιστήμιο Πειραιώς
Παναγιωτάκος Δημοσθένης, Χαροκόπειο Πανεπιστήμιο
Πλαγιανάκος Βασίλειος, Πανεπιστήμιο Θεσσαλίας
Τριανταφύλλου Ιωάννης, Πανεπιστήμιο Θεσσαλίας

<u>Υπεύθυνος Προγράμματος</u>
Ηλιόπουλος Γεώργιος, Πανεπιστήμιο Πειραιώς

εργασίες

στα ελληνικά

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 31ου Πανελληνίου Συνεδρίου Στατιστικής (2018), σελ.027-041

Αυτοπαλίνδρομα Υποδείγματα Διακριτών Χρονοσειρών με Πεπερασμένο Εύρος Τιμών: Μια Συγκριτική Εφαρμογή σε Δεδομένα Πληθωρισμού των Χωρών-Μελών της Ευρωζώνης

Μ. Αναστασοπούλου
Α.Ρακιτζής
Πανεπιστήμιο Αιγαίου – Τμήμα Στατιστικής και Αναλογιστικών-Χρηματοοικονομικών Μαθηματικών
anastasopoulou@aegean.gr
arakitz@aegean.gr

ΠΕΡΙΛΗΨΗ

Στην παρούσα εργασία, γίνεται μια επισκόπηση των πιο γνωστών και συχνά χρησιμοποιούμενων Αυτοπαλίνδρομων Υποδειγμάτων Διακριτών Χρονοσειρών με Πεπερασμένο Εύρος Τιμών. Τα μοντέλα αυτά είναι το Διωνυμικό Αυτοπαλίνδρομο μοντέλο $1^{ης}$ τάξης (BAR(1)), το Βήτα-Διωνυμικό Αυτοπαλίνδρομο μοντέλο $1^{ης}$ τάξης (BBAR(1)) και το Διωνυμικό Υπό Συνθήκη Ετεροσκεδαστικό Αυτοπαλίνδρομο μοντέλο $1^{ης}$ τάξης (BINARCH(1)). Παρουσιάζονται οι βασικές ιδιότητες των παραπάνω τριών μοντέλων καθώς και μέθοδοι εκτίμησης των παραμέτρων τους. Δίδεται επίσης και μια συγκριτική εφαρμογή τους σε μηνιαία δεδομένα πληθωρισμού για τις χώρες μέλη της Ευρωζώνης. Στόχος είναι η μοντελοποίηση του μηνιαίου αριθμού των χωρών-μελών της Ευρωζώνης με πληθωρισμό που δεν υπερβαίνει το 2%, για το διάστημα 2000-2018. Για τη σύγκριση και την επιλογή του καταλληλότερου μοντέλου, χρησιμοποιούνται τα κριτήρια AIC και BIC.

Λέξεις Κλειδιά: Αυτοπαλίνδρομα Υποδείγματα, Διωνυμική Κατανομή, Διωνυμικός τελεστής λέπτυνσης, Μέθοδος μέγιστης πιθανοφάνειας, Πληθωρισμός Ευρωζώνης.

1. ΕΙΣΑΓΩΓΗ

Η μοντελοποίηση και η ανάλυση διακριτών χρονοσειρών έχουν αποκτήσει αυξημένο ενδιαφέρον κατά τη διάρκεια των τελευταίων δεκαετιών. Δεδομένα τα οποία σχετίζονται π.χ. με τον μηνιαίο αριθμό ατυχημάτων σε ένα εργοστάσιο, με το πλήθος των περιστατικών μιας συγκεκριμένης ασθένειας ανα γεωγραφική περιοχή, τον αριθμό των εταιρειών που πραγματοποιούν συναλλαγές στο χρηματιστήριο κ.λ.π., δύναται να μοντελοποιηθούν μέσω ενός υποδείγματος χρονοσειρών με ακέραιες τιμές. Επίσης, τέτοια δεδομένα μπορούν να προκύψουν από τη διακριτοποίηση χρονοσειρών με συνεχείς τιμές.

Ειδικότερα, τα τελευταία χρόνια έχει αναπτυχθεί η επιστημονική περιοχή που αφορά μοντέλα διακριτών χρονοσειρών με πεπερασμένο εύρος τιμών. Έτσι, αν X_t είναι η τιμή της χρονοσειράς τη χρονική στιγμή t, τότε (συνήθως) θεωρούμε ότι λαμβάνει τιμές σε ένα σύνολο της μορφής $\{0, 1, \ldots, n\}$ (δείτε π.χ. Weiß (2009a), Weiß and Kim (2013) και Weiß and Pollett (2014)).

Το πιο γνωστό μοντέλο αυτής της κατηγορίας είναι το Διωνυμικό Αυτοπαλίνδρομο Μοντέλο $1^{ης}$ τάξης (*Binomial Autoregressive model of 1st order*, Binomial AR(1) ή BAR(1)) το οποίο προτάθηκε από τον McKenzie (1985) και αποτελεί το διακριτό ανάλογο του συνήθους AR(1) μοντέλου για συνεχείς τυχαίες μεταβλητές (τ.μ.). Το μοντέλο BAR(p), $p > 1$, μελετήθηκε από τον Weiß (2009b).

Οι Weiß and Kim (2014) παρατήρησαν ότι η Διωνυμική κατανομή χαρακτηρίζεται από μια σταθερή σχέση μεταξύ μέσης τιμής και διασποράς (ανάλογη με αυτή στην περίπτωση της κατανομής Poisson), η οποία είναι αρκετά περιοριστική σε πρακτικές εφαρμογές. Έτσι, με σκοπό να μοντελοποιήσουν την υπερ-μεταβλητότητα (*overdispersion*) η οποία μπορεί να παρατηρηθεί σε δεδομένα διακριτών χρονοσειρών με πεπερασμένο εύρος τιμών, πρότειναν και μελέτησαν το Βήτα-Διωνυμικό Αυτοπαλίνδρομο Μοντέλο $1^{ης}$ τάξης (*Beta-Binomial Autoregressive model of 1st order*, Beta-Binomial AR(1) ή BBAR(1)), δίνοντας παράλληλα και μια επέκταση για αυτοπαλίνδρομα μοντέλα μεγαλύτερης τάξης (BBAR(p), $p > 1$).

Μια τρίτη προσέγγιση για την επέκταση του μοντέλου BAR(1), αποτελούν τα μοντέλα INGARCH (δείτε π.χ. Ferland et al. 2006). Για την περίπτωση διακριτών χρονολογικών δεδομένων με πεπερασμένο εύρος τιμών, οι Weiß and Pollett (2014) πρότειναν το Υπό Συνθήκη Ετεροσκεδαστικό Διωνυμικό Αυτοπαλίνδρομο μοντέλο $1^{ης}$ τάξης (*Binomial Autoregressive Conditionally Heteroskedastic model of 1st orde*r ή BINARCH(1)) ενώ οι Ristic et al. (2015) επέκτειναν τη μελέτη των Weiß and Pollett (2014), προτείνοντας και μελετώντας το μοντέλο BINARCH(p), $p > 1$.

Στην παρούσα εργασία, παρουσιάζονται συνοπτικά τα παραπάνω τρία μοντέλα $1^{ης}$ τάξης και δίνεται μια συγκριτική εφαρμογή τους σε μηνιαία δεδομένα πληθωρισμού των χωρών-μελών της Ευρωζώνης. Η διάρθρωση της εργασίας είναι η εξής: Στις Ενότητες 2, 3 και 4 παρουσιάζονται τα μοντέλα BAR(1), BBAR(1) και BINARCH(1), αντίστοιχα. Δίδονται οι βασικές τους ιδιότητες καθώς και μέθοδοι εκτίμησης των παραμέτρων τους. Στην Ενότητα 5 γίνεται εφαρμογή των παραπάνω μοντέλων σε μηνιαία δεδομένα πληθωρισμού για τις χώρες-μέλη της Ευρωζώνης, με σκοπό την επιλογή του καταλληλότερου μοντέλου μεταξύ των τριών. Τέλος, χρήσιμα συμπεράσματα συνοψίζονται στην Ενότητα 6.

2. ΤΟ ΜΟΝΤΕΛΟ BAR(1)

2.1 Εισαγωγή

Στην έως τώρα βιβλιογραφία, για τη μοντελοποίηση διακριτών χρονοσειρών X_t, $t \geq 1$, χρησιμοποιούνται διακριτά μοντέλα με άπειρο εύρος τιμών (π.χ. στο $\mathbb{N}_0 = \{1, 2, \ldots\}$ ή στο $\mathbb{N} = \{0, 1, 2, \ldots\}$). Όμως σε αρκετές περιπτώσεις, το εύρος τιμών είναι

πεπερασμένο, δηλ., $\{0,1,...,n\}$ με $n \in \mathbb{N}$ γνωστό. Το πιο συχνά χρησιμοποιούμενο μοντέλο για τη μοντελοποίηση μιας διακριτής τ.μ. με πεπερασμένο εύρος τιμών είναι αυτό της Διωνυμικής κατανομής $B(n,\pi)$, όπου $\pi \in (0,1)$ είναι η πιθανότητα επιτυχίας. Ο McKenzie (1985) πρότεινε το μοντέλο BAR(1), το οποίο βασίζεται στο διωνυμικό τελεστή λέπτυνσης (*binomial thinning operator*, Steutel and van Harn (1979)), ο οποίος συμβολίζεται με «$\circ$».

Πριν προχωρήσουμε στην παράθεση των ιδιοτήτων του μοντέλου BAR(1), πρέπει να αναφέρουμε ότι οι τελεστές λέπτυνσης είναι πιθανοτικοί τελεστές (*probabilistic operators*) οι οποίοι εφαρμόζονται σε τ.μ. και ορίζουν πράξεις μεταξύ τους. Πιο συγκεκριμένα, αν *N* είναι μια μη αρνητική διακριτή τ.μ. και $a \in (0,1)$ τότε με χρήση του διωνυμικού τελεστή λέπτυνσης ορίζεται η τ.μ.

$$a \circ N = \sum_{t=1}^{N} X_t \,, \tag{1}$$

όπου X_t, $t \geq 1$, είναι μια ακολουθία ανεξάρτητων και ισόνομων (*independent and identically distributed, iid*) τ.μ. Bernoulli, ανεξάρτητων από το *N*, με πιθανότητα επιτυχίας *a*. Η τ.μ. $a \circ N$ μετρά τον αριθμό των επιτυχιών σε ένα τυχαίο πλήθος δοκιμών Bernoulli όπου η πιθανότητα *a* παραμένει σταθερή σε όλο το πείραμα. Το πλήθος των δοκιμών Bernoulli καθορίζεται από την τιμή τα *N*. Έτσι, η τ.μ. $a \circ N$ ακολουθεί τη Διωνυμική κατανομή *B*(*N*, *a*). Αν υποθέσουμε ότι αυτές οι μετρήσεις αντιπροσωπεύουν το μέγεθος τα πληθυσμού, τότε ο τελεστής λέπτυνσης διαγράφει τυχαία μερικά από τα μέλη του πληθυσμού με στόχο τη συρρίκνωση του, έτσι ώστε να οδηγεί πάντα σε ακέραιες τιμές. Ο πιο γνωστός τελεστής λέπτυνσης είναι ο διωνυμικός. Στη συνέχεια δίνουμε τα βασικές ιδιότητες της τ.μ. $a \circ N$ (για λεπτομέρειες, δείτε Weiß (2008)):

$$E(a \circ N) = a \cdot \mu_N \,,\ V(a \circ N) = a^2 \cdot \sigma_N^2 + a(1-a) \cdot \mu_N \,,\ \mathrm{Cov}(a \circ N) = a \cdot \sigma_N^2 \,, \tag{2}$$

όπου $\mu_N = E(N)$ και $\sigma_N^2 = V(N)$.

2.2 Βασικές Ιδιότητες του Μοντέλου BAR(1)

Σύμφωνα με τον McKenzie (1985), το μοντέλο BAR(1) ορίζεται ως εξής:
Ορισμός 2.1: Έστω $n \in \mathbb{N}$. Οι παράμετροι του μοντέλου είναι οι π, ρ με $\pi \in (0,1)$, $\rho \in \left(\left[\max\left(-\frac{\pi}{1-\pi}, -\frac{1-\pi}{\pi} \right) \right], 1 \right)$ και ορίζεται από τη σχέση:

$$X_t = a \circ X_{t-1} + \beta \circ (n - X_{t-1}) \,,\ t \geq 1, \tag{3}$$

όπου $X_0 \sim B(n,\pi)$, $\beta = \pi(1-\rho)$ και $a = \beta + \rho$. Οι τ.μ. $a \circ X_{t-1}$ και $\beta \circ (n - X_{t-1})$ είναι αμοιβαία ανεξάρτητες ακολουθίες ανεξάρτητων τ.μ. Bernoulli με παραμέτρους *a* και β, αντίστοιχα. Επίσης, είναι ανεξάρτητες των X_s, $s < t$.

Η φυσική ερμηνεία του παραπάνω μοντέλου είναι η εξής: Έστω ένα σύστημα με *n* ανεξάρτητες μονάδες που βρίσκονται είτε στην κατάσταση 0 (π.χ. μη λειτουργία) είτε

στην κατάσταση 1 (π.χ. λειτουργία). Ο αριθμός των μονάδων που λειτουργούν τη χρονική στιγμή *t* είναι X_t και ίσος με το άθροισμα των παρακάτω ποσοτήτων: (i) του πλήθους των μονάδων που βρίσκονται στην κατάσταση 1 τη χρονική στιγμή *t* ενώ την προηγούμενη χρονική στιγμή ήταν στην κατάσταση 1 και (ii) του πλήθους των μονάδων που βρίσκονται στην κατάσταση 1 τη χρονική στιγμή *t* ενώ την προηγούμενη χρονική στιγμή ήταν στην κατάσταση 0.

Η πρώτη ποσότητα εκφράζεται από τον όρο $a \circ X_{t-1}$, όπου a είναι η πιθανότητα η μονάδα να παραμείνει στην κατάσταση 1 ενώ η δεύτερη από τον όρο $\beta \circ (n - X_{t-1})$, όπου β είναι η πιθανότητα η μονάδα να «μεταβεί» από την κατάσταση 0 στην κατάσταση 1.

Επίσης, το μοντέλο BAR(1) έχει ως περιθώρια κατανομή τη $B(n,\pi)$, δηλ. η $P(X_t = x)$ υπολογίζεται μέσω της συνάρτησης πιθανότητας της $B(n,\pi)$. Άρα, ισχύει ότι $E(X_t) = n\pi$ και $V(X_t) = n\pi(1-\pi)$. Αν συμβολίσουμε με $\rho(k) = Corr(X_t, X_{t-k})$ την αυτοσυσχέτιση *k*-τάξης, αυτή ισούται με

$$\rho(k) = Corr(X_t, X_{t-k}) = \rho^k, \quad k \geq 0. \tag{4}$$

Μία βασική ιδιότητα του μοντέλου BAR(1) είναι ότι η ακολουθία X_t, $t \geq 1$, είναι μία στάσιμη Μαρκοβιανή αλυσίδα με πιθανότητες μετάβασης $p_{k|l} = P(X_t = k \mid X_{t-1} = l)$, $k,l \in \{0,...,n\}$, οι οποίες δίνονται από τη σχέση:

$$p_{k|l} = \sum_{m=\max(0,k+l-n)}^{\min(k,l)} \binom{l}{m}\binom{n-l}{k-m} a^m (1-a)^{l-m} \beta^{k-m} (1-\beta)^{n-l+m-k} \tag{5}$$

Ακόμη, η υπό συνθήκη μέση τιμή και διασπορά έχουν την ιδιότητα να είναι γραμμικές ως προς τη X_{t-1} και δίνονται από τις παρακάτω σχέσεις:

$$E(X_t \mid X_{t-1}) = \rho X_{t-1} + n\beta \tag{6}$$

$$V(X_t \mid X_{t-1}) = \rho(1-\rho)(1-2\pi)X_{t-1} + n\beta(1-\beta). \tag{7}$$

2.3 Μέθοδοι Εκτίμησης Παραμέτρων του Μοντέλου BAR(1)

Στην πράξη οι πραγματικές τιμές των παραμέτρων *π, ρ* του μοντέλου BAR(1) δεν είναι γνωστές και πρέπει να εκτιμηθούν. Έστω ότι έχουμε στη διάθεσή μας ένα τμήμα (*segment*) $X_1,...,X_T$ με $T \in \mathbb{N}$ από τη στάσιμη διαδικασία του μοντέλου BAR(1).

Η βασική μέθοδος εκτίμησης είναι η μέθοδος της μέγιστης πιθανοφάνειας (*maximum likelihood estimation,* MLE), με τη συνάρτηση πιθανοφάνειας να γράφεται στην ακόλουθη μορφή:

$$L(\pi,\rho) = \binom{n}{x_1} \pi^{x_1} (1-\pi)^{n-x_1} \prod_{t=2}^{T} p_{x_{t-1}|x_t}, \tag{8}$$

όπου οι πιθανότητες μετάβασης $p_{x_{t-1}|x_t}$ δίνονται από τη σχέση (5). Για την εύρεση των εκτιμητών μέγιστης πιθανοφάνειας (ΕΜΠ) $\hat{\pi}_{ML}$, $\hat{\rho}_{ML}$, μεγιστοποιούμε τη συνάρτηση $l(\pi,\rho)=\log L(\pi,\rho)$ με κατάλληλες αριθμητικές μεθόδους (εδώ έχει χρησιμοποιηθεί η συνάρτηση optim() της R). Επίσης, τα τυπικά σφάλματα των εκτιμητών, προκύπτουν από την παρατηρούμενη πληροφορία του Fisher (*observed Fisher information*, δείτε Weiß (2009a)).

3. ΤΟ ΜΟΝΤΕΛΟ BBAR(1)

3.1 Εισαγωγή

Στην παρούσα Ενότητα θα δούμε μια επέκταση του μοντέλου BAR(1), η οποία είναι χρήσιμη όταν παρατηρείται ετερογένεια μεταξύ των n μονάδων ή όταν η μεταβλητότητα των δεδομένων παρουσιάζει σημαντική απόκλιση από την αντίστοιχη του διωνυμικού μοντέλου. Σε πολλές πρακτικές εφαρμογές, δεν μπορούμε να υποθέσουμε ότι η πιθανότητα π είναι η ίδια και για τις n μονάδες.

Για να ανιχνεύσουμε την ύπαρξη ετερογένειας χρησιμοποιούμε τον Διωνυμικό δείκτη μεταβλητότητας (*binomial index of dispersion, I_d*). Έστω τ.μ. X με τιμές στο σύνολο $\{0,...,n\}$, με μέση τιμή μ και διασπορά σ^2. Τότε ο δείκτης I_d ισούται με

$$I_d=\frac{n\sigma^2}{\mu(n-\mu)}\in(0,\infty), \tag{9}$$

και παρέχει μία ένδειξη του κατά πόσο διαφέρει η διακύμανση από τη μέση τιμή. Στην περίπτωση της Διωνυμικής κατανομής $B(n,\pi)$, είναι $I_d=1$ για κάθε $\pi\in(0,1)$. Επομένως, οποιαδήποτε τ.μ. με πεπερασμένο εύρος τιμών και $I_d>1$, παρουσιάζει υπερ-μεταβλητότητα (*extra-binomial variation*), δηλαδή μεγαλύτερη μεταβλητότητα σε σύγκριση με τη μεταβλητότητα της κατανομής $B(n, \pi)$.

Ένα εύχρηστο διακριτό μοντέλο πιθανοτήτων με πεπερασμένο εύρος τιμών για την περίπτωση τα υπερ-μεταβλητότητας, είναι αυτό της Βήτα-Διωνυμικής κατανομής (*Beta-Binomial distribution*, δείτε π.χ. Griffiths (1973)), Το μοντέλο με το οποίο θα ασχοληθούμε σε αυτή την ενότητα είναι το μοντέλο BBAR(1) το οποίο βασίζεται σε μία γενίκευση του διωνυμικού τελεστή λέπτυνσης «$\circ$» μέσω της Βήτα-Διωνυμικής κατανομής.

Ειδικότερα, έστω τ.μ. a_ϕ, ανεξάρτητη της X, η οποία ακολουθεί την κατανομή $Beta\left(\frac{a(1-\phi)}{\phi},\frac{(1-a)(1-\phi)}{\phi}\right)$ με $a,\phi\in(0,1)$. Τότε, η τ.μ. $a_\phi\circ X$ προκύπτει από τη X μέσω Βήτα-Διωνυμικής λέπτυνσης αν ο τελεστής «$\circ$» εφαρμόζεται ανεξάρτητα στις a_ϕ και X. Κατά συνέπεια, η (δεσμευμένη) τ.μ. $a_\phi\circ X\,|\,X=x$ ακολουθεί τη Βήτα-Διωνυμική κατανομή και άρα

$$E(a_\phi\circ X)=a\cdot E(X), \tag{10}$$

$$V(a_\phi \circ X) = a^2 \cdot V(X) + a(1-a)E(X) + \phi a(1-a) - E(X(X-1)). \qquad (11)$$

Σημειώνεται ότι για τιμές του ϕ κοντά στο μηδέν η συμπεριφορά του Βήτα Διωνυμικού Τελεστή λέπτυνσης προσεγγίζει αυτή του Διωνυμικού Τελεστή λέπτυνσης.

3.2 Βασικές Ιδιότητες του Μοντέλου BBAR(1)

Σύμφωνα με τους Weiß and Kim (2014), το μοντέλο BBAR(1) ορίζεται ως εξής:

Ορισμός 3.1: Έστω $n \in$. Οι παράμετροι του μοντέλου είναι οι π, ρ, ϕ με $\pi \in (0,1)$ $\rho \in \left(\left[\max\left(-\frac{\pi}{1-\pi}, -\frac{1-\pi}{\pi} \right) \right], 1 \right)$, $\phi \in (0,1)$ και ορίζεται από τη σχέση

$$X_t = a_\phi \circ X_{t-1} + \beta_\phi \circ (n - X_{t-1}), \quad t \geq 1, \qquad (12)$$

όπου $\beta = \pi(1-\rho)$ και $a = \beta + \rho$. Όλες οι τ.μ. a_ϕ, β_ϕ καθώς και οι αντίστοιχες λεπτύνσεις εκτελούνται ανεξάρτητα των υπολοίπων ενώ όλες οι τ.μ. a_ϕ, β_ϕ καθώς και οι αντίστοιχες λεπτύνσεις τη χρονική στιγμή t είναι ανεξάρτητες των X_s, για $s < t$.

Οι πιθανότητες μετάβασης $p_{k|l} = P(X_t = k \mid X_{t-1} = l)$, $k,l \in \{0,...,n\}$ για το μοντέλο BBAR(1), δίνονται από την ακόλουθη σχέση (Weiß and Kim (2014)):

$$p_{k|l} = \sum_{m=\max(0,k+l-n)}^{\min(k,l)} \binom{l}{m} \binom{n-l}{k-m} \frac{B\left(m + \frac{1-\phi}{\phi} a, l - m + \frac{1-\phi}{\phi}(1-a) \right)}{B\left(\frac{1-\phi}{\phi} a, \frac{1-\phi}{\phi}(1-a) \right)}$$

$$\times \frac{B\left(k - m + \frac{1-\phi}{\phi} \beta, n - l - k + m + \frac{1-\phi}{\phi}(1-\beta) \right)}{B\left(\frac{1-\phi}{\phi} \beta, \frac{1-\phi}{\phi}(1-\beta) \right)} \qquad (13)$$

Θα πρέπει να σημειωθεί ότι δεν μπορούμε να εξάγουμε κλειστό τύπο για τη στάσιμη περιθώρια κατανομή (όπως π.χ. στην περίπτωση του μοντέλου BAR(1), όπου η στάσιμη κατανομή είναι η $B(n, \pi)$). Για δεδομένες τιμές των π, ρ και ϕ, η κατανομή αυτή μπορεί να προσδιοριστεί αριθμητικά, χρησιμοποιώντας τις πιθανότητες μετάβασης.

Ειδικότερα, αν η αρχική παρατήρηση X_0 προέρχεται από τη στάσιμη περιθώρια κατανομή, τότε όλη η διαδικασία X_t, $t \geq 1$, είναι στάσιμη και άρα οι πιθανότητες $p_k = P(X_t = k)$ προκύπτουν ως η λύση του συστήματος εξισώσεων $\mathbf{Pp} = \mathbf{p}$, όπου $\mathbf{P}$ είναι ο $(n+1)\times(n+1)$ πίνακας των πιθανοτήτων μετάβασης $p_{k|l}$ και $\mathbf{p} = (p_0,..., p_n)$ είναι το διάνυσμα των περιθωρίων πιθανοτήτων.

Για τη δεσμευμένη μέση τιμή $E(X_t \mid X_{t-1})$ και τη δεσμευμένη διασπορά $V(X_t \mid X_{t-1})$ του μοντέλου BBAR(1), διαπιστώνουμε από τις παρακάτω σχέσεις ότι η πρώτη έχει την ιδιότητα της γραμμικότητας ως προς τη X_{t-1} (είναι η ίδια με αντίστοιχη για το μοντέλο BAR(1)), ενώ η $V(X_t \mid X_{t-1})$ δεν είναι γραμμική ως προς τη X_{t-1}:

$$E(X_t \mid X_{t-1}) = aX_{t-1} + \beta(n - X_{t-1}) - \rho X_{t-1} + n\beta, \tag{14}$$

$$\begin{aligned} V(X_t \mid X_{t-1}) = \phi(a(1-a) + \beta(1-\beta))X_{t-1}^2 + n\beta(1-\beta)(1+\phi(n-1)) \\ + X_{t-1}(\rho(1-\rho)(1-2\pi)(1-\phi) - 2n\beta(1-\beta)\phi). \end{aligned} \tag{15}$$

Η k-τάξης αυτοσυσχέτιση ισούται με $\rho(k) = \rho^k$.

Από τις ιδιότητες της δεσμευμένης μέσης τιμής και δεσμευμένης διασποράς μπορούμε να εξάγουμε τύπους για τη μέση τιμή και διασπορά της στάσιμης κατανομής του μοντέλου BBAR(1). Για τη μέση τιμή μ έχουμε ότι

$$\mu = E(X_t) = E(E(X_t \mid X_{t-1})) = \rho \cdot \mu + n\beta, \tag{16}$$

απ' όπου επιλύοντας ως προς μ, λαμβάνουμε $\mu = n \cdot \beta/(1-\rho) = n\pi$. Ομοίως, για τη διασπορά σ^2, έχουμε ότι (δείτε π.χ. Weiß and Kim (2014))

$$\sigma^2 = n\pi(1-\pi) \cdot \frac{(1-\phi)(1+\rho) + n\phi(1-2\pi(1-\pi)(1-\rho))}{(1-\phi)(1+\rho) + \phi(1-2\pi(1-\pi)(1-\rho))}. \tag{17}$$

Από την εξίσωση (17) παρατηρούμε ότι η ποσότητα $n\pi(1-\pi)$ είναι η διωνυμική διασπορά και το πηλίκο που απομένει, καθορίζει την απόκλιση μεταξύ πραγματικής μεταβλητότητας και μεταβλητότητας του διωνυμικού μοντέλου.

3.3 Μέθοδοι Εκτίμησης Παραμέτρων του Μοντέλου BBAR(1)

Έστω $X_1, ..., X_T$ ένα τμήμα που προέρχεται από τη στάσιμη διαδικασία του μοντέλου BBAR(1). Για την εκτίμηση των παραμέτρων π, ρ και ϕ μπορούμε να χρησιμοποιήσουμε τη μέθοδο της Δεσμευμένης Μέγιστης Πιθανοφάνειας (*Conditional Maximum Likelihood*, CML). Ειδικότερα, για δεδομένη τιμή x_1, ο λογάριθμος της δεσμευμένης πιθανοφάνειας δίδεται από τη σχέση

$$l(\pi, \rho, \phi) = \log P(X_2 = x_2, ..., X_T = x_T \mid X_1 = x_1) = \sum_{i=1}^{T-1} \log p_{x_{i+1}|x_i}, \tag{18}$$

όπου οι δεσμευμένες πιθανότητες $p_{x_{i+1}|x_i}$, $i = 1, 2, ..., T-1$, υπολογίζονται μέσω της (13). Η εύρεση των ΕΜΠ για τις παραμέτρους π, ρ και ϕ γίνεται με χρήση αριθμητικών μεθόδων ενώ με αντίστοιχο τρόπο, όπως και στην περίπτωση του μοντέλου BAR(1), μπορούμε να εκτιμήσουμε τα τυπικά σφάλματα των εκτιμητών.

4. ΤΟ ΜΟΝΤΕΛΟ BINARCH(1)

4.1 Εισαγωγή

Σε αυτή την Ενότητα θα παρουσιάσουμε μία ακόμη επέκταση του μοντέλου BAR(1), το μοντέλο BINARCH(1), το οποίο είναι ένα απλό και κατάλληλο υπόδειγμα για τη μοντελοποίηση διακριτών δεδομένων, με πεπερασμένο εύρος τιμών. Αν και αποτελεί διακριτό ανάλογο του συνήθους μοντέλου ARCH (Engle (1982)), με μορφή παρόμοια των μοντέλων ARMA, εντούτοις αποτελεί ειδική περίπτωση του μοντέλου *density-dependent* BAR(1) (Weiß and Pollett (2014)), το οποίο με τη σειρά του αποτελεί γενίκευση του μοντέλου BAR(1), μιας και επιτρέπει στα a, β (δείτε εξίσωση (3)) να εξαρτώνται από το X_t. Επιπλέον, το μοντέλο BINARCH(1) μπορεί να χρησιμοποιηθεί για τη μοντελοποίηση διακριτών χρονοσειρών με πεπερασμένο εύρος τιμών, τα οποία παρουσιάζουν είτε υπέρ- είτε υπό-μεταβλητότητα (over- and under-dispersed data). Στη συνέχεια παρουσιάζουμε τις βασικές ιδιότητές του καθώς και μια μέθοδο εκτίμησης των παραμέτρων του.

4.2 Βασικές Ιδιότητες του Μοντέλου BINARCH(1)

Σύμφωνα με τους Ristic et al. (2015) (δείτε επίσης Weiß and Pollett (2014)) το μοντέλο BINARCH(1) είναι ένα μοντέλο διακριτών χρονοσειρών X_t, $t \geq 1$, με

$$X_t \mid F_{t-1} \sim B(n, \alpha_t), \tag{19}$$

όπου $n \in \mathbb{N}_0$, F_{t-1} το σ-πεδίο που παράγεται από τις τ.μ. $X_{t-1}, X_{t-2}, \ldots$ και παραμέτρους $\alpha_t = a_0 + (a_1/n)X_{t-1}$, όπου $a_0 > 0$ και $a_1 \geq 0$.

Από την (19), έπεται ότι

$$P(X_t \mid F_{t-1}) = \binom{n}{x} \alpha_t^x (1-\alpha_t)^{n-x}, \quad x = 0,1,\ldots,n, \tag{20}$$

και άρα η δεσμευμένη μέση τιμή και η δεσμευμένη διασπορά είναι ίσες με

$$E(X_t \mid F_{t-1}) = n\alpha_t, \quad V(X_t \mid F_{t-1}) = n\alpha_t\,(1-\alpha_t). \tag{21}$$

Άμεσα, έπεται ότι $V(X_t \mid F_{t-1}) < E(X_t \mid F_{t-1})$. Επίσης, από τις ιδιότητες της δεσμευμένης μέσης τιμής και δεσμευμένης διασποράς, έπεται η μέση τιμή και η διασπορά της στάσιμης κατανομής του μοντέλου BINARCH(1) (για περισσότερες λεπτομέρειες δείτε Weiß and Pollett (2014), Ristic et al. (2015)):

$$\mu \equiv E(X_t) = na_0/(1-a_1), \quad \sigma^2 \equiv V(X_t) = \frac{\mu(1-\mu/n)}{1-(1-1/n)a_1^2}. \tag{22}$$

Η k-τάξης αυτοσυσχέτιση ισούται με $\rho(k) = a_1^k$, $k \geq 1$.

4.3 Μέθοδοι Εκτίμησης Παραμέτρων του Μοντέλου BINARCH(1)

Έστω $X_1, \ldots, X_T$, ένα τμήμα που προέρχεται από τη στάσιμη διαδικασία του BINARCH(1) μοντέλου. Σύμφωνα με της Ristic et al. (2015) μπορούμε να εκτιμήσουμε τις παραμέτρους του μοντέλου χρησιμοποιώντας τη μέθοδο CML. Η συνάρτηση του λογαρίθμου της πιθανοφάνειας δίνεται από τη σχέση

$$l(a_0, a_1) = \sum_{t=2}^{T} \left\{ \log \binom{n}{x_t} + x_t \log \alpha_t + (n - X_t) \log(1 - \alpha_t) \right\}. \qquad (23)$$

Χρησιμοποιώντας κατάλληλες αριθμητικές μεθόδους, μεγιστοποιούμε την $l(a_0, a_1)$ και υπολογίζουμε τους ΕΜΠ των a_0, a_1.

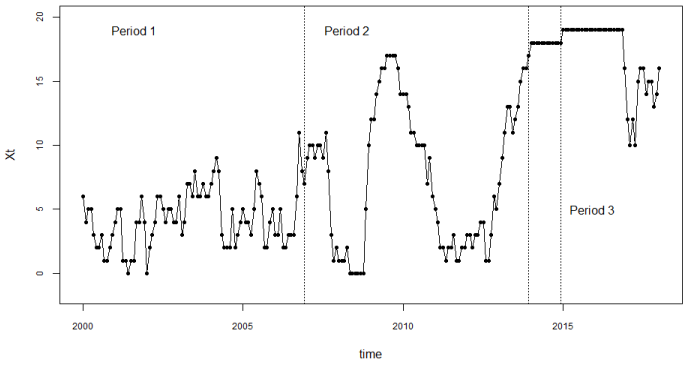
5. ΕΦΑΡΜΟΓΗ ΣΕ ΠΡΑΓΜΑΤΙΚΑ ΔΕΔΟΜΕΝΑ

5.1 Εισαγωγή

Στα κράτη μέλη της Ευρωπαϊκής Ένωσης (ΕΕ), οι μεταβολές στις τιμές του καταναλωτή (ειδικότερα του πληθωρισμού) μετριούνται με όρους του Εναρμονισμένου Δείκτη Τιμών Καταναλωτή (ΕνΔΤΚ). Η επιδίωξη πληθωρισμού της τάξεως του 2% ή μικρότερου ήταν κεντρικό στοιχείο της νομισματικής αρχιτεκτονικής της Ευρωζώνης (ΕΖ). Εξ' ορισμού, σταθερότητα τιμών στη ζώνη του ευρώ προσδιορίζεται ποσοτικά με βάση τον ΕνΔΤΚ και υφίσταται όταν ο πληθωρισμός είναι εγγύτερα στο 0% παρά στο 2% (δείτε π.χ. https://www.ecb.europa.eu/mopo/strategy/pricestab/html/index.en.html).

Στη διάθεσή μας έχουμε μηνιαία δεδομένα πληθωρισμού των χωρών-μελών της ΕΖ ξεκινώντας από τον Ιανουάριο 2000 μέχρι τον Ιανουάριο 2018. Στόχος είναι να αναλύσουμε πόσες από αυτές επέδειξαν σταθερότητα στις τιμές των αγαθών/υπηρεσιών με βάση τον ΕνΔΤΚ. Για το σκοπό αυτό χρησιμοποιούμε τους μηνιαίους ρυθμούς πληθωρισμού που δημοσιεύονται από την Ευρωπαϊκή Στατιστική Υπηρεσία (δείτε http://appsso.eurostat.ec.europa.eu/nui/submitViewTableAction.do). Στην Εικόνα 1 δίνεται η γραφική παράσταση των δεδομένων για την περίοδο 2000-2018. Στον κάθετο άξονα δίνεται το πλήθος των χωρών-μελών της ΕΖ με μηνιαίο πληθωρισμό μικρότερο από 2%.

Εικόνα 1. *Γραφική παράσταση των δεδομένων για τη χρονική περίοδο 2000-2018*



Μέχρι τις 31/12/2013, οι χώρες-μέλη της ΕΖ ήταν 17, στη συνέχεια έγιναν 18 (προστέθηκε η Λετονία την 1/1/2014) ενώ πλέον είναι 19 αφού από 1/1/2015 προστέθηκε και η Λιθουανία. Σημειώνεται επίσης ότι όπως και στην εργασία των Weiß and Kim (2014), έτσι και εδώ χρησιμοποιήθηκαν δεδομένα πληθωρισμού από τις χώρες-μέλη της ΕΖ κατόπιν διακριτοποίησης, προκειμένου να τονιστεί η

ικανότητα των μοντέλων BBAR(1), BINARCH(1) στη μοντελοποίηση δεδομένων διακριτών χρονοσειρών με πεπερασμένο εύρος τιμών, στα οποία παρουσιάζεται ετερογένεια ή/και υπερ-μεταβλητότητα.

Στη συνέχεια, θεωρούμε 3 συγκεκριμένες χρονικές περιόδους, οι οποίες αναλύονται ξεχωριστά. Η $1^{η}$ περίοδος είναι από 1/1/2000 έως 31/12/2006. Σύμφωνα με τους Weiß and Kim (2014), οι οποίοι ανέλυσαν τα δεδομένα της $1^{η}$ περιόδου, αυτή θεωρείται αρκετά σταθερή αφού δεν υπάρχουν συμβάντα τα οποία να επηρεάζουν τις οικονομίες των χωρών-μελών της ΕΖ, εκτός από την ετερογένεια την οποία αυτές παρουσιάζουν. Παρά ταύτα, η οικονομική ανάπτυξη στην επταετία 2000-06 (σε όρους ΑΕΠ) ήταν ίσως η υψηλότερη της τελευταίας εικοσαετίας.

Η $2^{η}$ περίοδος είναι από 1/1/2007 έως και 31/12/2013, στην οποία κυριαρχεί η διεθνής χρηματοοικονομική κρίση του 2007-08. Η ύφεση που ακολούθησε περιόρισε τη συνολική ζήτηση και τις τιμές στην ΕΕ. Αυτό είχε ως αποτέλεσμα λίγες (ή και καμία) από τις 17 χώρες-μέλη της ΕΖ να έχουν σταθερά χαμηλές τιμές σε ότι αφορά το μηνιαίο πληθωρισμό. Παρά ταύτα, λόγω της αύξησης των τιμών σε τρόφιμα και ενέργεια (τέλος του 2010), η εικόνα αναστράφηκε και παρατηρήθηκε καθοδική τάση σε ότι αφορά τον αριθμό των χωρών-μελών της Ευρωζώνης με μηνιαίο πληθωρισμό κάτω από το 2%, στο τέλος του 2011.

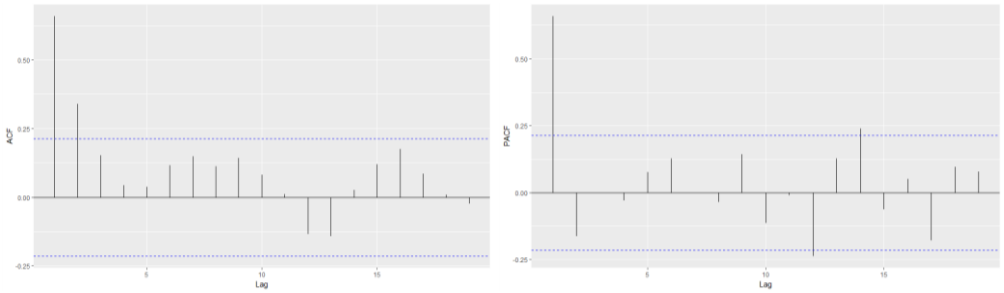
Τέλος, η $3^{η}$ περίοδος είναι από 1/1/2015 έως 1/1/2018, κατά την οποία είναι εμφανές ότι όλες οι χώρες-μέλη της ΕΖ έχουν σταθερά μηνιαίο πληθωρισμό κάτω από το 2% μέχρι και τα μέσα του 2016. Εκ των πραγμάτων, το πρόγραμμα ποσοτικής χαλάρωσης που εφαρμόστηκε από την Ευρωπαϊκή Κεντρική Τράπεζα (ΕΚΤ), τόνωσε τη συνολική ζήτηση και επέδρασε θετικά στις τιμές. Κατά συνέπεια, η σταθερότητα που παρατηρείται (μη υπέρβαση του κατωφλίου του 2% στον πληθωρισμό) οφείλεται στο μείγμα πολιτικής που ακολουθήθηκε από την ΕΖ και την ΕΚΤ.

Στόχος λοιπόν είναι να δούμε την προσαρμογή των μοντέλων BAR(1), BBAR(1) και BINARCH(1) στα δεδομένα των 3 αυτών περιόδων. Σημειώνεται πως από την ανάλυση απουσιάζει το έτος 2014 αφού το πλήθος των διαθέσιμων παρατηρήσεων είναι μόλις 12 και με όλες τις τιμές ίσες με 18, όσες και οι χώρες-μέλη της ΕΖ για αυτή την περίοδο.

5.2 Ανάλυση χρονοσειράς για την $1^{η}$ χρονική περίοδο

Τα διαθέσιμα δεδομένα αυτής της περιόδου αποτελούνται από 84 μηνιαίες παρατηρήσεις από τα οποία προκύπτει ότι $\bar{x} = 4.2738$, $s^2 = 4.9241$ και $I_d = 1.5391 > 1$, το οποίο αποτελεί ένδειξη υπερ-μεταβλητότητας σε σύγκριση με τη Διωνυμική κατανομή. Στο επόμενο σχήμα δίνουμε τις γραφικές παραστάσεις των συναρτήσεων της αυτοσυσχέτισης (ACF) και μερικής αυτοσυσχέτισης (Partial ACF, δείτε επίσης Weiß and Kim (2014)). Αν και από τη μορφή των αυτοσυσχετίσεων, διαφαίνεται ότι θα μπορούσε να χρησιμοποιηθεί και μοντέλο υψηλότερης τάξης (π.χ. $2^{ης}$), στη συνέχεια θα προχωρήσουμε με την προσαρμογή των μοντέλων $1^{ης}$ που έχουμε ήδη παρουσιάσει, συνεχίζοντας τη μελέτη των Weiß and Kim (2014).

Εικόνα 2. *Γραφικές παραστάσεις των ACF, Partial ACF, 1η Περίοδος*



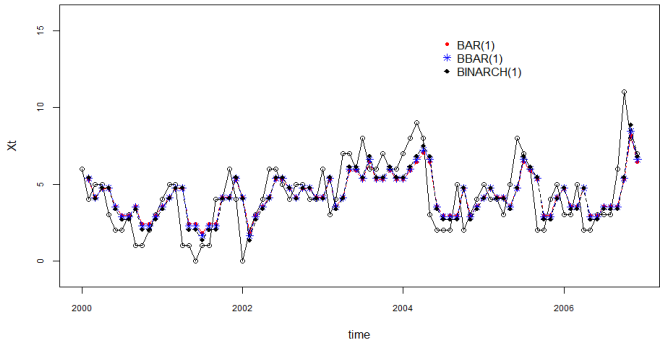
Στον Πίνακα 1 δίνονται οι εκτιμήσεις των παραμέτρων κάθε μοντέλου (χρησιμοποιήθηκε η μέθοδος της μέγιστης πιθανοφάνειας), οι εκτιμήσεις των τυπικών σφαλμάτων (τιμή στην παρένθεση) καθώς και οι τιμές των κριτηρίων AIC και BIC. Στα ήδη υπάρχοντα αποτελέσματα των Weiß and Kim (2014), προστέθηκε η προσαρμογή του μοντέλου BINARCH(1).

Πίνακας 1. *Σύγκριση Μοντέλων για την 1η Χρονική περίοδο (2000-2006)*

Μοντέλο	$\hat{\pi}_{ML}$	$\hat{\rho}_{ML}$	$\hat{\varphi}_{ML}$	$\hat{\alpha}_{0_{ML}}$	$\hat{a}_{1_{ML}}$	**AIC**	**BIC**
BAR(1)	0.255 (0.022)	0.578 (0.058)	-	-	-	327.3	332.1
BBAR(1)	0.255 (0.028)	0.621 (0.063)	0.037 (0.028)	-	-	322.5	329.8
BINARCH(1)	-	-	-	0.0792 (0.0212)	0.684 (0.084)	**318.4**	**323.3**

Από τις τιμές των AIC, BIC διαπιστώνουμε ότι το μοντέλο BINARCH(1) είναι αυτό που προσαρμόζεται καλύτερα στα δεδομένα της $1^{ης}$ περιόδου. Οι εκτιμήσεις των π, ρ στα μοντέλα BAR(1) και BBAR(1) είναι πάρα πολύ κοντά ενώ η αυτοσυσχέτιση (παράμετροι ρ και a_1) εκτιμάται ότι είναι σχεδόν η ίδια και στα 3 μοντέλα. Αυτό μπορούμε να το εξακριβώσουμε και από την επόμενη γραφική παράσταση που δίνει τις προβλέψεις των υποδειγμάτων για τη χρονική περίοδο 2000-2006. Με άσπρο σύμβολο, είναι οι τιμές X_t για την $1^η$ περίοδο.

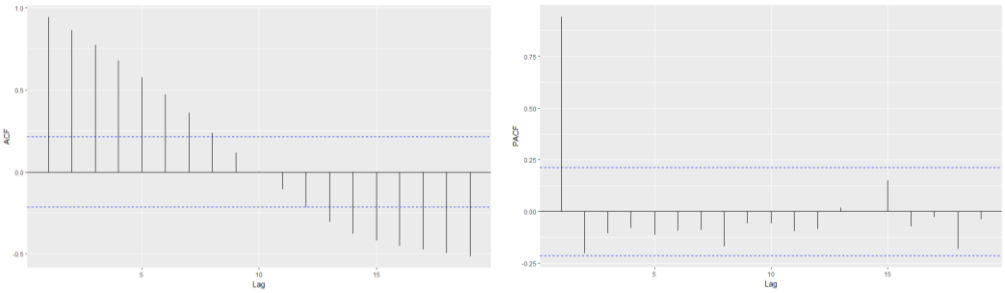
Εικόνα 3. *Προβλέψεις μοντέλων για την 1η χρονική περίοδο 2000-2006*



5.3 Ανάλυση χρονοσειράς για τη $2^{η}$ χρονική περίοδο

Στη $2^{η}$ χρονική περίοδο, από τις 84 διαθέσιμες μηνιαίες παρατηρήσεις προκύπτει ότι $\bar{x} = 7.5952$, $s^2 = 31.28$ και $I_d = 7.4443 > 1$, το οποίο αποτελεί ένδειξη υπερ-μεταβλητότητας σε σύγκριση με τη Διωνυμική κατανομή. Στην Εικόνα 4 δίνονται οι γραφικές παραστάσεις των ACF και PACF.

Εικόνα 4. *Γραφικές παραστάσεις των ACF, Partial ACF, $2^{η}$ Περίοδος*



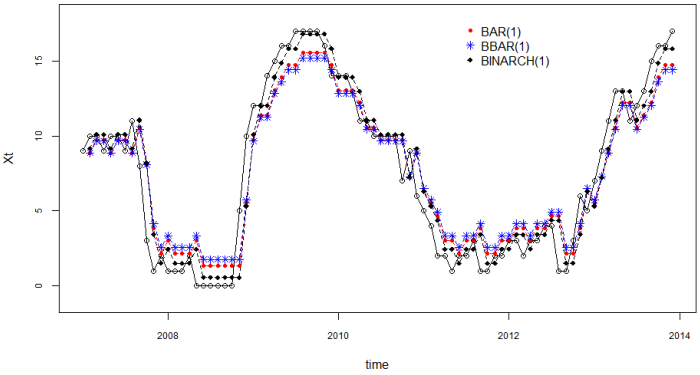
Από τη μορφή της ACF, παρατηρούμε ότι θα έπρεπε να χρησιμοποιηθεί κάποιο από τα προτεινόμενα μοντέλα με τάξη μεγαλύτερη του 1 ή ακόμη και διαφορετικής μορφής π.χ. Binomial INGARCH(*p*, *q*) (Ristic et al. (2015)) ή BARMA(*p*, *q*) (Al-Osh and Alzaid (1991)). Λόγω του ότι στην προηγούμενη χρονική περίοδο, χρησιμοποιήθηκαν μοντέλα $1^{ης}$ τάξης και επειδή η συγκεκριμένη εργασία επεκτείνει την εργασία των Weiss and Kim (2014) στις αμέσως επόμενες χρονικές περιόδους, θα συνεχίσουμε με την προσαρμογή των μοντέλων BAR(1), BBAR(1) και BINARCH(1). Η προσαρμογή μοντέλων μεγαλύτερης τάξης είναι ήδη σε εξέλιξη.

Ακολουθώντας την ίδια μεθοδολογία με πριν, προσαρμόσαμε τα τρία υποδείγματα στα δεδομένα και εκτιμήσαμε τις παραμέτρους. Τα αποτελέσματα δίνονται στον Πίνακα 2, απ' όπου παρατηρούμε ότι το BINARCH(1) μοντέλο είναι το καταλληλότερο· οι τιμές των κριτηρίων AIC και BIC είναι μικρότερες σε σχέση με τα άλλα δύο υποδείγματα. Επίσης, οι εκτιμήσεις των π, ρ στα μοντέλα BAR(1) και BBAR(1) είναι πάρα πολύ κοντά (κυρίως για το π) ενώ η αυτοσυσχέτιση (παράμετροι ρ και a_1) εκτιμάται ότι είναι σχεδόν η ίδια για τα μοντέλα BAR(1), BBAR(1) αλλά φαίνεται να διαφέρει στο μοντέλο BINARCH(1). Η επόμενη γραφική παράσταση δίνει τις προβλέψεις των υποδειγμάτων για τη χρονική περίοδο 2007-2013.

Πίνακας 2. *Σύγκριση Μοντέλων για την 2η Χρονική περίοδο (2007-2013)*

Μοντέλο	$\hat{\pi}_{ML}$	$\hat{\rho}_{ML}$	$\hat{\varphi}_{ML}$	$\hat{\alpha}_{0_{ML}}$	$\hat{a}_{1_{ML}}$	**AIC**	**BIC**
BAR(1)	0.481 (0.043)	0.837 (0.018)	-	-	-	381.6	386.4
BBAR(1)	0.493 (0.061)	0.789 (0.036)	0.206 (0.053)	-	-	324.8	329.6
BINARCH(1)	-	-	-	0.032 (0.021)	0.955 (0.084)	**297.1**	**301.9**

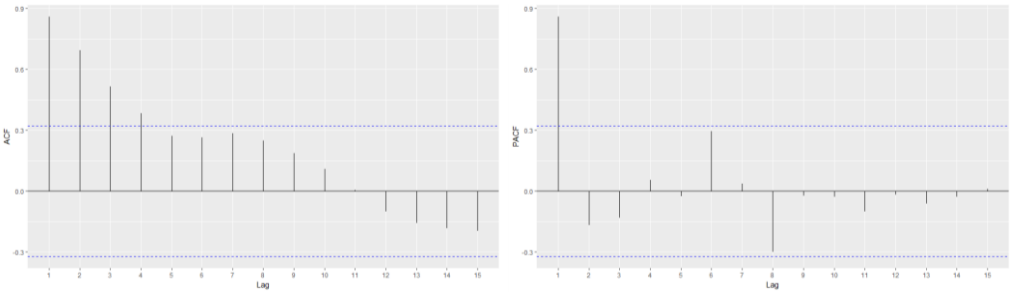
Εικόνα 5. *Προβλέψεις μοντέλων για τη 2η χρονική περίοδο 2007-2013*



5.4 Ανάλυση χρονοσειράς για την 3η χρονική περίοδο

Στην 3η χρονική περίοδο, από τις 37 διαθέσιμες μηνιαίες παρατηρήσεις προκύπτει ότι $\bar{x} = 17.0541$, $s^2 = 8.0526$ και $I_d = 4.6103 > 1$, το οποίο αποτελεί ένδειξη υπερ-μεταβλητότητας σε σύγκριση με τη Διωνυμική κατανομή. Στο επόμενο σχήμα δίνουμε τις γραφικές παραστάσεις των ACF και PACF.

Εικόνα 6. *Γραφικές παραστάσεις των ACF, Partial ACF, 3η Περίοδος*



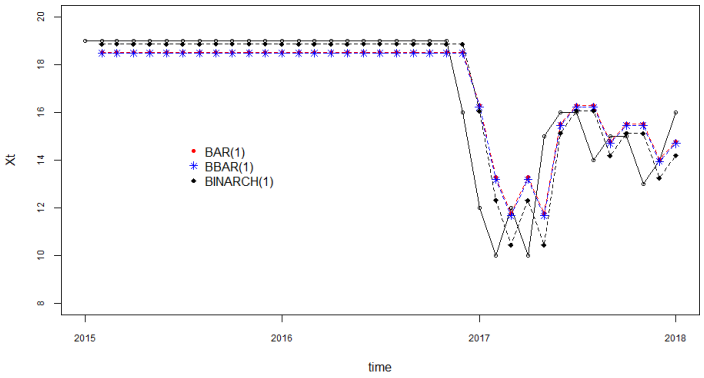
Όπως και για τα δεδομένα της 2ης περιόδου, από τη μορφή της ACF παρατηρούμε ότι θα έπρεπε να χρησιμοποιηθεί κάποιο από τα προτεινόμενα μοντέλα με τάξη μεγαλύτερη του 1 (πιθανά 2ης ή 3ης). Για τους ίδιους λόγους, παραθέτουμε στην Πίνακα 3 την προσαρμογή των μοντέλων BAR(1), BBAR(1) και BINARCH(1) στα δεδομένα της 3ης περιόδου ενώ είναι ήδη σε εξέλιξη σχετική μελέτη που θα αφορά μοντέλα μεγαλύτερης τάξης.

Πίνακας 3. *Σύγκριση Μοντέλων για την 3η Χρονική περίοδο (2015-2018)*

Μοντέλο	$\hat{\pi}_{ML}$	$\hat{\rho}_{ML}$	$\hat{\varphi}_{ML}$	$\hat{\alpha}_{0_{ML}}$	$\hat{a}_{1_{ML}}$	**AIC**	**BIC**
BAR(1)	0.900 (0.027)	0.749 (0.054)	-	-	-	114.0	117.2
BBAR(1)	0.887 (0.052)	0.757 (0.078)	0.165 (0.076)	-	-	87.3	92.17
BINARCH(1)	-	-	-	0.055 (0.021)	0.936 (0.084)	**84.06**	**87.2**

Από τον Πίνακα 3 παρατηρούμε ότι το BINARCH(1) μοντέλο είναι το καταλληλότερο αφού οι τιμές των κριτηρίων AIC και BIC είναι μικρότερες σε σχέση με τα άλλα δύο υποδείγματα. Οι εκτιμήσεις των π, ρ στα μοντέλα BAR(1) και BBAR(1) είναι πάρα πολύ κοντά ενώ η αυτοσυσχέτιση (παράμετροι ρ και a_1) εκτιμάται ότι είναι σχεδόν η ίδια για τα μοντέλα BAR(1), BBAR(1) αλλά φαίνεται να διαφέρει στο μοντέλο BINARCH(1). Κάτι αντίστοιχο είχε παρατηρηθεί και στα δεδομένα για τη $2^{η}$ χρονική περίοδο. Η επόμενη γραφική παράσταση δίνει τις προβλέψεις των υποδειγμάτων για την $3^{η}$ χρονική περίοδο.

Εικόνα 7. *Προβλέψεις μοντέλων για την 3η χρονική περίοδο 2015-2018*



6. ΣΥΜΠΕΡΑΣΜΑΤΑ

Στην παρούσα εργασία, παρουσιάστηκαν τα 3 πιο συχνά χρησιμοποιούμενα μοντέλα $1^{ης}$ τάξης για διακριτές χρονοσειρές με πεπερασμένο εύρος τιμών. Για κάθε μοντέλο, δόθηκαν οι βασικές του ιδιότητες και ο τρόπος εκτίμησης των παραμέτρων του. Στη συνέχεια, τα μοντέλα προσαρμόστηκαν σε δεδομένα που αφορούν το πλήθος των χωρών της ΕΖ με μηνιαίο πληθωρισμό κάτω από το 2%. Επιλέχθηκαν 3 διαφορετικές χρονικές περίοδοι στο διάστημα 2000-2018, λόγω του διαφορετικού αριθμού χωρών-μελών της ΕΖ.

Από την ανάλυση προέκυψε ότι (i) το μοντέλο BINARCH(1) έχει την καλύτερη προσαρμογή και στις 3 περιόδους, έναντι των μοντέλων BAR(1) και BBAR(1) και (ii) αν και από προηγούμενες μελέτες, φαίνεται πως στην περίοδο 2000-2006 είναι λογικό να χρησιμοποιηθούν μοντέλα $1^{ης}$ τάξης, εντούτοις, στις πιο πρόσφατες περιόδους κάτι τέτοιο δε δικαιολογείται από τη μορφή των ACF και PACF. Κατά συνέπεια, ως θέμα για μελλοντική έρευνα, πρέπει να εξεταστεί η προσαρμογή μοντέλων μεγαλύτερης τάξης ή/και διαφορετικής μορφής.

ABSTRACT

In this work, we present the most widely used autoregressive models for integer-valued time series with finite range. These are models are the Binomial Autoregressive of 1^{st} order model (BAR(1)), the Beta-Binomial Autoregressive of 1^{st} order model (BBAR(1)) and the Binomial Autoregressive Conditionally Heteroscedastic model of 1^{st} order (BINARCH(1)). We present their basic properties

as well as methods for estimating their parameters. Also, we present a real application of these models on monthly inflation data for the Eurozone members. The aim is to find the best model for the monthly number of Eurozone members with a lower 2% monthly inflation. The AIC and BIC are used as model selection criteria.

Ευχαριστίες: Ευχαριστούμε τον/την ανώνυμο/η κριτή για τα σχόλιά του/της, τα οποία βελτίωσαν το περιεχόμενο της παρούσας εργασίας.

ΑΝΑΦΟΡΕΣ

Al-Osh M.A. and Alzaid A.A. (1991). Binomial autoregressive moving average models. *Communications in Statistics-Stochastic Models*, **7**, 261-282.

Engle R.F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, **50**, 987-1007.

Ferland R., Latour A. and Oraichi D. (2006). Integer-Valued GARCH Process, *Journal of Time Series Analysis*, **27**, 923-942.

Griffiths D.A. (1973). Maximum likelihood estimation for the beta-binomial distribution and an application to the household distribution of the total number of cases of a disease. *Biometrics*, **29**, 637-648.

McKenzie E. (1985). Some simple models for discrete variate time series. *Water Resources*, **21**, 645-650.

Ristic M., Weiß C.H. and Janjic A. (2015). A Binomial Integer-Valued ARCH Model, *The International Journal of Biostatistics*, **12**, 1-21.

Steuetel F.W. and van Harn K. (1979). Discrete analogues of self-decomposability and stability, *Annals of Probability*, **7**, 893-899.

Weiß C.H. (2008). Thinning operations for modeling time series of count – a survey. *Advances in Statistical Analysis*, **92**, 319-341.

Weiß C.H. (2009a). Montitoring correlated processes with binomial marginal, *Journal of Applied Statistics*, **36**, 399-414.

Weiß C.H. (2009b). A new class of autoregressive models for time series of binomial counts, *Communications in Statistics–Theory and Methods*, **38**, 447-460.

Weiß C.H. and Kim H.Y. (2013). Parameter estimation for binomial AR(1) models with application in finance and industry. *Statistical Papers*, **47**, 563-590.

Weiß C.H. and Kim H.Y. (2014). Diagnosing and modeling extra-binomial variation for time-dependent counts, *Applied Stochastic Models in Business and Industry*, **30**, 519-656.

Weiß C.H. and Pollett P.K. (2014). Binomial autoregressive processes with density-dependent thinning, *Journal of Time Series Analysis*, **35**, 115-132.

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 31ου Πανελληνίου Συνεδρίου Στατιστικής (2018), σελ.042-052

ΠΙΝΑΚΕΣ ΟΓΚΟΥ ΔΑΣΙΚΩΝ ΣΥΣΤΑΔΩΝ ΜΕ ΤΗΝ BITTERLICH ΔΕΙΓΜΑΤΟΛΗΨΙΑ

Α. Γεωργάκης, Ι. Μέλφος, Γ. Σταματέλλος
Τμήμα Δασολογίας και Φυσικού Περιβάλλοντος, ΑΠΘ
arisgeorg@for.auth.gr, melfosiv@for.auth.gr, stamatel@for.auth.gr

ΠΕΡΙΛΗΨΗ

Στη διαχείριση των δασικών οικοσυστημάτων ο όγκος των δένδρων είναι ένα από τα χαρακτηριστικά που παρουσιάζουν μεγαλύτερο επιστημονικό και πρακτικό ενδιαφέρον. Η αποτελεσματική εκτίμηση του όγκου των δασικών συστάδων βασίζεται κυρίως στο χαμηλό κόστος συλλογής των δεδομένων πεδίου με ικανοποιητικό επίπεδο ακρίβειας. Στοιχεία της εκτίμησης αποτελούν οι μέθοδοι δειγματοληψίας και η μέτρηση του όγκου των δένδρων. Σκοπός της εργασίας είναι η κατασκευή αποτελεσματικών μοντέλων πρόβλεψης του όγκου, τα οποία μπορούν να εφαρμοσθούν στη διαχείριση των δασικών οικοσυστημάτων. Τα στοιχεία πάρθηκαν από την ορεογραφική μονάδα "Βαθύ" του Πανεπιστημιακού Δάσους Περτουλίου. Εφαρμόστηκε η δειγματοληψία Bitterlich, στην οποία τα δένδρα επιλέγονται με πιθανότητα ανάλογη της κυκλικής τους επιφάνειας. Επιπροσθέτως, η συγκεκριμένη δειγματοληψία είναι αποτελεσματική κόστους. Δοκιμάστηκαν μοντέλα που έχουν προταθεί στη βιβλιογραφία και τελικά επιλέχτηκαν αυτά με την απλούστερη μαθηματική έκφραση, με κριτήριο την ικανότητα πρόβλεψης. Συζητούνται, βασικές πλευρές κατασκευής των μοντέλων σε σχέση με την εφαρμογή τους στην συστάδα και γενικότερα στη διαχείριση των δασικών οικοσυστημάτων στην παρούσα κοινωνικο-οικονομική συγκυρία.

Λέξεις κλειδιά: Διαχείριση δασικού οικοσυστήματος, μαζοπίνακες συστάδας, επιφάνεια Bitterlich, σφάλμα πρόβλεψης, διασταυρωμένη επικύρωση, Δάσος Περτουλίου

1. ΕΙΣΑΓΩΓΗ

Η απογραφή δασοβιομετρικών χαρακτηριστικών και ειδικά του ξυλώδη όγκου σε επίπεδο δασικής συστάδας αποτελεί το πρώτο βήμα αλλά και τη βασική πηγή πληροφοριών για την ορθολογική διαχείριση ενός δασικού οικοσυστήματος (Dau and Dawaki, 2015, Kershaw et al., 2016). Η ποσοτική εκτίμηση του ξυλώδη όγκου αποτελεί απαραίτητη πληροφορία τόσο για τον τρέχοντα χειρισμό όσο και για τον μελλοντικό σχεδιασμό ενός δασικού οικοσυστήματος. Για τον παραπάνω σκοπό γίνεται μεγάλη ερευνητική προσπάθεια για την εύρεση βέλτιστων μεθόδων απογραφής και δειγματοληψίας, καθώς και για την δημιουργία "εργαλείων" που θα συμβάλουν στην αποτελεσματική εκτίμηση των παραμέτρων του οικοσυστήματος

(χαμηλό κόστος και ικανοποιητική ακρίβεια) όπως αναφέρεται και από τους Van Laar και Akça (2007), καθώς και από τους Kershaw Jr et al. (2016). Τέτοιου είδους "εργαλεία" αποτελούν τα μοντέλα (εξισώσεις) και οι μαζοπίνακες (πίνακες όγκου) δασικών συστάδων.

Οι μαζοπίνακες συστάδων είναι πρακτικά και αποτελεσματικά εργαλεία για την εκτίμηση του υπέργειου έμφλοιου ξυλώδη όγκου του κορμού των δένδρων (ξυλαπόθεμα) στη χωρική μονάδα. Το βασικό τους πλεονέκτημα είναι το μειωμένο κόστος απογραφής των δασικών πόρων. Όταν διατίθενται περιορισμένοι οικονομικοί πόροι λήψης δεδομένων πεδίου, λόγω και των ιδιαίτερων κοινωνικοοικονομικών συνθηκών της χώρας μας, τότε ενδείκνυνται η εφαρμογή των μαζοπινάκων συστάδας και κυρίως των μαζοπινάκων με επιφάνειες Bitterlich. Προϋπόθεση της εφαρμογής τους είναι να γίνει αποδεκτή η ακρίβεια των εκτιμήσεων που παρέχουν.

Οι μαζοπίνακες συστάδας έχουν ως εξαρτημένη μεταβλητή τον όγκο (V), ενώ ως ανεξάρτητες μεταβλητές την κυκλική επιφάνεια (G), το μέσο ύψος (Dube et al., 2017) ή το κυριαρχούν ύψος (Shiver and Brister, 1992, Næsset and Tveite, 1999) ή το μέσο ύψος του Lorey (h_L) (Μάτης, 2004, Van Laar and Akça, 2007) ανά δειγματοληπτική επιφάνεια. Οι παραπάνω ανεξάρτητες μεταβλητές συσχετίζονται ικανοποιητικά με τον όγκο (Machado et al., 2008). Επίσης, στην βιβλιογραφία (Shiver and Brister, 1992, Næsset and Tveite, 1999) αναφέρεται ότι μπορεί να εισαχθεί και η ποιότητα τόπου (site index) ως επιπλέον ανεξάρτητη μεταβλητή εκτίμησης του όγκου. Η ποιότητα τόπου αναφέρεται στην παραγωγική δυνατότητα ενός δασικού οικοσυστήματος και εκφράζεται με την δυναμικότητα του σε παραγωγή ξυλώδη όγκου (Μάτης, 2004). Η εκτίμηση των ποιοτήτων τόπου μιας δασοσυστάδας είναι απόρροια του συνδυασμού βιοτικών (βλάστηση) και κυρίως αβιοτικών χαρακτηριστικών όπως το κλίμα, το ανάγλυφο και το έδαφος (Aertsen et al. 2010).

Οι μαζοπίνακες γενικά αναφέρονται σε μεμονωμένα δένδρα, ενώ οι μαζοπίνακες συστάδων αναφέρονται σε δειγματοληπτικές επιφάνειες (χωρικές μονάδες), δηλαδή σε ομάδες δένδρων. Μαζοπίνακες συστάδας κατασκευάστηκαν για διάφορα δασοπονικά είδη και σε διάφορες χώρες, όπως είναι η Αυστραλία από το 1945 (Van Laar and Akça, 2007), η Νορβηγία από το 1959 (Næsset, 1995, Næsset and Tveite, 1999), οι Ηνωμένες Πολιτείες Αμερικής (Shepperd et al., 1984) από το 1981, και η Ιταλία από το 1980, οι οποίοι εφαρμόζονται και στις μέρες μας (Corona et al., 2009). Στην Ελλάδα, έως τώρα, δεν έχουν εφαρμοστεί στην πράξη, καθώς η έρευνά τους περιορίζεται σε επίπεδο μίας μεταπτυχιακής διατριβής (Χατζηλαζάρου, 2002). Οι μαζοπίνακες συστάδων εφαρμόστηκαν τόσο σε φυσικά δασικά οικοσυστήματα (πρεμνοφυή και σπερμοφυή), όπως είναι της λεύκης (Χατζηλαζάρου, 2002), της ερυθρελάτης (Næsset, 1995) και της αραουκάριας (Machado et al., 2008) όσο και σε τεχνητά δάση όπως και σε φυτεία ευκαλύπτου (Shiver and Brister, 1992). Οι Shiver and Brister (1992) επισημαίνουν ότι η χρήση τους ενδείκνυται σε αμιγή δάση και σε ομήλικα δάση, όπως είναι τα τεχνητά, σε φυτείες ή σε πρεμνοφυή δάση.

Σκοπός της εργασίας είναι η κατασκευή μοντέλων παλινδρόμησης, όπως είναι οι μαζοπίνακες συστάδας, με τους οποίους θα υπολογίζεται αποτελεσματικά (μικρό

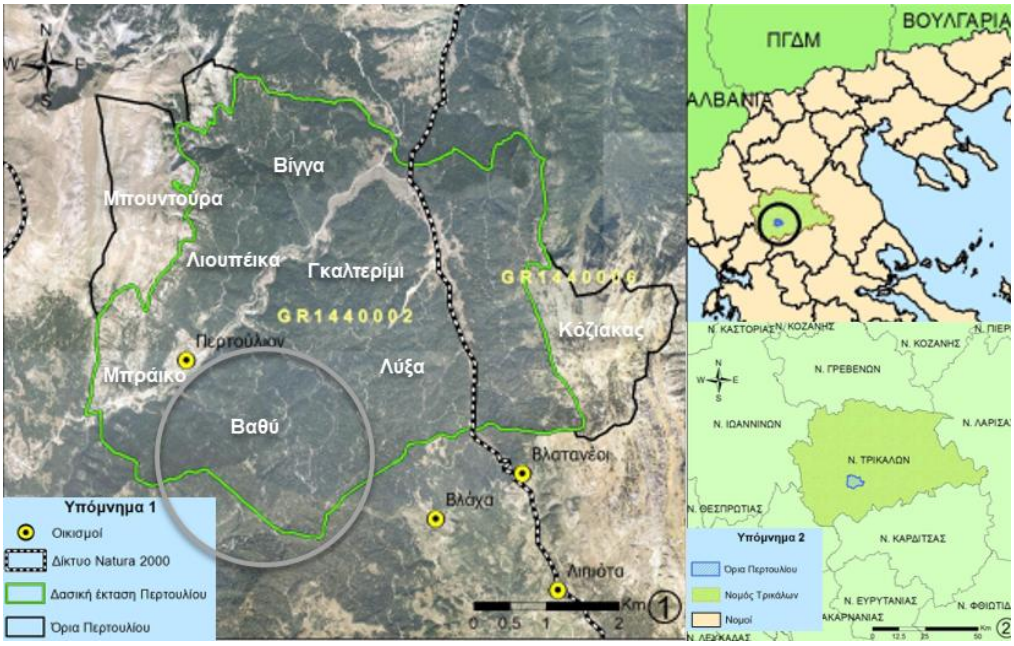
κόστος σε ορισμένη ακρίβεια) ο όγκος των δασικών συστάδων. Στην επόμενη ενότητα της εργασίας περιγράφονται τα δεδομένα και οι μέθοδοι, στη συνέχεια δίνονται τα αποτελέσματα της έρευνας και μια συζήτηση αυτών. Στην τελευταία ενότητα συνοψίζονται τα συμπεράσματα της έρευνας.

2. ΔΕΔΟΜΕΝΑ ΚΑΙ ΜΕΘΟΔΟΣ

2.1 Δεδομένα

Τα στοιχεία πάρθηκαν από την ορεογραφική μονάδα "Βαθύ" του Πανεπιστημιακού Δάσους Περτουλίου (ΠΔΠ). Το ΠΔΠ (Εικόνα 1) βρίσκεται στην κεντρική Πίνδο σε γεωγραφικό πλάτος 39°32'28"Β και γεωγραφικό μήκος 21°27'57"Α. Το υπερθαλάσσιο ύψος κυμαίνεται μεταξύ 1100 και 1700 m περίπου. Η έκταση του είναι 3.296,59 ha, εκ των οποίων τα 2.361,83 ha είναι δασικές εκτάσεις (Τ.Δ.Δ.Π.Δ., 2008). Το δάσος είναι ουσιαστικά αμιγές και καλύπτεται από υβριδογενή ελάτη (*Abies borisii-regis* Mattf), η οποία είναι αυτοφυής σε όλη την κεντρική Πίνδο και αποτελεί αυτόχθονο είδος στη χώρα μας.

Εικόνα 1. *Περιοχή μελέτης – Πανεπιστημιακό Δάσος Περτουλίου (διαμόρφωση χάρτη, Γεωργάκης, 2012)*



Στην παρούσα έρευνα εφαρμόστηκε η απλή τυχαία δειγματοληψία με επανάθεση, ελήφθησαν 196 δειγματοληπτικές επιφάνειες Bitterlich και μετρήθηκαν συνολικά 1225 δένδρα**.** Σε κάθε δειγματοληπτική επιφάνεια μετρήθηκαν η κυκλική επιφάνεια (*G*) με συντελεστή αρίθμησης 4 (BAF=4 m^2ha^{-1}) και το αριθμητικό μέσο ύψος των δένδρων σε κάθε σημείο δειγματοληψίας (ύψος Lorey - h_L σε m). Η ογκομέτρηση

του κάθε δένδρου έγινε με τον τύπο του Huber (Kershaw et al., 2016), ενώ όλες οι μετρήσεις έγιναν με το ρελασκόπιο Bitterlich (Bitterlich, 1984).

2.2 Μέθοδος

Εξετάστηκαν 12 μοντέλα (εξισώσεις/μοντέλα 1-12), τα οποία προτείνονται στη βιβλιογραφία (Corona et al., 2009, Machado et al., 2008). Η εκτίμηση των παραμέτρων των μοντέλων έγινε με τη μέθοδο των ελαχίστων τετραγώνων, μια μέθοδος που χρησιμοποιείται ευρέως για την κατασκευή μαζοπινάκων συστάδας (Næsset and Tveite, 1999). Στα μοντέλα εξαρτημένη μεταβλητή είναι ο όγκος (*V*) και ανεξάρτητες μεταβλητές η κυκλική επιφάνεια (*G*), το ύψος Lorey (h_L) και το $h_L^{-1/2}$. Εφαρμόστηκε η ανάλυση γραμμικής παλινδρόμησης και η μέθοδος εκτίμησης των ελαχίστων τετραγώνων. Στον έλεγχο των υπολοίπων (σφαλμάτων) για την επικύρωση των μοντέλων, εξετάστηκαν η ανεξαρτησία, η κανονικότητα και η ομοιογένεια της διακύμανσης τους ως προς τις προβλεπόμενες τιμές. Πρόβλημα ήταν η ύπαρξη ανομοιογένειας στη διακύμανση το οποίο αντιμετωπίστηκε με τη μέθοδο των σταθμισμένων ελαχίστων τετραγώνων, με βάρη τα οποία ελαχιστοποιούν την συνάρτηση πιθανοφάνειας και παρατίθενται στις εξισώσεις των μοντέλων: 3, 4, 7, 8, 11 και 12. Δεν υπήρξε πρόβλημα ακραίων τιμών και πολυσυγραμμικότητας στην εκτίμηση των μοντέλων παλινδρόμησης.

Στην αξιολόγηση της αποτελεσματικότητας των μοντέλων, από τα οποία θα κατασκευαστούν οι μαζοπίνακες, χρησιμοποιήθηκαν ως κριτήρια: ο συντελεστής προσδιορισμού για την πρόβλεψη (R^2_{pre}), όπως αναφέρεται από τους Ryan et al. (2013), η τετραγωνική ρίζα του μέσου τετραγωνικού σφάλματος για την πρόβλεψη (*RMSE*) και η μέση απόκλιση του σφάλματος (*MAE*), όπως αναφέρεται από τους Cort και Kenji (2005), τους Chai και Draxler (2014) και τον Brassington (2017). Στον υπολογισμό αυτών των κριτηρίων χρησιμοποιήθηκε επαναληπτικά με χωρισμό σε τμήματα η διασταυρωμένη επικύρωση (repeated k-fold cross validation), αξιοποιώντας το πακέτο "caret" της R (Kuhn, 2008).

Η επιλογή της διασταυρωμένης επικύρωσης έγινε με βάση την απλότητα και την καθολικότητα της χρήσης στην αξιολόγηση και επιλογή ενός μοντέλου (Arlot and Celisse, 2010). Ως μέθοδος διαχωρισμού των δεδομένων στη διασταυρωμένη επικύρωση επιλέχτηκε η k-fold (k=10). Αυτή αποφεύγει το μειονέκτημα της πιθανής μεγάλης διακύμανσης της μεθόδου leave-one-out (Kuhn and Johnson, 2013) ενώ το δείγμα εκτίμησης των μοντέλων παραμένει μεγάλο (~180). Τα δεδομένα χωρίστηκαν τυχαία σε δέκα επιμέρους τμήματα. Τα εννέα τμήματα χρησιμοποιήθηκαν ως δεδομένα εκπαίδευσης (Training Set) και το ένα τμήμα ως δεδομένα ελέγχου (Test Set). Τα μοντέλα κατασκευάστηκαν από ένα πλήθος 176 δειγματοληπτικών μονάδων και η πρόβλεψη τους από 20 δειγματοληπτικές μονάδες σε κάθε επανάληψη. Οι επαναλήψεις που έγιναν ήταν 50. Το υπολογιστικό μέρος της έρευνας, το οποίο περιλαμβάνει την εκτίμηση, την εγκυρότητα και την αξιολόγηση των μοντέλων, έγινε με τη γλώσσα R (James et al., 2013, Kuhn and Johnson, 2013). Στη συνέχεια ακολουθούν τα αποτελέσματα της έρευνας και γίνεται συζήτηση τους.

$$\hat{V} = \beta_o + \beta_1 G h_L \quad (1)$$

$$\hat{V} = \beta_1 G h_L \quad (2)$$

$$\hat{V} = \beta_o + \beta_1 G h_L \quad \text{με βάρος } W_i = (G h_L)^{(-1,3)} \quad (3)$$

$$\hat{V} = \beta_1 G h_L \quad \text{με βάρος } W_i = (G h_L)^{(-1,3)} \quad (4)$$

$$\hat{V} = \beta_o + \beta_1 G \sqrt{h_L} \quad (5)$$

$$\hat{V} = \beta_1 G \sqrt{h_L} \quad (6)$$

$$\hat{V} = \beta_o + \beta_1 G \sqrt{h_L} \quad \text{με βάρος } W_i = (G\sqrt{h_L})^{-1} \quad (7)$$

$$\hat{V} = \beta_1 G \sqrt{h_L} \quad \text{με βάρος } W_i = (G\sqrt{h_L})^{-1} \quad (8)$$

$$\hat{V} = \beta_o + \beta_1 G \quad (9)$$

$$\hat{V} = \beta_1 G \quad (10)$$

$$\hat{V} = \beta_o + \beta_1 G \quad \text{με βάρος } W_i = G^{-1} \quad (11)$$

$$\hat{V} = \beta_1 G \quad \text{με βάρος } W_i = G^{-1} \quad (12)$$

3. ΑΠΟΤΕΛΕΣΜΑΤΑ ΚΑΙ ΣΥΖΗΤΗΣΗ

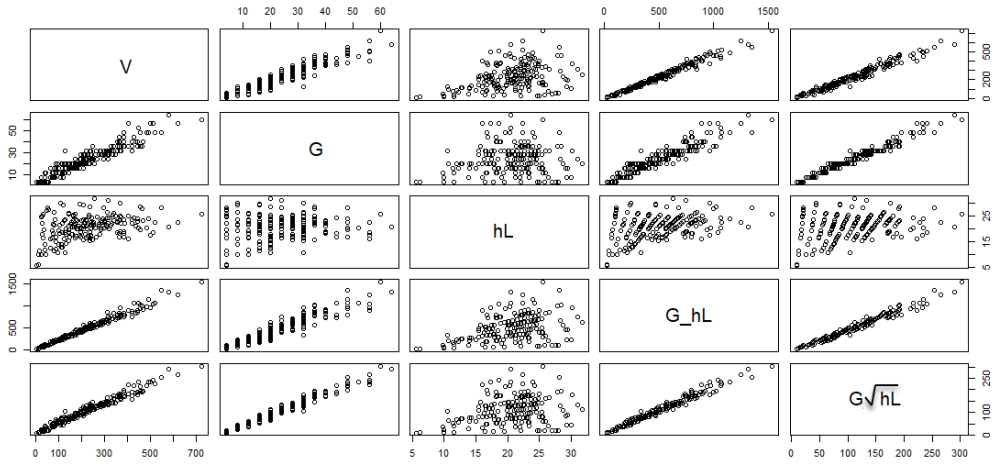
Στον Πίνακα 1 παρουσιάζεται η περιγραφική στατιστική των μεταβλητών ανά δειγματοληπτική επιφάνεια. Ο μέσος όρος των δένδρων ανά δειγματοληπτική επιφάνεια είναι 6,25. Οι κατανομές των μεταβλητών έχουν μικρή θετική ασσυμετρία. Στην Εικόνα 2 παρουσιάζεται ένας πίνακας με τα διαγράμματα διασποράς των μεταβλητών, όπου φαίνονται οι συσχετίσεις μεταξύ των μεταβλητών. Επίσης, παρατηρείται μια αύξηση της μεταβλητότητας του όγκου καθώς αυξάνονται οι τιμές των ανεξάρτητων μεταβλητών. Τα θηκογράμματα των μεταβλητών και των συνδυασμών τους, όπως αναλύονται στην παλινδρόμηση, παρουσιάζονται στην Εικόνα 3. Στον έλεγχο των σφαλμάτων διαπιστώθηκε ετεροσκεδαστικότητα, όπως φαίνεται στην Εικόνα 4 (αριστερό μέρος, μοντέλα 1^o και 5^o), η οποία διορθώθηκε με την εφαρμογή των σταθμισμένων ελαχίστων τετραγώνων (δεξιό μέρος, μοντέλα: 3^o και 7^o). Τα βάρη φαίνονται δίπλα στα αντίστοιχα μοντέλα: 3, 4, 7, 8, 11, και 12. Οι τιμές των κριτηρίων αξιολόγησης των μοντέλων παρουσιάζονται στον Πίνακα 2.

Πίνακας 1. *Περιγραφική στατιστική των δεδομένων ανά δειγματοληπτική επιφάνεια Bitterlich, όπου: z ο αριθμός των δένδρων, G η κυκλική επιφάνεια σε m^2, h_L το μέσο ύψος του Lorey σε m και V ο όγκος σε m^3*

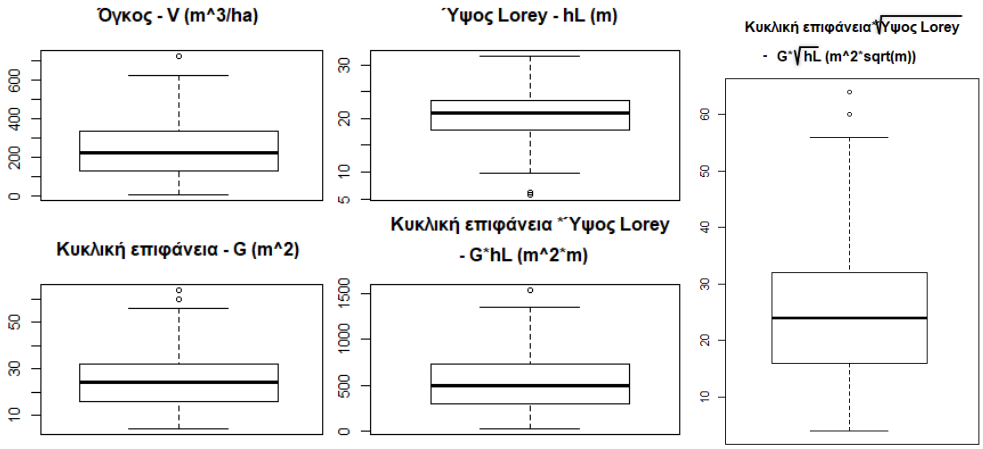
Εκτιμώμενες μεταβλητές	Μέσος όρος	Τυπικό σφάλμα μέσου	Ελάχιστο	Μέγιστο	Συντελεστής μεταβλητότητας (CV%)	Λοξότητα (skewness)
z	6,25	0,23	1,00	16,00	51,94	0,465
G	25,00	0,93	4,00	64,00	51,94	0,465
h_L	20,54	0,33	5,7	31,66	22,61	0,465
V	237,68	9,81	4,86	722,69	57,81	0,499

Η περαιτέρω αξιολόγηση της αποτελεσματικότητας των μοντέλων έγινε με τη χρήση τριών βασικών κριτήριων, με το *RMSE* (τετραγωνική ρίζα του μέσου τετραγωνικού σφάλματος) της πρόβλεψης, το *MAE* (μέσο απόλυτο σφάλμα) της πρόβλεψης και τον δείκτη R^2_{pre} (συντελεστής προσδιορισμού) της πρόβλεψης. Όλες οι τιμές των κριτηρίων προέρχονται από διασταυρωμένη επικύρωση. Οι διαφορές των δεικτών με ή χωρίς σταθερό όρο είναι πολύ μικρές και για τα τρία κριτήρια. Στον Πίνακα 2 μπορεί κανείς να παρατηρήσει τις τιμές των αποτελέσματα των κριτηρίων για τα 8 διαφορετικά μοντέλα πρόβλεψης.

Εικόνα 2. *Πίνακας με τα διαγράμματα διασποράς των μεταβλητών, όπου: V ο όγκος, G η κυκλική επιφάνεια Bitterlich, hL το μέσο ύψος του Lorey, G_hL το γινόμενο της G επί το μέσο ύψος του Lorey, G_RhL το γινόμενο της G επί την τετραγωνική ρίζα του μέσου ύψους του Lorey – όπως υπολογίστηκαν στις επιφάνειες Bitterlich*



Εικόνα 3. *Θηκογράμματα των μεταβλητών και συνδυασμών τους όπως περιλαμβάνονται στα γραμμικά μοντέλα*

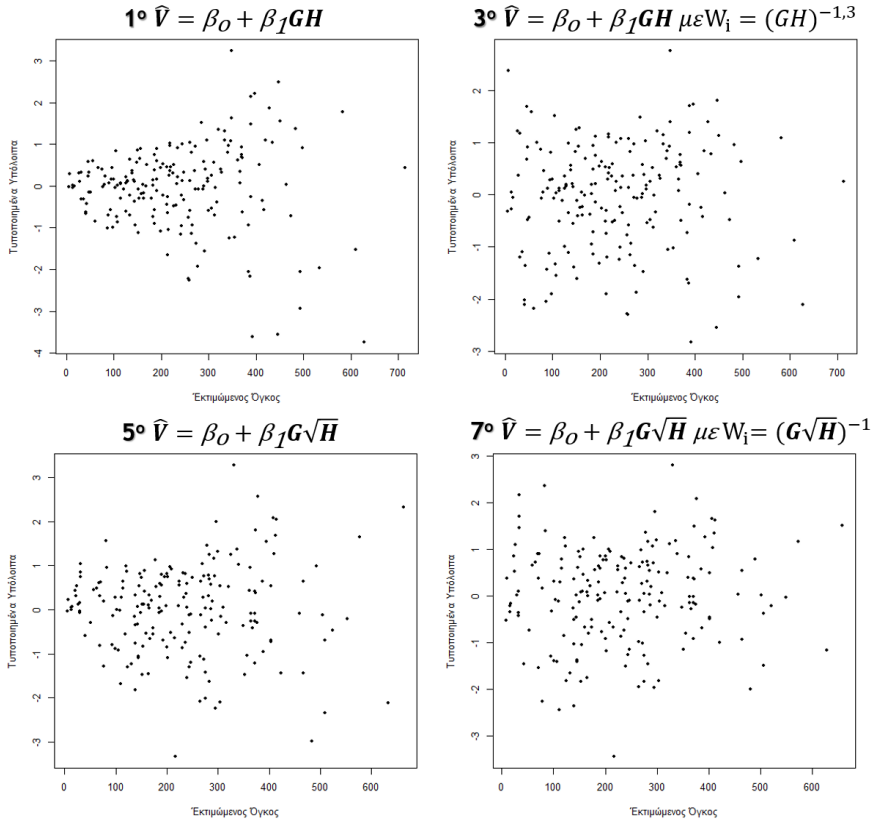


Μεταξύ των μοντέλων που προβλέπουν καλύτερα νέα δεδομένα επιλέχτηκαν ως καλύτερα εκείνα που έχουν σταθερό όρο και αντιστοιχούν στις εξισώσεις 3, 7 και 11, όπως συνοψίζονται στον Πίνακα 3. Το 3^{ο} μοντέλο θεωρείται πιο σταθερό στις προβλέψεις του γιατί είναι μικρότερα τα τυπικά σφάλματα των συντελεστών του β_0 και β_1. Στο 11^{ο} μοντέλο έχουμε αρκετά υψηλότερες τιμές στα κριτήρια *RMSE* και *MAE* σε σχέση με το 3^{ο} και 7^{ο} γιατί έχει παραληφθεί η ανεξάρτητη μεταβλητή του ύψους.

Πίνακας 2. *Κριτήρια αξιολόγησης των μοντέλων, όπου:* $\hat{V}$ *η εκτίμηση του όγκου, G η κυκλική επιφάνεια Bitterlich,* Gh_L *το γινόμενο του G επί το μέσο ύψος του Lorey (*h_L*),* $G\sqrt{h_L}$ *το γινόμενο του G επί την τετραγωνική ρίζα του (*h_L*)*

Μοντέλα	R^2_{pre}	*RMSE*	*MAE*
3^{ο} $\hat{V} = -5,409 + 1,158Gh_L$	0,977	21,397	15,687
4^{ο} $\hat{V} = 0,461G\,h_L$	0,978	21,677	15,947
7^{ο} $\hat{V} = -12,181 + 2,205G\sqrt{h_L}$	0,965	26,434	20,352
8^{ο} $\hat{V} = 2,098G\sqrt{h_L}$	0,965	27,498	21,097
11^{ο} $\hat{V} = -10,782 + 9,939G$	0,888	47,133	36,766
12^{ο} $\hat{V} = 9,507G$	0,888	47,243	36,693

Εικόνα 4. *Διασπορά των τυποποιημένων σφαλμάτων ως προς τον εκτιμώμενο όγκο, στα αρχικά μοντέλα (1ο και 5ο στο αριστερό μέρος) και στα σταθμισμένα μοντέλα (3ο και 7ο στο δεξιό μέρος), όπου: V ο όγκος, G η κυκλική επιφάνεια Bitterlich, H το μέσο ύψος του Lorey*



Το εκατοστιαίο απόλυτο σφάλμα στο 3ο μοντέλο είναι 6,6% και στο 7ο είναι 8,56% (Πίνακας 3). Ο Evert (1983), κατά την προσωπική του άποψη, κάνει αποδεκτή μια απόκλιση της τάξης του 10% μεταξύ των εκτιμωμένων και παρατηρούμενων τιμών, συνεπώς φαίνεται να γίνονται αποδεκτά τα μοντέλα 3 και 7 με βάση το παραπάνω όριο. Το 11ο μοντέλο έχει αντίστοιχο σφάλμα 15,32%, αλλά έχει μόνο μια ανεξάρτητη μεταβλητή (*G*) και είναι αποτελεσματικό κόστους. Έχοντας υπόψη ότι η συλλογή δεδομένων πεδίου είναι αρκετά κοστοβόρα για την δημιουργία των μαζοπινάκων συστάδας, είναι ενδιαφέρον να ερευνηθεί μελλοντικά η επίδραση του μεγέθους του δείγματος στο σφάλμα πρόβλεψης τέτοιων μοντέλων.

Η περίοδος που διατρέχουμε λόγω της οικονομικής κρίσης και η οποία χαρακτηρίζεται από οικονομική δυσπραγία, είχε ως αποτέλεσμα την μείωση των δημοσίων επενδύσεων και στον κλάδο της δασοπονίας (μειωμένα κονδύλια για δασικές μελέτες και κατασκευές, μείωση προσωπικού). Λόγω αυτής της κατάστασης, οι μαζοπίνακες δασικών συστάδων θα μπορούσαν να αποτελέσουν ένα γρήγορο και

οικονομικό "εργαλείο" απογραφής των δασικών πόρων, συμβάλλοντας στην ορθολογική διαχείριση των δασικών οικοσυστημάτων.

Πίνακας 3. *Συνοπτικός πίνακας παρουσίασης των τελικά επιλεγόμενων μοντέλων, όπου:* $\hat{V}$ *ο όγκος, G η κυκλική επιφάνεια Bitterlich,* Gh_L *το γινόμενο της G επί το μέσο ύψος του Lorey (*h_L*),* $G\sqrt{h_L}$ *το γινόμενο της G επί την τετραγωνική ρίζα του* h_L

Μοντέλα		β_0	β_1	*RMSE%*	*MAE%*
3ο $\hat{V} = \beta_o + \beta_1 Gh_L =$ $-5,409 + 1,158 Gh_L$	Συντελεστές	5,049	0,468	9,01	6,60
	Τυπικό σφάλμα	1,159	0,004		
7ο $\hat{V} = \beta_o + \beta_1 G\sqrt{H} =$ $-12,181 + 2,205 G\sqrt{H}$	Συντελεστές	-12,181	2,205	11,13	8,56
	Τυπικό σφάλμα	2,376	0,027		
11ο $\hat{V} = \beta_o + \beta_1 G =$ $-10,781 + 9,939 G$	Συντελεστές	-10,781	9,939	19,87	15,32
	Τυπικό σφάλμα	4,474	0,224		

4. ΣΥΜΠΕΡΑΣΜΑΤΑ

Το "καλύτερο" μοντέλο ως προς την ικανότητα πρόβλεψης είναι το 3ο μοντέλο $\hat{V} = -5,409 + 1,158 Gh_L$ με βάρος $W_i = (GH)^{(-1,3)}$, έχοντας το μικρότερο σφάλμα πρόβλεψης (δείκτες *RMSE* και *MAE*) καθώς και την μεγαλύτερη προσαρμογή σε νέα δεδομένα (R^2_{pre}). Ενδιαφέρον για την πράξη θα είχε η διερεύνηση της επίδρασης του μεγέθους του δείγματος στην παραπάνω κατασκευή των μοντέλων.

ABSTRACT

In the management of forest ecosystems, the volume of trees is one of the characteristics which is of greater scientific and practical interest. The effective estimation of the volume of the forest stands is based mainly in the low cost of collection of field data in the context of acceptable level accuracy. Elements of the estimation are the sampling methods and the mensuration of trees. The objective of this work is the creation of effective statistical prediction models of stand volume, where can be applied in the management of forest ecosystems. The data was taken from the orographic unit of "Bathy" at the University Forest of Pertouli. Bitterlich sampling was applied, where trees selected with probability proportional to the basal area. Additionally, this specific sampling is a cost-effective method. Various models that encounter in the bibliography were checked and validated, while at the end selected with the simplest mathematical expression, with criterion the prediction efficiency. The main aspects of modeling are discussed in relation to their application to the forest stand and more generally to the management of forest ecosystems in the current socio-economic context.

Keywords: Forest Management, stand volume tables, Bitterlich plot, prediction error, cross-validation, Forest of Pertouli

Ευχαριστίες: Ως πρώτο συμβαλλόμενο μέλος θα ήθελα να ευχαριστήσω την ΓΓΕΤ καθώς και το ΕΛΙΔΕΚ και να επισημάνω από πλευράς μου ότι: "This research has been financially supported by General Secretariat for Research and Technology (GSRT) and the Hellenic Foundation for Research and Innovation (HFRI) (Scholarship Code: 1319)".
Ιδιαίτερες ευχαριστίες οφείλονται στον κ. Παπαδόπουλο Γεώργιο, καθηγητή στο Γεωπονικό Πανεπιστήμιο Αθηνών, διότι οι ουσιαστικές και λεπτομερείς παρατηρήσεις του είχαν ως αποτέλεσμα την σημαντική βελτίωση της εργασίας.

ΑΝΑΦΟΡΕΣ

Aertsen, Wim, Vincent Kint, Jos van Orshoven, Kürşad Özkan, and Bart Muys. (2010). Comparison and ranking of different modelling techniques for prediction of site index in Mediterranean mountain forests, *Ecological Modelling*, 221: 1119-30.

Bitterlich, W. (1984). The relascope idea. Relative measurements in forestry. *Farnham Royal*: Commonwealth Agricultural Bureaux.

Chai, T., and R. R. Draxler. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature, *Geoscientific model development*., 7: 1247-50.

Corona, P., Cibella, R., Pizzurro, G. and Veca, D. S. l. M. (2009). Stand volume tables for forest stands in Sicily. *Italia Forestale e Montana*, 64(4), 249-261. doi:10.4129/IFM.2009.4.03

Cort, J. Willmott, and Matsuura Kenji. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance, *Climate Research*, 30: 79-82.

Brassington, Gary. (2017). "Mean absolute error and root mean square error: which is the better metric for assessing model performance?" In *EGU General Assembly Conference Abstracts*, 3574.

Dau, J. H., A., M., & Dawaki, S. A. (2015). Role of Forest Inventory in Sustainable Forest Management: A Review. *International Journal of Forestry and Horticulture (IJFH)*, 1(2), 33-40.).

Dube, T., Sibanda, M., Shoko, C. and Mutanga, O. (2017). Stand-volume estimation from multi-source data for coppiced and high forest Eucalyptus spp. silvicultural systems in KwaZulu-Natal, South Africa. *ISPRS Journal of Photogrammetry and Remote Sensing*, 132, 162-169. doi:https://doi.org/10.1016/j.isprsjprs.2017.09.001

Evert, F. (1983). An Equation for Estimating Total Volume of Both Stands and Single Trees of Black Spruce. *The Forestry Chronicle*, 59(1), 26-29. doi:10.5558/tfc59026-1

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning with applications in R: *Springer New York Heidelberg Dordrecht London*.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). The elements of statistical learning: Data Mining, Inference, and Prediction (Second Edition Vol. 1): *Springer series in statistics New York*.

Kershaw Jr, J. A., Ducey, M. J., Beers, T. W., & Husch, B. (2016). Forest Mensuration: *John Wiley & Sons*.

Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling (Vol. 26). *New York, NY: Springer*.

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. 2008, 28(5), 26. doi:10.18637/jss.v028.i05

Machado, S. d. A., Aguiar, L. P., Figueiredo Filho, A. and Koehler, H. S. (2008). Stand volume modelling of Mimosa scabrella benth. from Curitiba metropolitan region. Revista Árvore, 32, 465-478.

Næsset, E. (1995). Stand volume functions for picea abies in western Norway. *Scandinavian Journal of Forest Research,* 10(1-4), 42-50. doi:10.1080/02827589509382866

Næsset, E. and Tveite, B. (1999). Stand Volume Functions for Picea abies in Eastern, Central and Northern Norway. *Scandinavian Journal of Forest Research*, 14(2), 164-174. doi:10.1080/02827589950152890

Ryan, A. G., Montgomery, D. C., Peck, E. A., & Vining, G. G. (2013). Solutions Manual to Accompany Introduction to Linear Regression Analysis (5 ed.): *Wiley*.

Shepperd, W. D., Mowrer, H. T., Forest, R. M. and Station, R. E. (1984). Whole Stand Volume Tables for Quaking Aspen in the Rocky Mountains: *USDA Forest Service, Rocky Mountain Forest and Range Experiment Station*.

Shiver, B. D. and Brister, G. H. (1992). Tree and stand volume functions for Eucalyptus saligna. *Forest Ecology and Management*, 47(1), 211-223. doi:https://doi.org/10.1016/0378-1127(92)90275-E

Van Laar, A. and Akça, A. (2007). Forest Mensuration (Vol. 13): Springer Science & Business Media.

Γεωργάκης, Α. (2012). Σύγκριση του Βαθμού Συγκόμωσης με Επίγειες Μετρήσεις και με Μεθόδους Τηλεπισκόπησης στο Πανεπιστημιακό Δάσος Περτουλίου. Εργασία στο πλαίσιο των μαθημάτων Γεωγραφικά Συστήματα Πληροφοριών και Τηλεπισκόπισης του Π.Μ.Σ. *Σχολή Δασολογίας και Φυσικού Περιβάλλοντος*.

Μάτης, Κ. (2004). Δασική Βιομετρία ΙΙ. Δεντρομετρία. Θεσσαλονίκη: *Πήγασος*.

Τ.Δ.Δ.Π.Δ. (2008). Διαχειριστικό Σχέδιο Πανεπιστημιακού Δάσους Περτουλίου 2009-2018: *Ταμείο Διοικήσεως και Διαχειρίσεως Πανεπιστημιακών Δασών (Τ.Δ.Δ.Π.Δ.)*.

Χατζηλαζάρου, Γ. (2002). Μαζοπίνακας συστάδας για τη δρυ του δάσους του δήμου Κάτω Κλεινών νομού Φλώρινας. (Μεταπτυχιακή διατριβή), *Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης*.

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 31ου Πανελληνίου Συνεδρίου (2018), σελ. 053−062

ΕΠΕΚΤΕΤΑΜΕΝΟ ΦΙΛΤΡΟ KALMAN ΜΕ ΜΗ-ΑΡΝΗΤΙΚΟΥΣ ΠΕΡΙΟΡΙΣΜΟΥΣ ΚΑΙ ΧΡΗΣΗ ΤΕΤΡΑΓΩΝΙΚΗΣ ΜΟΡΦΗΣ

Ο. Θεοδοσιάδου[1], Γ. Τσακλίδης[1]
[1]Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
{outheod, tsaklidi}@math.auth.gr

ΠΕΡΙΛΗΨΗ

Στην παρούσα εργασία κατασκευάζεται ένα μη-γραμμικό μοντέλο χώρου καταστάσεων, όπου το διάνυσμα κατάστασης υπόκειται σε μη-αρνητικούς περιορισμούς. Στόχο αποτελεί η εκτίμηση των θετικών και αρνητικών αλμάτων των αποδόσεων μετοχών, τα οποία μοντελοποιούνται ως κρυφές τυχαίες μεταβλητές. Δεδομένου ότι η εξίσωση των παρατηρήσεων (αποδόσεις) εκφράζεται ως η διαφορά μεταξύ των αλμάτων για κάθε χρονική στιγμή, οι εκτιμώμενες τιμές των αλμάτων πρέπει να είναι μη-αρνητικές. Για τον λόγο αυτόν, στην εξίσωση των καταστάσεων, η οποία περιγράφει τη δυναμική εξέλιξη του διδιάστατου διανύσματος των αλμάτων, υιοθετούνται δυο θετικά οριμένες τετραγωνικές μορφές. Καθότι το προτεινόμενο μοντέλο είναι μη-γραμμικό, χρησιμοποιείται το επεκτεταμένο φίλτρο Kalman 2ης τάξης για τη διαδικασία εκτίμησης. Ακολούθως, ελέγχεται η προσαρμογή του μοντέλου σε εμπειρικά δεδομένα.

Λέξεις κλειδιά: επεκτεταμένο φίλτρο Kalman 2ης τάξης, τετραγωνική μορφή, άλματα αποδόσεων μετοχών.

1. ΕΙΣΑΓΩΓΗ

Η υιοθέτηση των αλμάτων για τη μοντελοποίηση των τιμών χρηματοοικονομικών μεγεθών είναι πλέον ευρέως διαδεδομένη, καθώς έχει αποδειχθεί ότι μοντέλα όπως των Black-Scoles (1973), τα οποία προτάσσουν την εξέλιξη των τιμών κατά συνεχή τρόπο, δε μπορούν να αποδώσουν ικανοποιητικά τις ιδιότητές τους. Η άποψη αυτή ενισχύεται και από μια πληθώρα εργασιών όπου γίνεται χρήση των αλμάτων για την επίτευξη καλής προσαρμογής σε εμπειρικές χρονοσοσειρές αποδόσεων (βλ. για παράδειγμα Pan, 2002 και Eraker, 2004). Μια λεπτομερής εισαγωγή στη χρήση των αλμάτων στη χρηματοοικονομική παρουσιάζεται για παράδειγμα στα συγγράμματα Cont and Tankov (2004) και Schoutens (2003).

Στην εργασία Polimenis (2012), τα άλματα θεωρούνται κρυφές τυχαίες μεταβλητές, δηλαδή δεν αποτελούν παρατηρήσιμες ποσότητες, και χωρίζονται σε δυο κατη-

γορίες: Τα θετικά άλματα (άλματα προς τα πάνω), τα οποία οφείλονται στην άφιξη θετικών ειδήσεων στην αγορά, και τα αρνητικά άλματα (άλματα προς τα κάτω), τα οποία οφείλονται στην άφιξη αρνητικών ειδήσεων.

Στην παρούσα εργασία στόχος είναι η εκτίμηση των θετικών και των αρνητικών αλμάτων των αποδόσεων μετοχών. Καθότι τα άλματα μοντελοποιούνται ως κρυφές τυχαίες μεταβλητές, κατασκευάζεται ένα κατάλληλο μη-γραμμικό μοντέλο χώρου καταστάσεων. Η χρήση τέτοιου είδους μοντελοποίησης στη χρηματοοικονομική είναι διαδεδομένη για την εκτίμηση παραμέτρων, οι οποίες είναι μεταβλητές στο χρόνο (βλ. για παράδειγμα Koopman et al., 2010 και Wells, 1996). Στο προτεινόμενο μοντέλο η εξίσωση των παρατηρήσεων (αποδόσεις) δίνεται ως διαφορά των αλμάτων για κάθε χρονική στιγμή t. Συνεπώς, οι εκτιμήσεις των αλμάτων πρέπει να είναι μη-αρνητικές. Για τον λόγο αυτόν, στην εξίσωση των καταστάσεων (άλματα), που περιγράφει τη δυναμική εξέλιξη του διδιάστατου διανύσματος των αλμάτων, υιοθετούνται δυο θετικά ορισμένες τετραγωνικές μορφές.

Για τη διαδικασία εκτίμησης των αλμάτων χρησιμοποιείται το επεκτεταμένο φίλτρο Kalman 2ης τάξης. Στην προκειμένη περίπτωση το γραμμικό φίλτρο Kalman (Kalman, 1960) δεν μπορεί να χρησιμοποιηθεί, καθότι το προτεινόμενο μοντέλο είναι μη-γραμμικό. Μετά την εκτίμηση των αλμάτων, εξετάζεται η προσαρμογή του μοντέλου σε εμπειρικά δεδομένα. Για τον σκοπό αυτόν, χρησιμοποιούνται οι αποδόσεις του δείκτη NASDAQ για τη χρονική περίοδο 2006-2008 (τα δεδομένα αντλήθηκαν από την ιστοσελίδα finance.yahoo.com).

Η εργασία οργανώνεται ως εξής:
Στην ενότητα 2 κατασκευάζεται το μη-γραμμικό μοντέλο χώρου καταστάσεων που χρησιμοποιείται για την εκτίμηση των αλμάτων. Στην ενότητα 3 περιγράφεται η διαδικασία εκτίμησης των αλμάτων με τη χρήση του επεκτεταμένου φίλτρου Kalman 2ης τάξης, και εξάγονται οι αναδρομικές σχέσεις που θα χρησιμοποιηθούν. Στην ενότητα 4 γίνεται εφαρμογή της προτεινόμενης μοντελοποίησης στις αποδόσεις του δείκτη NASDAQ, όπου και εξετάζεται η προσαρμογή του μοντέλου στα εμπειρικά δεδομένα. Τέλος, στην ενότητα 5 παραθέτονται τα συμπεράσματα που προέκυψαν από την προτεινόμενη μοντελοποίηση, και γίνεται αναφορά σε μελλοντικές τροποποιήσεις.

2. ΤΟ ΜΟΝΤΕΛΟ

Θεωρούμε ότι η απόδοση R τη χρονική στιγμή t δίνεται από τη σχέση

$$R_t = X_t - Y_t + e_t \tag{1}$$

όπου,

- X_t: το θετικό άλμα των αποδόσεων τη χρονική στιγμή t,
- Y_t: το αρνητικό άλμα των αποδόσεων τη χρονική στιγμή t,
- e_t: ο θόρυβος για τον οποίο θεωρούμε ότι $e_t \sim N(0, V)$.

Τα άλματα X_t, Y_t μοντελοποιούνται ως κρυφές τυχαίες μεταβλητές, δηλαδή είναι μη-παρατηρήσιμες ποσότητες. Προκειμένου να εκτιμηθούν αυτές οι καταστάσεις,

κατασκευάζεται ένα μοντέλο χώρου καταστάσεων (*state space model*), όπου το διάνυσμα κατάστασης αποτελείται από τις δύο συνιστώσες των αλμάτων. Επιπλέον, τα εκτιμώμενα άλματα που θα προκύψουν, πρέπει να ικανοποιούν τη σχέση (1) υπό τον περιορισμό $X_t, Y_t \geq 0$, t=1,2,....

Για τους παραπάνω λόγους, η εξίσωση των καταστάσεων που θα χρησιμοποιηθεί θα πρέπει να έχει κατάλληλη μορφή, ώστε τα εκτιμώμενα άλματα που θα προκύπτουν να είναι μη-αρνητικά για κάθε χρονική στιγμή t.

Για τον σκοπό αυτόν θεωρούμε ότι η εξίσωση των καταστάσεων δίνεται από τη σχέση

$$\left.\begin{aligned} X_t &= (\mathbf{z}_{t-1} + \mathbf{w}_{t-1})^T \mathbf{G^{(1)}} (\mathbf{z}_{t-1} + \mathbf{w}_{t-1}) = f_1(\mathbf{z}_{t-1}, \mathbf{w}_{t-1}) \\ Y_t &= (\mathbf{z}_{t-1} + \mathbf{w}_{t-1})^T \mathbf{G^{(2)}} (\mathbf{z}_{t-1} + \mathbf{w}_{t-1}) = f_2(\mathbf{z}_{t-1}, \mathbf{w}_{t-1})) \end{aligned}\right\} \tag{2}$$

ή ισοδύναμα

$$\mathbf{z}_t = \sum_{k=1}^{2} \phi_k (\mathbf{z}_{t-1} + \mathbf{w}_{t-1})^T \mathbf{G^{(k)}} (\mathbf{z}_{t-1} + \mathbf{w}_{t-1}) \tag{3}$$

όπου, $\mathbf{z}_t = \begin{bmatrix} X_t \\ Y_t \end{bmatrix}$ το διάνυσμα κατάστασης, $\mathbf{w_t}$ θόρυβος για τον οποίο θεωρούμε ότι $\mathbf{w_t} \sim N(\mathbf{0}, \mathbf{Q} = \begin{bmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{bmatrix})$ και $\mathbf{G}^{(k)}, k = 1, 2$ συμμετρικός και θετικά ορισμένος (2×2) πίνακας. Δηλαδή, ισχύουν οι σχέσεις

$$g_{11}^{(k)} > 0 \quad \text{και} \quad g_{11}^{(k)} g_{22}^{(k)} - (g_{12}^{(k)})^2 > 0 \ , \ k = 1, 2.$$

Το ϕ_k είναι ένα (2×1) διάνυσμα-στήλη, όπου τα στοιχεία του είναι μηδενικά, εκτός από το k-στοιχείο, το οποίο είναι ίσο με 1.

Η εξίσωση των παρατηρήσεων δίνεται από τη σχέση (1), ή ισοδύναμα σε μορφή πινάκων

$$R_t = \mathbf{H z_t} + e_t \ , \tag{4}$$

όπου $\mathbf{H}$= $\begin{bmatrix} 1 & -1 \end{bmatrix}$. Επιπλέον, θεωρούμε ότι οι θόρυβοι $\mathbf{w}_t$ και e_t είναι ανεξάρτητοι μεταξύ τους, δηλαδή $\mathbb{E}(e_k \mathbf{w}_j^T) = \mathbf{0}$.

Στο μοντέλο (3)-(4) υιοθετείται η χρήση των θετικά ορισμένων τετραγωνικών μορφών για την περιγραφή της δυναμικής εξέλιξης των αλμάτων, προκειμένου να διασφαλιστεί η μη-αρνητικότητα των εκτιμήσεων. Επιπλέον, ο θόρυβος στη σχέση (3) είναι πολλαπλασιαστικός, και όχι προσθετικός.

3. ΕΚΤΙΜΗΣΗ ΤΟΥ ΔΙΑΝΥΣΜΑΤΟΣ ΚΑΤΑΣΤΑΣΗΣ

Σημειώνεται ότι το προτεινόμενο μοντέλο (3)-(4) για την εκτίμηση των αλμάτων, είναι μη-γραμμικό. Συνεπώς, δε μπορεί να χρησιμοποιηθεί ο επαναληπτικός αλγόριθμος του φίλτρου Kalman για την εκτίμηση του διανύσματος κατάστασης $\mathbf{z_t}$. Με

δεδομένο ότι η εξίσωση των καταστάσεων στη σχέση (3) δίνεται ως τετραγωνική μορφή, για την εκτίμηση των αλμάτων θα χρησιμοποιηθεί το επεκτεταμένο φίλτρο Kalman 2ης τάξης.

Για την εξαγωγή των επαναληπτικών σχέσεων του φίλτρου Kalman 2ης τάξης, θα χρησιμοποιηθούν οι παρακάτω συμβολισμοί:

- $\hat{\mathbf{z}}_t(-)$: η εκ των προτέρων εκτίμηση του $\mathbf{z}_t$, μη λαμβάνοντας υπόψη την παρατήρηση στον χρόνο t,
- $\hat{\mathbf{z}}_t(+)$: η εκ των υστέρων εκτίμηση του $\mathbf{z}_t$, λαμβάνοντας υπόψη την παρατήρηση στον χρόνο t,
- $\mathbf{P}_t(-)$, $\mathbf{P}_t(+)$: οι (2×2) πίνακες διασπορών-συνδιασπορών των σφαλμάτων εκτίμησης για τις εκτιμήσεις $\hat{\mathbf{z}}_t(-)$ και $\hat{\mathbf{z}}_t(+)$, αντίστοιχα. Δηλαδή ισχύει,

$$\mathbf{P}_t(-) = \mathbb{E}[(\mathbf{z}_t - \hat{\mathbf{z}}_t(-))(\mathbf{z}_t - \hat{\mathbf{z}}_t(-))^T] \quad \text{και} \quad \mathbf{P}_t(+) = \mathbb{E}[(\mathbf{z}_t - \hat{\mathbf{z}}_t(+))(\mathbf{z}_t - \hat{\mathbf{z}}_t(+))^T] \, .$$

Επιπλέον, χρησιμοποιώντας το ανάπτυγμα Taylor 2ης τάξης της σχέσης (3) γύρω από το σημείο $(\hat{\mathbf{z}}_{t-1}(+), \mathbf{0})$, προκύπτει ότι:

$$\begin{aligned}
\mathbf{z}_t^k =& f_k(\hat{\mathbf{z}}_{t-1}(+), \mathbf{0})) \\
&+ (\frac{df_k(\hat{\mathbf{z}}_{t-1}(+), \mathbf{0}))}{dz})^T(\mathbf{z}_{t-1} - \hat{\mathbf{z}}_{t-1}(+)) + (\frac{df_k(\hat{\mathbf{z}}_{t-1}(+), \mathbf{0}))}{dw})^T \mathbf{w}_{t-1} \\
&+ \frac{1}{2}(\mathbf{z}_{t-1} - \hat{\mathbf{z}}_{t-1}(+))^T \frac{d^2 f_k(\hat{\mathbf{z}}_{t-1}(+), \mathbf{0}))}{dz^2}(\mathbf{z}_{t-1} - \hat{\mathbf{z}}_{t-1}) \\
&+ \frac{1}{2}\mathbf{w}_{t-1}^T \frac{d^2 f_k(\hat{\mathbf{z}}_{t-1}(+), \mathbf{0}))}{dw^2}\mathbf{w}_{t-1} \\
&+ (\mathbf{z}_{t-1} - \hat{\mathbf{z}}_{t-1}(+))^T \frac{d^2 f_k(\hat{\mathbf{z}}_{t-1}(+), \mathbf{0})}{dzdw}\mathbf{w}_{t-1}, \quad k = 1, 2
\end{aligned} \tag{5}$$

όπου οι συναρτήσεις f_k, $k = 1, 2$ δίνονται από τη σχέση (2).

Επομένως, παίρνοντας τις μέσες τιμές στη σχέση (5) προκύπτουν οι επαναληπτικές σχέσεις για την εκ των προτέρων εκτίμηση του $\mathbf{z}_t$ και του αντίστοιχου πίνακα διασπορών-συνδιασπορών των σφαλμάτων της εκτίμησης (στάδιο πρόβλεψης) :

$$\begin{aligned}
\hat{\mathbf{z}}_t^k(-) =& f_k(\hat{\mathbf{z}}_{t-1}(+), \mathbf{0})) + \frac{1}{2}\,\mathrm{tr}(\frac{d^2 f(\hat{\mathbf{z}}_{t-1}(+), \mathbf{0})}{dz^2}\mathbf{P}_{t-1}(+)) \\
&+ \frac{1}{2}\,\mathrm{tr}(\frac{d^2 f(\hat{\mathbf{z}}_{t-1}(+), \mathbf{0})}{dw^2}\mathbf{Q}), \quad k = 1, 2
\end{aligned}$$

ενώ τα στοιχεία του πίνακα $\mathbf{P}_t(-)$ δίνονται από τη σχέση,

$$\begin{aligned}\mathbf{P}_t^{km}(-) =&(\frac{df_k(\hat{\mathbf{z}}_{t-1}(+),\mathbf{0})}{dz})^T\mathbf{P}_{t-1}(+)\frac{df_m(\hat{\mathbf{z}}_{t-1}(+),\mathbf{0}))}{dz}\\ &+(\frac{df_k(\hat{\mathbf{z}}_{t-1}(+),\mathbf{0})}{dw})^T\mathbf{Q}\frac{df_m(\hat{\mathbf{z}}_{t-1}(+),\mathbf{0})}{dw}\\ &+\frac{1}{2}\operatorname{tr}(\frac{d^2f_k(\hat{\mathbf{z}}_{t-1}(+),\mathbf{0})}{dz^2}\mathbf{P}_{t-1}(+)\frac{d^2f_m(\hat{\mathbf{z}}_{t-1}(+),\mathbf{0})}{dz^2}\mathbf{P}_{t-1}(+))\\ &+\frac{1}{2}\operatorname{tr}(\frac{d^2f_k(\hat{\mathbf{z}}_{t-1}(+),\mathbf{0})}{dw^2}\mathbf{Q}\frac{d^2f_m(\hat{\mathbf{z}}_{t-1}(+),\mathbf{0})}{dw^2}\mathbf{Q}),\quad k,m=1,2\end{aligned}$$

όπου με tr(.) συμβολίζεται το ίχνος πίνακα. Στην προκειμένη περίπτωση, μετά από υπολογισμούς βασισμένους στη σχέση (3) προκύπτει ότι:

$$\hat{\mathbf{z}}_t^k(-)=\hat{\mathbf{z}}_{t-1}(+)^T\mathbf{G^{(k)}}\hat{\mathbf{z}}_{t-1}(+)+\operatorname{tr}(\mathbf{G}^k\mathbf{P}_{t-1}(+))+\operatorname{tr}(\mathbf{G}^k\mathbf{Q}),\quad k=1,2\,, \tag{6}$$

$$\begin{aligned}\mathbf{P}_t^{km}(-) =&4\hat{\mathbf{z}}_{t-1}(+)^T\mathbf{G^{(k)}}\mathbf{P}_{t-1}(+)\mathbf{G^{(m)}}\hat{\mathbf{z}}_{t-1}(+)+4\hat{\mathbf{z}}_{t-1}(+)^T\mathbf{G^{(k)}}\mathbf{Q}\mathbf{G^{(m)}}\hat{\mathbf{z}}_{t-1}(+)\\ &+2\operatorname{tr}(\mathbf{G^{(k)}}\mathbf{P}_{t-1}(+)\mathbf{G^{(m)}}\mathbf{P}_{t-1}(+))+2\operatorname{tr}(\mathbf{G^{(k)}}\mathbf{Q}\mathbf{G^{(m)}}\mathbf{Q}),\ k,m=1,2\,.\end{aligned} \tag{7}$$

Αναφορικά με τις εκ των υστέρων εκτιμήσεις (στάδιο επικαιροποίησης), οι επαναληπτικές σχέσεις προκύπτουν από την υπόθεση ότι οι τ.μ. $\mathbf{z}_t$ και R_t έχουν από κοινού κανονική κατανομή με βάση τη σχέση

$$\begin{bmatrix}\mathbf{z}_t\\ R_t\end{bmatrix}\sim N(\begin{bmatrix}\hat{\mathbf{z}}_t(-)\\ \mathbf{H}\hat{\mathbf{z}}_t(-)\end{bmatrix},\begin{bmatrix}\mathbf{P}_t(-) & \mathbf{P}_t(-)\mathbf{H}^T\\ \mathbf{H}\mathbf{P}_t(-)^T & \mathbf{H}\mathbf{P}_t(-)\mathbf{H}^T+V\end{bmatrix}).$$

και τη χρήση του παρακάτω Λήμματος (βλ. για παράδειγμα Kotz et al., 2000):

Λήμμα 3..1. *Ας είναι* $\mathbf{x}$, $\mathbf{y}$ *δυο τυχαίες μεταβλητές με κανονική από κοινού κατανομή όπου για τη μέση τιμή και τη συνδιασπορά ισχύουν:*

$$\mathbb{E}(\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix})=\begin{bmatrix}\mu_x\\ \mu_y\end{bmatrix}\quad \textit{και}\quad \mathbf{\Sigma}=\begin{bmatrix}\mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12}\\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22}\end{bmatrix}.$$

Τότε, η δεσμευμένη κατανομή της $\mathbf{x}$ *με δεδομένη την* $\mathbf{y}$ *είναι κανονική, δηλαδή,* $(\mathbf{x}|\mathbf{y})\sim N(\mu',\mathbf{\Sigma}')$*, όπου:*

$$\mu'=\mu_x+\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}(\mathbf{y}-\mu_y)\quad \textit{και}\quad \mathbf{\Sigma}'=\mathbf{\Sigma}_{11}-\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}\,.$$

□

Συνεπώς, προκύπτουν οι αναδρομικές σχέσεις,

$$\hat{\mathbf{z}}_t(+)=\hat{\mathbf{z}}_t(-)+\mathbf{K}_t(R_t-\mathbf{H}\hat{\mathbf{z}}_t(-))\,, \tag{8}$$

$$\mathbf{P}_t(+)=(\mathbf{I}-\mathbf{K}_t\mathbf{H})\mathbf{P}_t(-)\,, \tag{9}$$

όπου $\mathbf{K}_t = \mathbf{P}_t(-)\mathbf{H}^T(\mathbf{H}\mathbf{P}_t(-)\mathbf{H}^T + V)^{-1}$.

Επομένως, χρησιμοποιώντας τις αναδρομικές σχέσεις (6)-(9), μπορούμε να εκτιμήσουμε τα θετικά και αρνητικά άλματα των αποδόσεων που προκύπτουν με βάση το μοντέλο (3)-(4) για κάθε χρονική στιγμή t.

4. Εφαρμογή: Εκτίμηση των αλμάτων για τον δείκτη NASDAQ

Στην παρούσα ενότητα γίνεται εκτίμηση των «κρυφών» αλμάτων για τις αποδόσεις του δείκτη NASDAQ, που αφορούν στην τριετία 2006-2008 (755 μετρήσεις), χρησιμοποιώντας το μοντέλο (3)-(4). Αρχικά, πρέπει να εκτιμηθεί το συνόλο $\phi = (\mathbf{G}^{(1)}, \mathbf{G}^{(2)}, \sigma_x^2, \sigma_y^2, V)$ των παραμέτρων του μοντέλου (9 παράμετροι). Για τον σκοπό αυτόν, χρησιμοποιείται η μέθοδος της Μέγιστης Πιθανοφάνειας, λαμβάνοντας υπόψη ότι η δεσμευμένη κατανομή της τ.μ. R_t με δεδομένη την $\mathbf{z}_t$ είναι κανονική για κάθε t, δηλαδή

$$R_t \mid \mathbf{z}_t \sim N(\mathbf{H}\hat{\mathbf{z}}_t(-), \mathbf{H}\mathbf{P}_t(-)\mathbf{H}^T + V) \ .$$

Επομένως, ο Λογάριθμος της Πιθανοφάνειας δίνεται από τη σχέση

$$LogL(R_1, \ldots, R_n) = -n/2 \log(2\pi) - 0.5 \sum_{t=1}^{n} (log(|\omega_t|) + u_t^T \omega_t^{-1} u_t) \ , \qquad (10)$$

όπου,

$$u_t = R_t - \mathbf{H}\hat{\mathbf{z}}_t(-) \quad \text{και} \quad \omega_t = \mathbf{H}\mathbf{P}_t(-)\mathbf{H}^T + V \ .$$

Πρόκειται για μια διαδικασία βελτιστοποίησης, όπου καλύπτεται ένα κατάλληλο (με βάση τις τιμές που επιτρέπεται να πάρουν οι παράμετροι) χωρίο τιμών για τις παραμέτρους του συνόλου ϕ, υπολογίζεται η τιμή της συνάρτησης (10) για κάθε σύνολο τιμών αυτού του χωρίου, και επιλέγεται εκείνο το σύνολο που παράγει τη μεγαλύτερη τιμή για τη συνάρτηση (10). Εν συνεχεία, το σύνολο αυτό τοποθετείται ως αρχικό σημείο στο αλγόριθμο βελτιστοποίησης που χρησιμοποιεί μια μέθοδο τύπου Newton για την εύρεση του μεγίστου της (10), και καταγράφεται η 9-άδα τιμών που αντιστοιχεί στο μέγιστο. Στην προκειμένη περίπτωση εξετάστηκαν συνολικά 3386888 9-άδες τιμών.

Οι εκτιμήσεις των παραμέτρων του συνόλου ϕ που προκύπτουν μεγιστοποιώντας τη συνάρτηση (10) για τον δείκτη NASDAQ είναι:

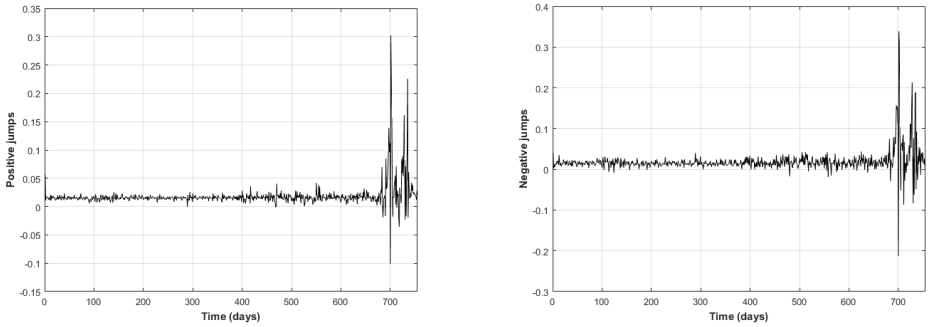
$$\mathbf{G}^{(1)} = \begin{bmatrix} 5.4741 & -2.8498 \\ -2.8498 & 7.3474 \end{bmatrix}, \mathbf{G}^{(2)} = \begin{bmatrix} 7.4368 & 1.4909 \\ 1.4909 & 2.8304 \end{bmatrix}$$

και

$$\sigma_x^2 = 0.9897x10^{-3}, \quad \sigma_y^2 = 0.86281x0^{-3}, \quad V = 4.961x10^{-11} \ .$$

Σημειώνεται ότι η βελτιστοποίηση έγινε ώστε οι (2×2) πίνακες $\mathbf{G}^{(1)}, \mathbf{G}^{(2)}$, να είναι θετικά ορισμένοι, δηλαδή επιλέχθηκαν συμμετρικοί, με θετικά διαγώνια στοιχεία και ορίζουσες θετικές, ενώ η διαδικασία βελτιστοποίησης πραγματοποιήθηκε στο περιβάλλον του υπολογιστικού πακέτου MATLAB 2017a.

Σχήμα 1: *(α) Εκτιμώμενα θετικά άλματα των αποδόσεων του δείκτη NASDAQ για την περίοδο 2006-2008, (β) Εκτιμώμενα αρνητικά άλματα των αποδόσεων του δείκτη NASDAQ για την περίοδο 2006-2008.*



Έχοντας πλέον εκτιμήσει τις παραμέτρους του μοντέλου (3)-(4) και χρησιμοποιώντας τις επαναληπτικές σχέσεις (6)-(9), προκύπτουν οι εκτιμήσεις των θετικών και αρνητικών αλμάτων των αποδόσεων του δείκτη NASDAQ, όπως φαίνονται στο Σχήμα 1.

Στο Σχήμα 1 παρατηρείται ότι υπάρχουν (εκ των υστέρων) εκτιμήσεις αλμάτων οι οποίες είναι αρνητικές. Πιο συγκεκριμένα, από την προτεινόμενη μέθοδο εκτίμησης προέκυψαν 36 ζεύγη αλμάτων από τα 755, όπου $X_t < 0$ ή $Y_t < 0$. Δηλαδή, παρά το γεγονός ότι στο μοντέλο (3)-(4) υιοθετείται θετικά ορισμένη τετραγωνική μορφή για την εκτίμηση του διανύσματος κατάστασης $\mathbf{z}_t$ και αυτό δίνει μη-αρνητικές εκ των προτέρων εκτιμήσεις, για τις εκ των υστέρων εκτιμήσεις εν τέλει προκύπτουν και κάποιες αρνητικές τιμές (ποσοστό 4,8% επί του συνόλου). Αυτό συμβαίνει λόγω της μορφής της σχέσης (8), καθώς η εκ των υστέρων εκτίμηση του $\mathbf{z}_t$ δίνεται ως γραμμικός συνδυασμός του $\hat{\mathbf{z}}_t(-)$ και της παρατήρησης R_t. Σημειώνεται ότι οι εκ των προτέρων εκτιμήσεις των αλμάτων (προβλέψεις) είναι μη-αρνητικές για κάθε t, καθότι προκύπτουν από τη σχέση (6), που αποτελείται από μη-αρνητικούς όρους.

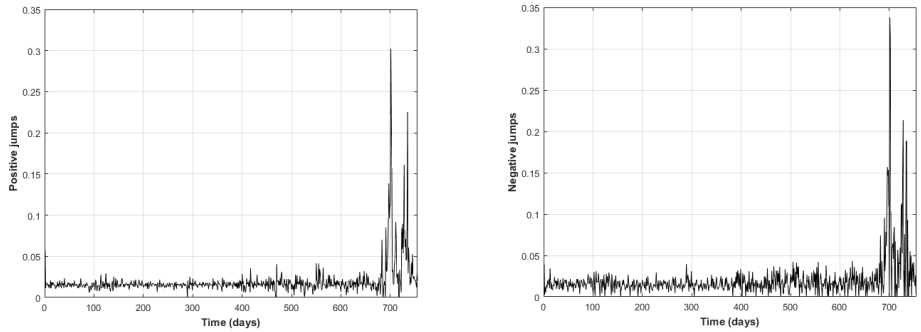
Προκειμένου οι (εκ των υστέρων) εκτιμήσεις των αλμάτων να ικανοποιούν τον περιορισμό $X_t, Y_t \geq 0$, για κάθε t, «διορθώνουμε» τις αρνητικές εκτιμήσεις που προέκυψαν, παίρνοντας ως «διορθωμένες» εκτιμήσεις $\hat{\tilde{\mathbf{z}}}_t(+)$ αυτές που προκύπτουν από την ελαχιστοποίηση της συνάρτησης

$$\hat{\tilde{\mathbf{z}}}_t(+) = \underset{z}{\operatorname{argmin}} (\mathbf{z}_t - \hat{\mathbf{z}}_t(+))^T \mathbf{P}_t(+)^{-1} (\mathbf{z}_t - \hat{\mathbf{z}}_t(+)) \ , \tag{11}$$

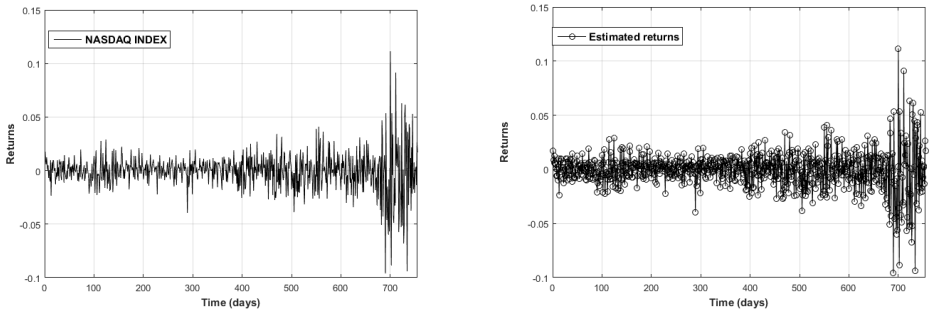
$$\text{με τον περιορισμό } \mathbf{z}_t \geq 0 \ .$$

Οι νέες εκτιμώμενες χρονοσειρές των αλμάτων που προκύπτουν δίνονται στο Σχήμα 2, όπου παρατηρείται ότι όλες οι τιμές είναι μη-αρνητικές. Οι «διορθωμένες» εκτιμήσεις (των 36 ζευγών αλμάτων) δεν ενσωματώθηκαν στον αλγόριθμο του φίλτρου,

Σχήμα 2: *(α) Εκτιμώμενα θετικά άλματα των αποδόσεων του δείκτη NASDAQ για την περίοδο 2006-2008 (μετά τη διόρθωση), (β) Εκτιμώμενα αρνητικά άλματα των αποδόσεων του δείκτη NASDAQ για την περίοδο 2006-2008 (μετά τη διόρθωση).*



Σχήμα 3: *(α) Ημερήσιες αποδόσεις του δείκτη NASDAQ για την περίοδο 2006-2008, (β) Εκτιμώμενες αποδόσεις του δείκτη NASDAQ για την περίοδο 2006-2008.*



δηλαδή περιορίστηκαν στο χρονικό βήμα που εμφανίστηκε η αρνητική εκτίμηση και δε χρησιμοποιήθηκαν στις επόμενες χρονικές στιγμές. Αντίστοιχες τεχνικές «διόρθωσης» των εκτιμήσεων που προκύπτουν με τη χρήση αλγορίθμων Kalman, όταν το διάνυσμα κατάστασης υπόκειται σε ισοτικούς ή ανισοτικούς περιορισμούς περιγράφονται στις εργασίες Simon (2010), Simon and Simon, (2010) και Gupta and Hauser (2007). Επιπλέον, η προσαρμογή της διαφοράς των αλμάτων για κάθε χρόνο t στις εμπειρικές αποδόσεις του δείκτη NASDAQ είναι πολύ ικανοποιητική, όπως φαίνεται στο Σχήμα 3, όπου τα δυο γραφήματα είναι σχεδόν πανομοιότυπα. Η καλή προσαρμογή επιβεβαιώνεται από την πολύ χαμηλή τιμή του μέσου τετραγωνικού σφάλματος (MSE), όπου $MSE = 4.97 \times 10^{-16}$.

5. ΣΥΜΠΕΡΑΣΜΑΤΑ

Προτείνεται ένα μη-γραμμικό μοντέλο χώρου καταστάσεων έχοντας ως σκοπό την εκτίμηση των θετικών και αρνητικών αλμάτων των αποδόσεων μετοχών. Τα άλματα

μοντελοποιούνται ως κρυφές τυχαίες μεταβλητές, και υπόκεινται σε μη-αρνητικούς περιορισμούς. Για τον λόγο αυτόν, η συνάρτηση μετάβασης που υιοθετείται στην εξίσωση των καταστάσεων είναι μια θετικά ορισμένη τετραγωνική μορφή. Ακολούθως, υπολογίζονται οι αναδρομικοί τύποι για την εκτίμηση του διανύσματος κατάστασης στο στάδιο της πρόβλεψης με τη χρήση του αναπτύγματος Taylor 2ης τάξης, και στο στάδιο της επικαιροποίησης με τη χρήση του Λήμματος 3.1.

Στις εκτιμώμενες χρονοσειρές των αλμάτων παρουσιάζεται ένα μικρό ποσοστό (4,8%) αρνητικών εκτιμήσεων, το οποίο «διορθώνεται» με την ελαχιστοποίηση της συνάρτησης (11), υπό μη-αρνητικούς περιορισμούς. Εν τέλει, η εκτιμώμενη χρονοσειρά της διαφοράς των θετικών και αρνητικών αλμάτων που προκύπτει μετά από τη διόρθωση προσαρμόζεται πολύ ικανοποιητικά στις αποδόσεις του δείκτη NASDAQ για την περιόδο 2006-2008, όπως παρατηρείται στο Σχήμα 3, και επιβεβαιώνεται από την πολύ χαμηλή τιμή του MSE (4.97×10^{-16}).

Στο προτεινόμενο μοντέλο, οι εκ των προτέρων εκτιμήσεις (προβλέψεις) των αλμάτων, που δίνονται από τη σχέση (6), ικανοποιούν τους μη-αρνητικούς περιορισμούς για κάθε t. Αναφορικά με τις εκ των υστέρων εκτιμήσεις του διανύσματος κατάστασης, μπορεί σε επόμενη μελέτη να υιοθετηθεί μια μη γραμμική σχέση ως εξίσωση (8), προκειμένου να μηδενιστεί το ποσοστό των αρνητικών εκτιμήσεων. Αυτό θα οδηγήσει στη δημιουργία νέων επαναληπτικών σχέσεων για το στάδιο της πρόβλεψης και της επικαιροποίησης.

ABSRTACT

An appropriate non-linear discrete-time state space model is proposed for the estimation of the two-sided jumps of asset returns, which are modeled as hidden random variables. The model adopts a positive definite quadratic form for the evolution of the hidden jump components, due to the non negativity restrictions that they must fulfil. The recursive relations of the estimation procedure are calculated via the 2nd order Kalman filter. Finally, the fitting of the difference between the two-sided jumps to the empirical returns of NASDAQ index during 2006-2008 period is examined.

ΑΝΑΦΟΡΕΣ

Black, F. and Scholes, M., (1973). The Pricing of Options and Corporate Liabilities. *J. Polit. Econ.* **81**, 637-659.

Cont, R. and Tankov, P. (2004). *Financial Modelling with Jump Processes,* Chapman & Hall/CRC.

Eraker, B. (1976). Do Stock Prices and Volatility Jump? Reconciling Evidence from Spot to Option Prices. *J. Finance* **59**, 1367-1403.

Gupta, N. and Hauser, R. (2007). Kalman Filtering with Equality and Inequality State Constraints. arXiv:0709.2791.

Kalman, R. (1960). A new approach to linear filtering and prediction problems. *J. Basic Engineering (Transactions of the ASME Series D* **82**, 35-45.

Kotz, S., Balakrinshnan, N. and Johnson, N.L. (2000). *Continuous Multivariate Distributions*, New York: John Wiley.

Koopman, S.J., Mallee, M. and Wel, M. (2010). Analysing the Term Structure of Interest Rates using the Dynamic Nelson-Siegel Model with Time-Varying Parameters. *J. of Business and Economic Statistics* **28**, 329-343.

Pan, J. (2002). The jump-risk premia implicit in options: Evidence from an integrated time-series study. *J. Financ. Econ.* **63**, 3-50.

Polimenis, V. (2012). Information arrival as price jumps. *Optimization* **61**, 1179-1190.

Schoutens, W. (2003). *Lévy Processes in Finance: Pricing Financial Derivatives,* Wiley.

Simon, D. (2010). Kalman filtering with state constraints: a survey of linear and nonlinear algorithms. *IET Control Theory Appl.* **4**, 1303-1318.

Simon, D and Simon D.L. (2010). Constrained Kalman filtering via density function truncation for turbofan engine health estimation. *Int. J. of Systems Science* **41**, 159-171.

Wells, C. (1996). *The Kalman filter in Finance,* Springer.

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 31ου Πανελληνίου Συνεδρίου Στατιστικής (2018), σελ. 063–067

ΕΛΕΓΧΟΙ F ΟΜΟΣΚΕΔΑΣΤΙΚΟΤΗΤΑΣ

Θεόφιλος Κάκουλλος
Τμήμα Μαθηματικών, Πανεπιστήμιο Αθηνών
tkakoul@math.uoa.gr

ΠΕΡΙΛΗΨΗ

Στη σύντομη αυτή ανασκόπηση ελέγχων ομοσκεδαστικότητας, επισημαίνεται ότι, **ακριβείς έλεγχοι** F είναι γνωστοί μόνο για δυο τυχαίες μεταβλητές x και y, για ανεξάρτητες με το γνωστό F, τον λόγο των δειγματικών διασπορών, και για συσχετισμένες ($\rho \neq 0$) βάσει του ελέγχου Pitman-Morgan, μέσω ελέγχου ανεξαρτησίας. Υπό τη γενικότερη k-διάστατη κανονική (πολυκανονικότητα), για $k \geq 3$ δεν είναι δυνατή η κατασκευή ελέγχου τύπου Pitman-Morgan. Έλεγχοι ομοσκεδαστικότητας για k ανεξάρτητες μεταβλητές (k ομάδες ή πληθυσμούς) δεν υπάρχουν. Εδώ παραθέτουμε δύο εύχρηστους, του Bartlett και του Levene.

Λέξεις Κλειδιά: Ομοσκεδαστικότητα, Έλεγχοι F, Pitman-Morgan, Bartlett, Levene.

1. ΕΙΣΑΓΩΓΗ

Είναι πολύ γνωστός ο έλεγχος F της υπόθεσης ομοσκεδαστικότητας δύο ανεξαρτήτων κανονικών τυχαίων μεταβλητών (τμ) x, y

$$H_0 : \sigma_1^2 = \sigma_2^2 \tag{1}$$

επί τη βάσει τυχαίου δείγματος (τδ) $x_{n_1} \ldots, x_{n_1}$ από την $N(\mu_1, \sigma_1^2)$ και τδ $y_{n_1}, \ldots, y_{n_2}$ από την $N(\mu_2, \sigma_2^2)$ με το κριτήριο

$$F = s_1^2/s_2^2 \sim F_{n_1-1, n_2-1}, \tag{2}$$

τον λόγο των δειγματικών διασπορών s_1^2, s_2^2 των σ_1^2, σ_2^2.

Σημ.: α) Για ευκολία, χρησιμοποιούμε μόνο μικρά γράμματα, τόσο για τις τυχαίες μεταβλητές, όσο και τις τιμές του δείγματος.

β) Το $X \sim t_n$ σημαίνει ότι η τμ X ακολουθεί την t_n κατανομή.

Για έλεγχο της (1) επί τη βάσει δείγματος n ζευγών (x_i, y_i), $i = 1, \ldots, n$, από διδιάστατη κανονική του (x, y) με άγνωστο συντελεστή συσχέτισης ρ, οι Pitman (1939) και Morgan (1939) κατέληξαν στο κριτήριο

$$t^* = \frac{\sqrt{n-2}\, r^*}{\sqrt{1 - r^{*2}}} \sim t_{n-2}, \tag{3}$$

όπου r^* είναι ο δειγματικός συντελεστής συσχέτισης, υπολογιζόμενος από τα ζεύγη (x_i^*, y_i^*) όπου

$$x_i^* = x_i + y_i \;\; \text{και} \;\; y_i^* = x_i - y_i, \;\; i = 1, \ldots, n \tag{4}$$

είναι ανεξάρτητα ($\rho = 0$), αν και μόνον αν (ανν) $\sigma_1^2 = \sigma_2^2$.

Ο μεν Pitman κατέληξε στο t^*, παρατηρώντας ότι τα

$$u_i = \frac{x_i}{\sigma_1} + \frac{y_i}{\sigma_2}, \;\; v_i = \frac{x_i}{\sigma_1} - \frac{y_i}{\sigma_2}, \;\; i = 1, \ldots, n$$

είναι ανεξάρτητα ($\rho = 0$), και από τη γνωστή t κατανομή του $r(u_i, v_i)$ έδωσε και διαστήματα εμπιστοσύνης για τον λόγο $\omega = \sigma_1^2/\sigma_2^2$, άρα και υπό την $H_0 : \sigma_1^2 = \sigma_2^2$. Ο Morgan όμως πρώτα υπολόγισε το κριτήριο του λόγου πιθανοφανειών λ για τον έλεγχο της (1) και άρα το r^* και το t^*. Κάνοντας χρήση της σχέσης $t - F_{n,n} \equiv F_{(n)}$, Cacoullos (1965), από το $t^* \sim t_{n-2}$ προέκυψε, Cacoullos (2001), το

$$F^* \equiv \frac{1 + r^*}{1 - r^*} \sim F_{(n-2)}, \tag{5}$$

2. ΕΛΕΓΧΟΙ ΟΜΟΣΚΕΔΑΣΤΙΚΟΤΗΤΑΣ ΥΠΟ ΠΟΛΥΚΑΝΟΝΙΚΟΤΗΤΑ

Στη σύντομη αυτή ανασκόπηση των ελέγχων ομοσκεδαστικότητας, εξετάζουμε την

$$H_0^* : \sigma_1^2 = \cdots = \sigma_k^2, \;\; \sigma_i^2 = Var(x_i), \;\; i = 1, \ldots, k \tag{6}$$

όταν το τυχαίο διάνυσμα $\boldsymbol{x} = (x_1, \ldots, x_k)$ ακολουθεί την k-διάστατη κανονική κατανομή $\boldsymbol{\mu}, \boldsymbol{\Sigma}$, όπου ο $k \times k$ πίνακας διασποράς $\boldsymbol{\Sigma} = (\sigma_{ij})$ της $\boldsymbol{x} = (x_1, \ldots, x_i)$ είναι αντιστρέψιμος ($|\boldsymbol{\Sigma}| \neq 0$) και

$$\sigma_{ii} = \sigma_i^2, \;\; \sigma_{ij} \neq 0, \;\; i \neq j, \;\; \text{για όλα τα} \;\; i, j = 1, \ldots, k. \tag{7}$$

Έστω τώρα τδ $\boldsymbol{x}_\alpha = (x_{1\alpha}, \ldots, x_{k\alpha})$, $\alpha = 1, \ldots, n$, από την $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Κατ΄ αναλογίαν του r^* της (3) μπορεί να υπολογισθεί ένας αντίστοιχος δειγματικός συντελεστής r_{ij}^* από τα ζεύγη $(x_{ij}^*(\alpha), y_{ij}^*(\alpha))$ όπου

$$x_{ij}^*(\alpha) = x_{i\alpha} + x_{j\alpha}, \;\; y_{ij}^*(\alpha) = x_{i\alpha} - x_{j\alpha}, \;\; \alpha = 1, \ldots, n$$

και τα $x_{ij}^*(\alpha)$, $y_{ij}^*(\alpha)$ είναι ανεξάρτητα ανν ισχύει η

$$H_{ij} : \sigma_i^2 = \sigma_j^2, \tag{8}$$

έτσι ώστε ο έλεγχος της (8) να ανάγεται στον έλεγχο ανεξαρτησίας των $x_{ij}^*(\alpha)$, και $y_{ij}^*(\alpha)$, άρα της H_{ij}, καταλήγοντας στον έλεγχο Pitman-Morgan με

$$t_{ij}^* = \frac{\sqrt{n-2}\, r_{ij}^*}{\sqrt{1 - r_{ij}^{*2}}} \sim t_{n-2} \tag{9}$$

και τον ισοδύναμο έλεγχο F με

$$F_{ij}^* = \frac{1 + r_{ij}^*}{i - r_{ij}^*} \sim F_{(n-2)}, \tag{10}$$

όπως στις (3) και (5).

Παρατηρούμε, λοιπόν, ότι η αποδοχή μιας H_{ij} μέσω του αντίστοιχου r_{ij}^* ή F_{ij}^* ισχύει χωριστά κάθε ζεύγος (i, j), $i, j = 1, \ldots, k$, συνεπαγόμενη συγχρόνως και την αποδοχή της ανεξαρτησίας των x_i και x_j. Αλλά, δυστυχώς, ακόμα κι αν δεχόμασταν όλες τις H_{ij}, χωριστά για κάθε (i, j), αυτό δεν θα αρκούσε για την αποδοχή της $H_0^a st$, ότι και οι k διασπορές σ_i^2 είναι ίσες μεταξύ τους. Διότι, απλούστατα τα r_{ij}^* δεν είναι ανεξάρτητα, άρα ούτε τα t_{ij}^* ή τα F_{ij}^*. Για παράδειγμα, και για $k = 3$, έχουμε τους $r_{12}^*, r_{13}^*, r_{23}^*$ (με αντίστοιχες $t^* \sim t_{n-2}$, $F^* \sim F_{(n-2)}$ κατανομές και μάλιστα η αποδοχή δύο H_{ij}, π.χ. των H_{12} και H_{13} δεν συνεπάγεται την ομοσκεδαστικότητα

$$\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma^2, \tag{11}$$

διότι τα δύο εμπλεκόμενα r_{12}^*, r_{13}^*, **δεν είναι ανεξάρτητα** (εξαρτώμενα από τα $x_{1\alpha}$). Επομένως, δεν διαθέτουμε κριτήριο-συνάρτηση των r_{12}^*, r_{13}^* για έλεγχο της

$$H_{123} : H_{12} \quad \text{και} \quad H_{13}, \tag{12}$$

μη γνωρίζοντας την από κοινού κατανομή των μη ανεξάρτητων r_{12}^*, r_{13}^*. Το συμπέρασμα λοιπόν είναι ότι για $k \geq 3$, δεν φαίνεται να είναι δυνατή η αναγωγή ελέγχου της (6) σε έλεγχο ανεξαρτησίας δύο συσχετισμένων κανονικών τμ, άρα και κατασκευή ελέγχου τύπου Pitman-Morgan.

Στο πλαίσιο αυτό, οι Mathai & Moschopoulos (2001) απέδειξαν ότι για $k \geq 3$ υπό τις (6) και (7) δεν υπάρχουν δύο ανεξάρτητοι γραμμικοί συνδυασμοί των $x_1, \ldots, x_k$, αλλά στην περίπτωση ομοσυσχετικότητας, δηλαδή

$$\sigma_{ij} = \rho\sigma^2, \quad i \neq j, \quad i, j = 1, \ldots, k, \quad \left(-\frac{1}{k-1} < \rho < 1 \right),$$

π.χ. για $k = 3$, τα u και v,

$$u = x_1 + x_2 + x_3, \quad v = x_1 + x_2 - 2x_3$$

είναι ανεξάρτητα ανν ισχύει η (12), και επομένως ισχύει ο έλεγχος Pitman-Morgan, μέσω του r_{uv} από τα ζεύγη $(u_1, v_1), \ldots (u_n, v_n)$, άρα και οι έλεγχοι $t_{uv} \sim t_{n-2}$ και $F_{uv}^* \sim F_{(n-2)}$.

Για την υπόθεση (6) και $k \geq 3$ δεν φαίνεται να υπάρχει ακριβής έλεγχος F (ή ισοδύναμος, μέσω β-κατανομής) .

3. ΕΛΕΓΧΟΙ ΟΜΟΣΚΕΔΑΣΤΙΚΟΤΗΤΑΣ k ΟΜΑΔΩΝ

Όπως σημειώσαμε εισαγωγικά, ακριβείς έλεγχοι της

$$H_0' : \sigma_1^2 = \cdots = \sigma_k^2 = \sigma^2, \quad \sigma_i^2 = Var(y_i), \quad i = 1, \ldots, k, \tag{13}$$

όπου η τμ y_i αντιπροσωπεύει τον πληθυσμό ή ομάδα G_i, $i = 1, \ldots, k$, δεν έχουν εμφανισθεί στη στατιστική βιβλιογραφία. Επειδή όμως είναι πολύ χρήσιμοι στη στατιστική μεθοδολογία και εφαρμογές, ιδιαίτερα στην Ανάλυση Διασποράς (ANOVA) και Ανάλυση Παλινδρόμησης (Regression Analysis), γενικότερα στα Γραμμικά Μοντέλα (Linear Models), όπου, κατά κανόνα, υποτίθεται ομοσκεδαστικότητα, εδώ παραθέτουμε τους πιο γνωστούς και εύχρηστους προσεγγιστικούς ελέγχους της (13), βάσει k τδ, το $(y_{i1}, \ldots, y_{in_i})$ από την G_i, $i = 1, \ldots, k$, $j = 1, \ldots, n_i$.

Ο Bartlett (1937), για τον έλεγχο της (13) πρότεινε το

$$\chi^2 = \frac{(N-k)\ln(s_p^2) - \sum_{i=1}^{k}(n_i - 1)\ln(S_i^2)}{1 + \frac{1}{3(k-1)}\Big(\sum_{i=1}^{k}\Big(\frac{1}{n_i - 1}\Big) - \frac{1}{N-k}\Big)} \tag{14}$$

όπου s_p^2 η **ενωμένη** ή **κοινή** (pooled) δειγματική διασπορά

$$N = \sum_{i=1}^{k} n_i, \quad s_p^2 = \frac{1}{N-k}\sum_{i=1}^{k}(n_i - 1)S_i^2, \quad S_i^2 = \sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i)^2,$$

και έδωσε την προσέγγιση $\chi^2 \sim \chi_{k-1}^2$, ώστε η (13) να απορρίπτεται όταν $\chi^2 > \chi_{k-1}^2(\alpha)$ το άνω α σημαντικότητας σημείο της χ_{k-1}^2.

Ο έλεγχος του Bartlett, είναι τροποποίηση του ελέγχου λόγου πιθανοφανειών υπό κανονικότητα με σκοπό την καλύτερη χ_{k-1}^2 προσέγγιση. Συνιστάται, λοιπόν, όταν τα δεδομένα είναι κοντά στην κανονικότητα, αλλά δεν είναι ευαίσθητος, όχι τόσο ανθεκτικός (robust) έναντι αποκλίσεων από την κανονικότητα.

Εναλλακτικός έλεγχος του χ^2 της (14) είναι του Levene (1960), ανθεκτικός σε αποκλίσεις από την κανονικότητα, και βασίζεται στο κριτήριο

$$W = \frac{(N-k)\sum_{i=1}^{k} n_i(\overline{z}_i - \overline{z})^2}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{n_i}(z_{ij} - \overline{z}_i)^2}, \quad \overline{z} = \frac{1}{k}\sum_{i=1}^{k}\overline{z}_i, \quad \overline{z} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i} z_{ij} \tag{15}$$

όπου $z_{ij} = |y_{ij} - \overline{y}_i|$, $\overline{y}_i = \dfrac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$. Ο προσεγγιστικός έλεγχος Levene απορρίπτει την (13) όταν $W > F_{k-1,n-k}(\alpha)$, όπου $F_{k-1,n-k}(\alpha)$ το άνω α σημείο της $F_{k-1,n-k}$.

ABSTRACT

In this short review of tests of homoscedasticity, it is pointed out that **exact** F (or equivalent) tests are known only for two normal rv's, x and y, if independent using the well-known F, the ratio of the sample variances, and if correlated ($\rho \neq 0$) using the Pitman-Morgan test. Under a k-variate normal, it is shown that for $k \geq 3$ it is not possible to construct Pitman-Morgan type tests. For the k-population homoscedasticity hypothesis, Bartlett's and Levene's approximate tests are given.

Ευχαριστίες: Ευχαριστώ τον συνάδελφο του Πανεπιστημίου Πειραιά κ. Γ. Ηλιόπουλο, ο οποίος μου ανέφερε την εργασία των Mathai & Moschopoulos.

REFERENCES

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Statistical Society, Series A* **60**, 268-282.

Cacoullos, T. (1965). A relation between t- and F-distributions. *Journal of the American Statistical Association* **60**, 179-182.

Cacoullos, T. (2001). The F-test of homoscedasticity for correlated normal variables. *Statistics and Probability Letters* **60 (1)**, 1-3.

Levene, H. (1960). Robust tests for equality of variances. *Contributions to Probability and Statistics*, Essays in Honor of Harold Hotelling (Ingram Olkin, Harold Hotelling, et al, eds), Stanford University Press, 278-297.

Mathai, A. M. and Moschopoulos, P. G. (2001). Testing Hypotheses on Variance in the Presence of Correlations. *Probability and Statistical Models with Applications*, 403-418. (Ch. A. Charalambides and M. V. Koutras, N. Balakrishnan, eds. Chapman & Hall).

Morgan, W. A. (1939). A test for the significance of the difference between the two variances in a sample from a normal bivariate population. *Biometrika* **31**, 13-19.

Pitman, E. J. G. (1939). A note on normal correlation. *Biometrika* **31**, 9-12.

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 31ου Πανελληνίου Συνεδρίου Στατιστικής (2018), σελ.068-077

Η ΕΝΝΟΙΑ ΤΟΥ ΤΥΧΑΙΟΥ ΚΑΙ ΤΗΣ ΠΙΘΑΝΟΤΗΤΑΣ 600 π.Χ.-1800 μ.Χ.

Σ. Κουνιάς
Ομότιμος καθηγητής ΕΚΠΑ
skounias@math.uoa.gr

ΠΕΡΙΛΗΨΗ

Στην εργασία αυτή γίνεται αναφορά στην εξέλιξη της έννοιας του τυχαίου και της πιθανότητας στο διάστημα 600 π.Χ-1800 μ.Χ . Από σημαντικούς στατιστικούς επισημαίνεται ότι η έννοια αυτή δεν αναπτύχθηκε όπως οι άλλες επιστήμες στην αρχαία Ελλάδα . Αυτό αποδίδεται στην ισχυρή επίδραση της θρησκείας, που απέδιδαν στις θεότητες όσα δεν ήταν συγκεκριμένα και κατανοητά. Πέρασαν 2000 χρόνια για να γίνουν τα πρώτα βήματα στον τομέα αυτό.

Λέξεις Κλειδιά: Αρχαίοι Έλληνες Μαθηματικοί, τυχερά παιχνίδια, 16ος αιώνας.

1. ΕΙΣΑΓΩΓΗ

Οι αρχαίοι Έλληνες είχαν κάνει μεγάλες προόδους στα μαθηματικά. Προηγήθηκαν όλων των χωρών της δύσης, όμως δεν είχαν εξετάσει την έννοια του τυχαίου και δεν είχαν ασχοληθεί με την πιθανότητα των αποτελεσμάτων στα τυχερά παιχνίδια.

Οι σπουδαιότεροι με συμβολή στα μαθηματικά ήταν,

- Θαλής ο Μιλήσιος (640-580 π.Χ),
- Πυθαγόρας ο Σάμιος (585-496 π.Χ.),
- Ευκλείδης (300 π.Χ.-;), θεωρείται ο πατέρας της Γεωμετρίας,
- Αρχιμήδης ο Συρακούσιος (287 -212 π.Χ.),
- Διόφαντος ο Αλεξανδρεύς (290-210 π. Χ.), ο πατέρας της Άλγεβρας.
- Ερατοσθένης (276-194 π.Χ), μέτρησε με ακρίβεια το μήκος της περιφέρειας τη γης.

- Πλάτωνας (427-347 π.Χ), μαθητής του Σωκράτη και δάσκαλος του Αριστοτέλη, ίδρυσε την ακαδημία το 387 π.Χ., δίδαξε μαθηματικά ως κλάδο της φιλοσοφίας και είχε πολλούς μαθητές ικανούς στα μαθηματικά, όπως ο Εύδοξος (390-337 π. Χ.). Η ακαδημία του διατηρήθηκε σχεδόν δέκα αιώνες, είναι το αρχαιότερο ίδρυμα ανώτερης εκπαίδευσης.

1.1 **Η έννοια του τυχαίου και της πιθανότητας.**

Η Πιθανότητα άρχισε ως εμπειρική επιστήμη και κατέληξε σε Μαθηματική επιστήμη. Με την εξέλιξη της έννοιας του τυχαίου και της Πιθανότητας ασχολήθηκαν γνωστοί επιστήμονες όπως,

M.G. Kendal (1956), P.C. Mahalanobis (1957), S. Sambursky (1956), F.N. David (1955). Ο Finley (1970) αναφέρει ότι οι Έλληνες ενδιαφέρονταν, σε όλα τα θέματα, για την ποιότητα και όχι για την ποσότητα, το ίδιο και ο Karl Marx στο Kapital (1867, Vol.I) που είχε μελετήσει αρχαίους Έλληνες συγγραφείς.

- Σωκράτης (469-399 π.Χ.): Τα μη αναμενόμενα γεγονότα στη ζωή μπορεί να μην οφείλονται στις θεότητες αλλά σε άλλες αιτίες που εμείς δεν γνωρίζουμε.

Δικάσθηκε σε θάνατο (με κώνειο) και πολλοί θεωρούν ότι το κατηγορητήριο ήταν ότι δικάσθηκε διότι διέφθειρε τους νέους, ό,τι και να σημαίνει αυτό. Αυτή η αντίληψη είναι παραπλανητική διότι εκκλησία και πολιτεία δεν αναφέρουν ότι η κατηγορία ήταν ότι διαφθείρει τους νέους διδάσκοντά τους να μην πιστεύουν στους θεούς (Πλάτων, η απολογία του Σωκράτη σελ. 41).

- Αριστοτέλης (384-322 π.Χ.): Το τυχαίο είναι αντίθετο με τη σκέψη και τη λογική και οι αιτίες του δεν είναι δυνατόν να ερμηνευτούν.
- Ιπποκράτης (460 -370 π.Χ.): Πατέρας της ιατρικής. Τα πειράματα δεν χρειάζονται διότι οι θεοί γνωρίζουν το αποτέλεσμα του πειράματος.

Οι αρχαίοι Έλληνες δεν είχαν ξεκαθαρίσει την έννοια του τυχαίου, δεν συνέδεσαν την πιθανότητα με τα μαθηματικά, θεωρούσαν ότι οι θεοί καθορίζουν τις τύχες μας (μοίρα).

Ό,τι δεν μπορούσαν να ερμηνεύσουν, το απέδιδαν στις θεότητες, για κάθε θέμα υπήρχε η αντίστοιχη θεότητα. Αυτό είχε διάρκεια 2000 χρόνια.
Ο Mahalanobis (1957) συνδέει την ινδική φιλοσοφία Jaina (500-400 π.Χ.) με τις πιθανότητες.

Η έννοια της σχετικής συχνότητας έγινε γνωστή μετά το 1400 μ.Χ, (Με την επανάληψη πολλές φορές του πειράματος η σχετική συχνότητα πλησιάζει την πιθανότητα του ενδεχόμενου (Pearson and Kendall 1970). Οι παίχτες όμως τυχερών παιχνιδιών είχαν εμπειρία από τη συχνότητα εμφάνισης των ενδεχομένων.

1.2 **Blaise Pascal (1623-1662) , Pierre de Fermat (1601-1665).**

Ο Pascal ήταν παιδί θαύμα, η μητέρα του πέθανε όταν ο Blaise ήταν 3 ετών, ο πατέρας του ήταν έμπορος και ερασιτέχνης μαθηματικός. Επέβλεψε την κατ' οίκον εκπαίδευση του υιού του, του δίδαξε τα στοιχεία του Ευκλείδη. Ο Blaise 12 ετών απέδειξε ότι το άθροισμα των γωνιών ενός τριγώνου είναι 2 ορθές και 16 ετών έκανε πραγματεία για τις κωνικές τομές. Σε μεγάλη ηλικία ασχολήθηκε με τη θεολογία και τα μαθηματικά.

Ο Fermat ήταν νομικός και ερασιτέχνης μαθηματικός, παντρεύτηκε δύο φορές, είχε 5 παιδιά. Έθεσε το πρόβλημα: Δεν υπάρχουν τρεις θετικοί ακέραιοι a, b, c που ικανοποιούν τις σχέσεις

$$a^n + b^n = c^n,\ n \geq 3$$

Για n=2 η λύση είναι οι Πυθαγόρειες τριάδες (3,4,5),(5,12,13),…

Μετά από 358 χρόνια έδωσε τη λύση το 1994 ο Βρετανός μαθηματικός Sir Andrew Wallis (1953-) έλαβε αρκετά βραβεία και το βραβείο Abel το 2016 (650000 euros). Δεκάδες εργασιών έχουν δημοσιευθεί σε αυτό το πρόβλημα, είναι γνωστό ως "Fermat last theorem".

Ένα άλλο πρόβλημα , που είχε απασχολήσει τους μαθηματικούς, χωρίς επιτυχία, ήταν:

Παίζουν οι α και β ένα παιχνίδι και κερδίζει το ποσό Χ όποιος κερδίσει πρώτος ν παιχνίδια. Διακόπτεται ο αγώνας, του α μένουν ακόμη να κερδίσει k παιχνίδια και του β μένουν να κερδίσει r παιχνίδια.

Πως θα μοιραστεί το ποσό Χ στους α και β;

Θεωρείται ότι οι δύο παίκτες είναι ισοδύναμοι και ότι τα παιχνίδια είναι ανεξάρτητα μεταξύ τους, η πιθανότητα να κερδίσει ένα παιχνίδι ο α είναι ίση με ½, το ίδιο για τον β. Αν P(A) είναι η πιθανότητα να κερδίσει τελικά ο α, τότε η σωστή λύση είναι:

$$P(A)=\frac{1}{2^{k+r-1}}\sum_{i=k}^{k+r-1}\binom{k+r-1}{i}=\sum_{j=k}^{k+r-1}\binom{j-1}{k-1}\frac{1}{2^j}$$

Ο Pascal έδωσε την πρώτη λύση και ο Fermat τη δεύτερη. Είναι ισοδύναμες.

Οι λύσεις περιγράφηκαν με λόγια και παραδείγματα, τους τύπους έδωσε ο Montmort το 1713. Η σκέψη του Pascal ήταν: Σε k+r-1 παιχνίδια θα έχει οριστεί ο νικητής. Κατά Fermat ο α κερδίζει στο παιχνίδι k ή στο k+1,…, ή στο k+r-1.

Παράδειγμα 1. k=5, r=2, κερδίζει ο α όταν,

- Pascal: αααααα, αααααβ, αααaβα, αααβαα, ααβαaα, αβαaaα, βαaaaα
- Fermat: αααaα, αααaβα, αααβαα, ααβαaα, αβαaaα, βαaaaα
- $P(A)=7/64$.

Περισσότερα περιέχονται στις εργασίες του Χ. Χαραλαμπίδη, στα πρακτικά του ΕΣΙ 2017, 2018. Είναι λάθος να θεωρούμε ότι η θεωρία Πιθανοτήτων άρχισε από τους Pascal-Fermat διότι είχαν προηγηθεί άλλοι, όπως Cardano και Galileo-Galilei όπως αναφέρεται παρακάτω.

2. ΕΡΕΥΝΗΤΕΣ ΠΟΥ ΑΣΧΟΛΗΘΗΚΑΝ ΜΕ ΤΟ ΠΡΟΒΛΗΜΑ ΤΩΝ PASCAL-FERMAT

- Luca de Bongo Pacioli (1445-1517):

Έγραψε το βιβλίο «Suma de arithmetica, geometria, proportioni et proportionalita».
Το 1475 έγινε καθηγητής Μαθηματικών στο παν/μιο της Perugia.
Για 7 χρόνια δίδαξε Γεωμετρία στον Leonardo da Vinci.
150 χρόνια πριν των Pascal και Fermat, το 1494 εξέτασε την περίπτωση του στοιχήματος με $\nu=6, k=1, r=4$. Έδωσε τη (λάθος) λύση, το στοίχημα να μοιραστεί ανάλογα με τις νίκες των α, β.

- Gerolano Cardano (1501-1576).

Παράνομος γιος, γεννήθηκε στην Pavia και γράφτηκε στο Πανεπιστήμιο της Pavia, έλαβε διδακτορικό στην ιατρική το 1525 από το Πανεπιστήμιο της Padua, ο πατέρας του ήθελε να σπουδάσει ο γιος του νομικός. Ήταν δύστροπος, φανατικός παίκτης τυχερών παιχνιδιών, παντρεμένος και είχε τρία παιδιά.

Το 1543 έγινε καθηγητής ιατρικής στο πανεπιστήμιο της Pavia και το 1562 έγινε καθηγητής ιατρικής στο πανεπιστήμιο Bologna. Ήταν ο σημαντικότερος μαθηματικός της εποχής του και περιζήτητος γιατρός.

Το 1539, στο βιβλίο του Practica Arithmetice et Mensurandi Singularis, απέρριψε τη λύση του Pacioli ως περιλαμβάνουσα λάθη.

Ο ίδιος πρότεινε το στοίχημα να μοιραστεί ανάλογα με: $r(r+1)/2,\ k(k+1)/2$ χωρίς απόδειξη, αυτό δημοσιεύθηκε στο *15* σελίδων βιβλίο του για τα τυχερά παιχνίδια Liber de Ludo Aleae, 1663.

Αναφέρει ότι το συμμετρικό ζάρι με 6 πλευρές είναι ποιο δίκαιο.

Δημοσίευσε 191 εργασίες, γνώριζε το νόμο των μεγάλων αριθμών, χρησιμοποίησε πρώτος τους αρνητικούς αριθμούς, F. David (1955) και στα βιβλία F. David (1962), O. Ore, (1963).

- Niccolo Fontana (Tartaglia) (1499-1557):

Μετάφρασε τα Στοιχεία του Ευκλείδη, από τα ελληνικά στα ιταλικά το 1543, ήταν η πρώτη μετάφραση σε Ευρωπαϊκή γλώσσα.
Έδωσε στον Cardano τη λύση της εξίσωσης τρίτου βαθμού.
1556 -1560 εκδόθηκε το βιβλίο του General Trattato di Numeri et Measure, ήταν το καλύτερο στα μαθηματικά τον $16^{ο}$ αιώνα. Απέρριψε τη λύση του Pacioli ως «ούτε δικαιολογημένη ούτε καλή», η δική του λύση όπως και άλλων σύγχρονων Ιταλών μαθηματικών ήταν λανθασμένες. Έδωσε τον όγκο του τετράεδρου.
Το 1542 τραυματίστηκε σοβαρά στο σαγόνι από έναν στρατιώτη, όταν ο βασιλιάς Louis XII κατέλαβε τη Βενετία. Μετά δεν μιλούσε καθαρά. Είχε διαφορές με τον Cardano για την πρωτοτυπία της λύσης της εξίσωσης τρίτου βαθμού.

- Galileo-Galilei (1564-1642):

1592 (28 ετών) καθηγητής μαθηματικών-Φυσικής στο πανεπιστήμιο της Piza.

1609 κατασκεύασε τηλεσκόπιο και ασχολήθηκε με την αστρονομία.

Του έθεσαν το πρόβλημα, ρίχνουμε τρία ζάρια μια φορά, ποια είναι η πιθανότητα να έλθει άθροισμα 11 ή 12. Οι περιπτώσεις είναι,

i. 641,632,551,542,533,443, για το 11, τότε 6+6+3+6+3+3=27
ii. 651,642,633,552,543,444 για το 12, τότε 6+6+3+3+6+1=25

Έδωσε τη σωστή λύση (27+25)/216=52/216. Το ίδιο συμβαίνει με τα αθροίσματα *10 ή 9.* Τη δημοσίευσε στο βιβλίο του Sopra le Scoperte dei Dadi. Το ίδιο πρόβλημα είχε λύσει ο Cardano έναν αιώνα νωρίτερα.

Οι παίκτες τυχερών παιχνιδιών γνώριζαν εμπειρικά σε ποιο άθροισμα θα στοιχηματίσουν στο παιχνίδι της ρίψης 3 κύβων.

1632 (68 ετών) στο βιβλίο του γράφει, η γη κινείται γύρω από τον ήλιο. Ο Πάπας τον πέρασε ιερά εξέταση, αρνήθηκε όλες τις ανακαλύψεις του και φυλακίσθηκε δια βίου.

- Thomas Cataker (1574-1674):

Βρετανός, γνώριζε άριστα τα ελληνικά. Το1619 έγραψε το βιβλίο: Of the Nature and Use of Lots, στο οποίο αναφέρει ότι οι φυσικοί νόμοι και όχι οι θεϊκοί καθορίζουν ποιος κερδίζει στα τυχερά παιχνίδια.

- Jacob Bernoulli (1655-1705):

Ο πρώτος γιος της οικογένειας Ελβετών μαθηματικών. Οι αριθμοί Bernoulli φέρουν το όνομά του. Έλαβε πτυχίο Θεολογίας, ήταν η επιθυμία του πατέρα του, αυτός όμως ασχολήθηκε με τα μαθηματικά. Δημοσίευσε εργασίες σε μαθηματικά, πιθανότητες, μηχανική.

1687, καθηγητής Μηχανικής, πανεπιστήμιο Basel Ελβετίας, 32 ετών.

Το βιβλίο του Ars Conjectandy δημοσιεύθηκε το 1713. Ασχολήθηκε με τη συνδυαστική, το λογισμό μεταβολών, τις πιθανότητες. Τον διαδέχθηκε στην έδρα το 1705 ο μικρότερος αδελφός του Johann (1667-1748), ο οποίος απέδειξε το νόμο των μεγάλων αριθμών και με τον αδελφό του είχαν διαφωνίες στα μαθηματικά.

- Daniel Bernoulli (1700-1782):

Γιος του Johann Bernoulli, 13 ετών επήγε στο πανεπιστήμιο Basel να σπουδάσει φιλοσοφία και λογική, το 1916 πήρε το πτυχίο του, μετά σπούδασε ιατρική και 21 ετών έλαβε το διδακτορικό στην ιατρική.

1725-1734, καθηγητής μαθηματικών στο πανεπιστήμιο της Αγίας Πετρούπολής στη Ρωσία. Έκανε σημαντική εργασία στις πιθανότητες. Η κατανομή Bernoulli πήρε το όνομά του. Ο Leonard Euler πήγε εκεί το 1927 και συνεργάστηκε με τον Daniel στα μαθηματικά. Εφάρμοσε στατιστική στην κίνηση των υγρών.
1750-1776, καθηγητής Φυσικής, πανεπιστήμιο Basel.

- Christiaan Huygens (1629-1695):

Ολλανδός μαθηματικός, αστρονόμος, φυσικός.
Ανακάλυψε το ρολόι εκκρεμές και την κυματική θεωρία του φωτός.
1657, έγραψε το πρώτο επιστημονικό βιβλίο για τις πιθανότητες.

Πίστευε ότι δεν γνωρίζουμε τίποτα με βεβαιότητα αλλά όλα με πιθανότητα.

- Pierre-Simon Laplace (1749 - 1827):

Το 1774, 1776 δημοσίευσε εργασίες στις πιθανότητες.

Μέλος της ακαδημίας επιστημών της Γαλλίας, Γερουσιαστής.

1812, έγραψε δύο τόμους Analytical Theory of Probabilities. Έδωσε τον ορισμό της πιθανότητας.

Ορισμός. Η πιθανότητα ενός ενδεχομένου είναι το πηλίκο των ευνοϊκών γι' αυτό περιπτώσεων προς τον αριθμό των δυνατών περιπτώσεων, όταν τίποτα δεν μας κάνει να πιστεύουμε ότι κάποια από τις περιπτώσεις μπορεί να εμφανιστεί συχνότερα από οποιαδήποτε άλλη, έτσι όλες οι περιπτώσεις αυτές να θεωρούνται εξίσου δυνατές (ισοπίθανες).

Αυτό τον ορισμό χρησιμοποιούμε και σήμερα για δειγματοχώρους με πεπερασμένο αριθμό ισοπίθανων σημείων.

Έγραψε 5 τόμους Mechanique celeste, τις έκανε δώρο στον Ναπολέοντα.

- Abraham de Moivre (1667-1774):

Μελέτησε ελληνικά για 4 χρόνια. Μαθηματικός και αστρονόμος. Ασχολήθηκε με πιθανότητες, και στατιστική, το 1718 έγραψε το βιβλίο: The doctrine of Chances.
Κατασκεύασε τον πίνακα της κανονικής κατανομής, στους Stirling και De Moivre αποδίδεται ο τύπος για την προσέγγιση του n. Δημοσίευσε αρκετές εργασίες στα μαθηματικά και τις πιθανότητες. Έζησε κάνοντας φροντιστήρια σε μαθητές.
Το 1685 φυλακίστηκε στη Γαλλία ως προτεστάντης, μετά την αποφυλάκισή του διέφυγε στο Λονδίνο, έγινε φίλος με τον Isaak Newton και τον αστρονόμο Edmund Halley.

- Carl Friedrich Gauss (1777-1855):

Θεωρείται ο σπουδαιότερος μαθηματικός από την αρχαιότητα. Έκανε σημαντική συμβολή στη θεωρία αριθμών, άλγεβρα, γεωφυσική, μηχανική, αστρονομία, στατιστική.
Καθηγητής αστρονομίας, πανεπιστήμιο Gottingen (1807), 30 ετών.

Χρησιμοποίησε τη μέθοδο ελαχίστων τετραγώνων για να ελαχιστοποιήσει τα σφάλματα μέτρησης (1809). Είχε 6 παιδιά, οι γονείς του ήταν φτωχοί, η μητέρα του ήταν αγράμματη, δεν ήταν θρήσκος.

- Adrian-Marie Legendre (1752-1833):

Δημοσίευσε τη μέθοδο ελαχίστων τετραγώνων το 1805. Ο Gauss είπε ότι χρησιμοποιούσε τη μέθοδο από το 1794-1795. Ακολούθησε αλληλογραφία Laplace-Lagrange (1992), κατέληξαν στα δικαστήρια για την πρωτοτυπία.

3.ΕΠΙΛΟΓΗ ΠΡΟΣΩΠΙΚΟΥ ΜΕ ΚΛΗΡΩΣΗ

Αν και οι αρχαίοι Έλληνες δεν ασχολήθηκαν με την έννοια του τυχαίου και της πιθανότητας είχαν επιλέξει να διορίζουν σε υπεύθυνες θέσεις, με κλήρωση, προσωπικό που είχε τα προσόντα.

- Σόλων ο Αθηναίος (640-560 π.Χ.):

Φιλόσοφος, ποιητής, ένας από τους επτά σοφούς της εποχής του.

Το 594/3 π.Χ., εκλέχτηκε άρχων της Αθήνας.

Εισήγαγε τη σεισάχθεια, δηλαδή κατάργησε τα χρέη των φτωχών και αγροτών, διότι όσοι χρωστούσαν γίνονταν τότε δούλοι. Υποχρέωσε όλους να ασχολούνται με τα κοινά. Περιόρισε την κοινωνική αδικία, έκανε και άλλες αλλαγές.

Χώρισε τους Αθηναίους σε 4 κοινωνικές τάξεις (φυλές).

Όλοι οι άρχοντες, στρατηγοί, δικαστές και οι υποψήφιοι για διάφορες θέσεις εκλέγονταν με κλήρο από ένα σώμα 500 (πεντακοσιομέδιμνων).

Ίδρυσε τη βουλή των 400, 100 από κάθε φυλή, με ετήσια θητεία και οι βουλευτές λογοδοτούσαν στην εκκλησία του δήμου.

Ακολούθησε η 30/ετής τυραννία του Πεισίστρατου (540 π.Χ-510 π.Χ.).

- Κλεισθένης ο Αθηναίος (570 π.Χ.-):

Ανέλαβε το 510 π.Χ. Το 508 εξορίσθηκε και το 507 επανήλθε στη Αθήνα θριαμβευτής. Μετέτρεψε τις 4 φυλές σε 10, κάθε χρόνο εκλέγονταν από κάθε φυλή 50 βουλευτές και έγινε η βουλή των 500.

Ο επώνυμος άρχοντας εκλεγόταν από τους 500 με κλήρο και άλλαζε κάθε μέρα. Επίσης οι στρατηγοί και άλλοι υπάλληλοι εκλέγονταν με κλήρο.

Το 507 μ.Χ. ο Κλεισθένης εξαφανίστηκε και είναι άγνωστο πότε πέθανε.

- Εφιάλτης (Άγνωστη η καταγωγή του)

Το 462 π.Χ. εκδηλώθηκε η αντιπαράθεση αριστοκρατικών και δημοκρατικών στην Αθήνα και το κίνημα του Εφιάλτη για την κατάλυση της εξουσίας των γαιοκτημόνων ανέλαβε την εξουσία του δήμου.

Ο Εφιάλτης ήταν φτωχός και άσημος και είναι άγνωστη η καταγωγή του.
Έδωσε πολιτικά δικαιώματα στους Αθηναίους, κυβερνούσαν αυτοί που επιλέγονταν με κλήρο, δολοφονήθηκε το 459 π.Χ.
Τον τέταρτο αιώνα οι στρατηγοί, δικαστές, εφοριακοί κλπ, επιλέγονταν με κλήρο .
Η εγκαθίδρυση της Δημοκρατίας στην Αθήνα, επί Σόλωνα, Κλεισθένη και Εφιάλτη είναι, χωρίς αμφιβολία, ένα από τα σημαντικότερα γεγονότα στην ιστορία της ανθρωπότητας.

4. **ΤΥΧΕΡΑ ΠΑΙΧΝΙΔΙΑ**

Αν και δεν ασχολήθηκαν στην αρχαιότητα με τις πιθανότητες και το τυχαίο, ασχολήθηκαν με τα τυχερά παιχνίδια. Είναι άγνωστο από πότε οι άνθρωποι άρχισαν να ασχολούνται με τυχερά παιχνίδια.
Τυχερά παιχνίδια υπήρχαν από την ομηρική εποχή, κατά τον Ηρόδοτο μας τα γνώρισε η Αίγυπτος. Στη διάρκεια του Τρωϊκού πολέμου οι στρατιώτες έπαιζαν τυχερά παιχνίδια για να περνάει ο χρόνος τους.
Χρησιμοποιούσαν τον αστράγαλο από σώμα ανθρώπου ή ζώου. Έχει τέσσερις πλευρές που τις ονόμαζαν 4, 3, 1, 6 και από αργότερες μελέτες η πιθανότητα εμφάνισης κάθε πλευράς είναι 0.4,0.4,0.1,0.1. Συνήθως έριχναν 4 αστράγαλους μαζί και η πιθανότητα να εμφανιστεί η άνω πλευρά και στους 4 αστράγαλους έχει υπολογιστεί σε 1/39. Ο κύβος ή ζάρι εμφανίστηκε αργότερα, οι αρχαιολόγοι βρήκαν το αρχαιότερο ζάρι στο Ιράκ και στην Ινδία. Η F. N. David (1955) δοκίμασε τα ζάρια που φυλάσσονται στο Μουσείο του Λούβρου και του Λονδίνου και παρουσιάζει πίνακες με τις αποδόσεις τους.
Τα παιχνίδια με χαρτιά (τράπουλες) άρχισαν μετά την ανακάλυψη της τυπογραφίας (1450 μ.Χ.)

ABSTRACT

In this manuscript a reference is made to the notion of randomness and probability in the time interval l 600 π.Χ. -1800 μ.Χ. From known statisticians comments were made that this notion was not developed in Ancient Greece, as in other scientific areas. One answer is that the ancient Greeks were mainly interested on everything for the quality and not the quantity. Another explanation is that this is due to the strong influence of religious beliefs that whatever is not understood is attributed to deities. It took two thousand years until progress was made to this scientific area.

ΑΝΑΦΟΡΕΣ

David, F. N., (1955) Dicing and Gaming (A note on the history of Probability). *Biometrika Vol. 42, 1-15.*

David, F.N., (1962) Games, Gods and Gambling. London, Charles Griffin and Co.

Kendall, M.G.,(1956) Studies in the history of Probability and Statistics, *Biometrika, vol. 43, parts 1,2.*

Finley, M.I., (1970) Aristotle and Economic Analysis, *Past and Present, Vol.47,1, May 1970,σελ.3-25. Oxford University Press.*

Mahalanobis, P. C. (1957). The foundations of statistics. *Sankhya 18, 183-194.*

Ore, O. (1963) *Cardano, the Gambling Scholar*, Princeton University Press.

Πλάτων. *Η απολογία του Σωκράτη, ΚΑΚΤΟΣ 1992*.

Pearson, E. S., and Kendall, M.G., (1970) *Studies in the History of Statistics and Probability*. Charles Griffin and Co.

Sambursky, S., et al. (1956) The physical world of the Greeks. *Journal of Philosophy Vol. 54, 25, pp.813- 822.*

Wallis, Andrew. (1995) Modular Elliptic Curves and Fermat's Last Theorem. *Annals of Mathematics, Vol. 141, No. 3, pp. 443-551.*

Χαραλαμπίδης, Χ., (2016) Αναθεώρημα του ισοδύναμου των δύο λύσεων των Pascal και Fermat στο κλασικό πρόβλημα της διανομής μεριδίων. *Πρακτικά Ελληνικού Στατιστικού Ινστιτούτου (ΕΣΙ), σελ.203-209.*

Χαραλαμπίδης, Χ., (2017) Γένεση Θεωρίας Πιθανοτήτων: Cardano εναντίον Pascal και Fermat. *Πρακτικά Ελληνικού Στατιστικού Ινστιτούτου (ΕΣΙ).*

Ευχαριστίες Ευχαριστώ τους δύο κριτές για τα σχόλιά τους, βελτίωσαν το περιεχόμενο της εργασίας.

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 31ου Πανελληνίου Συνεδρίου Στατιστικής (2018), σελ.078-089

ΔΗΜΙΟΥΡΓΙΑ ΝΕΑΣ ΔΙΔΙΑΣΤΑΤΗΣ ΚΑΤΑΝΟΜΗΣ ΜΕ ΓΑΜΜΑ ΚΑΙ ΒΗΤΑ ΠΕΡΙΘΩΡΙΕΣ ΜΕ ΤΗ ΜΕΘΟΔΟ ΤΗΣ TRIVARIATE REDUCTION

Μ.Β. Κούτρας, Θ.Κ. Κοκκινιώτη
Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης, Πανεπιστήμιο Πειραιά
mkoutras@unipi.gr, theodora_kokkinioti@hotmail.gr

ΠΕΡΙΛΗΨΗ

Στη διεθνή βιβλιογραφία υπάρχουν διάφοροι τρόποι δημιουργίας διδιάστατων κατανομών, ένας εκ των οποίων είναι η μέθοδος της trivariate reduction. Σύμφωνα με αυτή τη μέθοδο, ξεκινώντας από τρεις ανεξάρτητες τυχαίες μεταβλητές και παίρνοντας κάποιον κατάλληλο μετασχηματισμό αυτών ώστε να δημιουργηθεί συσχέτιση, προκύπτουν δύο νέες τυχαίες μεταβλητές οι οποίες είναι εξαρτημένες. Οι περιθώριες των διδιάστατων κατανομών που δημιουργούνται με αυτή τη μέθοδο, ανήκουν συνήθως στην ίδια οικογένεια μονοδιάστατων κατανομών. Στην παρούσα εργασία θα παρουσιασθεί μία νέα διδιάστατη κατανομή, η οποία έχει κατασκευαστεί με τη μέθοδο της trivariate reduction, όμως η επιλογή των μετασχηματισμών έγινε, ώστε οι περιθώριες να διαφέρουν και, συγκεκριμένα, να ακολουθούν η μία την κατανομή Γάμμα και η άλλη την κατανομή Βήτα.

Λέξεις Κλειδιά: trivariate reduction, διδιάστατες κατανομές, στοχαστική εξάρτηση

1. ΕΙΣΑΓΩΓΗ

Οι τρόποι δημιουργίας διδιάστατων κατανομών ποικίλλουν (Balakrishnan and Lai, 2009) και έχουν ιδιαίτερη σημασία στην επιστημονική κοινότητα καθώς χρησιμοποιούνται όταν χρειάζεται να μοντελοποιήσουμε δύο χαρακτηριστικά τα οποία έχουν κάποιο είδος εξάρτησης. Μερικές από αυτές τις μεθόδους είναι η μέθοδος της trivariate reduction, η μέθοδος δημιουργίας από κοινού συνάρτησης πυκνότητας με τη χρήση δεσμευμένων κατανομών, η χρήση των Copulas (Nelsen, 2006) κ.ά. Αυτό που παρατηρείται στις περισσότερες μεθόδους (εξαιρούνται τα Copulas), είναι ότι οι διδιάστατες κατανομές που κατασκευάζονται έχουν ίδιες περιθώριες κατανομές, γεγονός το οποίο περιορίζει τη χρησιμότητά τους, αφού αρκετά συχνά στην πράξη δεν παρατηρείται το φαινόμενο αυτό.

Υπάρχει, επομένως, αρκετό ενδιαφέρον να βρεθούν μοντέλα τα οποία θα προσαρμόζονται στην περίπτωση που τα υπό εξέταση χαρακτηριστικά ακολουθούν διαφορετικές κατανομές. Ποια κατανομή μπορεί να χρησιμοποιηθεί, αν, για παράδειγμα, το ένα χαρακτηριστικό ήταν ο χρόνος ζωής ενός μηχανήματος και το άλλο χαρακτηριστικό εξέφραζε ένα ποσοστό; Ένα γνωστό παράδειγμα από τη διεθνή βιβλιογραφία είναι το παράδειγμα με τη διάρκεια των ξηρασιών (Nadarajah, 2009). Στο παράδειγμα αυτό παρουσιάστηκε μία κατανομή με διαφορετικές περιθώριες κατανομές, όπου η μία μεταβλητή εξέφραζε το πλήθος των ξηρασιών και η δεύτερη μεταβλητή το ποσοστό των συμβάντων ξηρασίας.

Στην παρούσα εργασία θα παρουσιασθεί ένα καινούργιο μοντέλο στο οποίο οι περιθώριες έχουν κατασκευαστεί με τη μέθοδο της trivariate reduction και η μία θα ακολουθεί κατανομή Γάμμα ενώ η δεύτερη την κατανομή Βήτα. Η διαφορά με το μοντέλο του Nadarajah έγκειται στο γεγονός ότι έχουν εισαχθεί τρεις αρχικές τυχαίες μεταβλητές αντί για δύο, επομένως, το καινούργιο μοντέλο αναμένεται να είναι πιο ευέλικτο.

Ακολούθως, δίνεται η δομή της εργασίας. Στη δεύτερη ενότητα, παρουσιάζεται η μέθοδος της trivariate reduction, η οποία θα χρησιμοποιηθεί για την κατασκευή της καινούργιας κατανομής. Στην τρίτη ενότητα θα δοθούν εκφράσεις για τη συνάρτηση πυκνότητας και τη συνάρτηση κατανομής του νέου μοντέλου, τις δεσμευμένες συναρτήσεις πυκνότητας και κατανομής καθώς και τη δεσμευμένη μέση τιμή $E(X|Y=y)$. Τέλος, στην τέταρτη ενότητα αναφέρονται πρόσθετα ερευνητικά θέματα που έχουν ενδιαφέρον να εξετασθούν στο μέλλον.

2. ΜΕΘΟΘΟΣ ΤΗΣ TRIVARIATE REDUCTION

Όπως προδιαθέτει και το όνομά της, η μέθοδος της trivariate reduction χρησιμοποιείται ώστε από τρεις αρχικές μεταβλητές να καταλήξουμε σε δύο. Αρχικά έχουμε τρεις ανεξάρτητες τυχαίες μεταβλητές X_1, X_2, X_3 οι οποίες ακολουθούν κάποιες κατανομές με συναρτήσεις πυκνότητας f_1, f_2, f_3, αντίστοιχα. Κάνοντας τον εξής μετασχηματισμό:

$$X = g_1(X_1, X_2, X_3)$$
$$Y = g_2(X_1, X_2, X_3)$$

όπου g_1, g_2 κατάλληλες πραγματικές συναρτήσεις των τριών αρχικών μεταβλητών. Βασική προϋπόθεση για την εφαρμογή της μεθόδου είναι ότι τουλάχιστον ένα από τα X_1, X_2, X_3 είναι κοινό ώστε οι X, Y να έχουν κάποιο είδος εξάρτησης.

Ως παράδειγμα, ας υποθέσουμε ότι $X_1, X_2, X_3 \sim Gamma(\theta_i, \lambda),\ i = 1,2,3$. Δύο γνωστές στη διεθνή βιβλιογραφία κατανομές οι οποίες δημιουργήθηκαν με χρήση της παραπάνω μεθόδου είναι:

- η διδιάστατη Γάμμα κατανομή του Cheriyan (Cheriyan, 1941):

$$X = X_1 + X_3$$
$$Y = X_2 + X_3,$$

- η διδιάστατη κατανομή Βήτα του Jones (Jones, 2001, Olkin and Liu, 2003):

$$X = \frac{X_1}{X_1 + X_3}$$
$$Y = \frac{X_2}{X_2 + X_3}.$$

Η διδιάστατη Γάμμα κατανομή του Cheriyan είναι μια από τις πιο γνωστές διδιάστατες κατανομές με περιθώριες Γάμμα καθώς έχει μελετηθεί από πολλούς συγγραφείς και έχουν αποδειχθεί αρκετές ιδιότητές της. Επίσης, χρησιμοποιείται ευρέως γιατί ο τρόπος κατασκευής της διευκολύνει την πρακτική της εφαρμογή και έχει επεκταθεί σε $n > 2$ διαστάσεις. Ενδεικτικά αναφέρουμε το σύγγραμμα των Kotz, Balakrishnan and Johnson (2000).

Η διδιάστατη κατανομή Βήτα του Jones προϋποθέτει ότι η παράμετρος κλίμακος λ της Γάμμα κατανομής των αρχικών τυχαίων μεταβλητών πρέπει να είναι κοινή, ώστε οι περιθώριες να ακολουθούν κατανομή Βήτα. Στην εργασία του Jones (2001) παρουσιάζεται η σχέση της πολυδιαστατης αλλά και της διδιάστατης συνάρτησης πυκνότητας, ενώ οι Olkin and Liu (2003) επικεντρώθηκαν κυρίως στη διδιάστατη περίπτωση και στις ιδιότητές της.

Τα συγκεκριμένα παραδείγματα έχουν επιλεχθεί ανάμεσα σε άλλα καθώς αποτέλεσαν έμπνευση για τη δημιουργία της νέας κατανομής.

3. ΔΗΜΙΟΥΡΓΙΑ ΝΕΑΣ ΚΑΤΑΝΟΜΗΣ

Έστω $X_1, X_2, X_3 \sim Gamma(\theta_i, \lambda),\ i = 1,2,3$ τρεις ανεξάρτητες τυχαίες μεταβλητές. Η νέα κατανομή η οποία θα μελετηθεί είναι η διδιάστατη κατανομή του ζεύγους (X, Y) όπου τα X, Y ορίζονται μέσω των σχέσεων

$$X = X_1 + X_3 \sim Gamma(\theta_1 + \theta_3, \lambda)$$
$$Y = \frac{X_2}{X_2 + X_3} \sim B(\theta_2, \theta_3).$$

Στη συνέχεια παρουσιάζονται κάποια από τα θεωρητικά αποτελέσματα της καινούργιας κατανομής και δίνονται οι αποδείξεις, όπου αυτό κρίνεται αναγκαίο.

3.1 Από κοινού συνάρτηση πυκνότητας και από κοινού συνάρτηση κατανομής

Για το υπόλοιπο της εργασίας θα συμβολίζεται με ${}_1F_1$ η συνάρτηση του Kummer (Confluent Hypergeometric function), η οποία ορίζεται από τη σχέση

$${}_1F_1(a; b; z) = \sum_{n=0}^{\infty} \frac{(a)_n}{(b)_n} \frac{z^n}{n!}, \qquad (1)$$

όπου με $(x)_k = x(x+1)\cdots(x+k-1)$ συμβολίζονται τα ανοδικά παραγοντικά.

ΘΕΩΡΗΜΑ 1. Η από κοινού συνάρτηση πυκνότητας των X, Y δίνεται από τη σχέση

$$f(x,y) = \frac{x^{\theta_1+\theta_2+\theta_3-1} e^{-\frac{x}{\lambda}}}{\lambda^{\theta_1+\theta_2+\theta_3} B(\theta_2,\theta_3)\Gamma(\theta_1+\theta_2+\theta_3)} \frac{y^{\theta_2-1}}{(1-y)^{\theta_2+1}} \times$$

$$\times \ {}_1F_1\left(\theta_2+\theta_3;\theta_1+\theta_2+\theta_3,-\frac{xy}{\lambda(1-y)}\right)$$

όπου $x>0, y\in(0,1)$.

Απόδειξη. Χρησιμοποιώντας το μετασχηματισμό

$$\begin{cases} x=x_1+x_3=g_1(x_1,x_3) \\ y=\dfrac{x_2}{x_2+x_3} \\ z=x_3 \end{cases} \Rightarrow \begin{cases} x_1=x-z=g_1^*(x,z) \\ x_2=\dfrac{yz}{1-y}=g_2^*(y,z) \\ x_3=z \end{cases}$$

η Ιακωβιανή ορίζουσα που προκύπτει είναι ίση με

$$J=\frac{z}{(1-y)^2}>0.$$

Επομένως, η από κοινού συνάρτηση πυκνότητας των X,Y,Z θα δίνεται από τη σχέση

$$f_{X,Y,Z}(x,y,z)=|J|f_{X_1,X_2,X_3}(g_1^*(x,z),g_2^*(y,z),z)=$$
$$=\frac{z}{(1-y)^2}f_1(g_1^*(x,z))f_2(g_2^*(y,z))f_3(z),$$

όπου $x,z\in(0,\infty)$ με $z<x, y\in(0,1)$, και η από κοινού συνάρτηση πυκνότητας των X,Y μπορεί να υπολογισθεί από την έκφραση

$$f_{X,Y}(x,y)=\int_0^x \frac{z}{(1-y)^2}f_1(x-z)f_2\left(\frac{yz}{1-y}\right)f_3(z)dz=$$

$$=\frac{e^{-x/\lambda}}{\lambda^{\theta_1+\theta_2+\theta_3}\Gamma(\theta_1)\Gamma(\theta_2)\Gamma(\theta_3)}\frac{y^{\theta_2-1}}{(1-y)^{\theta_2+1}}\int_0^x (x-z)^{\theta_1-1}z^{\theta_2+\theta_3-1}e^{-yz/(\lambda(1-y))}dz,$$

όπου $x\in(0,\infty)$ και $y\in(0,1)$. Το τελευταίο υπολογίζεται με χρήση της σχέσης 3.383(1) των Gradshteyn and Ryzhik (2007), σύμφωνα με την οποία ισχύει

$$\int_0^u x^{\nu-1}(u-x)^{\mu-1}e^{\beta x}dx=B(\mu,\nu)u^{\mu+\nu-1}\ {}_1F_1(\nu;\mu+\nu;\beta u),$$

για $\nu,\mu>0$.

Αντικαθιστώντας την έκφραση αυτή στη σχέση

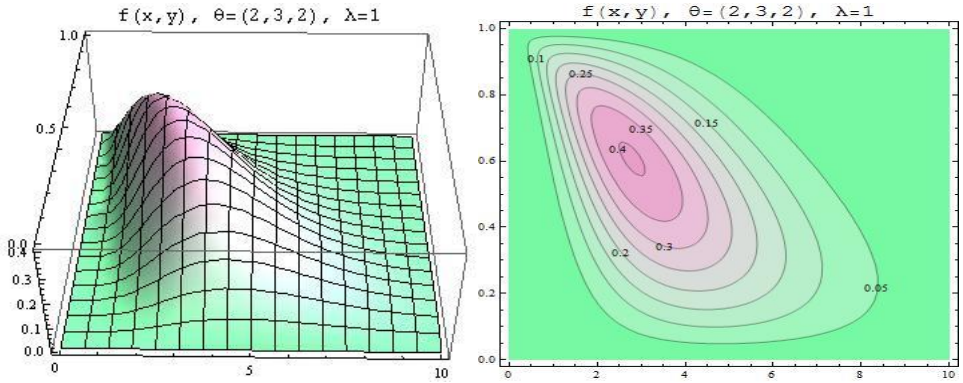
$$\int_0^x (x-z)^{\theta_1-1}z^{\theta_2+\theta_3-1}e^{-yz/(\lambda(1-y))}dz$$

προκύπτει το ζητούμενο.

Παρακάτω δίνονται ένα γράφημα της από κοινού συνάρτησης πυκνότητας $f(x,y)$ της νέας κατανομής και το αντίστοιχο contour plot. Θεωρούμε ότι οι αρχικές τυχαίες

μεταβλητές ακολουθούν την κατανομή Γάμμα με $\lambda = 1, \boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3) = (2,3,2)$. Η επιλογή του $\lambda = 1$ έγινε για απλοποίηση της από κοινού συνάρτησης πυκνότητας.

Σχήμα 1. *Από κοινού συνάρτηση πυκνότητας και το αντίστοιχο Contour Plot*



ΘΕΩΡΗΜΑ 2. Η από κοινού συνάρτηση κατανομής των X, Y δίνεται από τη σχέση

$$F(x,y) = \sum_{k=0}^{\infty} \frac{(-1)^k}{k!} \frac{\Gamma(\theta_2 + \theta_3 + k)}{\Gamma(\theta_2)\Gamma(\theta_3)\Gamma(\theta_1 + \theta_2 + \theta_3 + k)(\theta_2 + k)} \left(\frac{y}{1-y}\right)^{k+\theta_2} \times$$

$$\times \gamma\left(\frac{x}{\lambda}, \theta_1 + \theta_2 + \theta_3 + k\right)$$

όπου $x > 0, y \in (0,1)$.

Απόδειξη. Αναπτύσσοντας ως σειρά την υπεργεωμετρική συνάρτηση (1) και χρησιμοποιώντας τη σχέση

$$(x)_n = \frac{\Gamma(x+n)}{\Gamma(x)}, x > 0, x + n > 0, n \in \mathbb{R},$$

προκύπτει

$$F(x,y) = \int_0^x \int_0^y f(s,t)dtds$$

$$= \sum_{k=0}^{\infty} \frac{(-1)^k}{k!} \frac{\Gamma(\theta_2 + \theta_3 + k)}{\Gamma(\theta_2)\Gamma(\theta_3)\Gamma(\theta_1 + \theta_2 + \theta_3 + k)\lambda^{\theta_1+\theta_2+\theta_3+k}} \times$$

$$\times \int_0^x s^{\theta_1+\theta_2+\theta_3+k-1} e^{-s/\lambda} \left(\int_0^y \frac{t^{\theta_2+k-1}}{(1-t)^{\theta_2+k+1}} dt \right) ds.$$

Χρησιμοποιώντας τη μη πλήρη συνάρτηση Γάμμα, η οποία ορίζεται ως εξής

$$\gamma(x,a) = \int_0^x t^{a-1} e^{-t} dt$$

προκύπτει η ζητούμενη σχέση.

3.2 Οι δεσμευμένες συναρτήσεις πυκνότητας και οι δεσμευμένες συναρτήσεις κατανομής

Αρχικά δίνονται οι σχέσεις για τις δεσμευμένες συναρτήσεις πυκνότητας.

ΘΕΩΡΗΜΑ 3. Η δεσμευμένη συνάρτηση πυκνότητας της Y δοθέντος ότι $X = x$ δίνεται από τη σχέση

$$f_{Y|X}(y|x) = \frac{\Gamma(\theta_1 + \theta_3)}{B(\theta_2 + \theta_3)\Gamma(\theta_1 + \theta_2 + \theta_3)} \left(\frac{x}{\lambda}\right)^{\theta_2} \frac{y^{\theta_2 - 1}}{(1-y)^{\theta_2 + 1}} \times$$
$$\times \ {}_1F_1\left(\theta_2 + \theta_3; \theta_1 + \theta_2 + \theta_3; -\frac{xy}{\lambda(1-y)}\right),$$

όπου $x > 0, 0 < y < 1$.

Απόδειξη. Σύμφωνα με τον ορισμό της δεσμευμένης συνάρτησης πυκνότητας ισχύει

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)}$$

και, δεδομένου ότι $X \sim Gamma(\theta_1 + \theta_3)$, το ζητούμενο προκύπτει άμεσα.

ΘΕΩΡΗΜΑ 4. Η δεσμευμένη συνάρτηση πυκνότητας της X δοθέντος $Y = y$ δίνεται από τη σχέση

$$f_{X|Y}(x|y) = \frac{x^{\theta_1 + \theta_2 + \theta_3 - 1} e^{-x/\lambda}}{\Gamma(\theta_1 + \theta_2 + \theta_3)\lambda^{\theta_1 + \theta_2 + \theta_3}} (1-y)^{-\theta_2 - \theta_3} \times$$
$$\times \ {}_1F_1\left(\theta_2 + \theta_3; \theta_1 + \theta_2 + \theta_3; -\frac{xy}{\lambda(1-y)}\right),$$

όπου $x > 0, 0 < y < 1$.

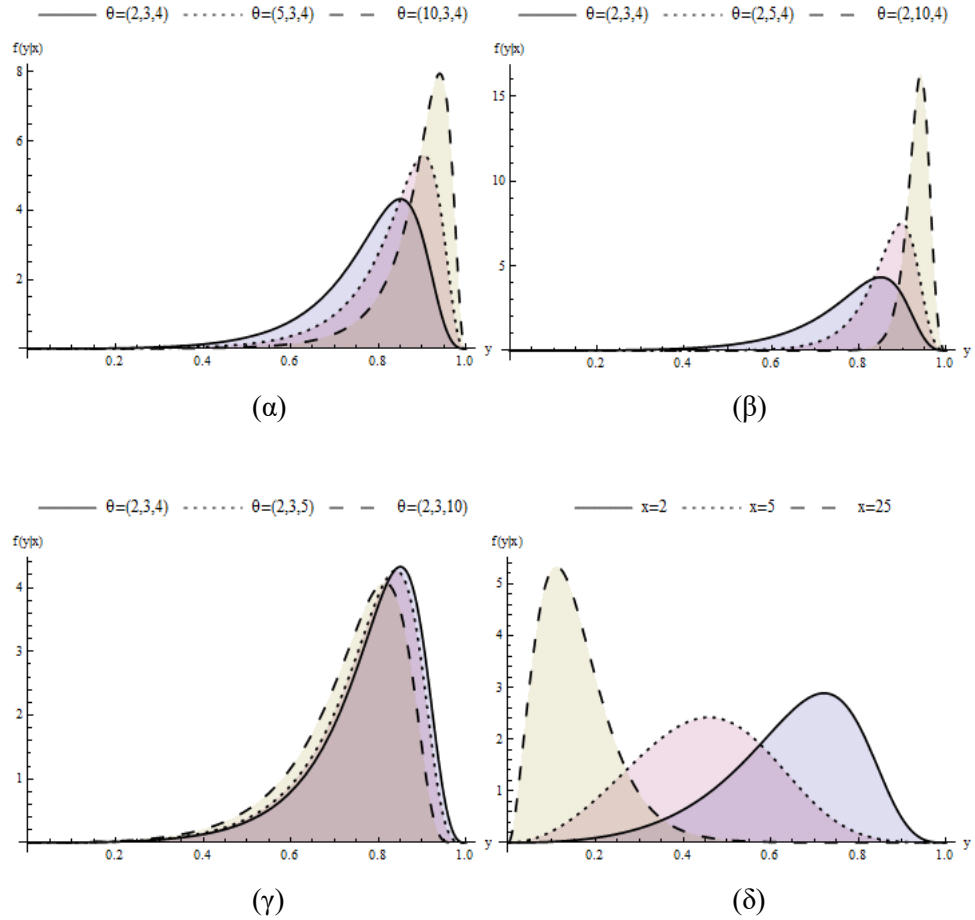
Απόδειξη. Σύμφωνα με τον ορισμό της δεσμευμένης συνάρτησης πυκνότητας ισχύει

$$f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)}$$

και, δεδομένου ότι $Y \sim B(\theta_2, \theta_3)$, το ζητούμενο προκύπτει άμεσα.

Παρακάτω δίνονται ορισμένα γραφήματα για τις δεσμευμένες συναρτήσεις πυκνότητας. Αρχικά δίνονται τα γραφήματα για την $f_{Y|X}(y|x)$. Για τα τρία πρώτα γραφήματα έχουμε θεωρήσει ότι $x = \lambda$, καθώς απλοποιείται η συνάρτηση πυκνότητας και χρησιμοποιούνται διαφορετικές τιμές για τις παραμέτρους $\theta_i, i = 1,2,3$, ενώ για το τέταρτο θεωρούμε σταθερά τα θ_i και αλλάζουμε την τιμή του x.

Σχήμα 2. *Γραφήματα για την* $f_{Y|X}(y|x)$



Στο γράφημα (α) παρατηρούμε ότι, αυξάνοντας την παράμετρο θ_1 και κρατώντας τις άλλες δύο παραμέτρους σταθερές, η κατανομή μετατοπίζεται προς τα δεξιά. Επομένως, παρουσιάζει πιο έντονη αρνητική ασυμμετρία. Επίσης, παρατηρούμε ότι η κατανομή γίνεται πιο λεπτόκυρτη και η τιμή της συνάρτησης στην κορυφή της αυξάνει.

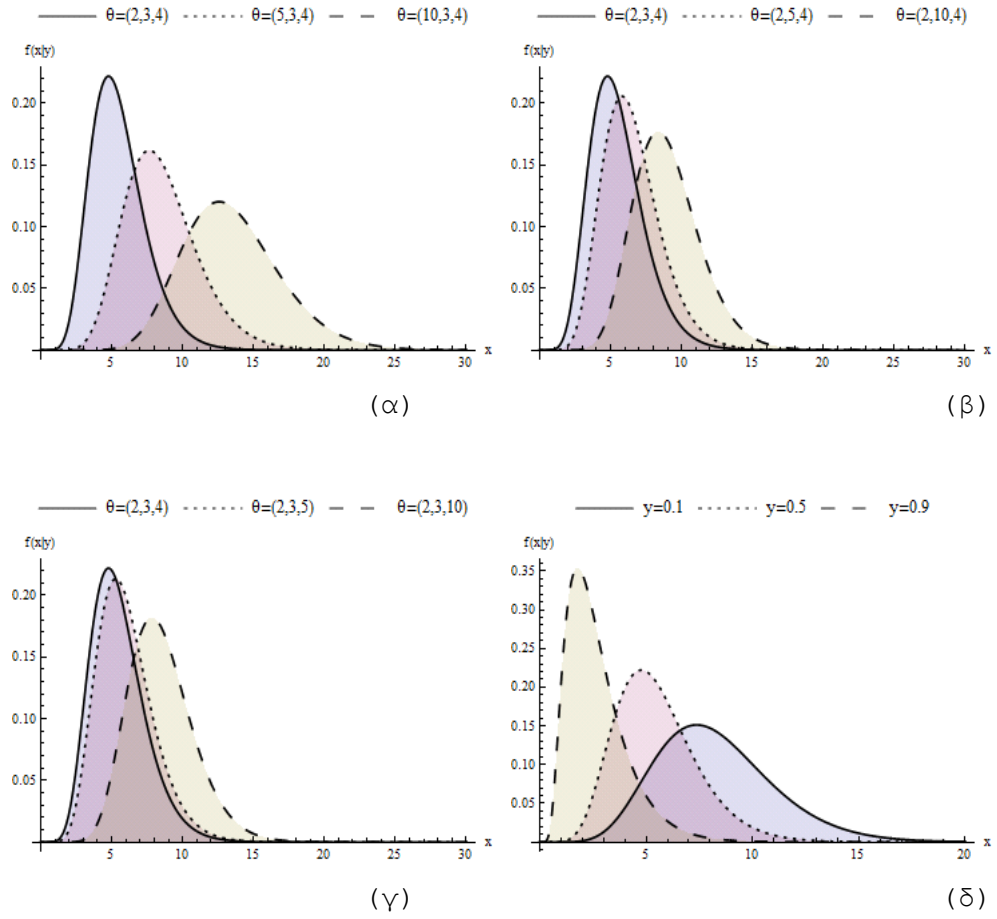
Στο γράφημα (β) παρατηρούμε παρόμοια συμπεριφορά της κατανομής με το πρώτο γράφημα. Αυξάνοντας, επομένως, την παράμετρο θ_2 κρατώντας τις άλλες δύο παραμέτρους σταθερές, η κατανομή έχει πιο έντονη αρνητική ασυμμετρία. Επίσης, γίνεται πιο λεπτόκυρτη και η τιμή της συνάρτησης στην κορυφή της κατανομής αυξάνει πιο γρήγορα.

Στο γράφημα (γ) παρατηρούμε ότι, αυξάνοντας την παράμετρο θ_3 και κρατώντας τις άλλες δύο παραμέτρους σταθερές, η κατανομή μετατοπίζεται ελαφρώς προς τα αριστερά. Παραμένει αρνητικά ασύμμετρη, αλλά ο ρυθμός της μετατόπισης καθώς και η τιμή της συνάρτησης στην κορυφή της κατανομής μεταβάλλονται με αργό ρυθμό.

Στο γράφημα (δ) έχουμε θεωρήσει ότι $\boldsymbol{\theta} = (2,3,4), \lambda = 1$ και αυξάνουμε την τιμή του x. Αυτό που παρατηρείται είναι ότι η κατανομή από θετικά ασύμμετρη φαίνεται να γίνεται συμμετρική και, αυξάνοντας και άλλο την x, γίνεται αρνητικά ασύμμετρη.

Στη συνέχεια δίνονται γραφήματα και για την $f_{X|Y}(x|y)$. Όπως και στα προηγούμενα γραφήματα, στα τρία πρώτα μεταβάλλονται οι παράμετροι $\theta_1, \theta_2, \theta_3$ θεωρώντας $y = 0.5$ και $\lambda = 1$, ενώ στο τέταρτο γράφημα θεωρούνται σταθερά τα θ_i, ενώ αλλάζει η τιμή του y.

Σχήμα 3. *Γραφήματα για την* $f_{X|Y}(x|y)$



Στα γραφήματα (α), (β), (γ) παρατηρούμε μια μετατόπιση προς τα δεξιά της κατανομής, η οποία είναι θετικά ασύμμετρη. Στο γράφημα (α), στο οποίο αυξάνεται η παράμετρος θ_1 και διατηρούνται οι άλλες δύο σταθερές, παρατηρούμε ότι η κατανομή μετατοπίζεται με πιο γρήγορο ρυθμό σε σύγκριση με το δεύτερο και το τρίτο γράφημα, στα οποία έχουμε υποθέσει ότι οι θ_2, θ_3 αυξάνονται, αντίστοιχα, θεωρώντας τις εναπομείνασες παραμέτρους σταθερές.

Στο γράφημα (δ), έχουμε θεωρήσει ότι $\boldsymbol{\theta} = (2,3,4), \lambda = 1$ και παρατηρούμε ότι αυξάνοντας την y η κατανομή μετατοπίζεται προς τα αριστερά. Επίσης, ενώ γίνεται πιο λεπτόκυρτη, ενδέχεται για μεγαλύτερες τιμές των θ_i να γίνεται και αυτή συμμετρική. Παρ' όλα αυτά, για τις συγκεκριμένες τιμές των παραμέτρων η κατανομή είναι θετικά ασύμμετρη.

Εν συνεχεία, δίνονται οι σχέσεις για τις δεσμευμένες συναρτήσεις κατανομής.

ΘΕΩΡΗΜΑ 5. Η δεσμευμένη συνάρτηση κατανομής της Y δοθέντος $X = x$ δίνεται από τη σχέση

$$F_{Y|X}(y|x) =$$

$$= \sum_{k=0}^{\infty} \frac{(-1)^k}{k!} \frac{\Gamma(\theta_1 + \theta_3)\Gamma(\theta_2 + \theta_3 + k)}{\Gamma(\theta_2)\Gamma(\theta_3)\Gamma(\theta_1 + \theta_2 + \theta_3 + k)(\theta_2 + k)} \left(\frac{x}{\lambda}\right)^{\theta_2 + k} \left(\frac{y}{1-y}\right)^{\theta_2 + k},$$

όπου $x > 0, 0 < y < 1$.

Απόδειξη. Η δεσμευμένη συνάρτηση κατανομής της Y δοθέντος $X = x$ δίνεται από τη σχέση

$$F_{Y|X}(y|x) = \int_0^y f_{Y|X}(t|x)dt =$$

$$= \sum_{k=0}^{\infty} \frac{(-1)^k}{k!} \frac{\Gamma(\theta_1 + \theta_3)\Gamma(\theta_2 + \theta_3 + k)}{\Gamma(\theta_2)\Gamma(\theta_3)\Gamma(\theta_1 + \theta_2 + \theta_3 + k)} \left(\frac{x}{\lambda}\right)^{\theta_2 + k} \int_0^y \frac{t^{\theta_2 + k - 1}}{(1-t)^{\theta_2 + k + 1}} dt$$

και υπολογίζοντας το παραπάνω ολοκλήρωμα προκύπτει το ζητούμενο.

ΘΕΩΡΗΜΑ 6. Η δεσμευμένη συνάρτηση κατανομής της X δοθέντος $Y = y$ δίνεται από τη σχέση

$$F_{X|Y}(x|y) =$$

$$= \sum_{k=0}^{\infty} \frac{(-1)^k}{k!} \frac{\Gamma(\theta_2 + \theta_3 + k)}{\Gamma(\theta_2 + \theta_3)\Gamma(\theta_1 + \theta_2 + \theta_3 + k)} \frac{y^k}{(1-y)^{\theta_2 + \theta_3 + k}} \times$$

$$\times \gamma\left(\frac{x}{\lambda}, \theta_1 + \theta_2 + \theta_3 + k\right),$$

όπου $x > 0, 0 < y < 1$.

Απόδειξη. Η απόδειξη είναι παρόμοια με αυτήν του θεωρήματος 5.

Τέλος, δίνεται η δεσμευμένη μέση τιμή της X δοθέντος ότι $Y = y$.

ΘΕΩΡΗΜΑ 7. Η δεσμευμένη μέση τιμή της X δοθέντος ότι $Y = y$ δίνεται από τη σχέση

$$E(X|Y = y) = \sum_{k=0}^{\infty} \frac{(-1)^k}{k!} \frac{\lambda\Gamma(\theta_2 + \theta_3 + k)(\theta_1 + \theta_2 + \theta_3 + k)}{\Gamma(\theta_2 + \theta_3)} \frac{y^k}{(1-y)^{\theta_2+\theta_3+k}},$$

όπου $x > 0, 0 < y < 1$.

Απόδειξη. Η δεσμευμένη μέση τιμή δίνεται από τη σχέση

$$E(X|Y = y) = \int_{-\infty}^{\infty} x f_{X|Y}(x|y)dx =$$

$$\sum_{k=0}^{\infty} \frac{(-1)^k}{k!} \frac{\Gamma(\theta_2 + \theta_3 + k)}{\Gamma(\theta_2 + \theta_3)\Gamma(\theta_1 + \theta_2 + \theta_3 + k)} \frac{y^k}{\lambda^{\theta_1+\theta_2+\theta_3+k}(1-y)^{\theta_2+\theta_3+k}} \times$$

$$\times \int_0^{\infty} x^{\theta_1+\theta_2+\theta_3+k} e^{-x/\lambda} dx$$

και υπολογίζοντας το ολοκλήρωμα που εμφανίζεται στο άθροισμα προκύπτει άμεσα το ζητούμενο.

4. ΜΕΛΛΟΝΤΙΚΗ ΕΡΕΥΝΑ

Στην εργασία αυτή ορίσαμε μια καινούργια διδιάστατη κατανομή με Γάμμα και Βήτα περιθώριες κατανομές και παρουσιάσαμε ορισμένα από τα θεωρητικά αποτελέσματα που έχουν αποδειχτεί για τη νέα κατανομή. Ολοκληρώνοντας τη θεωρία, ο επόμενος στόχος μας είναι η διενέργεια στατιστικής συμπερασματολογίας. Δοθέντος ότι η σχέση της συνάρτησης πυκνότητας είναι αρκετά πολύπλοκη, δε φαίνεται εφικτό να βρεθούν κλειστοί τύποι οπότε θα καταφύγουμε σε επαναληπτικές μεθόδους.

Συγκεκριμένα, για την εκτίμηση των παραμέτρων της κατανομής, έχουμε επιλέξει να χρησιμοποιήσουμε μια αριθμητική μέθοδο εκτίμησης, γνωστή και ως expectation – maximization algorithm ή εν συντομία EM-αλγόριθμος (Dempster et al. 1977). Η μέθοδος αυτή είναι μια επαναληπτική μέθοδος η οποία χρησιμοποιείται όταν δεν μπορούν να προσδιοριστούν σε κλειστή αναλυτική μορφή οι εκτιμητές μέγιστης πιθανοφάνειας. Στη νέα κατανομή είναι εμφανές πως είναι δύσκολη έως αδύνατη η εφαρμογή της μεθόδου της μέγιστης πιθανοφάνειας, γι'αυτό το λόγο εργαζόμαστε πάνω στον EM-αλγόριθμο. Στη συνέχεια, αφού υλοποιηθεί ο αλγόριθμος, απομένει η εφαρμογή της κατανομής σε πραγματικά δεδομένα, ενδεχομένως συγκρίνοντας τη με άλλα υποψήφια μοντέλα. Τέλος, έπειτα από τη συγκριτική μελέτη με άλλες κατανομές θα γίνει μια προσπάθεια γενίκευσης σε περισσότερες διαστάσεις.

ABSTRACT

In the statistics literature there are several methods of constructing bivariate distributions, one of which is the trivariate reduction method. According to this method, starting from three independent random variables we construct suitable transformations in order to introduce some sort of dependence for the two new random variables. The marginals of the bivariate distributions created by this method, usually belong to the same family of univariate distributions. In this work we introduce a new bivariate distribution, by the aid of the trivariate reduction method, with the transformations being selected so that the marginals differ and, specifically, one follows the Gamma distribution and the other the Beta distribution.

ΑΝΑΦΟΡΕΣ

Balakrishnan N. and Lai C.-D., (2009). *Continuous Bivariate Distributions*, 2nd edition, Springer, New York.

Cheriyan K.C., (1941). A bivariate correlated gamma-type distribution function, *Journal of the Indian Mathematical Society*, **5**, 133-144.

Dempster A.P., Laird N.M. and Rubin D.B., (1977). Maximum likelihood from Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society, Series B*, **39**, No. 1, 1-38.

Gradshteyn I.S. and Ryzhik I.M., (2007). *Table of Integrals, Series and Products*, 7th edition, Academic Press, Elsevier.

Jones M.C., (2001). Multivariate t and the beta distributions associated with the multivariate F distribution, *Metrika*, **54**, 215-231.

Kotz L., Balakrishnan N. and Johnson N.L., (2000). *Continuous Multivariate Distributions, Volume 1: Models and Applications*, 2nd edition, John Wiley & Sons, New York.

Nadarajah S., (2009). A bivariate distribution with gamma and beta marginals with application to drought data, *Journal of Applied Statistics*, **36**, No. 3, 277-301.

Nelsen R.B., (2006). *An Introduction to Copulas*, 2nd edition, Springer, NewYork.

Olkin I. and Liu R., (2003). A bivariate beta distribution, *Statistics and Probability Letters*, **62**, 407-412.

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 31ου Πανελληνίου Συνεδρίου Στατιστικής (2018), σελ.090-104

Μέθοδοι δημιουργίας νέων οικογενειών κατανομών και μελέτη των ιδιοτήτων τους

Μ. Κούτρας, Ν. Μαυρούτσος
Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης, Πανεπιστήμιο Πειραιώς
mkoutras@unipi.gr, nikos_mavroutsos@hotmail.com

Περίληψη

Η δημιουργία νέων οικογενειών κατανομών και η μελέτη των ιδιοτήτων τους παρότι αποτέλεσε αντικείμενο μελέτης διαφόρων ερευνητών για πολλά χρόνια εξακολουθεί να βρίσκεται στο επίκεντρο του ερευνητικού ενδιαφέροντος. Στην παρούσα εργασία θα εισάγουμε μια νέα οικογένεια κατανομών της οποίας η αθροιστική συνάρτηση κατανομής δίνεται από τον τύπο $F(x) = g^{-1}\big(R(x)\big)$, όπου g και R είναι συναρτήσεις με συγκεκριμένες ιδιότητες η καθεμιά. Επιπλέον για τη συγκεκριμένη οικογένεια κατανομών θα παρουσιαστούν μια σειρά από ιδιότητες που είναι χρήσιμες για την αναζήτηση και δημιουργία συναρτήσεων g και R οι οποίες θα μπορούν να χρησιμοποιηθούν για την κατασκευή νέων μονοδιάστατων κατανομών

Λέξεις Κλειδιά: Οικογένειες κατανομών, μονοδιάστατες συνεχείς κατανομές, συναρτήσεις μετασχηματισμού, συναρτήσεις–γεννήτορες.

1. ΕΙΣΑΓΩΓΗ

Οι κατανομές αποτελούν εργαλεία της στατιστικής που χρησιμοποιούνται συνήθως για να περιγράψουν φαινόμενα του πραγματικού κόσμου. Λόγω της χρησιμότητάς τους μελετώνται λεπτομερώς από τους ερευνητές δημιουργώντας έτσι νέες κατανομές ενώ η ανάγκη της Στατιστικής για την ανάπτυξη πιο ευέλικτων στατιστικών κατανομών είναι συνεχής. Μέχρι σήμερα πολλές γενικευμένες ομάδες κατανομών έχουν αναπτυχθεί και εφαρμοστεί για την περιγραφή διαφόρων φαινομένων. Οι Johnson et al. (1994) αναφέρουν ότι η χρήση κατανομών με τέσσερεις παραμέτρους είναι επαρκής για τις περισσότερες εφαρμογές αμφισβητώντας παράλληλα κάθε αισθητή βελτίωση από τη χρήση μιας πέμπτης ή έκτης παραμέτρου. Ωστόσο παραδέχτηκαν ότι είναι απαραίτητες τουλάχιστον τρεις παράμετροι.

Μια σύντομη ιστορική αναφορά στη δημιουργία νέων οικογενειών κατανομών θα μπορούσε να ξεκινήσει με τη μέθοδο που αναπτύχθηκε από τον Pearson. Η οικογένεια συνεχών κατανομών η οποία δημιουργήθηκε από τον Pearson (1895) είναι ένα σύστημα κατανομών με συναρτήσεις πυκνότητας πιθανότητας (σ.π.π.) $f(x)$ οι οποίες αποτελούν λύσεις της διαφορικής εξίσωσης

$$\frac{1}{f(x)}\frac{df(x)}{dx} = \frac{a+x}{b_0 + b_1 x + b_2 x^2}$$

όπου α, b_0, b_1, b_2 είναι παράμετροι. Με δεδομένο ότι το σχήμα της σ.π.π. $f(x)$, εξαρτάται από τις προηγούμενες παραμέτρους ο Pearson ομαδοποίησε τις κατανομές αυτές σε υποσύνολα τα οποία ονόμασε «τύπους». Κάθε τύπος αντιστοιχεί σε διαφορετική λύση της παραπάνω εξίσωσης.

Σύστημα συνεχών κατανομών διαφόρων σχημάτων αναπτύχθηκε και από τον Burr (1942). Στην περίπτωση αυτή το σύστημα κατανομών περιέχει συναρτήσεις κατανομών F που ικανοποιούν τη διαφορική εξίσωση

$$\frac{dF(x)}{dx} = F(x)\big(1 - F(x)\big)g(x)$$

όπου $0 \leq F \leq 1$ και g μια μη αρνητική συνάρτηση του x. Ο Burr για τη συγκεκριμένη εξίσωση πρότεινε 12 λύσεις που εξαρτώνται από την επιλογή της συνάρτησης g.

Ο Azzalini (1985) εισήγαγε την λοξή (skew) κανονική οικογένεια κατανομών. Υποθέτοντας ότι X και Y είναι δύο ανεξάρτητες τυχαίες μεταβλητές με σ.π.π. συμμετρικές γύρω από το μηδέν, τότε ορίζεται, για κάθε $\lambda \in \mathbb{R}$, η

$$2f_Y(y)F_X(\lambda y)$$

ως η σ.π.π. της παραπάνω οικογένειας κατανομών. Εάν ειδικότερα οι τυχαίες μεταβλητές X και Y ακολουθούν την τυποποιημένη κανονική κατανομή τότε η σ.π.π. της λοξής κανονικής οικογένειας κατανομών είναι

$$2\varphi(x)\Phi(\lambda x)$$

όπου φ και Φ η σ.π.π. και η αθροιστική συνάρτηση κατανομής αντίστοιχα της τυποποιημένης κανονικής κατανομής $N(0,1)$. Αν και στον τελευταίο τύπο η κατανομή χαρακτηρίζεται μόνο από την παράμετρο λ θα μπορούσαν να εισαχθούν οι παράμετροι θέσης και κλίμακας με τη χρήση του μετασχηματισμού $Y = \mu + \sigma X$.

Η οικογένεια των beta generated κατανομών είναι μια οικογένεια που ορίστηκε από τους Eugene et al. (2002) με τη χρήση της Βήτα κατανομής ως γεννήτορα. Μια τυχαία μεταβλητή X που ακολουθεί μια beta generated κατανομή έχει αθροιστική συνάρτηση κατανομής της μορφής:

$$G(x) = \int_0^{F(x)} b(t)dt \qquad (1.1)$$

όπου $b(t)$ είναι η σ.π.π. μιας τυχαίας μεταβλητής που ακολουθεί τη Βήτα κατανομή και $F(x)$ είναι η αθροιστική συνάρτηση κατανομής μιας οποιασδήποτε τυχαίας μεταβλητής. Η αντίστοιχη σ.π.π. της κατανομής δίνεται από τον τύπο

$$g(x) = \frac{1}{B(a,\beta)} f(x) F^{a-1}(x)\big(1 - F(x)\big)^{\beta-1}.$$

Μετά τη δημοσίευση των Eugene et al. (2002) μελετήθηκαν αρκετές beta generated κατανομές όπως η beta-Gumbel κατανομή από τους Nadarajah and Kotz (2004), η beta-exponential κατανομή από τους Nadarajah and Kotz (2005), η beta-Weibull κατανομή από τους Famoye et al. (2005) και άλλες.

Με την αντικατάσταση της σ.π.π. $b(t)$ της Βήτα κατανομής στον τύπο (1.1) από τη σ.π.π. $b(x) = \alpha\beta x^{a-1}(1-x^a)^{\beta-1}$, $x \in (0,1)$, της κατανομής Kumaraswamy (1980) οι Jones (2009) και Cordeiro και de Castro (2011) επέκτειναν την beta generated οικογένεια κατανομών και δημιούργησαν την Kumaraswamy οικογένεια γενικευμένων κατανομών (KW-G) με σ.π.π.

$$g(x) = a\beta f(x) F^{a-1}(x)(1 - F(x)^{\alpha})^{\beta-1}.$$

Υπάρχουν αρκετές KW-G κατανομές που έχουν μελετηθεί. Αυτές είναι η Kumaraswamy Weibull κατανομή από τους Cordeiro et al. (2010), η Kumaraswamy generalized Γάμμα κατανομή από τους de Castro et al. (2011) και άλλες.

Περισσότερες λεπτομέρειες σχετικά με τις οικογένειες κατανομών που περιεγράφηκαν αναφέρονται από τους Lee C. et. al. (2013).

Αντίστροφη διαδικασία από αυτή που χρησιμοποιήθηκε για τη δημιουργία της οικογένειας των beta generated κατανομών ανέπτυξαν οι Koutras et. al. (2014). Πιο συγκεκριμένα, δημιούργησαν την Πολυωνυμική Λογιστική κατανομή χρησιμοποιώντας ότι ο γνωστός logit μετασχηματισμός της αθροιστικής συνάρτησης F της Λογιστικής κατανομής δηλαδή η συνάρτηση $ln\frac{F(x)}{1-F(x)}$, είναι γραμμική συνάρτηση του x. Η ιδιότητα αυτή γενικεύτηκε με τη χρήση ενός logit μετασχηματισμού πολυωνυμικού τύπου και δημιουργήθηκε η Πολυωνυμική Λογιστική κατανομή. Προφανώς η συγκεκριμένη οικογένεια κατανομών, αν το πολυώνυμο είναι μια γραμμική συνάρτηση, οδηγεί στην απλή Λογιστική κατανομή με δύο παραμέτρους.

Στόχος του παρόντος άρθρου είναι να παρουσιάσει μια νέα τεχνική για τη δημιουργία οικογενειών συνεχών κατανομών. Το άρθρο είναι δομημένο ως εξής: στην Ενότητα 2 παρουσιάζεται η νέα τεχνική δημιουργίας οικογενειών συνεχών κατανομών. Στην Ενότητα 3 περιγράφονται μέθοδοι για τη δημιουργία συναρτήσεων μετασχηματισμού και συναρτήσεων γεννητόρων αν είναι γνωστές τουλάχιστον δύο συναρτήσεις από την κάθε περίπτωση. Στην Ενότητα 4 παρουσιάζεται μια ολοκληρωμένη διαδικασία δημιουργίας οικογενειών κατανομών συγκεκριμένου

είδους σύμφωνα με την τεχνική που περιγράφεται. Το άρθρο ολοκληρώνεται στην Ενότητα 5 με τα συμπεράσματα της μελέτης.

2. Μέθοδος για τη δημιουργία νέων οικογενειών συνεχών κατανομών

Όπως αναφέρθηκε στην προηγούμενη ενότητα η Πολυωνυμική Λογιστική κατανομή όπως ορίσθηκε από τους Koutras et. al. (2014) απαιτεί τη χρήση πολυωνυμικής συνάρτησης ως γεννήτορα και συγκεκριμένης συνάρτησης μετασχηματισμού. Η τεχνική που εφαρμόστηκε για τη δημιουργία της Πολυωνυμικής Λογιστικής κατανομής εγείρει το ακόλουθο ερώτημα. «Είναι δυνατόν να αναπτυχθούν υποθέσεις που θα ορίζουν μια συνάρτηση μετασχηματισμού και υποθέσεις που θα ορίζουν μια συνάρτηση γεννήτορα ώστε να χρησιμοποιούνται σε μια ορισμένη διαδικασία για τη δημιουργία νέων οικογενειών κατανομών;». Στην ενότητα αυτή το συγκεκριμένο ερώτημα θα απαντηθεί θετικά με την εισαγωγή μιας νέας τεχνικής η οποία ουσιαστικά γενικεύει την μέθοδο που χρησιμοποιήθηκε από τους Koutras et. al. (2014).

Η νέα αυτή τεχνική στηρίζεται αρχικά στην παρατήρηση ότι πολλές γνωστές κατανομές με α.σ.κ. $F(x)$ έχουν τη δυνατότητα να γραφούν στη μορφή

$$g\big(F(x)\big) = R(x)$$

όπου $R(x) = ax + b$ μια γραμμική συνάρτηση, a, b παράμετροι της κατανομής και g μια συνάρτηση που ορίζεται στο διάστημα $[0,1)$ ή $(0,1)$. Στον Πίνακα 1 που ακολουθεί παρουσιάζονται γνωστές κατανομές που μπορούν να γραφούν στη μορφή

$$g\big(F(x)\big) = ax + b$$

Πίνακας 1. *Γνωστές κατανομές που μπορούν να γραφούν στην μορφή* $g \circ F = R$

Κατανομή	Παράμετροι	$F(x)$	$g(x)$	a	b
Εκθετική $x \geq 0$	$\lambda > 0$	$1 - e^{-\lambda x}$	$ln\left(\frac{1}{1-x}\right)$	λ	0
Λογιστική $-\infty < x < \infty$	$-\infty < \mu < \infty$ $\sigma > 0$	$\frac{1}{1+e^{-\frac{x-\mu}{\sigma}}}$	$ln\left(\frac{x}{1-x}\right)$	$\frac{1}{\sigma}$	$-\frac{\mu}{\sigma}$
Gumbel $-\infty < x < \infty$	$-\infty < \mu < \infty$ $\sigma > 0$	$\exp(-e^{-\frac{x-\mu}{\sigma}})$	$-ln(-lnx)$	$\frac{1}{\sigma}$	$-\frac{\mu}{\sigma}$
Weibull $x \geq 0$	$\lambda > 0\ k > 0$	$1 - e^{-(\lambda x)^k}$	$\left(-\ln(1-x)\right)^{\frac{1}{k}}$	λ	0
Pareto $x \geq c$	$c > 0\ a > 0$	$1 - \left(\frac{c}{x}\right)^{\alpha}$	$(1-x)^{-\frac{1}{a}}$	$\frac{1}{c}$	0
Gompertz $x \geq 0$	$\eta > 0,\ b > 0$	$1 - \exp(-\eta\ (e^{bx} - 1))$	$ln\left(\frac{1}{\eta} ln\frac{1}{1-x} + 1\right)$	b	0
Lomax $x \geq 0$	$\lambda > 0\ a > 0$	$1 - \left(1 + \frac{x}{\lambda}\right)^{-\alpha}$	$(1-x)^{-\frac{1}{a}} - 1$	$\frac{1}{\lambda}$	0

Exponentiated Inverted Weibull $x > 0$	$\beta > 0\ \theta > 0$	$\left(exp\left(-x^{-\beta}\right)\right)^{\theta}$	$\left(-\frac{1}{\theta} lnx\right)^{-\frac{1}{\beta}}$	1	0
Log – Logistic $x \geq 0$	$\alpha > 0\ \beta > 0$	$\frac{x^{\beta}}{\alpha^{\beta} + x^{\beta}}$	$\left(\frac{x}{1-x}\right)^{\frac{1}{\beta}}$	1	0

Περισσότερες λεπτομέρειες σχετικά με τις κατανομές με α.σ.κ. $F(x)$ που αναφέρονται στον Πίνακα 1 κάποιος θα μπορούσε να βρει στα Balahrishnan και Nevzorov (2003), Flaih et. al. (2012), Kotz and Nadarajah (2000).

Γενικεύοντας την προηγούμενη παρατήρηση, η ανάπτυξη της τεχνικής για τη δημιουργία νέων οικογενειών κατανομών θα βασισθεί στη διερεύνηση των ιδιοτήτων που πρέπει να έχει μια συνάρτηση $g: A \to \mathbb{R}$ με $A = [0,1)$ ή $A = (0,1)$ και των υποθέσεων που πρέπει να ικανοποιεί μια συνάρτηση $R: B \to \mathbb{R}$ με $B = \mathbb{R}$ ή $B = [0, \infty)$ έτσι ώστε η συνάρτηση

$$F = g^{-1} \circ R \tag{2.1}$$

που ορίζεται στο σύνολο B, να αποτελεί μία συνάρτηση κατανομής.

Ορισμός 1: Μια τυχαία μεταβλητή X με στήριγμα το διάστημα $[0, \infty)$ θα λέμε ότι ακολουθεί την κατανομή $D_g^+(R)$ με συνάρτηση μετασχηματισμού την παραγωγίσιμη συνάρτηση $g: [0,1) \to \mathbb{R}$ και γεννήτορα την παραγωγίσιμη συνάρτηση $R: [0, \infty) \to \mathbb{R}$ αν και μόνο αν η αθροιστική συνάρτηση κατανομής της τυχαίας μεταβλητής X είναι της μορφής

$$F(x) = g^{-1}\big(R(x)\big)$$

και ισχύουν οι ακόλουθες ιδιότητες για τις συναρτήσεις g και R:

1. g γνησίως αύξουσα συνάρτηση για κάθε $x \in [0,1)$.
2. R γνησίως αύξουσα συνάρτηση για κάθε $x \in [0, \infty)$.
3. $\lim_{x \to 0^+} g(x) = g(0)$, $\lim_{x \to 1^-} g(x) = \infty$.
4. $R(0) = g(0)$, $\lim_{x \to \infty} R(x) = \infty$.

Για την τυχαία μεταβλητή X του προηγούμενου ορισμού θα γράφουμε συμβολικά $X \sim D_g^+(R)$. Στον Πίνακα 2 που ακολουθεί παρουσιάζονται, ενδεικτικά, συναρτήσεις που ικανοποιούν τις υποθέσεις 2 και 4 του ορισμού και θα μπορούσαν να χρησιμοποιηθούν ως γεννήτορες για τη δημιουργία οικογενειών κατανομών μέσω του προηγούμενου ορισμού.

Πίνακας 2. *Γεννήτορες για τη δημιουργία κατανομών* $D_g^+(R)$

Γραμμική	$R(x) = \alpha x \qquad \alpha > 0$
Power	$R(x) = \alpha x^b \qquad \alpha > 0, b > 1$
Λογαριθμική	$R(x) = aln\left(1 + \frac{x}{\lambda}\right),\ \alpha, \lambda > 0$

Εκθετική	$R(x) = a\left(e^{bx} - 1\right) \quad a, b > 0$
Εκθετική – Λογαριθμική	$R(x) = ln\left(ln\left(e + b(e^{cx} - 1)\right)\right),\ b > 0, c > 0$
Power - Λογαριθμική	$R(x) = a\ ln(1 + bx^c)\ a > 0, b > 0, c > 0$

Αντίστοιχη οικογένεια κατανομών με στήριγμα ολόκληρο το $\mathbb{R}$ μπορεί να κατασκευασθεί ως εξής:

Ορισμός 2: Μια τυχαία μεταβλητή X με στήριγμα το $\mathbb{R}$ θα λέμε ότι ακολουθεί την κατανομή $D_g(R)$ με συνάρτηση μετασχηματισμού την παραγωγίσιμη συνάρτηση $g: A \to \mathbb{R}$ με $A = [0,1)$ ή $A = (0,1)$ και γεννήτορα την παραγωγίσιμη συνάρτηση $R: \mathbb{R} \to \mathbb{R}$ αν και μόνο αν η αθροιστική συνάρτηση κατανομής της τυχαίας μεταβλητής X είναι της μορφής

$$F(x) = g^{-1}\left(R(x)\right)$$

και ισχύουν οι ακόλουθες ιδιότητες για τις συναρτήσεις g και R:

1. g γνησίως αύξουσα συνάρτηση για κάθε $x \in A$.
2. R γνησίως αύξουσα συνάρτηση για κάθε $x \in \mathbb{R}$
3. $\begin{cases} \lim_{x\to 0^+} g(x) = g(0)\ \ if \quad A = [0,1) \\ \qquad\qquad or \\ \lim_{x\to 0^+} g(x) = -\infty\ \ if \quad A = (0,1) \end{cases}$, $\lim_{x\to 1^-} g(x) = \infty$
4. $\lim_{x\to -\infty} R(x) = \lim_{x\to 0^+} g(x)$, $\lim_{x\to \infty} R(x) = \infty$

Για την τυχαία μεταβλητή X του προηγούμενου ορισμού θα γράφουμε συμβολικά $X \sim D_g(R)$.

Από τους δύο ορισμούς εύκολα προκύπτει ότι η συνάρτηση F λαμβάνει τιμές στο διάστημα $[0,1]$, είναι αύξουσα και παραγωγίσιμη. Με άλλα λόγια οι τέσσερεις συνθήκες από τους δύο ορισμούς εξασφαλίζουν ότι η συνάρτηση F είναι αθροιστική συνάρτηση κατανομής της τυχαίας μεταβλητής X. Επιπλέον, εύκολα διαπιστώνεται ότι οι υποθέσεις 1 και 3 σχετίζονται με τις συναρτήσεις μετασχηματισμού ενώ οι 2 και 4 με τους γεννήτορες και ότι το σύνολο που ορίζεται η τυχαία μεταβλητή X θα πρέπει να ταυτίζεται με το πεδίο ορισμού του γεννήτορα R.

3. Δημιουργία νέων οικογενειών κατανομών με χρήση των ήδη υπαρχουσών.

Εξετάζοντας κάποιος προσεκτικά τους ορισμούς 1 και 2 θα μπορούσε να διαπιστώσει ότι η αναζήτηση συναρτήσεων που θα έχουν τις ιδιότητες των g και R όπως αυτές περιεγράφηκαν, δεν είναι πάντα εύκολη. Για το σκοπό αυτό αναζητήθηκαν τρόποι δημιουργίας νέων συναρτήσεων που θα μπορούν να χρησιμοποιηθούν είτε ως γεννήτορες είτε ως συναρτήσεις μετασχηματισμού.

Κάνοντας χρήση ήδη γνωστών συναρτήσεων που έχουν τις επιθυμητές ιδιότητες στόχος είναι να μπορούν να δημιουργηθούν εύκολα, γρήγορα και με ασφάλεια νέες οικογένειες κατανομών όπως αυτές ορίστηκαν στην προηγούμενη παράγραφο. Οι τρόποι οι οποίοι αναλύονται στη συνέχεια διακρίνονται στις εξής δύο κατηγορίες:

- Μέθοδοι για τη δημιουργία γεννητόρων,
- Μέθοδοι για τη δημιουργία συναρτήσεων μετασχηματισμού.

Οι προτάσεις 1 – 5 που ακολουθούν περιγράφουν τρόπους όπου από γνωστές κατανομές του τύπου $D_g(R_1)$, και $D_g(R_2)$ (ή $D_g^+(R_1)$ και $D_g^+(R_2)$) θα μπορεί να ορισθεί μια νέα κατανομή της μορφής $D_g(R_3)$ (ή $D_g^+(R_3)$).

Πρόταση 1: Έστω ότι οι $D_g^+(R_1)$ και $D_g^+(R_2)$ είναι δύο κατανομές με γεννήτορες R_1, R_2 και την ίδια συνάρτηση μετασχηματισμού g τότε ο γραμμικός συνδυασμός R_3 των R_1, R_2 δηλαδή:

$$R_3(x) = aR_1(x) + bR_2(x)$$

μπορεί να χρησιμοποιηθεί ως γεννήτορας της κατανομής $D_g^+(R_3)$ όταν $a, b > 0$ και $a + b = 1$ έχοντας την ίδια συνάρτηση μετασχηματισμού με τις αρχικές.

Απόδειξη: Σύμφωνα με τις ιδιότητες 2 και 4 του Ορισμού 1 οι γεννήτορες R_1 και R_2 θα έχουν τις ακόλουθες ιδιότητες

- R_1 και R_2 είναι αύξουσες συναρτήσεις στο $[0, \infty)$,
- $\lim_{x\to\infty} R_1(x) = \lim_{x\to\infty} R_2(x) = \infty$ και
- $R_1(0) = R_2(0) = g(0)$

Επομένως, στην περίπτωση που $R_3(x) = aR_1(x) + bR_2(x)$ $a, b > 0$ είναι προφανές ότι η R_3 είναι αύξουσα συνάρτηση στο $[0, \infty)$.
Επίσης εύκολα αποδεικνύονται τα ακόλουθα:

- $\lim_{x\to\infty} R_3(x) = \infty$ και
- $R_3(0) = (a + b)g(0) = g(0)$ αφού $a, b > 0$ και $a + b = 1$.

Συνεπώς η $D_g^+(R_3)$ θα είναι μέλος της ίδιας οικογένειας κατανομών αφού για τη συνάρτηση R_3 ισχύουν οι ιδιότητες 2 και 4 του Ορισμού 1.

Πρόταση 2: Έστω ότι οι $D_g^+(R_1)$ και $D_g^+(R_2)$ είναι δύο κατανομές με μη αρνητικούς γεννήτορες R_1, R_2 και την ίδια συνάρτηση μετασχηματισμού g όπου $g(0) = 1$ ή $g(0) = 0$ τότε η συνάρτηση R_3 του γινομένου των R_1, R_2 δηλαδή $R_3(x) = R_1(x)R_2(x)$ μπορεί να χρησιμοποιηθεί ως γεννήτορας της κατανομής $D_g^+(R_3)$ η οποία έχει την ίδια συνάρτηση μετασχηματισμού με τις αρχικές.

Απόδειξη: Σύμφωνα με τις ιδιότητες 2 και 4 του Ορισμού 1 οι γεννήτορες R_1 και R_2 θα έχουν τις ακόλουθες ιδιότητες

- R_1 και R_2 είναι αύξουσες συναρτήσεις στο $[0, \infty)$,
- $\lim_{x\to\infty} R_1(x) = \lim_{x\to\infty} R_2(x) = \infty$ και
- $R_1(0) = R_2(0) = g(0) = 0$ ή 1

Επιπλέον, όταν $R_3(x) = R_1(x)R_2(x)$ και εφόσον R_1, R_2 είναι μη αρνητικές συναρτήσεις, είναι προφανές ότι η R_3 είναι αύξουσα συνάρτηση για κάθε $x \in [0, \infty)$. Επίσης εύκολα αποδεικνύονται τα ακόλουθα:

- $\lim_{x\to\infty} R_3(x) = \infty$ και
- $R_3(0) = [g(0)]^2 = g(0)$ αφού $g(0) = 1$ ή $g(0) = 0$

Συνεπώς η $D_g^+(R_3)$ θα είναι μέλος της ίδιας οικογένειας κατανομών αφού για τη συνάρτηση R_3 ισχύουν οι ιδιότητες 2 και 4 του Ορισμού 1.

Πρόταση 3: Έστω ότι οι $D_g^+(R_1)$ και $D_g^+(R_2)$ είναι δύο κατανομές με γεννήτορες R_1, R_2 και την ίδια συνάρτηση μετασχηματισμού g με $g(0) = 0$. Τότε οι συναρτήσεις:

- $R_3(x) = aR_1(x) + bR_2(x)$ $a, b > 0$ και (3.1)
- $R_4(x) = R_1(x)R_2(x)$ (R_1, R_2 μη αρνητικοί γεννήτορες) (3.2)

μπορούν να χρησιμοποιηθούν ως γεννήτορες δύο νέων κατανομών $D_g^+(R_3)$ και $D_g^+(R_4)$ οι οποίες έχουν ίδια συνάρτηση μετασχηματισμού με τις αρχικές.

Πρόταση 4: Έστω ότι οι $D_g(R_1)$ και $D_g(R_2)$ είναι δύο κατανομές με γεννήτορες R_1, R_2 και την ίδια συνάρτηση μετασχηματισμού g τότε η συνάρτηση R_3 ως γραμμικός συνδυασμός των R_1, R_2 δηλαδή:

$$R_3(x) = aR_1(x) + bR_2(x)$$

μπορεί να χρησιμοποιηθεί ως γεννήτορας της κατανομής $D_g(R_3)$ $\forall x \in \mathbb{R}$, η οποία έχει την ίδια συνάρτηση μετασχηματισμού με τις αρχικές.

Πρόταση 5 Έστω ότι οι $D_g(R_1)$ και $D_g(R_2)$ είναι δύο κατανομές με γεννήτορες R_1, R_2 και την ίδια συνάρτηση μετασχηματισμού g τότε η συνάρτηση $R_3 = R_1 \circ R_2$ μπορεί να χρησιμοποιηθεί ως γεννήτορας της κατανομής $D_g(R_3)$ $\forall x \in \mathbb{R}$, η οποία έχει την ίδια συνάρτηση μετασχηματισμού με τις αρχικές.

Οι αποδείξεις των προτάσεων 3 έως 5 προκύπτουν εύκολα και παρόμοια με τις αποδείξεις των προτάσεων 1 και 2.

Στην πρόταση 6 περιγράφεται η διαδικασία όπου από τις γνωστές οικογένειες του τύπου $D_{g_1}(R)$ και $D_{g_2}(R)$ (ή $D_{g1}^+(R)$ και $D_{g2}^+(R)$) μπορούν να ορισθούν οι οικογένειες της μορφής $D_{g_3}(R)$ (ή $D_{g_3}^+(R)$).

Πρόταση 6: Αν οι συναρτήσεις $g_1: A_1 \to \mathbb{R}$ και $g_2: A_2 \to \mathbb{R}$ με A_1, $A_2 = [0,1)$ ή $(0,1)$ είναι συναρτήσεις μετασχηματισμού δηλαδή ικανοποιούν τις υποθέσεις του ορισμού 1 ή του ορισμού 2 τότε και οι συναρτήσεις g_3 και g_4 με

- $g_3(x) = ag_1(x) + bg_2(x)$ $a, b > 0$ ή
- $g_4(x) = g_1(x)g_2(x)$

θα είναι συναρτήσεις μετασχηματισμού.

Απόδειξη: Εύκολα αποδεικνύεται ότι αν $A_1 = A_2 = [0,1)$, οι συναρτήσεις g_3 και g_4 ορίζονται στο $[0,1)$ και ικανοποιούν τις υποθέσεις 1 και 3 του ορισμού 1 οπότε μπορούν να χρησιμοποιηθούν ως συναρτήσεις μετασχηματισμού για τη δημιουργία οικογενειών κατανομών του τύπου $D_g^+(R)$.

Επίσης η συνάρτηση g_3 ικανοποιεί σε κάθε περίπτωση τις υποθέσεις 1 και 3 του Ορισμού 2 και μπορεί να χρησιμοποιηθεί ως συνάρτηση μετασχηματισμού για την δημιουργία οικογενειών κατανομών $D_g(R)$. Για τη συνάρτηση g_4 ικανοποιούνται οι υποθέσεις 1 και 3 του Ορισμού 2 στις παρακάτω περιπτώσεις:

- $g_1: [0,1) \to \mathbb{R}$ και $g_2: [0,1) \to \mathbb{R}$
- $g_1: (0,1) \to \mathbb{R}$ και $g_2: [0,1) \to \mathbb{R}$ (ή $g_1: [0,1) \to \mathbb{R}$ και $g_2: (0,1) \to \mathbb{R}$)

4. Εφαρμογή.

Στην ενότητα αυτή θα δημιουργηθεί ένα σύνολο από οικογένειες κατανομών του τύπου $D_g^+(R)$, με χρήση του τύπου (3.1) της Πρότασης 3 και γεννήτορες από τον Πίνακα 2. Στον Πίνακα 3 που ακολουθεί παρουσιάζονται οι γεννήτορες καθώς και οι κατανομές που τελικά δημιουργούνται. Στην τελευταία στήλη του πίνακα σημειώνονται γνωστές κατανομές που προκύπτουν για συγκεκριμένες τιμές των παραμέτρων.

Πίνακας 3. *Δημιουργία οικογενειών κατανομών του τύπου $D_g^+(R)$ με συνάρτηση μετασχηματισμού $g(x) = \left(-\ln(1-x)\right)^{\frac{1}{k}}$, $k \geq 1$*

Γεννήτορας *R*	**Αθροιστική Συνάρτηση κατανομής *F***	**Σχόλια**
$R(x) = \alpha x^b, \quad \alpha > 0, b \geq 1$	$F(x) = 1 - exp(-(ax^b)^k)$	
$R(x) = a(e^{bx} - 1), \ \alpha, b > 0$	$F(x) = 1 - exp\left(-\left(a(e^{bx} - 1)\right)^k\right)$	$k = 1$, Gompertz
$R(x) = aln\left(1 + \frac{x}{\lambda}\right), \alpha, \lambda > 0$	$F(x) = 1 - exp\left(-\left(aln\left(1 + \frac{x}{\lambda}\right)\right)^k\right)$	
$R(x) = ln\left(ln\left(e + b(e^{cx} - 1)\right)\right)$, $b, c > 0$	$F(x) = 1 - exp\left(-\left(ln\left(ln\left(e + b(e^{cx} - 1)\right)\right)\right)^k\right)$	
$R(x) = aln(1 + bx^c)$ $a > 0, b > 0, c > 0$	$F(x) = 1 - exp\left(-\left(aln(1 + bx^c)\right)^k\right)$	

Όπως μπορεί να παρατηρήσει κανείς από τον Πίνακα 3, με χρήση του γεννήτορα

$$R(x) = aln\left(1 + \frac{x}{\lambda}\right), \alpha, \lambda > 0$$

προκύπτει οικογένεια με τρεις παραμέτρους με συνάρτηση κατανομής

$$F(x) = 1 - exp\left(-\left(aln\left(1 + \frac{x}{\lambda}\right)\right)^k\right) \quad \alpha, \lambda > 0, k \geq 1$$

και αντίστοιχη σ.π.π.

$$f(x) = \frac{aexp\left(-\left(aln\left(1 + \frac{x}{\lambda}\right) + cx^b\right)^k\right) k\left(aln\left(1 + \frac{x}{\lambda}\right)\right)^{k-1}}{(\lambda + x)}$$

όπως αυτές απεικονίζονται στο Σχήμα 1 για συγκεκριμένες τιμές των παραμέτρων α, λ, k.

Για τη συνάρτηση μετασχηματισμού ισχύει $g(0) = 0$ οπότε σύμφωνα με την Πρόταση 1 μπορούμε να δημιουργήσουμε γεννήτορες με τη βοήθεια των σχέσεων (3.1) και (3.2). Πιο συγκεκριμένα, αν R_1, R_2 είναι συναρτήσεις – γεννήτορες από τον Πίνακα 3 και για $a = b = 1$ (τυχαία επιλογή για τα a, b) θα μπορούσαν να δημιουργηθούν $\binom{5}{2} = 10$ καινούριοι γεννήτορες από τη σχέση (3.1) και άλλοι 10 από τη σχέση (3.2).

Συνεπώς, αν

$$R_1(x) = cx^b, \qquad c > 0\,, b \geq 1$$

και

$$R_2(x) = aln\left(1 + \frac{x}{\lambda}\right), \alpha, \lambda > 0$$

και από τη σχέση (3.1) όπου $R_3(x) = R_1(x) + R_2(x)$ δημιουργείται ο γεννήτορας

$$R_3(x) = aln\left(1 + \frac{x}{\lambda}\right) + cx^b, \quad \alpha, \lambda, c > 0, b \geq 1$$

Στην περίπτωση αυτή δημιουργείται οικογένεια κατανομών με πέντε παραμέτρους με συνάρτηση κατανομής

$$F(x) = 1 - exp\left(-\left(aln\left(1 + \frac{x}{\lambda}\right) + cx^b\right)^k\right), \qquad \alpha, \lambda, c > 0, b \geq 1, , k \geq 1$$

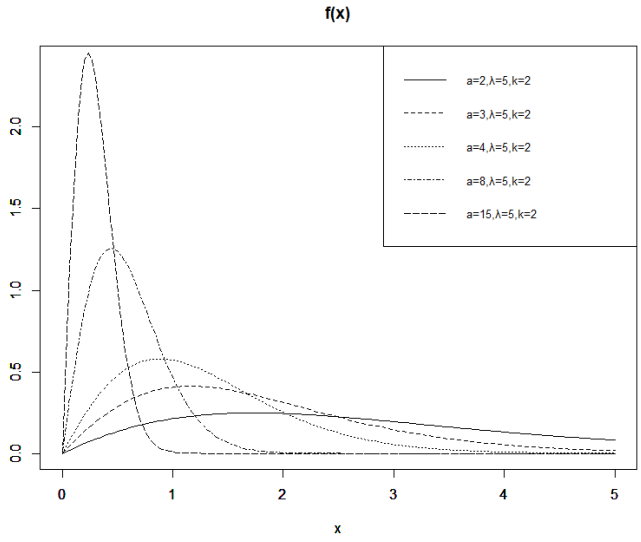
και αντίστοιχη σ.π.π. για $x > 0$

$$f(x) = \frac{exp\left(-\left(aln\left(1 + \frac{x}{\lambda}\right) + cx^b\right)^k\right) k\left(a\,x + b\,c\,x^b(\lambda + x)\right)\left(aln\left(1 + \frac{x}{\lambda}\right) + cx^b\right)^{k-1}}{x\,(\lambda + x)}$$

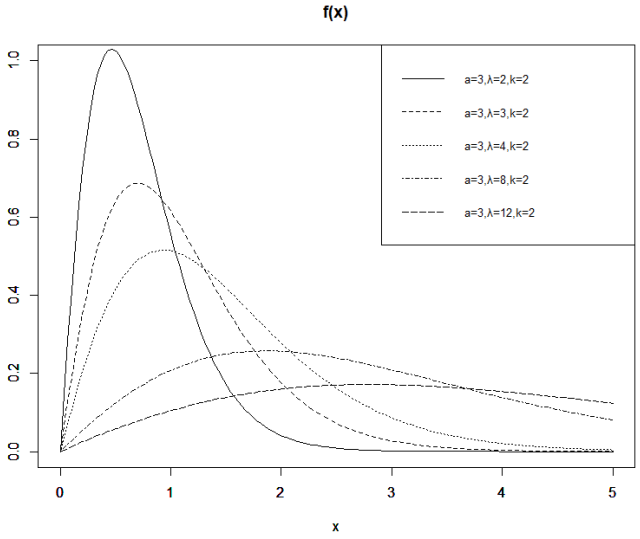
με τα αντίστοιχα γραφήματα να παριστάνονται στο Σχήμα 2 που ακολουθεί.

Σχήμα 1. *Γραφήματα της συνάρτησης πυκνότητας πιθανότητας της κατανομής*

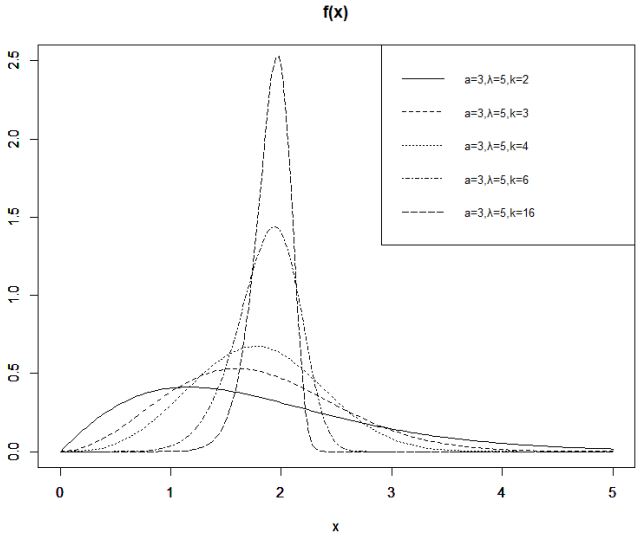
$$F(x) = 1 - exp\left(-\left(aln\left(1+\frac{x}{\lambda}\right)\right)^{k}\right) \quad \alpha, \lambda > 0, k \geq 1$$



(α)



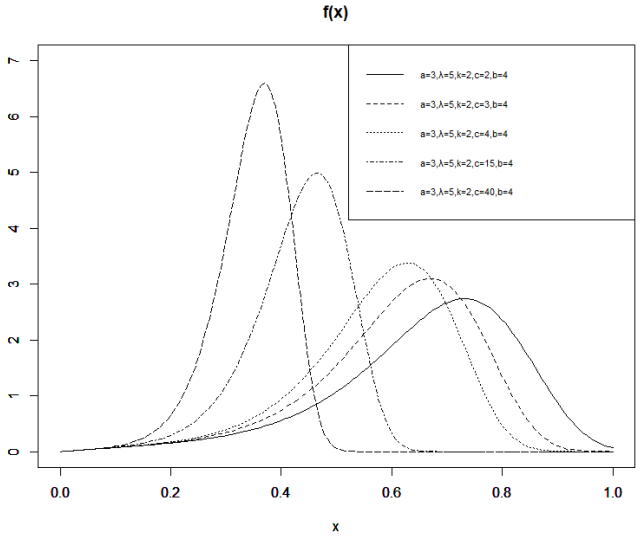
(β)



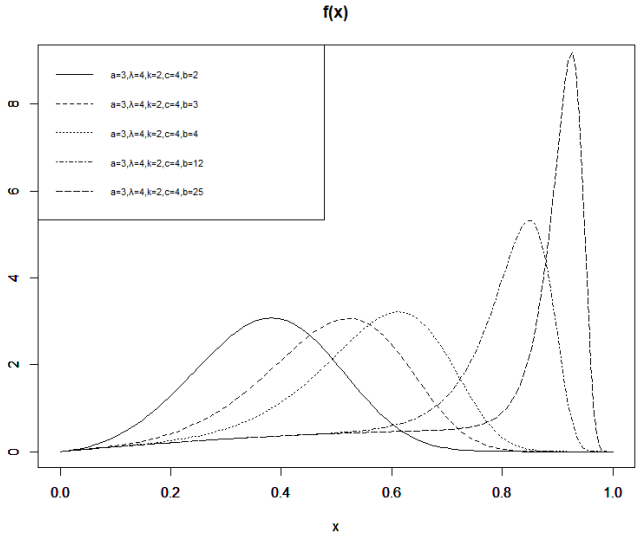
(γ)

Σχήμα 2. *Γραφήματα της συνάρτησης πυκνότητας πιθανότητας της κατανομής*

$$F(x) = 1 - exp\left(-\left(aln\left(1+\frac{x}{\lambda}\right)+cx^{b}\right)^{k}\right), \qquad \alpha,\lambda,c>0, b\geq 1,, k\geq 1$$



(α)



(β)

5. Συμπεράσματα.

Στο συγκεκριμένο άρθρο εισάγεται μια νέα μέθοδος για τη δημιουργία νέων οικογενειών κατανομών. Η τεχνική αυτή ορίζει νέες οικογένειες κατανομών με τη χρήση μίας συνάρτησης μετασχηματισμού και ενός γεννήτορα οι οποίες ικανοποιούν συγκεκριμένες υποθέσεις. Μπορούν να δημιουργηθούν δύο τύποι οικογενειών $D_g^+(R)$ και $D_g(R)$ ανάλογα με το αν η κατανομή ορίζεται στο διάστημα $[0,\infty)$ ή στο σύνολο $\mathbb{R}$ αντίστοιχα. Η εφαρμογή της σχέσης (2.1) ορίζει εν δυνάμει νέες κατανομές που

απαιτούν επιπλέον μελέτη αλλά και ήδη γνωστές κατανομές. Επίσης, αναπτύχθηκαν μέθοδοι για τη δημιουργία επιπλέον συναρτήσεων (Προτάσεις 1-6) που είναι απαραίτητες για την εφαρμογή της τεχνικής αυτής.

ABSTRACT

Although the generation of new families of distributions has been studied by several researchers for many years, is still an active research area. In this paper, considering suitable functions g and R, a new family of distributions, with cumulative distribution function $F(x) = g^{-1}(R(x))$, is introduced. In addition, for this family of distributions, a series of useful methods for the generation of new g and R functions is provided in order to construct more one-dimensional distributions.

ΑΝΑΦΟΡΕΣ

Azzalini, A. (1985), A class of distributions which includes the normal ones. Scand. J. Stat. 12, 171–178.

Balahrishnan N., Nevzorov V. B. (2003). *A Primer on Statistical Distributions*. New York: Wiley.

Burr, I.W. (1942). Cumulative frequency functions. Ann. Math. Stat. 13, 215–232

Cordeiro, G.M., de Castro, M. (2011). A new family of generalized distributions. J. Stat. Comput. Simul. 81(7), 883–898.

Cordeiro, G.M., Ortega, E.M.M., Nadarajah, S. (2010) The Kumaraswamy Weibull distribution with application to failure data. J. Frankl. Inst. 347, 1399–1429.

de Castro,M.A.R., Ortega, E.M.M., Cordeiro, G.M. (2011). The Kumaraswamy generalized gamma distribution with application in survival analysis. Stat. Methodol. 8(5), 411–433.

Eugene, N., Lee, C., Famoye, F. (2002). The beta-normal distribution and its applications. Commun. Stat. Theory Methods 31(4), 497–512.

Famoye, F., Lee, C., Olumolade, O. (2005). The beta-Weibull distribution. J. Stat. Theory Appl. 4(2), 121–136.

Flaih A, Elsalloukh H, Mendi E, Milanova M (2012). "The Exponentiated Inverted Weibull Distribution". *Applied Mathematics and Information Sciences*, Vol. 6, No. 2, 167-171.

Johnson, N.L., Kotz, S., Balakrishnan, N. (1994). *Continuous Univariate Distributions*, vol. 1, 2nd edn. Wiley, New York.

Jones, M.C. (2009). Kumaraswamys distribution: a beta-type distribution with tractability advantages. Stat. Methodol. 6, 70–81.

Kotz, S. & Nadarajah, S. (2000). *Extreme Value Distributions, Theory and Applications*. Imperial College Press, London.

Koutras, V. M., Drakos, K. and Koutras, M. V. (2014). A polynomial logistic distribution and its application in finance. Communications in Statistics: Theory and Methods 43, 2045–2065.

Kumaraswamy, P. (1980). A generalized probability density functions for double-bounded random processes. J. Hydrol. 46, 79–88.

Lee, C., Famoye, F., & Alzaatreh, A. (2013). Methods for generating families of univariate continuous distributions in the recent decades. *WIREs Computational Statistics*, 5(3), 219-238.

Nadarajah, S., Kotz, S. (2004). The beta Gumbel distribution. Math. Probl. Eng. 4, 323–332.

Nadarajah, S., Kotz, S. (2005). The beta exponential distribution. Reliab. Eng. Syst. Saf. 91(6), 689–697

Pearson, K. (1895). Contributions to the mathematical theory of evolution. II. Skew variation in homogeneous material. Philos. Trans. Royal Soc. Lond. A 186, 343–414.

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 31ου Πανελληνίου Συνεδρίου Στατιστικής (2018), σελ.105-119

ΕΦΑΡΜΟΓΗ ΤΗΣ ΠΟΛΥΩΝΥΜΙΚΗΣ ΛΟΓΙΣΤΙΚΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ ΚΑΙ ΤΕΧΝΙΚΩΝ ΟΜΑΔΟΠΟΙΗΣΗΣ ΣΤΙΣ ΔΙΑΤΡΟΦΙΚΕΣ ΣΥΝΗΘΕΙΕΣ ΤΩΝ ΕΛΛΗΝΩΝ ΕΦΗΒΩΝ

Ζ. Κυρανά, Κ. Γεροθανάση, Σ. Ζήμερας, Ε. Κυρανάς
[1]Τμήμα Στατιστικής και Αναλογιστικών-Χρηματοοικονομικών Μαθηματικών, Πανεπιστήμιο Αιγαίου
reniakyrana@gmail.com, zimste@aegean.gr
[2]Τμήμα Διατροφής και Διαιτολογίας, Αλεξάνδρειο Τ.Ε.Ι. Θεσσαλονίκης
{kgero, ekiranas}@nutr.teithe.gr

ΠΕΡΙΛΗΨΗ

Στην παρούσα εργασία διερευνάται η κατά το δυνατόν ακριβέστερη και πληρέστερη στατιστική επεξεργασία των διατροφικών συνηθειών 42.593 Ελλήνων εφήβων, που προέκυψε από τη συμπλήρωση ερωτηματολογίου. Η στατιστική ανάλυση πραγματοποιήθηκε με το στατιστικό πακέτο SPSS.

Αρχικά, περιγράφονται η μεθοδολογία της δειγματοληψίας, η τεχνική συμπλήρωσης των ερωτηματολογίων και τα γενικά χαρακτηριστικά του δείγματος.

Από την περιγραφική στατιστική ανάλυση των δημογραφικών και ατομικών χαρακτηριστικών των εφήβων φαίνεται ότι το ποσοστό των εφήβων με αποκλίνουσα κατάσταση θρέψης προσεγγίζει το 34% του δείγματος (20% με υπερβάλλον σωματικό βάρος και 14% λιποβαρείς).

Από την Παραγοντική Ανάλυση με τη μέθοδο των Κύριων Συνιστωσών, καθώς και τη μέθοδο του Ward της Ιεραρχικής Ανάλυσης Συστάδων προέκυψαν δύο ομαδοποιήσεις στην ομάδα του κρέατος θηλαστικών και των παρασκευών του. Κριτήριο ομαδοποίησης στην πρώτη συστάδα φαίνεται πως είναι η παραγωγή περισσότερο υγιεινών τροφίμων που παρασκευάζονται στο σπίτι και στα εστιατόρια, ενώ στη δεύτερη περιλαμβάνονται οι λιγότερο υγιεινές παρασκευές των ταχυφαγείων και των ταβερνών.

Από την εφαρμογή της Πολυωνυμικής Λογιστικής Παλινδρόμησης προέκυψαν τα "πιάτα" του κρέατος που επιδρούν αρνητικά στον Δείκτη Μάζας Σώματος των εφήβων και συνεπώς στην υγεία τους.

Λέξεις Κλειδιά: Στατιστική, Πολυμεταβλητή Ανάλυση, Ανάλυση Κύριων Συνιστωσών, Παραγοντική Ανάλυση, Ανάλυση σε Συστάδες, Λογιστική Παλινδρόμηση, Διατροφικές συνήθειες, Ερωτηματολόγιο, Έλληνες έφηβοι, ΔΜΣ

1. ΕΙΣΑΓΩΓΗ

Ο ρόλος της διατροφής στην ατομική υγεία, η οποία αποτελεί σημαντική προσωπική και κοινωνική αξία, είναι καθοριστικός (Goetzke et al, 2014). Η διατροφή κατά την παιδική και εφηβική ηλικία θεωρείται ιδιαίτερα κρίσιμη, επειδή συνδέεται με σοβαρές ασθένειες στην ενήλικη ζωή, ενώ πολλές από τις διατροφικές συνήθειες που διαμορφώνονται σε αυτές τις ηλικίες, διατηρούνται κατά βάση απαράλλακτες για το υπόλοιπο της ζωής.

Η εφηβεία ξεκινά και εξελίσσεται βαθμιαία περίπου από την ηλικία των 11 ετών και τελειώνει επίσης βαθμιαία με την έναρξη της ενήλικης ζωής από την ηλικία των 18 ως 21 ετών (Buckley, 2001). Αποτελεί μια αρκετά δύσκολη περίοδο, επειδή σηματοδοτεί μια σειρά από ριζικές αλλαγές, φυσικές (σωματικές και βιολογικές), κοινωνικές, συναισθηματικές και πνευματικές (γνωστική και ηθική ανάπτυξη). Οι ρυθμοί της σωματικής ανάπτυξης είναι ταχύτατοι, παρόμοιοι μ' αυτούς της βρεφικής ηλικίας. Εκδηλώνονται με μεταβολές στο βάρος, το ύψος, τη λιπώδη και την άλιπη μάζα, προκαλώντας αύξηση των αναγκών σε ενέργεια και εξειδικευμένα θρεπτικά συστατικά.

Οι διαιτητικές επιλογές διαδραματίζουν πολύ ουσιαστικό ρόλο και στην προσπάθεια αυτοπροσδιορισμού των νέων. Οι έφηβοι είναι συνήθως κάθετοι στην αποφυγή ή αποδοχή τροφίμων και συνήθως καταναλώνουν τεράστιες ή ελάχιστες ποσότητες γευμάτων (Neumark-Sztainer, 1999). Η παρατήρηση της "εικόνας του σώματος" αποτελεί απαραίτητη προϋπόθεση για νοητική δόμηση της εικόνας του εαυτού τους, ώστε να μπορέσουν να διαμορφώσουν της ατομική τους ταυτότητα. Η "εικόνα του σώματος" αποτελεί ουσιαστικά την εσωτερική απεικόνιση της εξωτερικής εμφάνισης (Thomson and Sherman, 1999).

Τα ανωτέρω οδήγησαν τους επιστήμονες στην προσέγγιση της διεξαγωγής εκτεταμένων επιδημιολογικών ερευνών, με τη συνδρομή και της επιστήμης της Στατιστικής. Οι στόχοι τους επικεντρώνονται κυρίως στη μελέτη των διατροφικών συνηθειών διαφόρων ομάδων του πληθυσμού, στην επίδραση διαφόρων παραγόντων/μεταβλητών (φύλο, ηλικία, μόρφωση, κοινωνικές και οικονομικές συνθήκες διαβίωσης κ.ά.) στις εν λόγω συνήθειες, καθώς και στη συσχέτιση αυτών με την υγεία.

Οι διατροφικές προτιμήσεις των εφήβων επηρεάζονται από πλέγμα παραγόντων, όπως το οικογενειακό περιβάλλον, η δομή και τα χαρακτηριστικά της οικογένειας, οι διαιτητικές συνήθειες των γονέων, ο βαθμός κοινωνικοποίησης, τα σωματικά πρότυπα-μοντέλα που επικρατούν, το σχολείο και οι συνομήλικοι, οι κοινωνικές και πολιτισμικές αξίες που υιοθετούνται, η διαφήμιση κ.ά. Οργανισμοί και Επιτροπές που ασχολούνται με τα τρόφιμα και τη διατροφή, όπως ο Παγκόσμιος Οργανισμός Γεωργίας και Τροφίμων (FAO), ο Παγκόσμιος Οργανισμός Υγείας (WHO), η Ευρωπαϊκή Επιτροπή για την Ασφάλεια των Τροφίμων (EFSA) κ.ά., συνιστούν την

ορθή στατιστική επεξεργασία και τεκμηρίωση της ακρίβειας κάθε επιδημιολογικής έρευνας και κάθε ποσοτικού προσδιορισμού.

2. ΕΡΩΤΗΜΑΤΟΛΟΓΙΟ ΚΑΙ ΜΕΘΟΔΟΛΟΓΙΑ ΣΥΜΠΛΗΡΩΣΗΣ

Το ερωτηματολόγιο, που αξιοποιείται στην παρούσα εργασία, σχεδιάστηκε στο Τμήμα Διατροφής και Διαιτολογίας του Αλεξάνδρειου ΤΕΙ Θεσσαλονίκης και εγκρίθηκε από τη Διεύθυνση Δευτεροβάθμιας Εκπαίδευσης του Υπουργείου Παιδείας και το Παιδαγωγικό Ινστιτούτο. Από την Εθνική Στατιστική Υπηρεσία Ελλάδος (ΕΣΥΕ) αναζητήθηκε ο συνολικός πληθυσμός των εφήβων της χώρας και η κατανομή ανά φύλο στο σύνολο της επικράτειας. Τα σχολεία στα οποία αναζητήθηκε το δείγμα επιλέχθηκαν από τους πίνακες Δευτεροβάθμιας Εκπαίδευσης (Υπουργείο Παιδείας), με κριτήρια την πληθυσμιακή κατανομή ανά δημοτική ενότητα (αστική, ημιαστική και αγροτική) και την γεωφυσική της μορφολογία (παραθαλάσσια, ενδοχώρα, πεδινή και ορεινή).

Η έρευνα διήρκησε τέσσερα έτη (2011-2015). Σκοπός της ήταν η εκτίμηση των διατροφικών συνηθειών των εφήβων της χώρας και η εκτίμηση της διατροφικής τους κατάστασης μέσω ενδεδειγμένων ανθρωπομετρικών δεικτών. Τα ερωτηματολόγια συμπληρώθηκαν από τους έφηβους μαθητές, υπό την επίβλεψη και καθοδήγηση ομάδων προπτυχιακών και μεταπτυχιακών φοιτητών του Τμήματος Διατροφής και Διαιτολογίας του Αλεξάνδρειου ΤΕΙ Θεσσαλονίκης.

Ο πληθυσμός του δείγματος ήταν N=42.593 έφηβοι. Επιχειρήθηκε η ισόποση λήψη του από μαθητές γυμνασίων και λυκείων, ώστε να επιτευχθεί η καλύτερη ηλικιακή κατανομή στο συνολικό δείγμα, καθώς και εκπροσώπηση των δύο φύλων που ανταποκρίνεται στις εκτιμήσεις της ΕΣΥΕ πανελλαδικά. Η καταχώρηση των δεδομένων από ολόκληρη την ελληνική επικράτεια σε ένα ενιαίο αρχείο excel, πραγματοποιήθηκε στο πλαίσιο υλοποίησης της παρούσας εργασίας.

3. ΠΑΡΟΥΣΙΑΣΗ ΚΑΙ ΣΤΑΤΙΣΤΙΚΗ ΑΝΑΛΥΣΗ ΤΟΥ ΔΕΙΓΜΑΤΟΣ

3.1 Περιγραφικά στατιστικά Δημογραφικών και Ατομικών χαρακτηριστικών των εφήβων

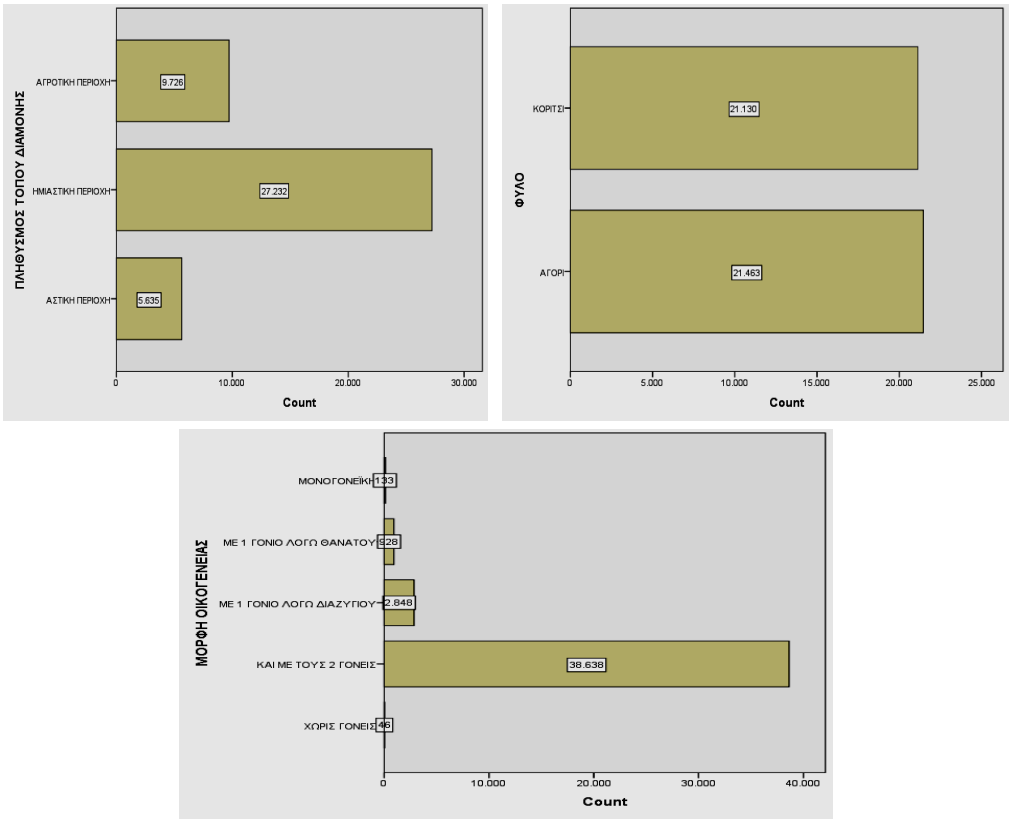
Οι μεταβλητές του δείγματος εφήβων που μελετήθηκαν είναι οι ακόλουθες:

ΠΟΣΟΤΙΚΕΣ ΜΕΤΑΒΛΗΤΕΣ	ΠΟΙΟΤΙΚΕΣ ΜΕΤΑΒΛΗΤΕΣ
Ηλικία	Φύλο
Ύψος	Μορφή οικογένειας
Βάρος	Πληθυσμιακή κατανομή
Δείκτης Μάζας Σώματος (ΔΜΣ)	

3.2 Ποιοτικές μεταβλητές του δείγματος

Στον *γράφημα 1* παρατίθενται τα ραβδογράμματα των ποιοτικών μεταβλητών του δείγματος που μελετήθηκαν:

- Από το ραβδόγραμμα της «πληθυσμιακής κατανομής» του δείγματος των εφήβων προκύπτει, πως ο πληθυσμός αυτών που διαμένουν σε αγροτικές περιοχές (22,8%) προσεγγίζει ποσοστιαία τον συνολικό πληθυσμό των αγροτικών περιοχών της χώρας (27,2% στην απογραφή του 2001), ενώ ο υπόλοιπος κατανέμεται σε ημιαστικές και αστικές περιοχές (Παπαδάκης κ.ά., 2012).
- Από το ραβδόγραμμα της μεταβλητής «φύλο» παρατηρείται, πως η κατανομή των εφήβων του δείγματος αντιστοιχεί με την πληθυσμιακή κατανομή των δύο φύλων σ' ολόκληρη τη χώρα.
- Στο ραβδόγραμμα της μεταβλητής «μορφή οικογένειας» των εφήβων παρατηρείται, πως η συχνότητα εμφάνισης πυρηνικών οικογενειών (με δύο γονείς) είναι εμφανώς η υψηλότερη, ακολουθούμενη από τη συχνότητα εμφάνισης οικογενειών με έναν γονέα λόγω διαζυγίου.



Γράφημα 1. *Ραβδογράμματα των ποιοτικών μεταβλητών*

3.3 Ποσοτικές μεταβλητές του δείγματος

Η παρουσίαση των ποσοτικών μεταβλητών με την εύρεση των περιγραφικών στατιστικών τους γίνεται στον *πίνακα 1*. Το επίπεδο σημαντικότητας για όλους τους στατιστικούς ελέγχους που ακολούθησαν είναι α=5%.

Πίνακας 1. *Περιγραφικά στατιστικά των ποσοτικών μεταβλητών*

Περιγραφικά Στατιστικά					
Ποσοτικές Μεταβλητές		Ηλικία (έτη)	Ύψος (m)	Βάρος (kg)	ΔΜΣ
N	Στατιστικό	42593	42593	42593	42593
Εύρος	Στατιστικό	10	,79	142	43,0
Ελάχιστο	Στατιστικό	11	1,24	25	12,2
Μέγιστο	Στατιστικό	21	2,03	167	55,2
Μέση τιμή	Στατιστικό	14,99	1,6619	62,11	22,316
Τυπική απόκλιση	Στατιστικό	1,798	,09402	14,306	4,0407
Διακύμανση	Στατιστικό	3,231	,009	204,661	16,328
Λοξότητα	Τυπικό σφάλμα	,012	,012	,012	,012
	Στατιστικό	-,066	,170	,890	1,149
Κυρτότητα	Τυπικό σφάλμα	,024	,024	,024	,024
	Στατιστικό	-,944	-,138	1,489	2,464
K-S Test	p-value	,0001	,0001	,0001	,0001

Ακολουθεί ο στατιστικός έλεγχος υποθέσεων των Kolmogorov-Smirnov, για διερεύνηση της ύπαρξης κανονικότητας στα δεδομένα. Οι τιμές p-value για όλες τις ποσοτικές μεταβλητές είναι μηδέν (<0,0001). Συνεπώς, τα δεδομένα δεν ακολουθούν την κανονική κατανομή. Γι' αυτό συνεχίζεται περαιτέρω η ανάλυσή τους με τον μη παραμετρικό έλεγχο των Mann-Whitney. Πρόκειται για τον αντίστοιχο μη παραμετρικό έλεγχο του t-test για δύο ανεξάρτητα δείγματα, που χρησιμοποιείται για να διαπιστωθεί αν δύο δείγματα προέρχονται από τον ίδιο πληθυσμό (*πίνακας 2*).

Πίνακας 2. *Έλεγχος διαφοροποίησης των ποσοτικών μεταβλητών (ηλικία, ύψος, βάρος και ΔΜΣ) με τις δύο υποκατηγορίες της κατηγορικής μεταβλητής «Φύλο»*

Στατιστικό τεστ	Ηλικία	Ύψος	Βάρος	ΔΜΣ
p-value	,0001	,0001	,0001	,0001

Ομαδοποιημένη Μεταβλητή: **Φύλο**

Προκύπτει p-value < 0,0001 < 0,05 = α. Συνεπώς υπάρχει διαφοροποίηση μεταξύ της ηλικίας, του ύψους, του βάρους και του ΔΜΣ με τις δύο υποκατηγορίες της κατηγορικής μεταβλητής «Φύλο».

Ακολουθεί ο έλεγχος των Kruskal-Wallis (*πίνακας 3*), που είναι ο αντίστοιχος μη παραμετρικός έλεγχος του t-test για k ανεξάρτητα δείγματα (ANOVA), για να ελεγχθεί αν k δείγματα προέρχονται από τον ίδιο πληθυσμό (Πραμαγγιούλης, 2008).

Πίνακας 3. *Έλεγχος διαφοροποίησης των ποσοτικών μεταβλητών (ηλικία, ύψος, βάρος και ΔΜΣ) με τις τρεις υποκατηγορίες της κατηγορικής μεταβλητής «Πληθυσμιακή κατανομή του τόπου διαμονής» (αστική, ημιαστική και αγροτική περιοχή)*

Στατιστικό τεστ	Ηλικία	Ύψος	Βάρος	ΔΜΣ
p-value	,0001	,0001	,0001	,0001

Ομαδοποιημένη Μεταβλητή: **Πληθυσμιακή κατανομή του τόπου διαμονής**

Επειδή p-value < 0,0001 < 0,05 = α, προκύπτει ότι υπάρχει διαφοροποίηση μεταξύ της ηλικίας, του ύψους, του βάρους και του ΔΜΣ με κάθε υποκατηγορία της μεταβλητής «πληθυσμιακή κατανομή του τόπου διαμονής».

3.4 Εκατοστιαία διακύμανση του ΔΜΣ των εφήβων ανά φύλο και ηλικία

Σύμφωνα με τον Παγκόσμιο Οργανισμό Υγείας (WHO), ο ΔΜΣ συσχετίζει το βάρος (kg) του σώματος με το ύψος (m) και δίνεται από τη σχέση: ΔΜΣ = Βάρος/Υψος^2. Πρόκειται για αριθμό που αξιοποιείται από τους επιστήμονες της υγείας, στην εκτίμηση της κατάστασης θρέψης των εφήβων και των ενηλίκων. Όταν:

- ΔΜΣ < 18,5 σημαίνει πως ο/η έφηβος είναι λιποβαρής.
- ΔΜΣ = 18,5 ως 24,9 ο/η έφηβος χαρακτηρίζεται νορμοβαρής.
- ΔΜΣ = 25,0 ως 29,9 ο/η έφηβος χαρακτηρίζεται υπέρβαρος.
- ΔΜΣ > 30 σημαίνει πως ο/η έφηβος είναι παχύσαρκος.

Πίνακας 4: *Εκατοστιαία διακύμανση του ποσοστού των λιποβαρών, νορμοβαρών, υπέρβαρων και παχύσαρκων εφήβων ανά ηλικία, συναρτήσει του φύλου*

Ηλικία εφήβων	Φύλο	Λιποβαρείς (<18.5)	Νορμοβαρείς (18.5-24.9)	Υπέρβαροι (25.0-29.9)	Παχύσαρκοι (>30.0)
11	Αγόρια	23.2	61.5	11.5	3.8
	Κορίτσια	24.1	62.1	10.3	3.5
12	Αγόρια	32.7	55.7	9.2	2.4
	Κορίτσια	26.9	58.5	8.7	5.9
13	Αγόρια	24.8	56.7	14.3	4.2
	Κορίτσια	21.4	61.5	13.5	3.6
14	Αγόρια	16.3	61.7	17.0	5.0
	Κορίτσια	15.5	66.7	13.5	4.3
15	Αγόρια	13.6	65.1	16.4	4.9
	Κορίτσια	11.8	70.2	13.2	4.8
16	Αγόρια	7.5	65.5	20.1	6.8
	Κορίτσια	9.4	71.6	14.6	4.4
17	Αγόρια	4.7	64.9	23.0	7.4
	Κορίτσια	7.8	71.3	16.0	4.9
18	Αγόρια	2.9	62.3	27.0	7.8
	Κορίτσια	9.2	69.1	16.9	4.8
Μέση % τιμή επί του συνόλου του δείγματος		14.0	64.9	15.6	4.3

Στον *πίνακα 4* παρατίθεται η εκατοστιαία διακύμανση του ποσοστού των εφήβων ανά ηλικία και φύλο στις ανωτέρω τέσσερεις κατηγορίες του ΔΜΣ. Από τα δεδομένα του

πίνακα 4 προκύπτει ότι, το σύνολο των υπέρβαρων και παχύσαρκων εφήβων στη χώρα μας προσεγγίζει το 20% του συνολικού πληθυσμού τους. Η διαπίστωση ότι 1 στους 5 έφηβους εμφανίζει υπερβάλλον βάρος είναι ανησυχητική, επειδή αυτό αποτελεί αφετηρία εκδήλωσης ποικίλων μεταβολικών διαταραχών και νοσημάτων (Levesque and Lamarche, 2008· WHO, 2009).

4. ΠΟΛΥΜΕΤΑΒΛΗΤΗ ΣΤΑΤΙΣΤΙΚΗ ΑΝΑΛΥΣΗ

Στο ερωτηματολόγιο που απαντήθηκε από τους έφηβους, περιλαμβάνεται ερώτημα που αφορά την εβδομαδιαία συχνότητα κατανάλωσης 140 διαφορετικών "πιάτων" ποικίλης σύνθεσης ή/και απλών τροφίμων. Στο πλαίσιο των εργασιών της παρούσας έρευνας, τα τρόφιμα αυτά ταξινομήθηκαν σε διάφορες ομάδες, με κριτήριο την ύπαρξη στη σύνθεσή τους βασικών τροφίμων της συνήθους διατροφής.

Για την στατιστικά ανάλυση που ακολουθεί επιλέχθηκε η ομάδα του κρέατος θηλαστικών και των διαφόρων παρασκευών του, με στόχο την εύρεση κριτηρίων ομαδοποίησης των εν λόγω τροφίμων.

4.1 Περιγραφικά Στατιστικά της ομάδας του κρέατος θηλαστικών και των παρασκευών του

Από τη μέση τιμή της εβδομαδιαίας συχνότητας κατανάλωσης (*πίνακας 5*) προκύπτει, πως τα πλέον αγαπημένα "πιάτα" των εφήβων, αλλά και των οικογενειών τους που τα εισάγουν στο τραπέζι, είναι το ψητό κρέας, το κρέας με λαχανικά και το κρέας με ζυμαρικά ή πιλάφι. Στα μικρότερης εβδομαδιαίας συχνότητας κατανάλωσης "πιάτα" ξεχωρίζουν το λουκάνικο σάντουιτς, οι παρασκευές εντοσθίων (συκώτι, κοκορέτσι, σπληνάντερο) και η κρεατόσουπα. Τέλος, από την διεξαγωγή του ελέγχου υποθέσεων των Kolmogorov-Smirnov προκύπτει ότι, τα δεδομένα δεν ακολουθούν την κανονική κατανομή.

Το ψητό κρέας και οι παρασκευές του κρέατος με λαχανικά, ζυμαρικά ή πιλάφι ξεπερνούν ως προς την εβδομαδιαία συχνότητα κατανάλωσης κρέατος θηλαστικών (κόκκινου κρέατος) τις διατροφικές συστάσεις. Από τον Εθνικό Διατροφικό Οδηγό (2014) προτείνονται για την εφηβική ηλικία 2 ως 3 μερίδες κρέατος/εβδομάδα, ενώ οι ποσότητες που καταναλώνονται από τους έφηβους είναι υπερδιπλάσιες, αν ληφθεί υπόψη και η παράλληλη κατανάλωση άλλων παρασκευών κόκκινου κρέατος (κεφτέδες, παστίτσιο, μουσακάς), κρέατος πτηνών, καθώς και βασικών παρασκευών των ταχυφαγείων (γύρος, σουβλάκι, λουκάνικο κ.ά.).

Βεβαίως το κρέας αποτελεί καλή πηγή πρωτεϊνών υψηλής βιολογικής αξίας (με υψηλή αμιναξία) και σημαντική πηγή σιδήρου, μαγνησίου και ψευδαργύρου, καθώς και βιταμινών του συμπλέγματος Β και βιταμίνης Ε. Ωστόσο, η αυξημένη κατανάλωση, ιδιαίτερα των επεξεργασμένων προϊόντων κρέατος, είναι επιβλαβής για την υγεία, επειδή η σύστασή τους σε αλάτι, συντηρητικά και κορεσμένα λιπαρά οξέα είναι αυξημένη.

Πίνακας 5. *Δεδομένα της εβδομαδιαίας κατανάλωσης κρέατος θηλαστικών και των παρασκευών του από το σύνολο του δείγματος των εφήβων στην Ελλάδα*

Είδος τροφίμου	N	Ελά-χιστο	Μέγι-στο	Μέση Τιμή	Τυπική Από-κλιση	Διακύ-μανση	K-S Test (p-value)
Κρέας ψητό	42573	0	14	1,81	1,632	2,663	,0001
Κρέας με λαχανικά	42582	0	14	1,31	1,476	2,180	,0001
Κρέας με ζυμαρικά ή πιλάφι	42586	0	14	1,24	1,394	1,943	,0001
Κεφτέδες/Σουτζουκάκια	42585	0	14	1,14	1,355	1,836	,0001
Γύρος χοιρινός μερίδα	42580	0	14	1,12	1,532	2,347	,0001
Παστίτσιο με κιμά	42579	0	14	1,11	1,337	1,787	,0001
Σουβλάκι μερίδα	42579	0	14	1,06	1,433	2,054	,0001
Γύρος χοιρινός σάντουιτς	42584	0	14	,98	1,455	2,117	,0001
Μπιφτέκι με ζυμαρικά/πιλάφι	42589	0	14	,77	1,224	1,498	,0001
Σουβλάκι σάντουιτς	42587	0	14	,74	1,315	1,729	,0001
Μουσακάς	42592	0	14	,69	1,201	1,443	,0001
Μπιφτέκι με λαχανικά	42585	0	14	,67	1,165	1,357	,0001
Λουκάνικο μερίδα	42582	0	14	,63	1,226	1,504	,0001
Κρεατόσουπα	42590	0	14	,61	1,150	1,323	,0001
Συκώτι/Κοκορέτσι/Σπληνάντερο	42583	0	14	,49	1,111	1,235	,0001
Λουκάνικο σάντουιτς	42591	0	14	,45	1,114	1,241	,0001

4.2 Παραγοντική Ανάλυση με τη μέθοδο της Ανάλυσης Κύριων Συνιστωσών

Η Παραγοντική Ανάλυση αποτελεί τεχνική μείωσης των διαστάσεων ενός συνόλου δεδομένων, που αποτελείται από έναν μεγάλο αριθμό συσχετισμένων μεταβλητών, διατηρώντας όσο το δυνατόν περισσότερη πληροφορία (Καρλής, 2005). Στην παρούσα εργασία έγινε εφαρμογή της Παραγοντικής Ανάλυσης με τη μέθοδο των Κύριων Συνιστωσών, λόγω μη ύπαρξης κανονικότητας στα δεδομένα, και με χρήση του Πίνακα Συσχέτισης.

Αρχικός στόχος ήταν η εκδίωξη από την ομάδα του κρέατος των μεταβλητών με τις χαμηλότερες βαρύτητες και επιδράσεις, έτσι ώστε να μείνουν μόνον εκείνες που ερμηνεύουν το μεγαλύτερο ποσοστό της συνολικής πληροφορίας, δηλαδή οι «καλύτερες» μεταβλητές. Όπως φαίνεται και στη συνέχεια, εκδιώχθηκαν οι μεταβλητές παρασκευές εντοσθίων (συκώτι, κοκορέτσι και σπληνάντερο), μουσακάς, παστίτσιο με κιμά, καθώς και η βασική μεταβλητή ψητό κρέας.

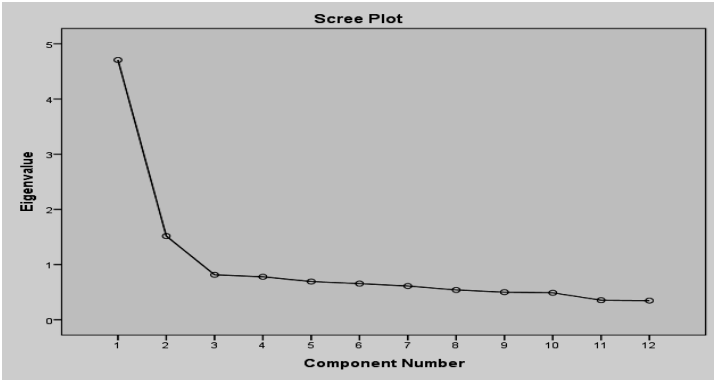
Ο δείκτης KMO ισούται με 0,892, γεγονός που υποδεικνύει ότι, η Παραγοντική Ανάλυση δίνει πολύ ικανοποιητικά αποτελέσματα. Ο έλεγχος σφαιρικότητας του Barlett, που αποφαίνεται για την παρουσία συσχετίσεων μεταξύ των μεταβλητών, δίνει p-value $< 0.001 < 0.05$, γεγονός που απορρίπτει την υπόθεση της μη ύπαρξης σημαντικών συσχετίσεων. Στον *πίνακα 6* παρουσιάζονται οι βαρύτητες των

μεταβλητών, δηλαδή η συνεισφορά καθεμιάς από τις μεταβλητές στο μέρος της συνολικής διακύμανσης που ερμηνεύεται από τους κοινούς παράγοντες (Καρλής, 2003).

Πίνακας 6. *Εταιρικότητες των «καλύτερων» μεταβλητών*

Είδος τροφίμου	Εταιρικότητες
Σουβλάκι σάντουιτς	,627
Σουβλάκι μερίδα	,595
Γύρος χοιρινός σάντουιτς	,577
Γύρος χοιρινός μερίδα	,571
Λουκάνικο σάντουιτς	,563
Λουκάνικο μερίδα	,547
Κρέας με ζυμαρικά ή πιλάφι	,524
Μπιφτέκι με λαχανικά	,491
Μπιφτέκι με ζυμαρικά ή πιλάφι	,467
Κρέας με λαχανικά	,454
Κεφτέδες ή Σουτζουκάκια	,405
Κρεατόσουπα	,402

Από τον *πίνακα 6* προκύπτει ότι, το σουβλάκι σάντουιτς ερμηνεύει το 62,7% της συνολικής πληροφορίας κ.ο.κ. Επιπλέον, υπάρχουν δυο παράγοντες, με ιδιοτιμές μεγαλύτερες της μονάδας, οι οποίοι εξηγούν το 51,863% της συνολικής διακύμανσης. Στο *γράφημα 2* φαίνεται να επιλέγονται τρεις παράγοντες, λόγω της κλίσης της ευθείας που προκύπτει μέχρι και τον τρίτο παράγοντα. Η ανάλυση θα συνεχιστεί με δύο παράγοντες, καθώς τα αποτελέσματα του πίνακα της συνολικής ερμηνευμένης διακύμανσης είναι πιο έγκυρα από το γράφημα των ιδιοτιμών ως προς τον αύξοντα αριθμό τους.



Γράφημα 2. *Γραφική παράσταση του αριθμού των παραγόντων σε φθίνουσα σειρά*

Η περιστροφή των αξόνων στοχεύει στο να αυξήσει την ερμηνευτική ικανότητα του παραγοντικού μοντέλου. Με τη μέθοδο Varimax επιχειρήθηκε η ελαχιστοποίηση του αριθμού των μεταβλητών που έχουν μεγάλες επιβαρύνσεις για κάθε παράγοντα. Χρησιμοποιείται όταν είναι επιθυμητό να εκτιμηθεί ποιες μεταβλητές επιδρούν σ'

έναν παράγοντα, αλλά και για να «ερμηνευθούν» οι παράγοντες (Γκίτσης, 2006). Οι ομάδες που προκύπτουν είναι τόσες, όσοι και οι παράγοντες που έχουν επιλεχθεί και είναι οι εξής:

- Ομάδα παρασκευών κρέατος = Κρέας με λαχανικά, Κρέας με ζυμαρικά ή πιλάφι, Κρεατόσουπα, Μπιφτέκι με λαχανικά, Μπιφτέκι με ζυμαρικά ή πιλάφι, Κεφτέδες ή Σουτζουκάκια.
- Ομάδα ειδών κρέατος = Σουβλάκι μερίδα, Σουβλάκι σάντουιτς, Γύρος χοιρινός μερίδα, Γύρος χοιρινός σάντουιτς, Λουκάνικο μερίδα, Λουκάνικο σάντουιτς.

Συνεπώς, το κρέας θηλαστικών και οι παρασκευές του μπορούν να ταξινομηθούν σε δύο ομάδες (εκτός του ψητού κρέατος που αποτελεί από μόνο του ιδιαίτερη ομάδα).

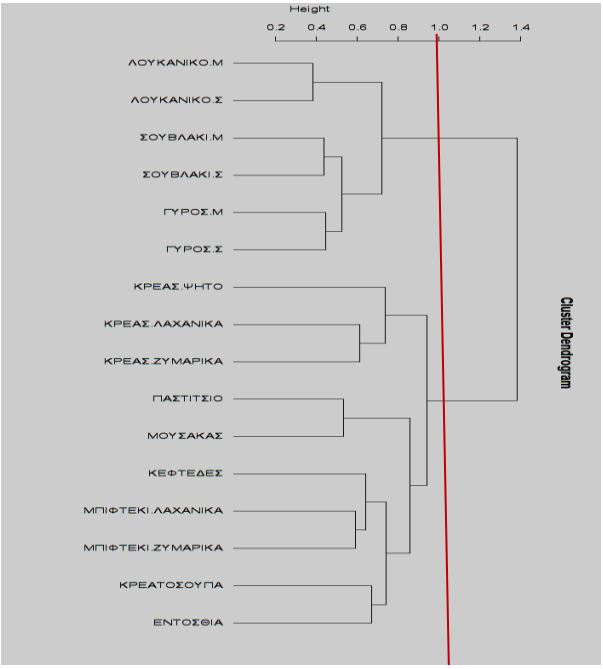
4.3 Ανάλυση κατά συστάδες: Η Ιεραρχική Μέθοδος Ward και το Δενδρόγραμμα στην ομάδα των ειδών κρέατος θηλαστικών και των παρασκευών του

Στην ιεραρχική ομαδοποίηση, ο αριθμός των ομάδων δεν είναι γνωστός από πριν. Οι μέθοδοι λειτουργούν ιεραρχικά, υπό την έννοια ότι ξεκινούν χρησιμοποιώντας κάθε παρατήρηση σε μια ομάδα και σε κάθε βήμα ενώνουν σε ομάδες τις παρατηρήσεις που βρίσκονται πιο κοντά. Ο αλγόριθμος που χρησιμοποιήθηκε είναι ο ακόλουθος: Αρχικά, δημιουργήθηκε ο πίνακας αποστάσεων για όλες τις παρατηρήσεις. Στη συνέχεια, εντοπίστηκε η μικρότερη απόσταση και ενώθηκαν οι παρατηρήσεις που την είχαν. Τέλος, επαναλήφθηκε η διαδικασία μέχρι την τοποθέτηση όλων των παρατηρήσεων σε μία ομάδα.

Το μέτρο ομοιότητας που χρησιμοποιήθηκε είναι η Ευκλείδεια απόσταση και η μέθοδος είναι αυτή του Ward (Ward linkage) (Καρλής, 2005).

Τα αποτελέσματα της ιεραρχικής μεθόδου αναπαριστώνται σ' ένα δισδιάστατο διάγραμμα, που είναι γνωστό ως δενδρόγραμμα (*γράφημα 3)*. Η μελέτη του δενδρογράμματος αποτελεί το βασικό εργαλείο για να προσδιοριστεί ο αριθμός των ομάδων που θα προκύψουν από την ανάλυση (Iliopoulou, 2015).

Η μεγάλη απόσταση μεταξύ των τροφίμων είναι αυτή που οδήγησε την κοπή του δενδρογράμματος (*γράφημα 3*) με μια γραμμή, η οποία οδηγεί στην ομαδοποίηση τροφίμων που έχουν μικρές αποστάσεις, δηλαδή μικρή ανομοιογένεια (κοινά χαρακτηριστικά) μεταξύ τους. Απομονώνει τρόφιμα με μεγαλύτερες αποστάσεις, δηλαδή μεγαλύτερη ανομοιογένεια από τα υπόλοιπα, τα οποία αποτελούν μεμονωμένες ομάδες, επειδή δεν εμφανίζουν κοινά χαρακτηριστικά με τα υπόλοιπα. Οι ομάδες που προκύπτουν είναι ίδιες με αυτές της Παραγοντικής Ανάλυσης.



Γράφημα 3. *Δενδρόγραμμα των μεταβλητών της ομάδας του κρέατος θηλαστικών και των παρασκευών του*

5. ΠΟΛΥΩΝΥΜΙΚΗ ΛΟΓΙΣΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

Η Πολυωνυμική Λογιστική Παλινδρόμηση είναι τεχνική σχεδιασμένη για την πραγματοποίηση ανάλυσης δεδομένων, που αφορούν τη μελέτη και την πρόβλεψη τιμών κάποιας κατηγορικής εξαρτημένης μεταβλητής με περισσότερες από δύο κατηγορίες (Agresti, 2001). Στην παρούσα μελέτη είναι ο ΔΜΣ με τις τέσσερεις κατηγορίες (Λιποβαρείς, Φυσιολογικοί, Υπέρβαροι και Παχύσαρκοι). Το μοντέλο αποτελείται από δεκαέξι ανεξάρτητες ποσοτικές μεταβλητές, που εκπροσωπούν την εβδομαδιαία συχνότητα κατανάλωσης κρεάτων θηλαστικών και των παρασκευών τους. Η έκφραση και η μορφή του υπό μελέτη μοντέλου είναι:

$$\boldsymbol{ln(odds)} = \boldsymbol{a} + \boldsymbol{b1}\, x1 + \boldsymbol{b2}\, x2 + \ldots + \boldsymbol{b16}\, x16$$

Αρχικά, προκύπτει το πλήρες μοντέλο με όλες τις ανεξάρτητες μεταβλητές. Εφαρμόζοντας τον στατιστικό έλεγχο του Wald εξετάζεται η στατιστική σημαντικότητα των ανεξάρτητων μεταβλητών της παλινδρόμησης και αφαιρούνται από το μοντέλο οι μη στατιστικά σημαντικές ανεξάρτητες μεταβλητές (κρέας με λαχανικά, γύρος μερίδα και σουβλάκι μερίδα και σάντουιτς). Το νέο περιορισμένο μοντέλο που προκύπτει θεωρείται το καλύτερο εκ των δύο, μέσω του κριτηρίου AIC (Akaike Information Criterion), αφού διαθέτει μικρότερη τιμή:

$$AICfull = 84870.35 > \boldsymbol{AICrestricted = 84856.59}$$

Η κατηγορία αναφοράς της εξαρτημένης μεταβλητής που επιλέχθηκε είναι οι έφηβοι με φυσιολογικό ΔΜΣ (Νορμοβαρείς), αφού εκτός του ότι αποτελούν την πλειονότητα του δείγματος, είναι και η πιο υγιεινή και επιθυμητή κατηγορία ΔΜΣ. Στόχος αυτής της ανάλυσης είναι η εύρεση των μορφών επεξεργασίας του κρέατος, που μπορούν να επηρεάσουν αρνητικά τον ΔΜΣ των εφήβων. Παράλληλα, επιτυγχάνεται η κατηγοριοποίηση αυτών σε υγιεινά και μη, με κριτήριο την επίδρασή τους στον ΔΜΣ και κατ' επέκταση στην υγεία των εφήβων. Παρακάτω παρουσιάζονται τα καλύτερα Πολυωνυμικά Λογιστικά μοντέλα Παλινδρόμησης που προκύπτουν για ολόκληρη την ελληνική επικράτεια:

1. ln(P(Λιποβαρείς)/P(Νορμοβαρείς))= -1,520 - 0,035*Κρέας Ψητό* ***- 0,067****Κρέας ζυμ.*

+ 0,046*Κρεατόσουπα* ***- 0,014****Μπιφτέκι λαχ .-* ***0,027****Μπιφτέκι ζυμ.+* ***0,045****Κεφτέδες*
- 0,009*Εντόσθια* ***+ 0,027****Γύρος μερ.* ***+ 0,120****Λουκάνικο μερ.* ***- 0,041****Λουκάνικο σάντ.*
+ 0,022*Παστίτσιο* ***- 0,037****Μουσακάς*

2. ln(P(Υπέρβαροι)/P(Νορμοβαρείς))= -1,359 + 0,020*Κρέας Ψητό* ***- 0,025****Κρέας ζυμ.*

+ 0,033*Μπιφτέκι λαχ.* ***- 0,029****Κεφτέδες* ***+ 0,029****Εντόσθια* ***- 0,044****Παστίτσιο*

3. ln(P(Παχύσαρκοι)/P(Νορμοβαρείς))= -2,650 + 0,045*Κρέας ζυμ.*
+ 0,070*Κρεατόσουπα* ***+ 0,176****Μπιφτέκι λαχ.* ***- 0,162****Μπιφτέκι ζυμ.* ***- 0,078****Παστίτσιο*
+ 0,065*Λουκάνικο σάντ.*

<u>Ερμηνεία των πολυωνυμικών λογιστικών μοντέλων:</u>

- Όταν η εβδομαδιαία συχνότητα κατανάλωσης ψητού κρέατος αυξάνεται κατά μία ημέρα, μειώνεται η πιθανότητα εμφάνισης λιποβαρών εφήβων και αυξάνεται η πιθανότητα εμφάνισης υπέρβαρων έναντι των νορμοβαρών.
- Αντίστοιχα, όταν αυξάνεται κατά μία ημέρα η κατανάλωσης κρέατος με ζυμαρικά, μειώνεται η πιθανότητα εμφάνισης λιποβαρών και υπέρβαρων εφήβων, ενώ είναι $e^{0,045}$=1,05 φορές πιο πιθανό οι έφηβοι να είναι παχύσαρκοι.
- Όταν η εβδομαδιαία συχνότητα κατανάλωσης κρεατόσουπας αυξάνεται κατά μία ημέρα, τότε είναι $e^{0,046}$=1,05 φορές πιο πιθανό ο έφηβος να είναι λιποβαρής και $e^{0,070}$=1,07 φορές να είναι παχύσαρκος, αντίστοιχα.
- Η συχνότερη κατανάλωση μπιφτεκιού με λαχανικά μειώνει την πιθανότητα εμφάνισης λιποβαρών εφήβων και αυξάνει την πιθανότητα εμφάνισης υπέρβαρων και παχύσαρκων, αντίστοιχα. Αντίθετα, η κατανάλωση μπιφτεκιού με ζυμαρικά μειώνει την πιθανότητα εμφάνισης λιποβαρών και παχύσαρκων εφήβων, έναντι των νορμοβαρών.

- Όσο συχνότερα οι έφηβοι καταναλώνουν κεφτέδες είναι $e^{0,045}$=1,05 φορές πιο πιθανό να είναι λιποβαρείς και μειώνεται η πιθανότητα να είναι υπέρβαροι.
- Η συχνότερη κατανάλωση εντοσθίων μειώνει την πιθανότητα εμφάνισης λιποβαρών εφήβων και αυξάνει την πιθανότητα εμφάνισης υπέρβαρων, αντίστοιχα.
- Επίσης, η κατανάλωση χοιρινού γύρου σε μερίδα αυξάνει την πιθανότητα εμφάνισης λιποβαρών εφήβων.
- Όταν οι έφηβοι καταναλώνουν μία φορά επιπλέον την εβδομάδα λουκάνικο μερίδα, τότε είναι $e^{0,120}$=1,13 φορές πιο πιθανό να είναι λιποβαρείς παρά νορμοβαρείς, ενώ όταν καταναλώνουν αντίστοιχα λουκάνικο σάντουιτς, μειώνεται η πιθανότητα να είναι λιποβαρείς και αυξάνεται η πιθανότητα εμφάνισης της εφηβικής παχυσαρκίας.
- Όταν η εβδομαδιαία κατανάλωση παστίτσιου αυξάνεται κατά μία μέρα την εβδομάδα, τότε είναι $e^{0,022}$=1,02 φορές πιο πιθανό οι έφηβοι να είναι λιποβαρείς, ενώ μειώνεται η πιθανότητα να είναι υπέρβαροι και παχύσαρκοι.
- Τέλος, όταν η εβδομαδιαία κατανάλωση μουσακά αυξάνεται κατά μία ημέρα την εβδομάδα, τότε μειώνεται η πιθανότητα οι έφηβοι να είναι λιποβαρείς παρά νορμοβαρείς.

6. ΣΥΖΗΤΗΣΗ - ΣΥΜΠΕΡΑΣΜΑΤΑ

Τα δεδομένα της παρούσας μελέτης διαφοροποιούνται από τα αντίστοιχα προηγούμενων μελετών σε επί μέρους περιοχές της χώρας, τα οποία εμφανίζουν τα ποσοστά υπερβάλλοντος σωματικού βάρους να ξεπερνούν το 21% και να προσεγγίζουν το 30% (Georgiadis and Nassis, 2007· Tzotzas et al, 2008· Patsopoulou et al, 2016). Αυτά βρέθηκαν υψηλότερα στα αγόρια (23,2%) σε σχέση με τα κορίτσια (18,4%), όπως έχει διαπιστωθεί και σε άλλες αντίστοιχες μελέτες (Tzotzas et al, 2008).

Εξαιρετικά ανησυχητική θεωρείται η αύξηση των εφήβων που βρίσκονται σε κατάσταση υποθρεψίας, πιθανά και λόγω της οικονομικής κρίσης που ταλανίζει τη χώρα μας. Τα ποσοστά, που είναι παρόμοια και για τα δύο φύλα, κυμαίνονται περί το 14% του συνολικού δείγματος των εφήβων, ενώ σε μελέτες του παρελθόντος δεν ξεπερνούσαν το 5% (Birbilis et al, 2013· Patsopoulou et al, 2016).

Οι ομαδοποιήσεις του κρέατος θηλαστικών και των παρασκευών του που προέκυψαν από την Παραγοντική Ανάλυση και τη μέθοδο του Ward είναι κοινές μεταξύ τους. Αυτή η κοινή στατιστική ομαδοποίηση των τροφίμων ώθησε στην αναζήτηση του διατροφικού κριτηρίου ταξινόμησης αυτών των προϊόντων σε δύο ομάδες. Προέκυψε ότι:
- Η ομάδα παρασκευών κρέατος περιλαμβάνει παρασκευές που υλοποιούνται κυρίως στο σπίτι και στα εστιατόρια. Πρόκειται για συνδυασμούς των πρωτεϊνών του κρέατος με υδατάνθρακες και διαιτητική ίνα, που προσεγγίζουν καλύτερα τις συστάσεις της διατροφικής πυραμίδας. Πρέπει κατά προτίμηση να επιλέγονται σύμφωνα με τον Εθνικό Διατροφικό Οδηγό (2014) και να καταναλώνονται με μεγαλύτερη εβδομαδιαία συχνότητα από τους έφηβους.

- Η ομάδα ειδών κρέατος περιλαμβάνει είδη επεξεργασμένου κρέατος, τα οποία προσφέρονται κυρίως σε ταχυφαγεία και ταβέρνες. Θεωρούνται περισσότερο επιβλαβή για την υγεία, λόγω της περιεκτικότητάς τους σε αλάτι, συντηρητικά και κορεσμένα λιπαρά οξέα. Καταναλώνονται από τους έφηβους με μικρότερη εβδομαδιαία συχνότητα σε σχέση με τις παρασκευές κρέατος της προηγούμενης ομάδας.

Από την Πολυωνυμική Λογιστική Παλινδρόμηση προέκυψε ότι, η αύξηση της συχνότητας κατανάλωσης τροφίμων όπως το ψητό κρέας, το κρέας με ζυμαρικά ή πιλάφι, η κρεατόσουπα, το μπιφτέκι με λαχανικά, τα εντόσθια, ο γύρος μερίδα, το λουκάνικο μερίδα και σάντουιτς και ο μουσακάς, επιδρά αρνητικά στον ΔΜΣ των εφήβων και κατ' επέκταση στην υγεία τους, επειδή συμβάλλει στην αυξημένη πιθανότητα εμφάνισης της εφηβικής παχυσαρκίας.

ABSTRACT

This paper explores the most accurate and complete statistical processing of the eating habits of 42,593 Greek adolescents, resulting from the completion of a questionnaire. The statistical analysis was performed with the statistical package SPSS.

Initially, the sampling methodology, the technical completion of the questionnaires and the general characteristics of the sample are described.

From the descriptive statistical analysis of the demographic and individual characteristics of Greek adolescents it is appeared that the percentage of them with divergent nutritional status is close to 34% of the sample (20% overweight and 14% underweight).

From the Factor Analysis with the Principal Component Method, as well as from the Ward Method of the Hierarchical Cluster Analysis, two clusters in the mammalian meat group and its preparations emerged. As clustering criterion in the first cluster appears to be the production of more healthy foods made at home and in restaurants, while the second one includes less healthy products of fast foods and taverns.

From Multinomial Logistic regression the meat "dishes" with negative impact on Body Mass Index of adolescents and therefore on their health are presented.

Key-Words: Statistics, Multivariate analysis, Principal Component Analysis, Factor Analysis, Cluster Analysis, Hierarchical Method, Dendrogram, Logistic Regression, Eating habits, Questionnaire, Greek adolescents, BMI

ΑΝΑΦΟΡΕΣ

Γκίτσης Σ. (2006). *Ανάλυση σε κύριες συνιστώσες και παραγοντική ανάλυση*, Διπλωματική εργασία, Πανεπιστήμιο Πατρών.

Καρλής Δ. (2003). *Πολυμεταβλητή Στατιστική Ανάλυση*, Αθήνα: Εκδόσεις Οικονομικού Πανεπιστημίου Αθηνών.

Καρλής Δ. (2005). *Πολυμεταβλητή Στατιστική Ανάλυση*, Αθήνα: Εκδόσεις Σταμούλη.

Παπαδάκης Μ., Κογεβίνας Μ. και Τριχόπουλος Δ. (2012). *Ο πληθυσμός της Ελλάδας: Θεωρήσεις-Προοπτικές-Προσανατολισμοί*, Αθήνα.

Agresti A. (2001). *Categorical Data Analysis*, 2nd edition, New Jersey: John Wiley & Sons.

Birbilis M., Moschonis G., Mougios V. and Manios Y. (2013). Obesity in adolescence is associated with perinatal risk factors, parental BMI and sociodemographic characteristics. *European Journal of Clinical Nutrition*, **67**, 115-121.

Buckley B.M. (2001). Healthy ageing: ageing safely. *European Heart Journal Supplements*, **3**, 6-10.

Georgiadis G. and Nassis G.P. (2007). Prevalence of overweight and obesity in a national representative sample of Greek children and adolescents. *European Journal of Clinical Nutrition*, **61**, 1072-1074.

Goetzke B., Nitzko S. and Spiller A. (2014). Consumption of organic and functional food. A matter of well-being and health? *Appetite*, **77C**, 94-103.

Levesque J. and Lamarche B. (2008). The metabolic syndrome: definitions, prevalence and management. *Journal of Nutrigenetics and Nutrigenomics*, **1**, 100-108.

Neumark-Sztainer D. (1999). Factors influencing food choices of adolescents: Findings from focus-group discussions with adolescents. *Journal of the American Dietetic Association*, **99(8)**, 929-937.

Patsopoulou A., Tsimtsiou Z., Katsioulis A., Rachiotis G., Malissiova E. and Hadjichristodoulou C. (2016). Prevalence and Risk Factors of Overweight and Obesity among Adolescents and Their Parents in Central Greece (FETA Project). *International Journal of Environmental Research and Public Health*, **13(1)**, 83-92.

Thomson R.A. and Sherman R.T. (1999). "Good Athlete" Traits and characteristics of Anorexia Nervosa: Are they similar? *The Journal of Treatment and Prevention*, **7(3)**, 181-190.

Tzotzas T., Kapantais E., Tziomalos K., loannidis I., Mortoglou A. and Bakatselos S. (2008). Epidemiological survey for the prevalence of overweight and abdominal obesity in Greek adolescents. *Obesity*, **16(7)**, 1718-1722.

WHO, ENHIS (2009). *Prevalence of overweight and obesity in children and adolescents*. Fact sheet 2.3.

Iliopoulou P. (2015). Πολυμεταβλητές Μέθοδοι Ανάλυσης. (https://repository.kallipos.gr/bitstream/11419/2065/1/02_chapter_06.pdf.)

Εθνικός Διατροφικός Οδηγός (2014). (http://www.diatrofikoiodigoi.gr/files/PDF/KIDS.pdf.)

Πραμαγγιούλης, Π. (2008). Οδηγός Ανάλυσης Δεδομένων με τη χρήση SPSS. (http://mde-lab.aegean.gr/files/docs/Odigos-SPSS-Pramaggioulis.pdf.)

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 31ου Πανελληνίου Συνεδρίου Στατιστικής (2018), σελ.120-131

ΣΥΓΚΡΙΣΗ ΜΟΝΤΕΛΩΝ ΟΓΚΟΥ ΓΡΑΜΜΙΚΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ ΜΕ ΑΝΤΙΣΤΟΙΧΑ ΝΕΥΡΩΝΙΚΩΝ ΔΙΚΤΥΩΝ ΓΙΑ ΔΑΣΟΒΙΟΜΕΤΡΙΚΑ ΔΕΔΟΜΕΝΑ ΤΟΥ ΔΗΜΟΤΙΚΟΥ ΔΑΣΟΥΣ ΝΑΟΥΣΑΣ

Σ. Μεσσαριτάκης[1], Α. Λέλης[2], Α. Γεωργάκης[1]
[1]Τμήμα Δασολογίας και Φυσικού Περιβάλλοντος, ΑΠΘ
spymess@yahoo.gr, arisgeorg@for.auth.gr
[2]Εξωτερικός συνεργάτης Τμήματος Δασολογίας και Φυσικού Περιβάλλοντος, ΑΠΘ
athalelis@gmail.com

ΠΕΡΙΛΗΨΗ

Τα δασοβιομετρικά δεδομένα και κυρίως τα δεδομένα ξυλώδη όγκου των δασικών πόρων, παρέχουν πρωταρχικής σημασίας πληροφορίες οι οποίες είναι απαραίτητες για τη διαχείριση των δασικών οικοσυστημάτων. Τα δεδομένα που χρησιμοποιήθηκαν για στατιστική ανάλυση και εντοπίζονται στο Δημοτικό Δάσος Νάουσας, αποτελούνται από μετρήσεις 250 κορμών δέντρων τα οποία επιλέχθηκαν με απλή τυχαία δειγματοληψία. Η εκτίμηση του όγκου πραγματοποιήθηκε αρχικά με τη χρήση εξισώσεων όγκου που προκύπτουν από την εφαρμογή απλής και πολλαπλής γραμμικής παλινδρόμησης. Στη συνέχεια πραγματοποιήθηκε η εκτίμηση όγκου με νευρωνικά δίκτυα, τα οποία ως σύγχρονα υπολογιστικά εργαλεία, αποδίδουν ικανοποιητικά σε παρόμοιες δασικές εφαρμογές. Στο σύνολο των δεδομένων και με ανεξάρτητες μεταβλητές τη στηθιαία διάμετρο και το ολικό ύψος προσαρμόζονται συνολικά 23 μοντέλα εξισώσεων. Αντίστοιχα εκπαιδεύτηκαν νευρωνικά δίκτυα διαφόρων αρχιτεκτονικών (επίπεδα, αριθμός νευρώνων ανά επίπεδο) με διαφορετικούς αλγορίθμους οπισθοδιάδοσης του λάθους. Από τα αποτελέσματα προκύπτει ότι από τα 23 μοντέλα μόνο τα μετασχηματισμένα ή σταθμισμένα μοντέλα δίνουν ικανοποιητικά ή καλά αποτελέσματα στην εκτίμηση του όγκου. Τα νευρωνικά δίκτυα από την άλλη, ειδικά όταν εκπαιδεύτηκαν με βαθιές αρχιτεκτονικές, εμφάνισαν γενικά μικρότερη απόκλιση κατά την εκτίμηση του όγκου των δέντρων. Συμπερασματικά, τα νευρωνικά δίκτυα ξεπερνούν γενικά την απόδοση των παραδοσιακών εξισώσεων-μοντέλων κατά την εκτίμηση του ξυλώδη όγκου των δέντρων της περιοχής.

Λέξεις κλειδιά: Μοντέλα παλινδρόμησης, Νευρωνικά δίκτυα, Δημοτικό Δάσος Νάουσας.

1. ΕΙΣΑΓΩΓΗ

Ο κορμός είναι ένα δυναμικό τμήμα του δένδρου που επηρεάζει τη συνολική ανάπτυξη, την οικονομική αξία καθώς και τη θέση του δένδρου ως μέλος μιας δασικής συστάδας. Ο όγκος του κορμού καθορίζει σε μεγάλο βαθμό την οικονομική αξία του δένδρου και κατ' επέκταση και του δάσους. Εξισώσεις με διάφορες ανεξάρτητες μεταβλητές έχουν εξεταστεί από πολλούς ερευνητές. Περισσότερο συνηθισμένες και πρακτικές είναι εξισώσεις που περιλαμβάνουν ως ανεξάρτητες μεταβλητές τη στηθιαία διάμετρο, το ολικό ύψος ή συνδυασμούς τους και προκύπτουν με εφαρμογή απλή και πολλαπλής γραμμικής παλινδρόμησης σε ένα σύνολο δασοβιομετρικών δεδομένων. Τα τελευταία χρόνια, μοντέλα που προέρχονται από την εφαρμογή των νευρωνικών δικτύων χρησιμοποιούνται στην εκτίμηση του όγκου κορμού και το πλεονέκτημά τους είναι ότι δεν εξετάζουν υποθέσεις για τα δεδομένα ή για τα υπόλοιπα της μεθόδου. Διάφοροι ερευνητές έχουν εξετάσει την εφαρμογή νευρωνικών δικτύων στην Δασοπονία (Diamantopoulou M., 2005, Maier, H.R. & Dandy G.C. 2000, M.P. Tiruveedhula, J. F. et.al., 2010)

Στην παρούσα εργασία εξετάζονται μοντέλα εκτίμησης του όγκου με τη μέθοδο της απλής και πολλαπλής γραμμικής παλινδρόμησης. Από τα μοντέλα αυτά όμως τα περισσότερα απορρίπτονται είτε για το υψηλό σφάλμα εκτίμησης είτε επειδή τα υπόλοιπα των εξισώσεων δεν ικανοποιούν τις υποθέσεις της μεθόδου των ελαχίστων τετραγώνων. Για το λόγο αυτό εξετάζουμε εάν η εφαρμογή νευρωνικών δικτύων και των μοντέλων που παράγονται από αυτή δίνουν καλύτερα αποτελέσματα αφού δε χρειάζεται να ικανοποιούν επιπρόσθετες υποθέσεις όπως στην περίπτωση της παλινδρόμησης.

Σκοπός της εργασίας είναι να βρεθεί εάν η καλύτερη εξίσωση εκτίμησης όγκου προκύπτει με την εφαρμογή απλής ή πολλαπλής γραμμικής παλινδρόμησης ή νευρωνικών δικτύων.

2. ΣΥΛΛΟΓΗ ΚΑΙ ΕΠΕΞΕΡΓΑΣΙΑ ΔΕΔΟΜΕΝΩΝ

Τα δεδομένα τα οποία χρησιμοποιήθηκαν στην εργασία προέρχονται από μετρήσεις σε κορμούς 250 δένδρων του Δημοτικού Δάσους Νάουσας (Εικόνα 1). Σε κάθε δένδρο μετρήθηκαν οι διάμετροι στο 1,3 μέτρα από το έδαφος και ανά δύο μέτρα δηλ. 3,3, 5,3, 7,3 κ.ο.κ.. Επίσης μετρήθηκαν οι διάμετροι στα σχετικά ύψη δηλ. στο 0,1, 0,3, 0,5, 0,7, και 0,9 του ολικού ύψους. Το ολικό ύψος (h) κάθε δένδρου μετρήθηκε με κατοπτρικό ρελασκόπιο (Philip M., 1994). Λαμβάνοντας υπόψη τα παραπάνω στοιχεία γίνεται τμηματικός υπολογισμός του συνολικού όγκου κάθε δένδρου με k αριθμό κορμοτεμαχίων από τον τύπο:

$$v = v_0 + \sum_{i=1}^{k} v_i + v_c \qquad (1)$$

Το v_0 αντιστοιχεί στον όγκο από το ύψος των 0,3 μέτρων μέχρι το στηθιαίο ύψος και υπολογίστηκε εφαρμόζοντας τον τύπο του Smalian:

$$v_o = \frac{\pi}{4}\frac{d_{0,3}^2 + d^2}{2} l \quad (2)$$

όπου $d_{0,3}$ η διάμετρος στο ύψος 0,3 μέτρα, d η στηθιαία διάμετρος και το μήκος του τμήματος δηλαδή 1 μέτρο.

Από το ύψος των 1,3 μέτρων ως το ύψος της τελευταίας μετρημένης διαμέτρου ο όγκος υπολογίστηκε εφαρμόζοντας τη μέθοδο τμηματικής ογκομέτρησης με τον τύπο του Smalian:

$$v_i = \frac{\pi}{4}\frac{d_0^2 + d_n{}^2}{2} l \quad (3)$$

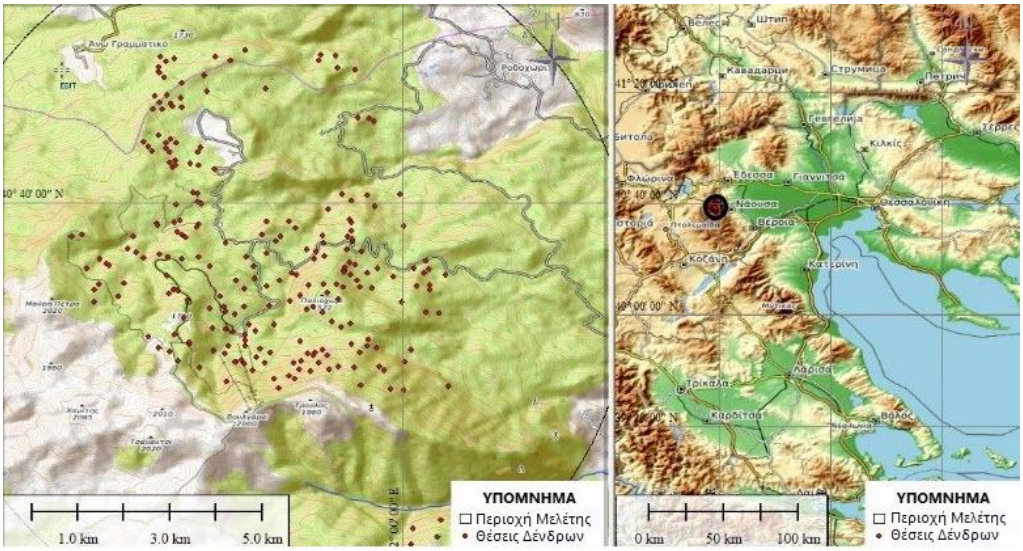
όπου d_0 η διάμετρος στη μεγάλη βάση του κορμοτεμαχίου i, d_n η διάμετρος στη μικρή βάση του κορμοτεμαχίου i και το μήκος του κορμοτεμαχίου.

Ο όγκος του κορυφοτεμαχίου υπολογίστηκε από το ύψος της τελευταίας μετρημένης διαμέτρου μέχρι την κορυφή με τον τύπο του κώνου:

$$v_c = \frac{1}{3}\frac{\pi}{4} d_c^2 l_c \quad (4)$$

όπου d_c η διάμετρος στη βάση του κορυφοτεμαχίου και το μήκος του κορυφοτεμαχίου.

Εικόνα 1. *Σημεία δειγματοληψίας εντός του Δημοτικού Δάσους Νάουσας*



3. ΜΕΘΟΔΟΛΟΓΙΑ

3.1 Εξισώσεις παλινδρόμησης

Στη συνέχεια επιλέχθηκαν 23 εξισώσεις οι οποίες χρησιμοποιούνται συχνά για την εκτίμηση του όγκου κορμού. Με τη χρήση του στατιστικού πακέτου SPSS (Ver.22) οι 23 εξισώσεις εξετάστηκαν ως προς τον συντελεστή προσδιορισμού (R^2), τη συγγραμμικότητα (VIF), την κανονικότητα των υπολοίπων. Οι εξισώσεις που εμφάνιζαν συγγραμμικότητα αποκλείστηκαν από τη διαδικασία του μετασχηματισμού γιατί η ύπαρξη συγγραμμικότητας επηρεάζει την αξιοπιστία των αποτελεσμάτων. Παρατηρήθηκε ότι σε όσες εξισώσεις η εξαρτημένη μεταβλητή ήταν ο όγκος η κατανομή των υπολοίπων δεν προσέγγιζε την κανονική κατανομή. Αντίθετα σε όσες εξισώσεις εξαρτημένη μεταβλητή ήταν ο νεπέριος λογάριθμος του όγκου ή ήταν σταθμισμένες με σταθμικό το $1/d^2$ ή το $1/d^2h$ η κατανομή των υπολοίπων προσέγγιζε την κανονική κατανομή. Επειδή οι εξαρτημένες μεταβλητές σε όλες τις εξισώσεις δεν είναι ίδιες, χρησιμοποιήθηκε ο δείκτης Furnival ώστε να είναι δυνατή η σύγκρισή μεταξύ τους. Έπειτα, ελέγχθηκε εάν οι πέντε καλύτερες εξισώσεις είναι στατιστικά σημαντικές και στη συνέχεια εφαρμόστηκαν κριτήρια αξιολόγησης αυτών. Με τη βοήθεια των κριτηρίων και την εμπειρία του ερευνητή θα εξαχθούν χρήσιμα συμπεράσματα σχετικά με τις εξισώσεις και θα γίνουν συγκρίσεις με την απόδοση που προκύπτει από τα νευρωνικά δίκτυα.

Ως πρώτο κριτήριο αξιολόγησης χρησιμοποιήθηκε το τετραγωνικό μέσο σφάλμα το οποίο υπολογίστηκε από τον τύπο:

$$MSE = \sum_{i=1}^{n} \frac{(v_i - \hat{v}_i)^2}{n} \tag{5}$$

Επόμενο κριτήριο το οποίο χρησιμοποιήσαν οι Farrar and Myrphy (1987), Li and Weiskittel (2010), χρήσιμο για τη σύγκριση εξισώσεων, είναι η τετραγωνική ρίζα του μέσου τετραγώνου των υπολοίπων (RMSE) το οποίο όσο μικρότερο είναι τόσο καλύτερη είναι η εξίσωση. Υπολογίστηκε με βάση τον τύπο:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(v_i - \hat{v}_i)^2}{n}} \tag{6}$$

Το τρίτο κριτήριο για τη σύγκριση εξισώσεων είναι το μέσο απόλυτο σφάλμα (mean absolute error) το οποίο χρησιμοποίησαν οι Ozcelik and Bal (2013), οι Tabacchi et al (2011), Li and Weiskittel (2010) και ισούται με:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} \left| v_i - \hat{v}_i \right| \tag{7}$$

Τέλος, ο συντελεστής προσδιορισμού R^2 υπολογίστηκε σύμφωνα με τον τύπο:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (\hat{v}_i - v_i)^2}{\sum_{i=1}^{n} (v_i - \overline{v})^2} \quad (8)$$

Στους παραπάνω τύπους όπου v_i είναι η *i* παρατηρηθείσα τιμή της εξαρτημένης μεταβλητής, $\hat{v}_i$ είναι η *i* εκτιμηθείσα τιμή της εξαρτημένης μεταβλητής, $\overline{v}$ ο αριθμητικός μέσος όρος των εκτιμηθεισών τιμήν και n είναι το μέγεθος του δείγματος. Ο συντελεστής προσδιορισμού χρησιμοποιείται ως κριτήριο για εξισώσεις με την ίδια εξαρτημένη μεταβλητή (Μάτης 1994).

3.2 Μεθοδολογία εκπαίδευσης νευρωνικών δικτύων

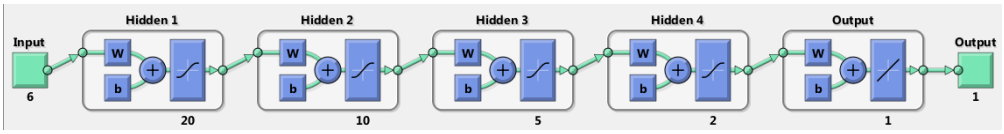
Στην παρούσα εργασία χρησιμοποιήθηκαν κατά κύριο λόγο βαθιές αρχιτεκτονικές των τριών ή και περισσοτέρων επιπέδων. Τα επιπλέον επίπεδα αποσπούν διαφορετικούς συνδυασμούς των δεδομένων, μοντελοποιώντας έτσι πολύπλοκα συστήματα πιο εύρωστα, χωρίς όμως η χρήση τους να περιορίζεται μόνο σε αυτά (Rumelhart et.al 1986). Στον αριθμό των επιπέδων σε κάθε περίπτωση καταλήξαμε με την τεχνική δοκιμής-λάθους, κάτι που ισχύει και για τον αριθμό νευρώνων που περιέχεται σε κάθε ένα από αυτά τα επίπεδα. Κριτήριο ήταν η ελαχιστοποίηση του λάθους τόσο στα δεδομένα εκπαίδευσης, όσο και στα δεδομένα επαλήθευσης, έτσι ώστε να αποφευχθεί η υπερπροσαρμογή (overfitting) και τα δίκτυα να μπορούν να γενικεύουν καλύτερα τις προβλέψεις τους. Στον πίνακα 1 φαίνονται συνοπτικά οι παράμετροι των δικτύων, ανάλογα με τις αντίστοιχες μεταβλητές εισόδου τους.

Πίνακας 1. *Στον πίνακα φαίνονται συνοπτικά οι αρχιτεκτονικές των δικτύων που εκπαιδεύτηκαν, καθώς και ο αλγόριθμος εκπαίδευσης που επιστρατεύτηκε για την εκπαίδευση*

Είσοδοι	**Αριθμός Επιπέδων**	**Αλγόριθμος Εκπαίδευσης**	**Εποχές Εκπαίδευσης**
$d_{1,3}$, h	1, 3, 4	SCG, LM	100-150
$d_{1,3}$	1, 3	SCG, LM	100
Relative d, h	1, 3, 4	SCG-LM	200
Relative d	1, 3, 4	SCG-LM	200

Η εικόνα 2 αντίστοιχα, δείχνει ένα παράδειγμα αρχιτεκτονικής δικτύου που εκπαιδεύτηκε.

Εικόνα 2. *Αρχιτεκτονική νευρωνικού δικτύου τεσσάρων επιπέδων με 6 εισόδους και μία έξοδο*



Ως αλγόριθμος οπισθοδιάδοσης του λάθους κατά την εκπαίδευση επιλέχθηκε αρχικά ο Levenberg-Marquardt (LM) (Levenberg, K., 1944), αλλά στη συνέχεια αντικαταστάθηκε με τον Scaled Conjugate Gradient (SCG) (Moller M. F. 1993). Η αντικατάσταση αυτή πραγματοποιήθηκε διότι ο LM, ενώ σε δίκτυα του ενός ή δύο επιπέδων εμφάνισε γενικά καλή απόδοση, όταν εφαρμόστηκε σε πιο βαθιά δίκτυα φάνηκε να αδυνατεί να προβλέψει ικανοποιητικά τον ζητούμενο όγκο. Ο SCG, όμοια με τον LM είναι ένας αλγόριθμος με αρκετά περιορισμένο αριθμό υπέρ-παραμέτρων που πρέπει να ρυθμιστούν από το χρήστη. Επίσης επιτυγχάνει ταχύτατη και εύρωστη σύγκλιση σε ολικό ελάχιστο και παρουσιάζει καλή απόδοση σε βαθιά δίκτυα.

Συνολικά εκπαιδεύτηκαν νευρωνικά δίκτυα για τέσσερις περιπτώσεις εισόδων. Αρχικά εκπαιδεύτηκαν δίκτυα με εισόδους που χρησιμοποιήθηκαν σε παλιότερες περιπτώσεις (Razi and Athappilly, 2005, Vahedi, 2016), τη στηθιαία διάμετρο ($d_{1,3}$) και το συνολικό ύψος των δέντρων (h), ενώ στη συνέχεια επιχειρήθηκε η εκπαίδευση δικτύων με μία μόνο είσοδο, αυτήν της στηθιαίας διαμέτρου. Τέλος για τον προσδιορισμό του όγκου εκπαιδεύτηκαν δίκτυα με εισόδους το ύψος και τις σχετικές διαμέτρους των δέντρων, δηλαδή τη διάμετρό τους σε 0,1, 0,3, 0,5, 0,7, 0,9 του ολικού ύψους.

Τα δεδομένα, πριν την εισαγωγή τους στα δίκτυα, κανονικοποιήθηκαν στο διάστημα (-1, 1) με βάση τον τύπο:

$$x_i = -1 + 2\frac{x_i - \min(x) - e}{\max(x) + e - \min(x) - e} \tag{9}$$

με x_i το εκάστοτε στοιχείο και e μία μικρή σταθερά, ώστε να μπορούν να γενικευτούν τα αποτελέσματα σε ψηλότερα ή χαμηλότερα δέντρα και ταυτόχρονα να μην υπάρχει στοιχείο εισόδου με τιμή 1 ή -1. Ο μετασχηματισμός αυτός πραγματοποιήθηκε καθώς τα νευρωνικά δίκτυα ανταποκρίνονται γενικά καλύτερα σε μικρές και συγκρίσιμες μεταξύ τους εισόδους (Ioffe and Szegedy, 2015). Επίσης, λειτουργεί καταλυτικά σε συνδυασμό με την υπερβολική εφαπτομένη (tanh), που χρησιμοποιήθηκε ως συνάρτηση ενεργοποίησης των νευρωνικών επιπέδων:

$$\tan h(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{10}$$

Ένας σημαντικός παράγοντας που επηρεάζει την ικανότητα γενίκευσης των αποτελεσμάτων των νευρωνικών δικτύων είναι ο διαχωρισμός του συνόλου των δεδομένων (dataset ή σετ δεδομένων) πριν από την εκπαίδευση. Στην παρούσα εργασία για την περίπτωση των νευρωνικών δικτύων εφαρμόστηκε η μέθοδος του hold-out cross-validation (επικυρωμένη διασταύρωση). Τα δεδομένα διαχωρίστηκαν σε ποσοστά: 70% σετ εκπαίδευσης, 15% σετ επαλήθευσης και 15% σετ ελέγχου με τυχαίο τρόπο Για να αποφευχθούν τυχόν πολώσεις στα δεδομένα κατά το διαχωρισμό, ο τελευταίος πραγματοποιήθηκε αρκετές φορές δίνοντας κάθε φορά διαφορετικά υπο-σετ δεδομένων. Το σετ εκπαίδευσης ήταν η πηγή μάθησης των δικτύων, το σετ επαλήθευσης χρησιμοποιήθηκε για ρυθμίσεις υπερ-παραμέτρων κατά τη διάρκεια της εκπαίδευσης και τέλος το σετ ελέγχου για τον τελικό έλεγχο απόδοσης των δικτύων σε ξένα για αυτά δεδομένα. Τα αποτελέσματα του πίνακα 3 δείχνουν τη μέση απόδοση των δικτύων με βάση αποκλειστικά το σετ ελέγχου.

4. ΑΠΟΤΕΛΕΣΜΑΤΑ

Για τις εξισώσεις παλινδρόμησης που παρουσίασαν καλές τιμές στο R^2, δεν εμφάνισαν συγγραμμικότητα και η κατανομή των υπολοίπων προσέγγιζε την κανονική κατανομή υπολογίστηκε ο δείκτης Furnival. Οι εξισώσεις 11-15 εμφάνισαν τη μικρότερη τιμή στο δείκτη Furnival. Για τις παραπάνω εξισώσεις στη συνέχεια εφαρμόστηκαν τα κριτήρια αξιολόγησης και προέκυψε ότι τα καλύτερα αποτελέσματα δίνει η εξίσωση 13 καθώς παρουσιάζει το μικρότερο RMSE, MSE και MAE καθώς και το μεγαλύτερο R^2. Αναλυτικά τα αποτελέσματα των κριτηρίων αξιολόγησης των πέντε καλύτερων εξισώσεων φαίνονται στον Πίνακα 2.

$$\ln v = b_0 + b_1 \ln(d^2 h) \qquad 0{,}065 \qquad (11)$$

$$\ln v = b_0 + b_1 \ln(dh) \qquad 0{,}077 \qquad (12)$$

$$\ln v = b_0 + b_1 \ln(d) + b_2 \ln h \qquad 0{,}064 \qquad (13)$$

$$\frac{v}{d^2 h} = b_0 + b_1 \frac{1}{d^2 h} \qquad 0{,}061 \qquad (14)$$

$$\frac{v}{d^2} = b_0 + b_1 h \qquad 0{,}065 \qquad (15)$$

Πίνακας 2. *Αποτελέσματα εξισώσεων παλινδρόμησης*

Εξίσωση	**RMSE**	**MSE**	**R^2**	**MAE**
11	7%	0,49%	0,9326	0,0488
12	7,51%	0,56%	0,9225	0,062
13	6,89%	0,47%	0,9348	0,0482
14	7,66%	0,59%	0,9194	0,0509

15	7,95%	0,63%	0,9132	0,0525

Αντίστοιχα, η απόδοση των νευρωνικών δικτύων παρουσιάζεται στον Πίνακα 3. Από τα αποτελέσματα φαίνεται η υπεροχή των πολλαπλών επιπέδων απέναντι σε μονο-επίπεδα νευρωνικά δίκτυα, ιδιαίτερα σε συνδυασμό με SCG αλγόριθμο οπισθοδιάδοσης. Ο LM αντίστοιχα με ίδιο αριθμό επιπέδων παρουσίασε γενικά πολύ μεγαλύτερο MSE και ως εκ τούτου περιοριστήκαμε στην εκπαίδευση μονο-επίπεδων δικτύων με χρήση LM, στα οποία και αποδίδει καλύτερα.

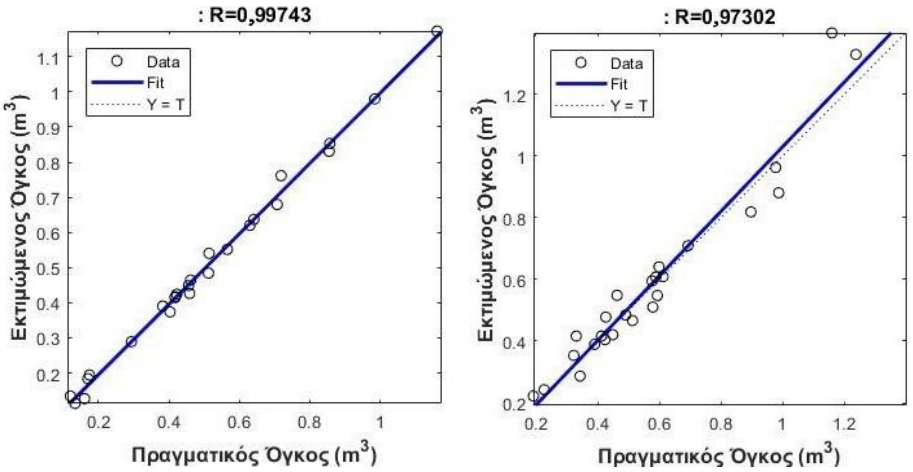
Πίνακας 3. *Αποτελέσματα νευρωνικών δικτύων. Η Αρχιτεκτονική του δικτύου αναφέρεται στον αριθμό νευρώνων των κρυφών επιπέδων του δικτύου. Ο αριθμός εισόδων καθορίζεται από το σετ εκπαίδευσης και το επίπεδο εξόδου έχει πάντα ένα νευρώνα.*

Αρχιτεκτονική	**Είσοδοι**	**Εποχές**	**Αλγόριθμος**	**RMSE**	**MSE**	**R^2**	**MAE**
20-10-5-2	$d_{1,3}$, h	200	SCG	7,56%	0,57%	0,9117	0,0531
10-5-2	$d_{1,3}$, h	200	SCG	5,65%	0,32%	0,9498	0,0415
10	$d_{1,3}$, h	150	LM	6,1%	0,37%	0,9396	0,0439
10-5-2	Relative d, h	200	SCG	4,65%	0,076%	0,9993	0,0200
20-10-5-2	Relative d, h	200	SCG	4,52%	0,06%	0,9994	0,0186
10	Relative d, h	100	LM	23,28%	0,053%	0,9491	0,0698
20-10-5-2	Relative d	200	SCG	8,8%	0,7%	0,9919	0,0650
10-5-2	Relative d	200	SCG	7,36%	0,54%	0,9947	0,0599
10	Relative d	100	LM	17,65%	3,12%	0,9697	0,1294
10-5-2	$d_{1,3}$	200	SCG	7,92%	0,63%	0,9077	0,0625
10	$d_{1,3}$	100	LM	7,53%	0,57%	0,9319	0,0632

Σε σχέση με τις εξισώσεις παλινδρόμησης, όπως φαίνεται στους πίνακες 2, 3, καθώς και στα διαγράμματα της εικόνας 3, τα νευρωνικά δίκτυα αποδίδουν αρκετά καλύτερα. Τα δίκτυα με ίδιες εισόδους με αυτές των εξισώσεων, τις $d_{1,3}$, h,

ξεπέρασαν γενικά την απόδοση των εξισώσεων κατά μέχρι και 1,24 RMSE. Με εισόδους δικτύων τις σχετικές διαμέτρους και το ύψος, η απόδοση των δικτύων παρουσίασε ακόμα μεγαλύτερη διαφορά από τις εξισώσεις. Αξιοσημείωτα είναι και τα αποτελέσματα των δικτύων με μία μόνο είσοδο, αυτήν της στηθιαίας διαμέτρου. Ανταγωνίζονται την απόδοση των εξισώσεων και ταυτόχρονα εξαλείφουν τη χρήση του ύψους, ως επιπλέον μέτρηση και ανεξάρτητη μεταβλητή.

Εικόνα 3. *Αριστερά, διάγραμμα παραδείγματος παλινδρόμησης πραγματικού με εκτιμώμενου όγκου με βάση τα νευρωνικά δίκτυα. Δεξιά παρουσιάζεται αντίστοιχο διάγραμμα με βάση τις εξισώσεις.*



5. ΣΥΜΠΕΡΑΣΜΑΤΑ

Σε αυτήν την εργασία εξετάστηκαν διαφορετικοί τρόποι προσέγγισης της εκτίμησης του όγκου δέντρων που ανήκουν στο Δημοτικό Δάσος Νάουσας. Με βάση τα συμπεράσματα της μελέτης το ζητούμενο προκύπτει με ικανοποιητική ακρίβεια με χρήση εξισώσεων παλινδρόμησης, αλλά τα νευρωνικά δίκτυα γενικά αποδίδουν πολύ καλύτερα. Ειδικά δε, με βαθιές αρχιτεκτονικές και όταν υπάρχουν περισσότερες διαθέσιμες μετρήσεις, η προσέγγιση όγκου πραγματοποιείται με εξαιρετική ακρίβεια. Ενδιαφέρον παρουσιάζουν τα αποτελέσματα των δικτύων με χρήση μόνο της στηθιαίας διαμέτρου ως είσοδο, ανταγωνίζονται αυτά των εξισώσεων παλινδρόμησης με εκπαίδευση ενός σχετικά απλού δικτύου. Τέλος, τα νευρωνικά δίκτυα εξάγουν τα συμπεράσματά τους απευθείας από τα δεδομένα, χωρίς να εισάγονται επιπλέον υποθέσεις και περιορισμοί από τον ερευνητή, κάτι που σε συνδυασμό με την απόδοσή τους τα καθιστά ένα ισχυρό εργαλείο. Συμπερασματικά, ανάλογα με την

πρακτική εφαρμογή και τα περιθώρια ακρίβειας-χρόνου προσφέρονται αρκετές προσεγγίσεις στον ερευνητή.

ABSTRACT

Forest attributes and more specifically timber volume data parameters, provide crucial information for the management of forest ecosystems. The dataset, which was randomly sampled for statistical analysis, includes measurements of 250 trees from the territory of the Municipal Forest of Naoussa. The volume estimation process was initially performed with equations based on applications of simple and multiple linear regression. Subsequently the tree volume was estimated with neural networks, which perform well in similar forest applications. Respectively neural networks of various architectures (layers, neurons per layer) and backpropagation algorithms were employed. Based on the results of our methodology, only transformed or weighted equation models produce satisfying results in the process of volume estimation. On the other hand, neural network performance on tree volume estimation, especially when trained with deeper architectures, surpasses regression equation models as their results diverge far less from the original volume.

Keywords: Regression models, Neural networks, Municipal Forest of Naoussa

Ευχαριστίες: Αρχικά να εκφράσουμε τις ευχαριστίες στον κ. Σταματέλλο Γεώργιο, καθηγητή ΑΠΘ, για τις συζητήσεις πτυχών της εργασίας. Επίσης, ως τρίτο συμβαλλόμενο μέλος θα ήθελα να ευχαριστήσω την ΓΓΕΤ καθώς και το ΕΛΙΔΕΚ και να επισημάνω από πλευράς μου ότι: «This research has been financially supported by General Secretariat for Research and Technology (GSRT) and the Hellenic Foundation for Research and Innovation (HFRI) (Scholarship Code: 1319)".

ΑΝΑΦΟΡΕΣ

Diamantopoulou, M. (2005). Artificial neural networks as an alternative tool to pine bark volume estimation. Computers and Electronics in Agriculture 48: 235-244.

Farrar, R. and Murphy, P. (1987). Taper functions for predicting product volume in natural shortleaf pines. USDA Research paper SO-234.

Fukushima, K. (1982). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. 36(4):193–2.

Ioffe, S.and Szegedy C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift, CoRR abs/1502.03167.

Levenberg, K. (1944). A method for the solution of certain non–linear problems in least squares. Quarterly of Applied Mathematics 2(2):164 – 168.

Li R. and Weiskittel A.R. (2010). Comparison of model form for estimating stem taper and volume in the primary conifer species of the North American Acadian Region. Annals of Forest Science 67. 17p.

Maier, H.R. and Dandy, G.C. (2000). Neural networks for the prediction and forecasting of water resources variables: A review of modelling issues and applications. Environ. Model. Softw. 15(1):101–124.

Μάτης, Κ. (1994). Μια μέθοδος κατάρτισης μαζοπινάκων δένδρου Επιστημονική Επετηρίδα Τμήματος Δασολογίας & Φυσικού Περιβάλλοντος, Τόμος ΛΖ, σελ. 317-336.

Moller, M. F. (1993). A scaled conjugate gradient algorithm for fast supervised learning. Neural Networks, 6:525—533.

Ozcelik, R. and Bal, C. (2013). Effects of adding crown variables in stem taper and volume predictions for black pine. Turkish Journal of Agriculture and Forestry 37:231-242.

Ozcelık, R. Diamantopoulou, M., Eker, M., and Gurlevık, N. (2017). Artificial Neural Network Models: An Alternative Approach for Reliable Aboveground Pine Tree Biomass Prediction. Forest Science 63(3):291-302.

Razi, M. A. and Athappilly, K. (2005). A comparative predictive analysis of neural networks (NNs), nonlinear regression and classification and regression tree (CART) models. Department of Business Information Systems, Haworth College of Business, Western Michigan University, Kalamazoo, MI 49008-3821, USA.

Rumelhart, D.E., Hinton, G.E., and R. J. Williams. (1986). "Learning Internal Representations by Error Propagation". David E. Rumelhart, James L. McClelland, and the PDP research group. (editors), Parallel distributed processing: Explorations in the microstructure of cognition, Volume 1: Foundation. MIT Press.

Tabacchi, G., Di Cosmo, L., and Gasparini, P. (2011). Aboveground tree volume and phytomass prediction equations for forest species in Italy, European Journal of Forest Research 130: 911-934.

Tiruveedhula, M.P., Fan,J., Sadasivuni, R.R., Durbha, S.S., and Evans, D.L. (2010). Comparative analysis of spectral unmixing and neural networks for estimating small diameter tree above-ground biomass in the State of Mississippi. P. 76–85 in Proc. of the 7th southern forestry and natural resources GIS conference, 2009 December 7–9. Warnell School of Forestry and Natural Resources, University of Georgia, Athens, GA.

Vahedi, A. A. (2016). Artificial neural network application in comparison with modeling allometric equations for predicting above-ground biomass in the

Hyrcanian mixed-beech forests of Iran. Research Institute of Forests and Rangelands, Agricultural Research Education and Extension Organization (AREEO), P.O. Box: 13185-116, Tehran, Iran.

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 31ου Πανελληνίου Συνεδρίου Στατιστικής (2018), σελ.132-145

ΕΚΤΙΜΗΣΗ ΑΠΟΘΕΜΑΤΩΝ ΜΕ ΧΡΗΣΗ ΜΕΘΟΔΩΝ KALMAN FILTER

Ι. Μπαντούνας, Γ. Πιτσέλης

Πανεπιστήμιο Πειραιώς

ibadounas@yahoo.gr & pitselis@unipi.gr

ΠΕΡΙΛΗΨΗ

Η εκτίμηση των αποθεμάτων μίας ασφαλιστικής εταιρίας στον κλάδο ζημιών αποτελεί μία άκρως σημαντική διαδικασία καθώς έχει ως σκοπό την οποιαδήποτε κάλυψη των μελλοντικών αξιώσεων από ασφαλιστικούς κινδύνους οι οποίοι έχουν ήδη επέλθει. Η διαδικασία εκτίμησης των αποθεμάτων περιλαμβάνει αρκετά περίπλοκες μεθόδους, οι οποίες πρέπει να συνδυάζονται με την κρίση και την εμπειρία του εκάστοτε αναλογιστή. Οι περισσότερες μέθοδοι αποθεματοποίησης απαιτούν την ύπαρξη παρελθοντικών στοιχείων της εταιρείας, πάνω στα οποία θα βασιστεί η εκτίμηση των μελλοντικών αποθεμάτων. Στην παρούσα εργασία παρουσιάζεται μία μέθοδος εκτίμησης των αποθεμάτων με χρήση του φίλτρου Kalman (Kalman Filter), έναν αλγόριθμο που εφαρμόζεται σε συστήματα που δέχονται εξωτερικές φυσικές διαταραχές (θορύβους) και έχει ως σκοπό τη δημιουργία μιας νέας εκτίμησης της κατάστασης του συστήματος, απαλλαγμένη από διαταραχές. Επιπλέον, επεκτείνουμε τον αλγόριθμο κάνοντάς τον ανθεκτικό (robust) σε ακραίες τιμές, οι οποίες αν αγνοηθούν θα υπερεκτιμήσουν το τελικό απόθεμα που απαιτείται για την ασφαλιστική υποχρέωση της εταιρίας.

Λέξεις Κλειδιά: Αποθεματοποίηση, Φίλτρο Kalman, Ανθεκτικός (Robust).

1. Εισαγωγή

Η αποθεματοποίηση ζημιών είναι μια σημαντική διαδικασία των ασφαλιστικών εταιριών με σκοπό την πρόβλεψη των απαιτούμενων κεφαλαίων για την κάλυψη μελλοντικών απαιτήσεων. Στην αναλογιστική βιβλιογραφία υπάρχουν πολλές εργασίες που περιγράφουν αναλυτικά στοχαστικές μεθόδους αποθεματοποίησης απαιτήσεων, μεταξύ των οποίων είναι τα άρθρα των Kremer (1982), Verrall (1994) καθώς και τα βιβλία των Taylor (2000), Wüthrich & Merz (2008) και των Radtke et al. (2012). Ένα από τα βασικά προβλήματα των βασικών τεχνικών εκτίμησης των αποθεμάτων μίας ασφαλιστικής εταιρίας (όπως είναι αυτή της μεθόδου Chain Ladder) είναι ότι δημιουργούν μεγάλες «σχέσεις - συσχετίσεις» μεταξύ των ετών που συνέβη το ατύχημα (ή των ετών που καταγράφηκε μία ζημιά). Αυτό συμβαίνει κυρίως διότι τα μοντέλα αυτά βασίζονται στην υπόθεση ότι οι παράγοντες εξέλιξης έχουν την ίδια μορφή για κάθε έτος ατυχήματος. Χρησιμοποιώντας μοντέλα χώρου καταστάσεων (state space models) δίνεται η δυνατότητα στους παράγοντες εξέλιξης να έχουν διαφορετική μορφή ανάλογα με το έτος ατυχήματος. Συγκεκριμένα, τα μοντέλα χώρου καταστάσεων είναι πολύ σημαντικά γιατί κυρίως μπορούν να δώσουν μία ποικιλία από διαφορετικά μοντέλα τα οποία είναι εύκολα ερμηνεύσιμα, αλλά και ταυτόχρονα οδηγούν σε συνεπείς εκτιμήσεις μέσω της μεθόδου του φίλτρου Kalman (Kalman Filter). Στις ασφαλίσεις ζημιών μπορούν να χρησιμοποιηθούν ως ένα εργαλείο ώστε να αυτοματοποιηθεί η διαδικασία αποθεματοποίησης.

Ένα μοντέλο χώρου καταστάσεων προσδιορίζεται από δύο εξισώσεις οι οποίες δηλώνουν την εξάρτηση που υπάρχει μεταξύ των παρατηρήσεων και της κατάστασης

- Η εξίσωση παρατήρησης (observation equation) ορίζεται ως

$$\mathbf{y}_t = \mathbf{F}_t\, \boldsymbol{\theta}_t + \mathbf{w}_t\,,\ t \geq 0,$$

όπου $\mathbf{y}_t$ είναι ένα διάνυσμα με τις παρατηρήσεις, $\boldsymbol{\theta}_t$ είναι ένα άγνωστο διάνυσμα κατάστασης, $\mathbf{F}_t$ είναι ένας γνωστός πίνακας, ο οποίος συνδέει το διάνυσμα κατάστασης $\boldsymbol{\theta}_t$ με τις παρατηρήσεις και τέλος $\mathbf{w}_t$ είναι τα πιθανά σφάλματα.

- Η εξίσωση κατάστασης (state equation) ορίζεται ως

$$\boldsymbol{\theta}_{t+1} = \mathbf{G}_{t+1}\boldsymbol{\theta}_t + \mathbf{u}_{t+1},\ t \geq 0,$$

η οποία χαρακτηρίζει το διάνυσμα $\boldsymbol{\theta}_t$ ως Μαρκοβιανή μορφή χρησιμοποιώντας τον πίνακα καταστάσεων μετάβασης $\mathbf{G}_{t+1}$. Όπως και η εξίσωση παρατήρησης έτσι και η εξίσωση κατάστασης συμπεριλαμβάνει ένα τυχαίο διάνυσμα $\mathbf{u}_t$ σφαλμάτων.

Το βασικό σημείο αυτής της προσέγγισης είναι η εξαγωγή συμπερασμάτων σχετικά με το διάνυσμα κατάστασης $\boldsymbol{\theta}_t$ ενώ έχει παρατηρηθεί ένα σύνολο δεδομένων $\boldsymbol{y_t}$. Γίνεται η υπόθεση ότι $\boldsymbol{y_t}$ είναι ένα διάνυσμα n×1, $\boldsymbol{\theta}_t$ είναι ένα διάνυσμα καταστάσεων r×1, $\mathbf{F}_t$ είναι πίνακας n×r και $\mathbf{G}_t$ είναι πίνακας r×r. Τα διανύσματα με τα σφάλματα $\mathbf{w}_t$ και $\mathbf{u}_t$ είναι διαστάσεων n×1 και r×1 με πίνακες διακύμανσης-συνδιακύμανσης

$$E\left[\mathbf{u}_t\,\mathbf{u}'_{t+k}\right] = \begin{cases} \mathbf{Q}_t, & k = 0, \\ 0, & \text{αλλού} \end{cases} \quad \text{και} \quad E\left[\mathbf{w}_t\,\mathbf{w}'_{t+k}\right] = \begin{cases} \mathbf{H}_t, & k = 0, \\ 0, & \text{αλλού}, \end{cases}$$

όπου $\mathbf{Q}_t$ και $\mathbf{H}_t$ είναι r×r και n×n πίνακες οι οποίοι εξαρτώνται από το χρόνο t. Επιπλέον γίνεται η υπόθεση $E\left[\mathbf{u}_t\,\mathbf{w}'_{t+k}\right] = 0, \forall t, k$.

Υποθέτοντας ότι τα σφάλματα ακολουθούν την κανονική κατανομή, για την εκτίμηση του διανύσματος $\boldsymbol{\theta}_t$ χρησιμοποιείται ο αλγόριθμος του φίλτρου Kalman.

2. Ο Αλγόριθμος του Φίλτρου Kalman (Kalman Filter)

Η εκτίμηση του διανύσματος $\boldsymbol{\theta}_t$ βασίζεται σε μία σειρά δεδομένων που έχουν παρατηρηθεί $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_t$. Ο αλγόριθμος του φίλτρου Kalman είναι ένα σύνολο από αναδρομικές σχέσεις στις οποίες γίνεται η εκτίμηση του διανύσματος $\boldsymbol{\theta}_t$ σύμφωνα με το διάνυσμα $\boldsymbol{\theta}_{t-1}$ και το νέο σετ δεδομένων $\mathbf{y}_t$. Τα βήματα του αλγορίθμου περιγράφονται παρακάτω:

Βήμα 1: $\hat{\mathbf{y}}_t = \mathbf{F}_t \hat{\boldsymbol{\theta}}_{t|t-1}$,

Βήμα 2: $\boldsymbol{\varepsilon}_t = \mathbf{y}_t - \hat{\mathbf{y}}_t$,

Βήμα 3: $\hat{\boldsymbol{\theta}}_{t|t} = \hat{\boldsymbol{\theta}}_{t|t-1} + \mathbf{P}_{t|t-1}\mathbf{F}_t \left(\mathbf{F}_t' \mathbf{P}_{t|t-1} \mathbf{F}_t + \mathbf{H}_t\right)^{-1} \boldsymbol{\varepsilon}_t$,

Βήμα 4: $\hat{\boldsymbol{\theta}}_{t+1|t} = \mathbf{G}_{t+1} \hat{\boldsymbol{\theta}}_{t|t}$,

Βήμα 5: $\mathbf{P}_{t|t} = \mathbf{P}_{t|t-1} - \mathbf{P}_{t|t-1}\mathbf{F}_t \left(\mathbf{F}_t' \mathbf{P}_{t|t-1} \mathbf{F}_t + \mathbf{H}_t\right)^{-1} \mathbf{F}_t' \mathbf{P}_{t|t-1}$,

Βήμα 6: $\mathbf{P}_{t+1|t} = \mathbf{G}_{t+1}\mathbf{P}_{t|t}\mathbf{G}'_{t+1} + \mathbf{Q}_{t+1}$.

Στον παραπάνω αλγόριθμο, $\hat{\boldsymbol{\theta}}_{t|t}$ είναι η εκτίμηση του $\boldsymbol{\theta}_t$ όταν είναι γνωστό το σύνολο των δεδομένων $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_t$ στο χρόνο t, ενώ $\mathbf{P}_{t|t}$ είναι ο πίνακας διακύμανσης - συνδιακύμανσης αυτής της εκτίμησης. Αντίστοιχα, το διάνυσμα $\hat{\boldsymbol{\theta}}_{t+1|t}$ και ο πίνακας $\mathbf{P}_{t+1|t}$ είναι η εκτίμηση του $\boldsymbol{\theta}_{t+1}$ και της διακύμανσης σύμφωνα με το ίδιο σύνολο δεδομένων σε χρόνο t. Στην πραγματικότητα ο αλγόριθμος δεν επανεκτιμά μόνο το διάνυσμα $\boldsymbol{\theta}_{t+1}$ αλλά και τον πίνακα διακύμανσης-συνδιακύμανσης αυτής της εκτίμησης. Στην περίπτωση που τα σφάλματα κατανέμονται κανονικά, είναι εφικτό να κατασκευαστούν (1-α)100% διαστήματα εμπιστοσύνης για τις παραμέτρους σύμφωνα με τους τύπους $\hat{\boldsymbol{\theta}}_{k,t+1|t} \pm t_a \sqrt{\mathbf{p}_{k,k,t+1|t}}$, όπου $\hat{\boldsymbol{\theta}}_{k,t+1|t}$ είναι το k στοιχείο του $\hat{\boldsymbol{\theta}}_{t+1|t}$, $\mathbf{p}_{k,k,t+1|t}$ είναι το (k,k) διαγώνιο στοιχείο του πίνακα $\mathbf{P}_{t+1|t}$ και t_a είναι το αντίστοιχο ποσοστημόριο.

3. Ο Ανθεκτικός (Robust) Αλγόριθμος Φίλτρου Kalman

Μία από τις υποθέσεις του αλγορίθμου του φίλτρου Kalman είναι η κανονικότητα των σφαλμάτων και τα περισσότερα μοντέλα βασίζονται σε αυτή την υπόθεση. Σε περίπτωση όμως ασυμμετρίας της κατανομής των σφαλμάτων είτε έκτοπων παρατηρήσεων (outliers), κάποιες υποθέσεις όπως η προηγούμενη παραβιάζονται, το οποίο σημαίνει ότι οι εκτιμήσεις των παραμέτρων, των διαστημάτων εμπιστοσύνης και άλλων στατιστικών ποσοτήτων είναι αναξιόπιστες (Cipra, & Romera, 1991). Στο Βήμα 3 του αλγορίθμου, η έλλειψη κανονικότητας των σφαλμάτων ή η ύπαρξη έκτοπων τιμών θα οδηγήσει σε λανθασμένες εκτιμήσεις για το $\boldsymbol{\theta}_{t|t}$ αλλά και για όλο το διάνυσμα $\boldsymbol{\theta}_{t+i|t+i}, \forall i > 0$. Για το λόγο αυτό, μία απλή ανθεκτική (robust) στρατηγική είναι να απαλειφθούν τέτοιες επιρροές αντικαθιστώντας τα σφάλματα $\boldsymbol{\varepsilon}_t$ με μία φραγμένη συνάρτηση $\psi\left(\boldsymbol{\varepsilon}_t\right)$. Αντί επομένως να χρησιμοποιηθούν τα σφάλματα $\boldsymbol{\varepsilon}_t$, εφαρμόζεται η συνάρτηση του Huber (Huber, 1981)

$$H_b\left(X\right) = X \cdot \min\left\{1, \frac{b}{\left\|X\right\|}\right\},$$

όπου $\left\| \ \right\|$ είναι η ευκλείδεια νόρμα και $0 < b < \infty$. Σε αυτή την περίπτωση, το Βήμα 3 του αλγορίθμου μετασχηματίζεται ως εξής:

$$\hat{\boldsymbol{\theta}}_{t|t} = \hat{\boldsymbol{\theta}}_{t|t-1} + \mathbf{P}_{t|t-1}\mathbf{F}_t\left(\mathbf{F}_t'\mathbf{P}_{t|t-1}\mathbf{F}_t + \mathbf{H}_t\right)^{-1} H_b\left(\boldsymbol{\varepsilon}_t\right),$$

όπου η σταθερά b είναι μία παράμετρος προς εκτίμηση. Ο αλγόριθμος μπορεί να γίνει ανθεκτικός στην περίπτωση που μπορούν να χρησιμοποιηθούν διαφορετικές συναρτήσεις αλλά παρόμοιες με αυτή του Huber (Pitselis et al., 2015, Cipra & Romera, 1991). Μία προσέγγιση για να επιλεγεί η σταθερά είναι χρησιμοποιώντας την αποδοτικότητα (efficiency) σε σχέση με το απλό μοντέλο του φίλτρου Kalman. Αν οι τιμές της σταθεράς b είναι μικρές, τότε η σχετική αποδοτικότητα παίρνει μεγάλες τιμές (Moore et. al, 1980).

4. Δημιουργία Μοντέλων Φίλτρου Kalman σε Τρίγωνα Ζημιών

Οι ζημιές μίας ασφαλιστικής εταιρίας στον κλάδο ζημιών μελετώνται κυρίως χρησιμοποιώντας τρίγωνα εξέλιξης ζημιών. Τη μορφή των δεδομένων δηλώνει ο Πίνακας 4. Συγκεκριμένα, οι γραμμές δηλώνουν το έτος που συνέβη το ατύχημα, ενώ οι στήλες δηλώνουν το χρόνο εξέλιξης της ζημιάς, δηλαδή το χρονικό διάστημα που χρειάστηκε ώστε να καταβληθεί αποζημίωση (Verrall, 1994).

Πίνακας 4: Τρίγωνο εξέλιξης αποθεμάτων

	Έτος εξέλιξης j				
Έτος ατυχήματος i	**1**	**2**		**n-1**	**n**
1	$C_{1,1}$	$C_{1,2}$	⋯	$C_{1,n-1}$	$C_{1,n}$
2	$C_{2,1}$	$C_{2,2}$		$C_{2,n-1}$	
⋮	⋮		⋰		
	⋮	$C_{n-1,2}$			
n	$C_{n,1}$				

Επίσης, η διαγώνιος δηλώνει το ημερολογιακό έτος. Η τιμή $C_{i,j}$ δηλώνει μία ζημιά που έγινε το έτος i αλλά πληρώθηκε το έτος j. Φυσικά, τα δεδομένα παρουσιάζονται ως τρίγωνο διότι οι γνωστές παρατηρήσεις είναι έως και την τελευταία διαγώνιο με στοιχεία $C_{n-i+1,i}, \forall i = 1,2,\ldots,n,$ τα οποία δηλώνουν το ποσό των ζημιών που έγιναν το έτος $n-i+1$και πληρώθηκαν το έτος $n,$ όπου n είναι το σύνολο των ετών που είναι γνωστά τα μεγέθη των ζημιών. Όλες οι υπόλοιπες τιμές, δηλαδή το τρίγωνο δεξιά κάτω, πρέπει να εκτιμηθούν διότι αποτελούν την εξέλιξη των ζημιών που έγιναν και πρόκειται να πληρωθούν σε μελλοντικό χρόνο. Συγκεκριμένα, θα πρέπει

να υπολογιστεί το απόθεμα $R_i = \sum_{n-i+2}^{n} C_{i,j}$ που αναμένεται να έχει μία ασφαλιστική εταιρία για κάθε έτος ατυχήματος $i = 2,\ldots,n$, με $R_1 = 0$.

Για τη μοντελοποίηση των ζημιών ενός τριγώνου χρησιμοποιείται το μοντέλο των λογαρίθμων των δεδομένων

$$\log\left(C_{i,j}\right) = \mu + \alpha_i + \beta_j + e_{i,j},$$

με μ>0, α_i μία παράμετρος που περιγράφει το έτος ατυχήματος και β_j μία παράμετρος που περιγράφει το έτος εξέλιξης της ζημιάς ενώ $e_{i,j}$ είναι τα σφάλματα που ακολουθούν την κανονική κατανομή με μέση τιμή 0 και κάποια σταθερή διακύμανση (De Jong and Zehnwirth, 1983). Για την εκτίμηση των αποθεμάτων, χρησιμοποιώντας το φίλτρο Kalman τα δεδομένα θα έχουν την ακόλουθη μορφή σύμφωνα με το ημερολογιακό έτος (βλέπε Verrall, 1994):

$$C_{1,1}, \begin{bmatrix} C_{1,2} \\ C_{2,1} \end{bmatrix}, \begin{bmatrix} C_{1,3} \\ C_{2,2} \\ C_{3,2} \end{bmatrix}, \cdots$$

Η αναπαράσταση του λογαριθμικού μοντέλου σε ένα Μοντέλο Χώρου Καταστάσεων είναι

$$\underbrace{\begin{pmatrix} \log\left(C_{1,t}\right) \\ \log\left(C_{2,t-1}\right) \\ \vdots \\ \log\left(C_{t,1}\right) \end{pmatrix}}_{\mathbf{y}_t} = \underbrace{\begin{pmatrix} 1 & 0 & \ldots & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 0 & \cdots & 0 & 0 \\ 1 & 0 & \cdots & 1 & 0 \end{pmatrix}}_{\mathbf{F}_t} \underbrace{\begin{pmatrix} \mu \\ \alpha_1 \\ \vdots \\ \vdots \\ \beta_t \end{pmatrix}}_{\boldsymbol{\theta}_t} + \underbrace{\begin{pmatrix} w_{1,t} \\ w_{2,t-1} \\ \vdots \\ \vdots \\ w_{t,1} \end{pmatrix}}_{\mathbf{w}_t},$$

με το διάνυσμα καταστάσεων $\boldsymbol{\theta}_t = \left(\mu, \alpha_1, \ldots, \beta_t\right)'$ να ικανοποιεί την εξίσωση

$$\boldsymbol{\theta}_{t+1} = \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}}_{\mathbf{G}_{t+1}} \boldsymbol{\theta}_t + \begin{bmatrix} 0 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \mathbf{u}_t ,$$

όπου $\mathbf{u}_t$ είναι ένα διάνυσμα τυχαίων σφαλμάτων. Να τονιστεί ότι οι διαστάσεις όλων των πινάκων μεταβάλλονται δυναμικά καθώς νέες πληροφορίες εισάγονται στο μοντέλο. Εφαρμόζοντας τον αλγόριθμο, προκύπτουν οι εκτιμήσεις των αγνώστων παραμέτρων ώστε να γίνει η τελική εκτίμηση του αποθέματος.

5. Αριθμητική εφαρμογή

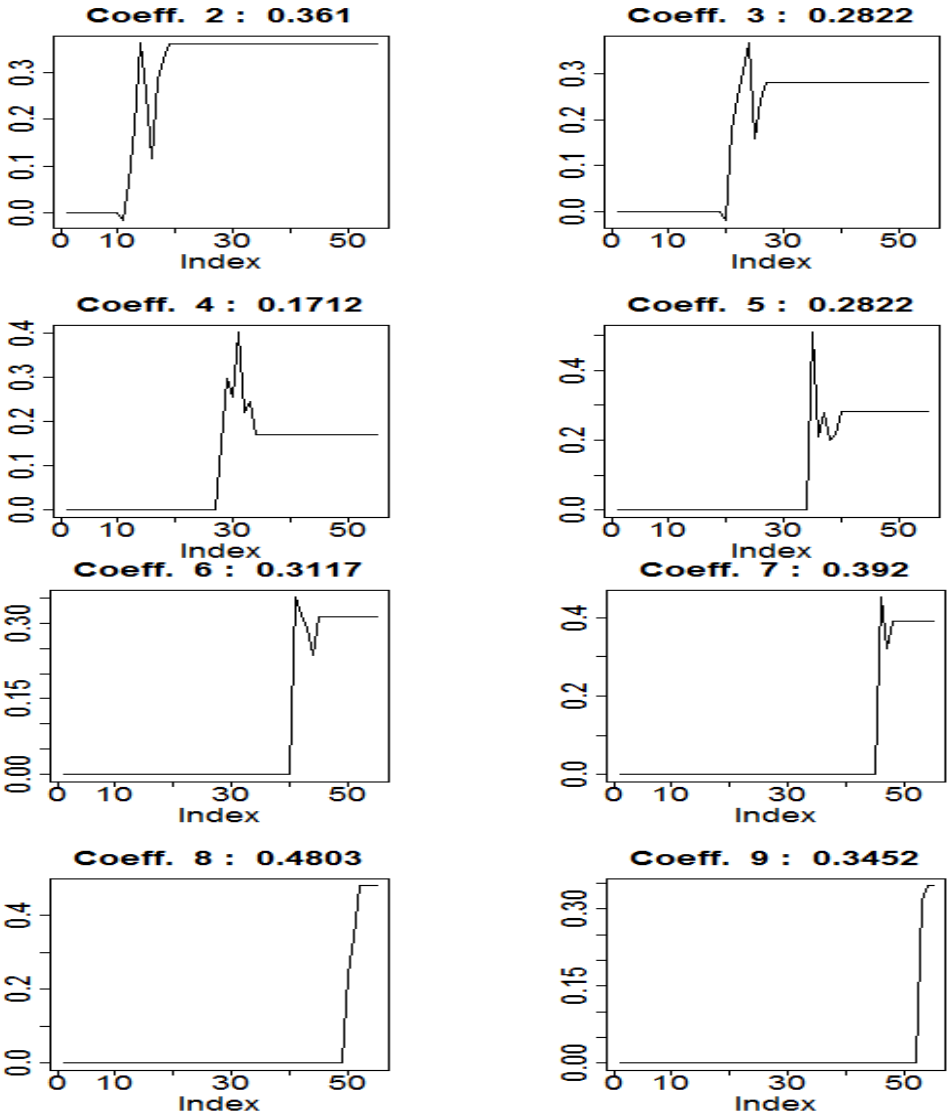
Για την εφαρμογή της μεθόδου θα χρησιμοποιήθηκε το πακέτο R και ένα ευρέως γνωστό σύνολο δεδομένων το οποίο δείχνει ο Πίνακας 5 (Taylor & Ashe, 1983). Στο συγκεκριμένο τρίγωνο εξέλιξης αποθεμάτων υπάρχουν δεδομένα για 10 έτη ατυχήματος καθώς και για 10 έτη εξέλιξης. Στόχος είναι η εκτίμηση του κάτω τριγώνου το οποίο θα οδηγήσει και στην εκτίμηση των αναγκαίων αποθεμάτων που καλείται να διαθέτει η εκάστοτε ασφαλιστική εταιρία ώστε να είναι φερέγγυα για τις μελλοντικές της υποχρεώσεις.

Πίνακας 5: Taylor and Ashe data (1983)

	1	2	3	4	5	6	7	8	9	10
1	357848	766940	610542	482940	527326	574398	146342	139950	227229	67948
2	352118	884021	933894	1183289	445745	320996	527804	266172	425046	
3	290507	1001799	926219	1016654	750816	146923	495992	280405		
4	310608	1108250	776189	1562400	272482	352053	206286			
5	443160	693190	991983	769488	504851	470639				
6	396132	937085	847498	805037	705960					
7	440832	847631	1131398	1063269						
8	359480	1061648	1443370							
9	376686	986608								
10	344014									

Εφαρμόζοντας τον αλγόριθμο του φίλτρου Kalman, λαμβάνουμε τις εκτιμήσεις του διανύσματος κατάστασης του μοντέλου (Πίνακας 6). Η Εικόνα 1 παρουσιάζει τη διαχρονική εκτίμηση κάποιων στοιχείων του διανύσματος κατάστασης (β_1 έως β_8) για τα οποία οι μεταβολές στις εκτιμώμενες τιμές είναι εμφανείς, καθώς εισάγονται νέες πληροφορίες στο τρίγωνο. Συγκεκριμένα, διαχρονικά εισάγονται νέες ζημιές στο τρίγωνο και επομένως επανεκτιμούνται οι παράμετροι του μοντέλου. Υπάρχουν περιπτώσεις όπου αρχικά κάποια από τα στοιχεία του διανύσματος κατάστασης δεν μπορούν να εκτιμηθούν, καθώς όμως εισέρχονται πληροφορίες στο μοντέλο τότε είναι εφικτό να εκτιμηθούν, αλλά και να επανεκτιμηθούν στα επόμενα βήματα του αλγορίθμου.

Εικόνα 1: Διαχρονική εκτίμηση στοιχείων του διανύσματος κατάστασης



Να τονιστεί ότι πρόκειται για ένα διάνυσμα, το οποίο περιέχει τις εκτιμήσεις των παραμέτρων και για τα έτη ατυχήματος (παράμετροι α_i), αλλά και για τα έτη εξέλιξης (παράμετροι β_i).

Πίνακας 6: Εκτιμήσεις διανύσματος κατάστασης

μ	**α1**	**α2**	**α3**	**α4**	**α5**	**α6**	**α7**	**α8**	
12.51	0.3610	0.2822	0.1712	0.2822	0.3118	0.392	0.4802	0.3452	
β0	**β1**	**β2**	**β3**	**β4**	**β5**	**β6**	**β7**	**β8**	**β9**

0.229	0.9111	0.9387	0.9649	0.3832	-0.005	-0.118	-0.439	-0.053	-1.393

Χρησιμοποιώντας τις εκτιμήσεις του διανύσματος κατάστασης εκτιμάται το απόθεμα για τα υπόλοιπα έτη. Ο Πίνακας 7 παρουσιάζει τις εκτιμήσεις με τα μελλοντικά αναμενόμενα αποθέματα. Να τονιστεί ότι οι εκτιμήσεις αυτές αποτελούν το αναμενόμενο απόθεμα που χρειάζεται να έχει η ασφαλιστική εταιρία ώστε να είναι συνεπής στις μελλοντικές της υποχρεώσεις, οι οποίες προέκυψαν από την εξέλιξη των ζημιών που έχουν ήδη γίνει.

Πίνακας 7: Τρίγωνο εξέλιξης ζημιών με εκτιμήσεις

	1	2	3	4	5	6	7	8	9	10
1	357848	766940	610542	482940	527326	574398	146342	139950	227229	67948
2	352118	884021	933894	1183289	445745	320996	527804	266172	425046	10927
3	290507	1001799	926219	1016654	750816	146923	495992	280405	379508	102650
4	310608	1108250	776189	1562400	272482	352053	206286	228731	340091	91988
5	443160	693190	991983	769488	504851	470639	351067	256033	380684	102968
6	396132	937085	847498	805037	705960	404479	362429	264319	393004	106300
7	440832	847631	1131398	1063269	646825	439787	394066	287392	427310	115580
8	359480	1061648	1443370	1268870	710402	483014	432799	315640	469311	126940
9	376686	986608	1090101	1120657	627422	426594	382246	278771	414492	112113
10	344014	973601	1001989	1030076	576709	392113	351349	256238	380990	103051

Σύμφωνα με τις εκτιμήσεις του διανύσματος κατάστασης και το μελλοντικό πίνακα σχεδιασμού προκύπτει ο Πίνακας 8 όπου παρουσιάζονται οι εκτιμήσεις για το τελικό απόθεμα για κάθε έτος ατυχήματος.

Πίνακας 8: Αθροιστικές ζημιές και εκτίμηση αποθεμάτων και τελικών πληρωθεισών ζημιών ανά έτος ατυχήματος

	1	2	3	4	5	6	7	8	9	10	Συνολικές Τελικές Πληρωθείσες ζημιές	Τελικό Εκτιμώμενο Απόθεμα
1	357848	1124788	1735330	2218270	2745596	3319994	3466336	3606286	3833515	3901463	3901463	0
2	352118	1236139	2170033	3353322	3799067	4120063	4647867	4914039	5339085	5450013	5350012	110928
3	290507	1292306	2218525	3235179	3985995	4132918	4628910	4909315	5288823	5391473	5391473	482158
4	310608	1418858	2195047	3757447	4029929	4381982	4588268	4816999	5157090	5249079	5249078	660811

5	443160	1136350	2128333	2897821	3402672	3873311	4224379	4480411	4861096	4964064	4964063	1090753
6	396132	1333217	2180715	2985752	3691712	4096191	4458621	4722940	5115945	5222245	5222243	1530533
7	440832	1288463	2419861	3483130	4129955	4569742	4963809	5251200	5678511	5794091	5794090	2310961
8	359480	1421128	2864498	4133368	4843771	5326785	5759584	6075224	6544536	6671476	6671474	3806978
9	376686	1363294	2453395	3574052	4201475	4628070	5010315	5289086	5703579	5815692	5815690	4452398
10	344014	1317616	2319606	3349683	3926393	4318507	4669856	4926095	5307085	5410136	5410130	5066122
										Σύνολο	**53769716**	**19511642**

Το ολικό απόθεμα εκτιμάται 19511642 ενώ το τυπικό σφάλμα εκτιμάται 6266024 (32.11%). Επίσης, η διακύμανση των σφαλμάτων του μοντέλου εκτιμάται 0.116217. Προφανώς, καθώς αυξάνεται διαχρονικά το έτος ατυχήματος, το συνολικό αναμενόμενο απόθεμα αυξάνεται. Να τονιστεί επίσης ότι γίνεται η υπόθεση ότι μετά τα 10 έτη δεν υπάρχει εξέλιξη των ζημιών, με αποτέλεσμα το πρώτο έτος ατυχήματος η εκτίμηση του αποθέματος να είναι μηδενική, όπως δείχνει ο Πίνακας 8.

Τέλος, εφαρμόζεται το μοντέλο με το ανθεκτικό φίλτρο Kalman για διάφορες τιμές της παραμέτρου b. Συγκεκριμένα, για διάφορες τιμές της παραμέτρου b εκτιμάται το απόθεμα, η διακύμανση των σφαλμάτων, η σχετική αποδοτικότητα σε σχέση με το απλό φίλτρο Kalman (δηλαδή ο λόγος των μέσων τετραγωνικών σφαλμάτων) καθώς επίσης και το ποσοστό του αποθέματος ως προς τη συνολική διακύμανση των αποθεμάτων. Όπως φαίνεται από τα αποτελέσματα, καθώς αυξάνεται η τιμή της σταθεράς b, μειώνεται το απόθεμα που απαιτείται και επίσης η αποδοτικότητα τείνει στο 1, καθώς σε αυτή την περίπτωση δεν εμφανίζεται κάποια ακραία τιμή και οι δύο αλγόριθμοι βγάζουν ίδια αποτελέσματα. Επίσης, μειώνεται και το ποσοστό του αποθέματος ως προς τη συνολική διακύμανση. Ο Πίνακας 9 παρουσιάζει τα αποτελέσματα της εφαρμογής του Ανθεκτικού φίλτρου Kalman στα δεδομένα.

Πίνακας 9: Αποτελέσματα Ανθεκτικού Φίλτρου Kalman

α/α	**--b--**	**Απόθεμα**	**σ^2**	**Αποδοτικότητα**	**%SE**
1	10.00	26165879	0.17	1.43	0.36
2	10.20	25527872	0.16	1.37	0.36
3	10.40	24919698	0.15	1.32	0.35

4	10.60	24339883	0.15	1.27	0.35
5	10.80	23787045	0.14	1.22	0.35
6	11.00	23259881	0.14	1.18	0.34
7	11.20	22757171	0.13	1.14	0.34
8	11.40	22277765	0.13	1.11	0.33
9	11.60	21820582	0.13	1.08	0.33
10	11.80	21384606	0.12	1.05	0.33
11	12.00	20968880	0.12	1.03	0.33
12	12.20	20572506	0.12	1.02	0.32
13	12.40	20194635	0.12	1.01	0.32
14	12.60	19834471	0.12	1.00	0.32
15	12.80	19511642	0.12	1.00	0.32
16	13.00	19511642	0.12	1.00	0.32

ABSTRACT

The estimate of the total loss reserve of an insurance company is a very important process as it intends to cover future claims from insurance risks that have already occurred. The procedure of loss reserving involves several complicated methods, which must be combined with the judgment and experience of the actuary. Most loss reserving methods require past claims data for the estimation of future reserves. In this work, we present a loss reserving method using Kalman Filter, an algorithm applied to systems that receive external physical disturbances (noises) and aims to create a new estimate of the state of the system without disturbances. In addition, we extend the algorithm by making it robust at extreme values, which, if ignored, will overestimate the final reserve required for the company's liability.

ΑΝΑΦΟΡΕΣ

Busse, M., Müller, U., & Dacorogna, M., (2010). Robust Estimation of Reserve Risk, *ASTIN Bulletin*, 40(2), 453-489.

Cipra, T., & Romera, R., (1991). Robust Kalman Filter and its application in time series analysis. *Kybernetika*, 27(6), 481-494.

Huber, P. J., (1981). *Robust Statistics*, Wiley, New York.

De Jong, P., & Zehnwirth, B., (1983). Claims reserving, state-space models and the Kalman filter. *Journal of the Institute of Actuaries*, 110, 157-181.

Kremer, E., (1982). IBNR-Claims and the two-way model of ANOVA. *Scandinavian Actuarial Journal*, 47–55.

Moore, J. B., & Anderson, B. D. O., (1980). Coping with singular transition matrices in estimation and control stability theory, *Int. J. Contr*., vol. 3.

Pitselis, G., (2008). Robust Regression Credibility: The Influence Function Approach, *Insurance Mathematics and Economics*, 288-300.

Pitselis, G., Grigoriadou, V., & Badounas., I., (2015). Robust loss reserving in a log-linear model. *Insurance: Mathematics and Economics*, 64, 14–27.

Radtke, M., Schmidt, K. D., & Schnaus, A., (2012). *Handbook on Loss Reserving, EAA Lecture Notes*, European Actuarial Academy, Springer.

Taylor, G., C. & Ashe, F.R., (1983). Second moments of estimates of outstanding claims. *Journal of Econometrics*, 23, 37–61.

Taylor, G., C., (2000). *Loss Reserving - An Actuarial Perspective*. Kluwer Academic Publishers, Norwell, Mass.

Verdonck, T., Van Wouwe, & M., Dhaene, J., (2009). A Robustification of the Chain-Ladder Method. *North American Actuarial Journal*, 13(2), 280-298.

Verrall, R. J., (1989). A state space representation of the chain-ladder linear model. *Journal of the Institute of Actuaries*, 116, 589-610.

Verrall, R. J., (1994). *Statistical methods for the chain-ladder technique*. Casualty Actuarial Society Forum, Spring 1994, 393-446.

Wüthrich, M. V., & Merz, M., (2008). *Stochastic Claims Reserving Methods in Insurance*, John Wiley.

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 31ου Πανελληνίου Συνεδρίου Στατιστικής (2018), σελ. 146−160

ΠΡΟΒΛΕΨΗ ΘΝΗΣΙΜΟΤΗΤΑΣ ΜΕ ΤΗ ΧΡΗΣΗ ΑΝΑΛΟΓΙΣΤΙΚΩΝ ΥΠΟΔΕΙΓΜΑΤΩΝ ΑΞΙΟΠΙΣΤΙΑΣ

Α. Μποζίκας, Γ. Πιτσέλης
Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης, Πανεπιστήμιο Πειραιώς
{bozikas, pitselis}@unipi.gr

ΠΕΡΙΛΗΨΗ

Στην εργασία αυτή προτείνεται το υπόδειγμα παλινδρόμησης αξιοπιστίας χαρτοφυλακίου ως μέθοδος μοντελοποίησης και πρόβλεψης των δεικτών θνησιμότητας ενός πληθυσμού. Τα αποτελέσματα από την προσαρμογή του υποδείγματος αυτού στα ελληνικά δεδομένα αξιολογούνται με τη χρήση μέτρων ακρίβειας σφάλματος, οδηγώντας στο συμπέρασμα ότι οι προβλέψεις που προκύπτουν μέσω αξιοπιστίας χαρτοφυλακίου είναι πιο ακριβείς σε σύγκριση με τα αντίστοιχα αποτελέσματα των υποδειγμάτων θνησιμότητας Lee-Carter και CBD.

Λέξεις Κλειδιά: Lee-Carter, CBD, Θεωρία Αξιοπιστίας Χαρτοφυλακίου, Παλινδρόμηση, Πρόβλεψη Θνησιμότητας, Ελλάδα.

1. ΕΙΣΑΓΩΓΗ

Τις τελευταίες δεκαετίες, η θνησιμότητα έχει μειωθεί σημαντικά σε όλες τις ηλικίες παγκοσμίως. Αυτό οφείλεται κυρίως στη συνεχή βελτίωση των συνθηκών διαβίωσης και στην ανάπτυξη της ιατρικής επιστήμης. Επιπλέον, μελέτες έχουν δείξει ότι το παγκόσμιο προσδόκιμο ζωής αυξήθηκε κατά τους τελευταίους δύο αιώνες, περίπου 40 έτη για τους άνδρες και 45 για τις γυναίκες (Oeppen and Vaupel, 2002). Όμως, η επικείμενη αύξηση της μακροζωίας δημιουργεί υψηλότερες οικονομικές ευθύνες για τα συνταξιοδοτικά ταμεία και αναπροσαρμογή των εισφορών των εργαζομένων και των εργοδοτών.

Τα τελευταία χρόνια έχουν αναπτυχθεί αρκετές μέθοδοι για την καταγραφή των τάσεων θνησιμότητας ενός πληθυσμού. Πιο συγκεκριμένα, η ανάπτυξη καινοτόμων τεχνικών μοντελοποίησης της θνησιμότητας προκάλεσε το ενδιαφέρον των αναλογιστών και των δημογράφων. Μία από τις πιο σημαντικές μεθόδους πρόβλεψης θνησιμότητας αποτελεί το υπόδειγμα Lee-Carter (Lee and Carter, 1992). Επίσης, μια αξιοσημείωτη παραλλαγή του υποδείγματος Lee-Carter, ειδικά σχεδιασμένο για μεγαλύτερες ηλικίες προτάθηκε από τους Cairns et al. (2006) και αναφέρεται στη βιβλιογραφία ως CBD.

Συνήθως, ένα πρόβλημα που παρατηρείται κατά τη διαδικασία μοντελοποίησης της θνησιμότητας είναι ότι για ορισμένες χώρες δεν υπάρχει επαρκής αριθμός δεδομένων

προσαρμογής. Ως εκ τούτου, οι υπάρχουσες μέθοδοι μοντελοποίησης δεν μπορούν να στηρίξουν τις προβλέψεις τους σε πληθυσμιακά δεδομένα περιορισμένης ιστορικής περιόδου. Έτσι, η έρευνα έχει στραφεί στην αναζήτηση εναλλακτικών μεθόδων μοντελοποίησης, οι οποίες μπορούν να άρουν τους παραπάνω περιορισμούς. Η θεωρία αξιοπιστίας χαρτοφυλακίου (credibility theory) που παραδοσιακά χρησιμοποιείται στον κλάδο γενικών ασφαλίσεων δύναται να άρει τους παραπάνω περιορισμούς και να χρησιμεύσει ως εργαλείο για την πρόβλεψη θνησιμότητας.

Ανατρέχοντας στη βιβλιογραφία, οι Hardy and Panjer (1998) ήταν οι πρώτοι που χρησιμοποίησαν τη θεωρία αξιοπιστίας ως μεθοδολογία για τον υπολογισμό των μέτρων που συνδέονται με τον κίνδυνο θνησιμότητας σε χαρτοφυλάκια ασφαλιστικών εταιρειών. Πολύ αργότερα, οι Salhi et al. (2016) πρότειναν μια προσεγγιστική μέθοδο αξιοπιστίας για την αναθεώρηση των παραμέτρων προσαρμογής της καμπύλης θνησιμότητας του Makeham, ενώ οι Schinzinger et al. (2016) παρουσίασαν μια μέθοδο αξιοπιστίας για τη μοντελοποίηση των ποσοστών βελτίωσης της θνησιμότητας σε διάφορους πληθυσμούς. Δύο αξιοσημείωτες εργασίες στη πρόβλεψη θνησιμότητας με χρήση κλασσικών υποδειγμάτων αξιοπιστίας δημοσιεύτηκαν πρόσφατα από τους Tsai and Lin (2017a, 2017b). Στην πρώτη εργασία τους εφάρμοσαν το υπόδειγμα αξιοπιστίας του Bülmann στα δεδομένα θνησιμότητας της Ιαπωνίας, του Ηνωμένου Βασιλείου και των Ηνωμένων Πολιτειών, ενώ στη δεύτερη, ενσωμάτωσαν το ίδιο υπόδειγμα αξιοπιστίας στις μεθόδους των Lee and Carter (1992) και Cairns et al. (2006) και στο γραμμικό υπόδειγμα των Tsai and Yang (2015) με σκοπό τη βελτίωση των προβλέψεων για τα δεδομένα του Ηνωμένου Βασιλείου, με εφαρμογές στα καθαρά ενιαία ασφάλιστρα προϊόντων ασφάλισης ζωής.

Ωστόσο, σε πολλές περιπτώσεις διαπιστώθηκε ότι τα εμπειρικά δεδομένα θνησιμότητας παρουσιάζουν μια έντονη πτωτική τάση κατά τη διάρκεια μιας παρατηρούμενης περιόδου. Σε τέτοιες περιπτώσεις προτιμάται η χρήση μεθόδων αξιοπιστίας χαρτοφυλακίου που λαμβάνουν υπόψιν αυτές τις τάσεις. Ως εκ τούτου, για τη μοντελοποίηση της θνησιμότητας προτείνουμε τη χρήση του υποδείγματος παλινδρόμησης αξιοπιστίας, αντί του υποδείγματος Bülmann που αξιοποιήθηκε από τους Schinzinger et al. (2016) και Tsai and Lin (2017a, 2017b).

Η εισαγωγή του υποδείγματος παλινδρόμησης στην αξιοπιστία χαρτοφυλακίου έγινε από τον Hachemeister (1975), ο οποίος πρότεινε ένα υπόδειγμα παλινδρόμησης για την εκτίμηση των αποζημιώσεων από σωματικές βλάβες σε ατυχήματα αυτοκινήτων για διάφορες πολιτείες των Ηνωμένων Πολιτειών. Για περισσότερες λεπτομέρειες σχετικά με το υπόδειγμα του Hachemeister, αναφερόμαστε στα βιβλία των Goovaerts et al. (1990) και Bühlmann and Gisler (2005).

Η εργασία μας αξιοποιεί τα πλεονεκτήματα της χρήσης μεθόδων αξιοπιστίας χαρτοφυλακίου στο υπόδειγμα παλινδρόμησης με στόχο τη βελτίωση των προβλέψεων θνησιμότητας, ειδικά για πληθυσμούς σχετικά περιορισμένου μεγέθους ή/και με περιορισμένα ιστορικά δεδομένα. Η δομή της εργασίας είναι η ακόλουθη. Η Ενότητα 2 παρέχει μια επισκόπηση των υποδειγμάτων θνησιμότητας Lee-Carter και CBD, μαζί με τις πιο ευρέως χρησιμοποιούμενες παραλλαγές εκτίμησης των παραμέτρων τους.

Η Ενότητα 3 προτείνει τη μέθοδο αξιοπιστίας χαρτοφυλακίου παλινδρόμησης για την εκτίμηση των μελλοντικών ποσοστών θνησιμότητας. Η εφαρμογή του υποδείγματος στα ελληνικά δεδομένα θνησιμότητας για άνδρες και γυναίκες παρουσιάζεται στην Ενότητα 4, όπου οι προβλέψεις μέσω αξιοπιστίας και μέσω των μεθόδων Lee-Carter, CBD αξιολογούνται συγκριτικά με τη χρήση διαφόρων μέτρων ακρίβειας σφάλματος. Τέλος, τα συμπεράσματα της εργασίας συγκεντρώνονται στην Ενότητα 5.

2. ΜΕΘΟΔΟΙ ΠΡΟΒΛΕΨΗΣ ΘΝΗΣΙΜΟΤΗΤΑΣ

Στην ενότητα αυτή, περιγράφονται συνοπτικά τα υποδείγματα Lee-Carter και CBD με τις πιο ευρέως χρησιμοποιούμενες μεθόδους εκτίμησης για τις παραμέτρους τους. Με $D(t,x)$ συμβολίζουμε τον παρατηρούμενο αριθμό θανάτων στην ηλικία x και το έτος t και με $E(t,x)$ τον μέσο πληθυσμό ηλικίας x κατά τη διάρκεια του έτους t (έκθεση σε κίνδυνο). Στη συνέχεια, τα ποσοστά θνησιμότητας για μια συγκεκριμένη ηλικία $m(t,x)$ λαμβάνονται από τον λόγο $D(t,x)/E(t,x)$, ενώ οι πιθανότητες θανάτου μπορούν να εξαχθούν από την ταυτότητα $q(t,x) = 1 - \ exp[-m(t,x)]$, η οποία υποδηλώνει την υπόθεση σταθερής έντασης θνησιμότητας για κάθε έτος πλήρους ηλικίας και για κάθε ημερολογιακό έτος.

2.1 Το Υπόδειγμα Lee-Carter

Το υπόδειγμα Lee-Carter (LC) συνδέει τον φυσικό λογάριθμο του ειδικού κατά ηλικία δείκτη θνησιμότητας $Y_{t,x} = \log m(t,x)$ με τις ακόλουθες παραμέτρους:

$$Y_{t,x} = \alpha_x^{(1)} + \alpha_x^{(2)}\kappa_t + \epsilon_{t,x} \ , \ \text{για } x = x_0, \ldots, x_{k-1} \ , \ t = t_0, \ldots, t_{n-1} \ , \qquad (1)$$

όπου $\alpha_x^{(1)}$ και $\alpha_x^{(2)}$ είναι οι παράμετροι ηλικίας και κ_t η παράμετρος περιόδου που εκφράζει το γενικό επίπεδο θνησιμότητας κατά το έτος t. Τα σφάλματα $\epsilon_{t,x}$ «υποθέτουμε» ότι ακολουθούν την κανονική κατανομή (το εάν την ακολουθούν ή όχι εξαρτάται από τα εμπειρικά δεδομένα), με μηδενική μέση τιμή και σταθερή διασπορά και αντανακλούν ειδικά αποτελέσματα ηλικίας και περιόδου που δε συλλαμβάνονται από το υπόδειγμα. Έτσι, οι παράμετροι του υποδείγματος LC εκτιμώνται με δύο μεθόδους, είτε με τη μέθοδο ελαχίστων τετραγώνων, είτε με τη μέθοδο μέγιστης πιθανοφάνειας.

Υποθέτοντας ότι τα σφάλματα είναι ανεξάρτητα και ομοσκεδαστικά με μηδενική μέση τιμή, οι Lee and Carter (1992) θεώρησαν δύο περιορισμούς $\sum_{x=x_0}^{x_{k-1}} \alpha_x^{(2)} = 1$, $\sum_{t=t_0}^{t_{n-1}} \kappa_t = 0$, ώστε να προκύψει μία μοναδική λύση, ελαχιστοποιώντας τη συνάρτηση τετραγωνικών σφαλμάτων:

$$f(\alpha_x^{(1)}, \alpha_x^{(2)}, \kappa_t) = \sum_{t,x} (Y_{t,x} - \alpha_x^{(2)}\kappa_t)^2 = \sum_{t,x} [\log m(t,x) - \alpha_x^{(1)} - \alpha_x^{(2)}\kappa_t]^2 \ ,$$

η οποία μας δίνει του παρακάτω εκτιμητές ελαχίστων τετραγώνων (ΕΤ):

$$\alpha_x^{(1)} = \frac{1}{t_{n-1} - t_0 + 1} \sum_{t=t_0}^{t_{n-1}} \log m(t,x) \ , \tag{2}$$

$$\kappa_t = \sum_{x=x_0}^{x_{k-1}} [\log m(t,x) - \alpha_x^{(1)}] \ , \tag{3}$$

$$\alpha_x^{(2)} = \frac{\sum_{t=t_0}^{t_{n-1}} [\log m(t,x) - \alpha_x^{(1)}]\kappa_t}{\sum_{t=t_0}^{t_{n-1}} \kappa_t^2} \ . \tag{4}$$

Για να επιτραπεί η περίπτωση ετεροσκεδαστικότητας των σφαλμάτων, οι Brouhns et al. (2002) υπέθεσαν ότι ο αριθμός των θανάτων $D(t,x)$ ακολουθεί την κατανομή Poisson με μέσο $m(t,x) \cdot E(t,x)$. Σύμφωνα με τη προσέγγιση αυτή, οι παράμετροι ηλικίας και περιόδου εκτιμώνται με τη μέθοδο μέγιστης πιθανοφάνειας (ΜΠ) μεγιστοποιώντας τη συνάρτηση:

$$\mathcal{L}\Big(D(t,x)\Big) = \sum_{x,t} D(t,x)[\alpha_x^{(1)} + \alpha_x^{(2)}\kappa_t + \log E(t,x)] - E(t,x)\exp(\alpha_x^{(1)} + \alpha_x^{(2)}\kappa_t) - \log(D(t,x)!) \ . \tag{5}$$

Μετά την επιλογή της μεθόδου εκτίμησης των παραμέτρων, ακολουθεί η προβολή των εκτιμήσεων για την παράμετρο περιόδου με τη χρήση ενός κατάλληλου υποδείγματος χρονολογικών σειρών. Οι Lee and Carter (1992) πρότειναν το υπόδειγμα του τυχαίου περιπάτου με μετατόπιση για να εξάγουν προβλέψεις που δίνονται από τη σχέση $\kappa_{t_{n-1}+h} = \kappa_{t_{n-1}} + \theta h,$ για αριθμό μελλοντικών ετών $h = 1, 2, 3, \ldots$. Έπειτα, οι προβλέψεις των κ_t χρησιμοποιούνται μαζί με τις εκτιμήσεις των παραμέτρων $\alpha_x^{(1)}$ και $\alpha_x^{(2)}$ για την εξαγωγή των μελλοντικών δεικτών θνησιμότητας, σύμφωνα με τη παρακάτω σχέση:

$$\widehat{Y}_{t_{n-1}+h,x} = \widehat{\alpha}_x^{(1)} + \widehat{\alpha}_x^{(2)}\widehat{\kappa}_{t_{n-1}+h} = \widehat{Y}_{t_{n-1},x} + (\widehat{\alpha}_x^{(2)}\widehat{\theta})h, \text{ για } h = 1, 2, 3, \ldots \tag{6}$$

2.2 Το Υπόδειγμα CBD

Το υπόδειγμα CBD συνδέει τον logit μετασχηματισμό της πιθανότητας θανάτου $Y_{t,x} = \operatorname{logit} q(t,x)$ με τις ακόλουθες παραμέτρους:

$$Y_{t,x} = \operatorname{logit} q(t,x) = \kappa_t^{(1)} + (x - \bar{x})\kappa_t^{(2)} + \epsilon_{t,x} \ , \tag{7}$$

όπου $\kappa_t^{(1)}$ και $\kappa_t^{(2)}$ είναι οι παράμετροι περιόδου, με $\bar{x}$ συμβολίζεται η μέση ηλικία των δεδομένων προσαρμογής, ενώ τα σφάλματα $\epsilon_{t,x}$ υποθέτουμε ότι ακολουθούν την κανονική κατανομή, με μέση τιμή μηδέν και σταθερή διασπορά. Οι εκτιμήσεις ελαχίστων τετραγώνων για τις παραμέτρους του υποδείγματος CBD δίνονται μέσω της συνάρτησης

$$f(\kappa_t^{(1)}, \kappa_t^{(2)}) = \sum_{t,x} [\text{logit}\, q(t,x) - \kappa_t^{(1)} - (x - \bar{x})\kappa_t^{(2)}]^2 \ ,$$

από την ελαχιστοποίηση της οποίας προκύπτουν οι εκτιμητές ΕΤ:

$$\kappa_t^{(1)} = \frac{1}{m} \sum_{x=x_0}^{x_{k-1}} \text{logit}\, q(t,x) \ , \tag{8}$$

$$\kappa_t^{(2)} = \frac{\sum_{x=x_0}^{x_{k-1}} (\text{logit}\, q(t,x) - \kappa_t^{(1)})(x - \bar{x})}{\sum_{x=x_0}^{x_{k-1}} (x - \bar{x})^2} \ . \tag{9}$$

Αξιοποιώντας τη σχέση $m(t,x) = -\log[1-q(t,x)]$ και υποθέτοντας ότι ο αριθμός των θανάτων $D(t,x)$ ακολουθεί την κατανομή Poisson με μέση τιμή $m(t,x) \cdot E(t,x)$, οι Cairns et al. (2009) εκτίμησαν τις παραμέτρους του υποδείγματος CBD μεγιστοποιώντας την ακόλουθη συνάρτηση πιθανοφάνειας:

$$\mathcal{L}\Big(D(t,x)\Big) = \sum_{x,t} D(t,x)[\log E(t,x) - \log(1-q(t,x))] + E(t,x)\log(1-q(t,x)) - \log(D(t,x)!) \ , \tag{10}$$

όπου

$$q(t,x) = \frac{\exp(\kappa_t^{(1)} + (x - \bar{x})\kappa_t^{(2)})}{1 + \exp(\kappa_t^{(1)} + (x - \bar{x})\kappa_t^{(2)})} \ .$$

Υποθέτοντας ότι οι εκτιμώμενες τιμές για τις παραμέτρους περιόδου είναι ανεξάρτητες μεταξύ τους, η προβολή τους πραγματοποιείται με τη χρήση ενός αντίστοιχου τυχαίου περιπάτου με μετατόπιση $\widehat{\theta}_i,\ i = 1, 2$ για την κάθε μία. Κατά συνέπεια, οι εκτιμήσεις των δεικτών θνησιμότητας για τα έτη $h = 1, 2, 3, \ldots$ προκύπτουν από τη σχέση:

$$\widehat{Y}_{t_{n-1}+h,x} = (\widehat{\kappa}_{t_{n-1}}^{(1)} + \widehat{\theta}_1 h) + (x - \bar{x})(\widehat{\kappa}_{t_{n-1}}^{(2)} + \widehat{\theta}_2 h) = \widehat{Y}_{t_{n-1},x} + [\widehat{\theta}_1 + (x - \bar{x})\widehat{\theta}_2]h \ . \tag{11}$$

Τέλος, μπορούμε εύκολα να παρατηρήσουμε ότι οι σχέσεις (6) και (11) είναι γραμμικές συναρτήσεις του ορίζοντα πρόβλεψης h, όπου οι τεταγμένες δίνονται από τις προσαρμοσμένες τιμές των δεικτών θνησιμότητας, ενώ η κλίση τους ισούται με το γινόμενο των παραμέτρων ηλικίας με τις σταθερές μετατόπισης.

3. ΠΡΟΒΛΕΨΗ ΘΝΗΣΙΜΟΤΗΤΑΣ ΜΕ ΤΗ ΧΡΗΣΗ ΤΟΥ ΥΠΟΔΕΙΓΜΑΤΟΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ ΣΤΗΝ ΑΞΙΟΠΙΣΤΙΑ

Τα εμπειρικά δεδομένα θνησιμότητας διαφόρων χωρών δείχνουν ότι τα ποσοστά θνησιμότητας σε κάθε ηλικία $x = x_0, \ldots, x_{k-1}$ μειώνονται γραμμικά με το χρόνο, λόγω εξωγενών παραγόντων (ιατρικοί, βιολογικοί, περιβαλλοντικοί ή και άλλοι μη ταυτοποιημένοι παράγοντες). Ειδικά στις μεγαλύτερες ηλικίες παρατηρούμε ότι τα ποσοστά θνησιμότητας έχουν βελτιωθεί σημαντικά τα τελευταία χρόνια. Ωστόσο,

αξίζει να αναφέρουμε ότι η θνησιμότητα ενδέχεται να μη παρουσιάζει βελτίωση σε όλο το εύρος παρατήρησης της. Πιο συγκεκριμένα, πρόσφατες μελέτες (Laliotis et al., 2016) παρουσιάζουν ενδείξεις για αύξηση της θνησιμότητας στην Ελλάδα από την έναρξη της οικονομικής κρίσης και έπειτα (μετά το έτος 2008). Συνεπώς, μας ενδιαφέρει μια δομή υποδείγματος που θα μπορεί να συλλάβει τις υποκείμενες τάσεις βελτίωσης και να περιγράφει την εξέλιξη της θνησιμότητας για κάθε ηλικία σε συνάρτηση με το χρόνο.

Υποθέτουμε ότι η εξαρτημένη μεταβλητή $Y_{t,x}$ αφορά σε έναν κατάλληλο μετασχηματισμό (log ή logit) ενός μέτρου θνησιμότητας $\left[m(t,x) \text{ ή } q(t,x)\right]$ για την ηλικία $x = x_0, \ldots, x_{k-1}$ (k ηλικίες) κατά το έτος $t = t_0, \ldots, t_{n-1}$ (n έτη). Επιπλέον θεωρούμε ότι A_x είναι μια τυχαία παράμετρος κινδύνου ηλικίας, $\boldsymbol{Y}_x = (Y_{t_0,x}, Y_{t_1,x}, \ldots, Y_{t_{n-1},x})^{'}$ είναι το αντίστοιχο διάνυσμα θνησιμότητας για την κάθε ηλικία και $\boldsymbol{Z}_x$ ο πίνακας σχεδιασμού (design matrix) των ανεξάρτητων μεταβλητών. Σημειώνεται ότι ο πίνακας σχεδιασμού μπορεί εν γένει να αποτελείται από διάφορες επεξηγηματικές μεταβλητές που αντανακλούν το προφίλ θνησιμότητας κάθε ηλικίας. Επί παραδείγματι, σε μία ιατρική μελέτη, η θνησιμότητα κάθε ηλικίας μπορεί να εξαρτάται από διάφορους παράγοντες, όπως τα γονιδιακά χαρακτηριστικά, τον τρόπο ζωής, τη διατροφή, την τοξικότητα του περιβάλλοντος, μια πιθανή μολυσματική αιτία (μικρόβια, παράσιτα, μύκητες) ή και άλλους οικονομικο-δημογραφικούς παράγοντες. Έτσι, το πλαίσιο της μελέτης μας επικεντρώνεται σε ένα ζεύγος $(A_x, \boldsymbol{Y}_x)$, το οποίο θα περιγράφει την εξέλιξη της θνησιμότητας για κάθε ηλικία x, σύμφωνα με τις παρακάτω υποθέσεις:

(i) Οι τυχαίες μεταβλητές (παράμετροι κινδύνου) $A_{x_0},\ A_{x_1}, \ldots,\ A_{x_{k-1}}$ είναι ανεξάρτητες και ισόνομες.

(ii) Τα ζεύγη $(A_{x_0}, \boldsymbol{Y}_{x_0})$, $(A_{x_1}, \boldsymbol{Y}_{x_1})$, $\ldots$, $(A_{x_{k-1}}, \boldsymbol{Y}_{x_{k-1}})$ είναι ανεξάρτητα μεταξύ τους.

(iii) Για κάθε $x = x_0, \ldots, x_{k-1}$ και δοθέντος A_x, τα διανύσματα $\boldsymbol{Y}_x$ είναι υπό δέσμευση ανεξάρτητα.

(iv) $E(\boldsymbol{Y}_x) = \boldsymbol{Z}_x \boldsymbol{\beta}(A_x)$, όπου $\boldsymbol{\beta}(A_x)$ είναι το διάνυσμα συντελεστών παλινδρόμησης διάστασης p και $\boldsymbol{Z}_x$ ο αντίστοιχος $n \times p$ πίνακας σχεδιασμού.

(v) $Cov(\boldsymbol{Y}_x | A_x) = \sigma^2(A_x) \boldsymbol{I}_n$, όπου $\boldsymbol{I}_n$ είναι ο ταυτοτικός πίνακας διάστασης $n \times n$.

Οι παράμετροι δόμησης του υποδείγματος ορίζονται ως:

$$\boldsymbol{b} = E[\boldsymbol{\beta}(A_x)],\ \boldsymbol{U} = Cov[\boldsymbol{\beta}(A_x)] \text{ και } s^2 = E[\sigma^2(A_x)]\ . \tag{12}$$

Ο εκτιμητής ελαχίστων τετραγώνων του $\boldsymbol{\beta}(A_x)$ δίνεται από τη σχέση:

$$\widehat{\boldsymbol{\beta}}_x = (\boldsymbol{Z}_x^{'} \boldsymbol{Z}_x)^{-1} \boldsymbol{Z}_x^{'} \boldsymbol{Y}_x\ ,$$

ενώ ο πίνακας διασποράς συνδιασποράς $Cov(\widehat{\boldsymbol{\beta}}_x | A_x) = \sigma^2(A_x)(\boldsymbol{Z}_x^{'} \boldsymbol{Z}_x)^{-1}$, με μέση τιμή $E[Cov(\widehat{\boldsymbol{\beta}}_x | A_x)] = E[\sigma^2(A_x)(\boldsymbol{Z}_x^{'} \boldsymbol{Z}_x)^{-1}] = s^2 (\boldsymbol{Z}_x^{'} \boldsymbol{Z}_x)^{-1}$.

3.1 Εκτίμηση την Παραμέτρων Δόμησης του Υποδείγματος Αξιοπιστίας

Σύμφωνα με τις παραπάνω υποθέσεις, ο βέλτιστος γραμμικός εκτιμητής αξιοπιστίας $\widehat{\boldsymbol{B}}_x = (\widehat{B}_{1x}, \cdots, \widehat{B}_{px})^{'}$ δίνεται από τη σχέση

$$\widehat{\boldsymbol{B}}_x = \widehat{\boldsymbol{K}}_x \, \widehat{\boldsymbol{\beta}}_x + (\boldsymbol{I} - \widehat{\boldsymbol{K}}_x) \, \widehat{\boldsymbol{b}} \, , \tag{13}$$

όπου $\widehat{\boldsymbol{K}}_x = \widehat{\boldsymbol{U}}[s^2(\boldsymbol{Z}_x^{'}\boldsymbol{Z}_x)^{-1} + \widehat{\boldsymbol{U}}]^{-1}$ είναι οι συντελεστές αξιοπιστίας. Για την εκτίμηση των παραμέτρων δόμησης s^2, $\boldsymbol{U}_x$ και $\boldsymbol{b}$ προτείνουμε τους ακόλουθους αμερόληπτους εκτιμητές ελάχιστης διασποράς (Hachemeister, 1975; De Vylder, 1978):

$$\widehat{s}^2 = \frac{1}{(x_{k-1} - x_0 + 1)(n-p)} \sum_{x=x_0}^{x_{k-1}} (\boldsymbol{Y}_x - \boldsymbol{Z}_x \widehat{\boldsymbol{\beta}}_x)^{'} (\boldsymbol{Y}_x - \boldsymbol{Z}_x \widehat{\boldsymbol{\beta}}_x) \, , \tag{14}$$

$$\widehat{\boldsymbol{U}} = \frac{1}{(x_{k-1} - x_0)} \sum_{x=x_0}^{x_{k-1}} \widehat{\boldsymbol{K}}_x (\widehat{\boldsymbol{\beta}}_x - \widehat{\boldsymbol{b}})(\widehat{\boldsymbol{\beta}}_x - \widehat{\boldsymbol{b}})^{'} \, , \tag{15}$$

$$\widehat{\boldsymbol{b}} = (\sum_{x=x_0}^{x_{k-1}} \widehat{\boldsymbol{K}}_x)^{-1} \sum_{x=x_0}^{x_{k-1}} \widehat{\boldsymbol{K}}_x \, \widehat{\boldsymbol{\beta}}_x \, . \tag{16}$$

Παρατηρούμε ότι στη σχέση εκτίμησης του $\widehat{\boldsymbol{U}}$, η προς εκτίμηση παράμετρος εμφανίζεται και στο δεξί μέρος της (15) μέσω του $\widehat{\boldsymbol{K}}_x$. Στις περιπτώσεις αυτές η λύση δίνεται με επαναλήψεις θέτοντας αρχικές συνθήκες.

Τέλος, ο εκτιμητής αξιοπιστίας των μελλοντικών δεικτών θνησιμότητας $\widehat{Y}_{t_{n-1}+h,x}$, $h = 1, 2, \ldots,$ (για το έτος t_n και μετά) μπορεί να γραφεί ως

$$\widehat{\boldsymbol{Y}}_{n+h,x} = \boldsymbol{Z}_{n+h,x} \, \widehat{\boldsymbol{B}}_x,$$

όπου με $\boldsymbol{Z}_{n+h,x}$ συμβολίζουμε τον πίνακα σχεδιασμού μελλοντικών χρονικών περιόδων.

3.2 Εκτίμηση Μελλοντικών Δεικτών Θνησιμότητας

Στην υποενότητα αυτή δείχνουμε πως τα ποσοστά θνησιμότητας για την κάθε ηλικία $x = x_0, \ldots, x_{k-1}$ και περίοδο $t = t_0, \ldots, t_{n-1}$ προσαρμόζονται στο υπόδειγμα αξιοπιστίας. Έτσι, οι δείκτες έως και το έτος t_{n-1} μπορούν να γραφούν ως $\widehat{\boldsymbol{Y}}_x = \boldsymbol{Z}_x \widehat{\boldsymbol{\beta}}_x$. Όπως επισημάνθηκε και πριν, ο πίνακας σχεδιασμού $\boldsymbol{Z}_x$ μπορεί να αποτελείται από διάφορες ανεξάρτητες μεταβλητές που αντανακλούν τους παράγοντες κινδύνου για την κάθε ηλικία x, αλλά λόγω έλλειψης ειδικών δεδομένων, υποθέτουμε ότι οι δείκτες $\boldsymbol{Y}_x$ για μια δεδομένη ηλικία x, εξαρτώνται μόνο από το ημερολογιακό έτος κατά το οποίο συμβαίνει ο θάνατος. Ως συνέπεια, θεωρούμε τις ίδιες ανεξάρτητες μεταβλητές για όλες τις ηλικίες, που μας οδηγεί στη χρήση του ίδιου πίνακα σχεδιασμού για κάθε μία από τις ηλικίες $x = x_0, \ldots, x_{k-1}$.

Έτσι, προσαρμόζοντας στο υπόδειγμα μας τα δεδομένα του διανύσματος $\widehat{\boldsymbol{Y}_x} = (Y_{t_0,x}, Y_{t_1,x}, \ldots, Y_{t_{n-1},x})'$, οι δείκτες θνησιμότητας για το επόμενο έτος μπορούν να εκτιμηθούν από τη σχέση:

$$\widehat{Y}_{t_{n-1}+1,x} = \widehat{B}_{1x} + \widehat{B}_{2x}\ (t_{n-1} - t_0 + 2) = \widehat{B}_{1x} + \widehat{B}_{2x}\ (n+1)\ , \qquad (17)$$

η οποία για κάθε επιλογή ηλικίας $x = x_0, \ldots, x_{k-1}$ έχει έχουν ως εξαρτημένη μεταβλητή τους αντίστοιχους λογαρίθμους των δεικτών θνησιμότητας και ως ανεξάρτητη μεταβλητή την χρονική περίοδο, προσομοιάζοντας σε ένα υπόδειγμα χρονολογικών σειρών γραμμικής τάσης. Ωστόσο στην εργασία αυτή επικεντρωνόμαστε στην εφαρμογή του γραμμικού υποδείγματος αξιοπιστίας για την εκτίμηση του βέλτιστου εκτιμητή $\widehat{\boldsymbol{B}}_x = (\widehat{B}_{1x}, \widehat{B}_{2x})'$.

Έτσι, για την προβολή των δεικτών θνησιμότητας $\widehat{Y}_{t_{n-1}+h,x}$, για $h = 2, 3, \ldots, H$ προτείνουμε μια στατική μέθοδο (Μέθοδος 1), όπου για τις προβλέψεις χρησιμοποιούνται τα αρχικά δεδομένα προσαρμογής της εξαρτημένης μεταβλητής και δύο δυναμικές μεθόδους (Μέθοδοι 2 και 3), όπου στην εξαρτημένη μεταβλητή ενσωματώνονται και οι προηγούμενες προβλέψεις για την εξαγωγή νέων προβλέψεων. Σημειώνεται ότι η εκτίμηση για τον πρώτο χρόνο $\widehat{Y}_{t_{n-1}+1,x} = \widehat{Y}_{t_n,x}$, για $h = 1$ που δίνεται από τη σχέση (17) είναι ίδια και για τις τρεις μεθόδους.

3.2.1 Μέθοδος 1: Διατήρηση Αρχικού Διαστήματος Προσαρμογής (ΑΔ)

Για κάθε ηλικία $x = x_0, \ldots, x_{k-1}$, η εκτίμηση των μελλοντικών δεικτών θνησιμότητας πραγματοποιείται γενικεύοντας τη σχέση (17) για $h = 2, 3, \ldots, H$, έτσι ώστε $\widehat{Y}_{t_{n-1}+h,x} = \widehat{B}_{1x} + \widehat{B}_{2x}\ (n+h)$, όπου οι εκτιμητές αξιοπιστίας των συντελεστών παλινδρόμησης προκύπτουν από τη σχέση (13). Σύμφωνα με αυτή τη μέθοδο, οι μελλοντικές εκτιμήσεις βασίζονται στα δεδομένα θνησιμότητας του αρχικού διαστήματος προσαρμογής.

3.2.2 Μέθοδος 2: Μετακίνηση Διαστήματος Προσαρμογής (ΜΔ)

Σύμφωνα με τη μέθοδο αυτή, οι προβλέψεις για το πρώτο έτος $\widehat{Y}_{t_{n-1}+1,x}$ ενσωματώνονται στα δεδομένα προσαρμογής, ενώ ταυτόχρονα αφαιρούνται οι δείκτες $Y_{t_0,x}$, με αποτέλεσμα το εύρος του διαστήματος προσαρμογής να παραμένει σταθερό, έτσι ώστε $\widehat{\boldsymbol{Y}_x} = (Y_{t_1,x}, \ldots, Y_{t_{n-1},x}, \widehat{Y}_{t_{n-1}+1,x})'$. Έτσι, μετά την εκτίμηση των παραμέτρων δόμησης, λαμβάνουμε την πρόβλεψη για το δεύτερο έτος $\widehat{Y}_{t_{n-1}+2,x}$. Ακολουθώντας την ίδια διαδικασία λαμβάνονται και οι $\widehat{Y}_{t_{n-1}+3,x}, \widehat{Y}_{t_{n-1}+4,x}, \ldots, \widehat{Y}_{t_{n-1}+H,x}$. Με τη μέθοδο αυτή οι μελλοντικές εκτιμήσεις βασίζονται σε πιο πρόσφατα δεδομένα.

3.2.3 Μέθοδος 3: Επέκταση Διαστήματος Προσαρμογής (ΕΔ)

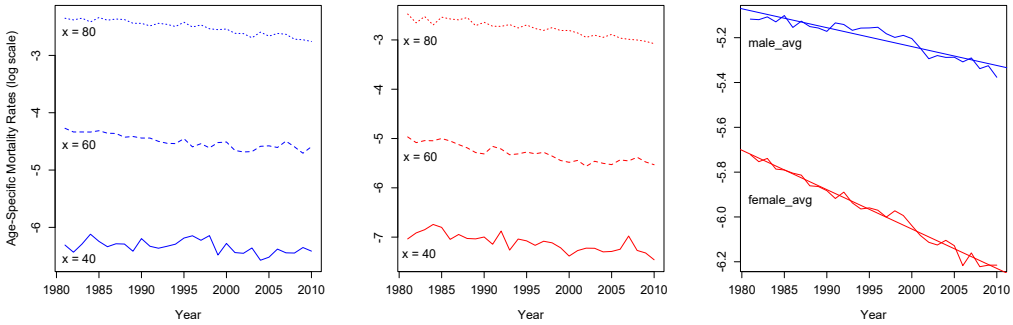
Τέλος στη μέθοδο αυτή, οι προβλέψεις για το πρώτο έτος $\widehat{Y}_{t_{n-1}+1,x}$ ενσωματώνονται στα δεδομένα προσαρμογής χωρίς όμως την αφαίρεση των $Y_{t_0,x}$, έτσι ώστε η εξαρτημένη μεταβλητή να διευρύνεται κατά μια παρατήρηση κάθε φορά ώστε $\widehat{\boldsymbol{Y}_x} = (Y_{t_0,x}, Y_{t_1,x}, \ldots, Y_{t_{n-1},x}, \widehat{Y}_{t_{n-1}+1,x})'$. Συνεπώς σε κάθε βήμα εκτίμησης, το υπόδειγμα μας προσαρμόζεται σε μια συνεχώς διευρυνόμενη (κατά μια παρατήρηση) εξαρτημένη μεταβλητή $\widehat{Y}_{t_{n-1}+2,x}$, $\widehat{Y}_{t_{n-1}+3,x}$, $\widehat{Y}_{t_{n-1}+4,x}, \ldots, \widehat{Y}_{t_{n-1}+H,x}$. Με τη μέθοδο αυτή οι μελλοντικές εκτιμήσεις βασίζονται τόσο στα αρχικά δεδομένα, όσο και σε πιο πρόσφατες εκτιμήσεις.

4. ΑΡΙΘΜΗΤΙΚΗ ΕΦΑΡΜΟΓΗ

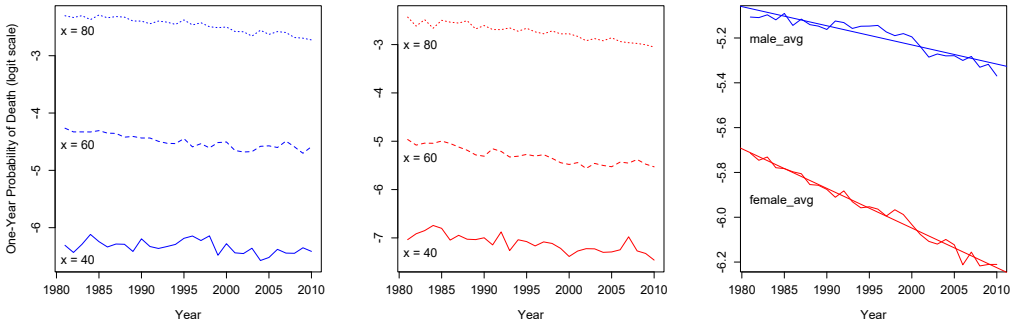
Στην ενότητα αυτή, προσαρμόζουμε τα δεδομένα θνησιμότητας του ελληνικού πληθυσμού στα υποδείγματα LC, CBD και στο γραμμικό υπόδειγμα αξιοπιστίας που προτείνουμε, ενώ οι προβλέψεις που λαμβάνουμε αξιολογούνται με τη χρήση δύο εναλλακτικών μέτρων ακρίβειας σφάλματος.

Η Ελλάδα είναι μια χώρα με περιορισμένο αριθμό επίσημων ιστορικών δεδομένων (1981-2013) και ενός σχετικά μικρού πληθυσμού (11,142,161), `http://www.worldometers.info/population/countries-in-the-eu-by-population/`. Τα ελληνικά δεδομένα είναι διαθέσιμα στη βάση δεδομένων Human Mortality Database (2017), ανά έτος, ηλικία και φύλο. Η εξέλιξη της θνησιμότητας στην Ελλάδα για την περίοδο 1981-2010 απεικονίζεται και για τα δύο φύλα μέσω του λογαρίθμου των ποσοστών θνησιμότητας $\log m(t,x)$ στο Σχήμα 1, καθώς και μέσω του logit μετασχηματισμού των πιθανοτήτων θανάτου $\operatorname{logit} q(t,x)$ στο Σχήμα 2. Εύκολα φαίνεται ότι και τα δύο μέτρα παρουσιάζουν μια γραμμικότητα στο χρόνο για διαφορετικές ηλικίες $x = 40,\ 60,\ 80$ ανδρών (αριστερά panels των Σχημάτων 1 και 2) και γυναικών (μεσαία panels των Σχημάτων 1 και 2). Επιπλέον παρατηρούμε ότι και για τα δύο φύλα, η μέση θνησιμότητα παρουσιάζει μια πτωτική τάση με την πάροδο του χρόνου (δεξιά panels των Σχημάτων 1 και 2). Η τάση αυτή, σε συνδυασμό με τον περιορισμένο αριθμό διαθέσιμων ιστορικών δεδομένων μπορεί να υποστηρίξει την εφαρμογή τεχνικών αξιοπιστίας.

Για την αριθμητική εφαρμογή που ακολουθεί, χρησιμοποιήθηκαν τα ποσοστά θνησιμότητας της Ελλάδας από το 1981 έως το 2010, για άνδρες και γυναίκες ηλικίας 15 έως 84 ετών. Η επιλογή αυτού του εύρους ηλικιών ακολουθείται και σε αντίστοιχες μελέτες από τους Tsai and Lin (2017a, 2017b), καθώς αντιστοιχεί στην ηλικία ενός νεαρού ενήλικα μέχρι το μέσο επίπεδο του προσδόκιμου ζωής στις ανεπτυγμένες χώρες. Τα υποδείγματα αξιοπιστίας, LC και CBD προσαρμόστηκαν στα ελληνικά δεδομένα χρησιμοποιώντας το στατιστικό λογισμικό R (R Core Team, 2017). Σημειώνουμε ότι στη σύγκριση αυτή χρησιμοποιήθηκαν μόνο υποδείγματα ηλικίας - περιόδου (age - period models). Για μια μελέτη σύγκρισης υποδειγμάτων στα ελληνικά δεδομένα που επιτρέπει και αποτελέσματα από παραμέτρους γενεάς (cohort effects), παραπέμπουμε στην εργασία των Bozikas and Pitselis (2018).



Σχήμα 1: Παρατηρούμενες τιμές $\log m(t,x)$ της περιόδου 1981-2010 στην Ελλάδα, για άνδρες (αριστερά panels) και γυναίκες (μεσαία panels) στην ηλικία των 40, 60 και 80. Οι μέσες τιμές των $\log m(t,x)$ για ηλικίες 15-84 ανδρών και γυναικών απεικονίζονται στα δεξιά panels, όπου η ευθεία γραμμή δείχνει την τάση της μείωσης στη θνησιμότητα για το κάθε φύλο.



Σχήμα 2: Παρατηρούμενες τιμές $\text{logit}\, q(t,x)$ της περιόδου 1981-2010 στην Ελλάδα, για άνδρες (αριστερά panels) και γυναίκες (μεσαία panels) στην ηλικία των 40, 60 και 80. Οι μέσες τιμές των $\text{logit}\, q(t,x)$ για ηλικίες 15-84 ανδρών και γυναικών απεικονίζονται στα δεξιά panels, όπου η ευθεία γραμμή δείχνει την τάση της μείωσης στη θνησιμότητα για το κάθε φύλο.

4.1 Αποτελέσματα Πρόβλεψης

Για την εφαρμογή υποθέτουμε ότι οι δείκτες $m(t,x)$, για κάθε δεδομένη ηλικία x, εξαρτώνται μόνο από το ημερολογιακό έτος του θανάτου. Κατά συνέπεια, ο πίνακας σχεδιασμού είναι ο $\boldsymbol{Z}_x{}^{(p=2)} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ 1 & 2 & \dots & n \end{pmatrix}'$ για $[x_0, x_{k-1}] = [15, 84]$ και $[t_0, t_{n-1}] = [1981, 2000]$. Για να διασφαλιστεί η ανθεκτικότητα του υποδείγματος ως προς τις αλλαγές στο εύρος προσαρμογής των δεδομένων, η ηλικία και η περίοδος διαχωρίστηκαν σε διάφορα υποσύνολα, τα οποία παρουσιάζονται στον Πίνακα 1.

Επίσης, προκειμένου να διατηρηθεί η γραμμικότητα σε κάθε αντίστοιχη περίοδο προσαρμογής, χρησιμοποιήθηκε ο λογαριθμικός μετασχηματισμός $Y_{t,x} = \log m(t,x)$

Πίνακας 1: Επιλεγμένες περίοδοι προσαρμογής και πρόβλεψης.

Ηλικίες	Περίοδος Προσαρμογής	Περίοδος Πρόβλεψης
$[x_0, x_{k-1}]$	$[t_0, t_{n-1}]$	$[t_{n-1}+1, t_{n-1}+H]$
$[15, 84]$	$[1981, 2000]$	$[2001, 2010]$
ή	$[1986, 2000]$	$[2001, 2010]$
$[55, 84]$	$[1991, 2000]$	$[2001, 2010]$

για τους δείκτες θνησιμότητας και ο logit μετασχηματισμός $Y_{t,x} = \text{logit}\, q(t,x) = \log \dfrac{q(t,x)}{1-q(t,x)}$ για τις πιθανότητες θανάτου. Σημειώνεται ότι εάν και η εξαρτημένη μεταβλητή λαμβάνεται είτε σε λογαρίθμους, είτε σε *logit* μετασχηματισμό, είναι πιθανό τα σφάλματα από την προσαρμογή των μοντέλων αξιοπιστίας (κατάλοιπα) να μην ακολουθούν σε ορισμένες περιπτώσεις την κανονική κατανομή. Αυτό δεν σημαίνει ότι το υπόδειγμα δεν έχει αυξημένη προβλεπτική ικανότητα, ωστόσο σημαίνει ότι ίσως δεν μπορεί να αξιολογηθεί η στατιστική σημαντικότητα των παραμέτρων. Στη συνέχεια, οι προβλέψεις αξιολογήθηκαν χρησιμοποιώντας το μέσο απόλυτο σφάλμα πρόβλεψης (MAFE), καθώς και τη ρίζα του μέσου τετραγωνικού σφάλματος πρόβλεψης (RMSFE), όπου μικρότερες τιμές υποδεικνύουν μία καλύτερη προβλεπτική απόδοση. Οι μέσες τιμές των MAFE και RMSFE δίνονται από τις σχέσεις:

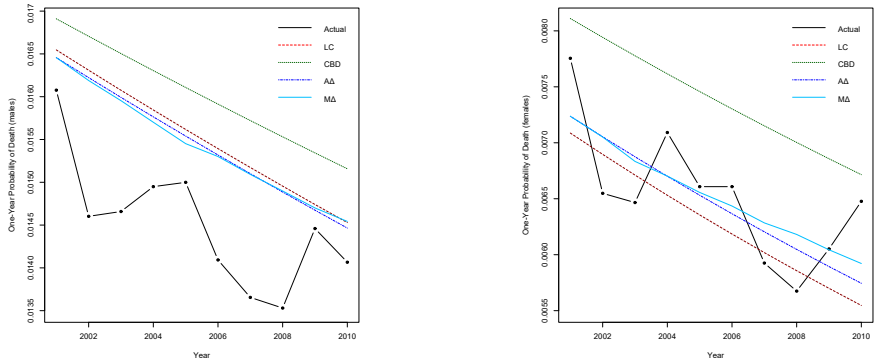
$$MAFE_{avg} = \frac{1}{H \times (x_{k-1} - x_0 + 1)} \sum_{h=1}^{H} \sum_{x=x_0}^{x_{k-1}} |\widehat{m}(t_{n-1}+h, x) - m(t_{n-1}+h, x)| \times 100\ , \qquad (18)$$

και

$$RMSFE_{avg} = \sqrt{\frac{1}{H \times (x_{k-1} - x_0 + 1)} \sum_{h=1}^{H} \sum_{x=x_0}^{x_{k-1}} \left[\widehat{m}(t_{n-1}+h, x) - m(t_{n-1}+h, x)\right]^2} \times 100\ . \qquad (19)$$

Στην περίπτωση που η εξαρτημένη μεταβλητή μας είναι η $Y_{t,x} = \text{logit}\, q(t,x)$, στις παραπάνω σχέσεις οι δείκτες θνησιμότητας $m(t,x)$ αντικαθίστανται από τις πιθανότητες θανάτου $q(t,x)$. Και για τα δύο φύλα, τα ποσοστιαία αποτελέσματα των μέτρων ακρίβειας σφάλματος MAFE και RMSFE παρουσιάζονται στον Πίνακα 2 και αφορούν στη περίοδο πρόβλεψης $[2001, 2010]$ για ηλικίες $[15, 84]$, για τις 2 μεθόδους εκτίμησης παραμέτρων (ΕΤ, ΜΠ) του υποδείγματος LC και τις 3 μεθόδους πρόβλεψης (ΑΔ, ΜΔ, ΕΔ) για το υπόδειγμα αξιοπιστίας με εξαρτημένη μεταβλητή την $Y_{t,x} = \log m(t,x)$. Τα αντίστοιχα αποτελέσματα για ηλικίες $[55, 84]$ παρουσιάζονται στον Πίνακα 3, συμπεριλαμβανομένων και των 2 μεθόδων εκτίμησης παραμέτρων (ΕΤ, ΜΠ) για το υπόδειγμα CBD, έχοντας όμως ως εξαρτημένη μεταβλητή την $Y_{t,x} = \text{logit}\, q(t,x)$. Σημειώνεται ότι το υπόδειγμα CBD συμπεριλαμβάνεται μόνο για συγκρίσεις στις ηλικίες $[55, 84]$, καθώς είναι ειδικά σχεδιασμένο για μεγαλύτερες ηλικίες.

Τα αποτελέσματα σύγκρισης δείχνουν ότι το υπόδειγμα αξιοπιστίας υπερέχει ως προς τα υποδείγματα LC και CBD και για τα δύο μέτρα σφάλματος. Οι τιμές με έντονη γραφή υποδηλώνουν τις χαμηλότερες τιμές σφάλματος κατά τις αντίστοιχες περιόδους προσαρμογής για κάθε φύλο, ενώ οι μέσες τιμές των υποπεριόδων δίνονται στις τελευταίες σειρές των υποπινάκων. Πιο συγκεκριμένα, για τις ηλικίες [15,84], οι μέθοδοι ΑΔ και ΜΔ παράγουν τα μικρότερα κατά μέσο όρο MAFE και RMSFE, ενώ για τις ηλικίες [55,84], η μέθοδος ΜΔ υπερέχει έναντι των άλλων μεθόδων, ως μία ένδειξη ότι οι προβλέψεις για μεγαλύτερες ηλικίες βασίζονται σε πιο πρόσφατες τάσεις θνησιμότητας. Επίσης, παρατηρούμε ότι επιλέγοντας το μικρότερο διάστημα ηλικιών προσαρμογής [55,84], τα σφάλματα γίνονται ολοένα και μεγαλύτερα. Αυτό οφείλεται στο ότι οι ποσότητες $|\widehat{m}(t_{n-1}+h,x)-m(t_{n-1}+h,x)|$ και $\left[\widehat{m}(t_{n-1}+h,x)-m(t_{n-1}+h,x)\right]^2$ στις σχέσεις (18) και (19), αντίστοιχα, μεγαλώνουν καθώς επιλέγουμε μεγαλύτερες ηλικίες x. Συνεπώς, οι τιμές $MAFE_{avg}$ και $RMSFE_{avg}$ για ηλικίες [55, 84] είναι μεγαλύτερες από τις αυτές των ηλικιών [15, 84]. Τέλος, στο Σχήμα 3 απεικονίζονται τα αποτελέσματα των προβλέψεων από την προσαρμογή των υποδειγμάτων LC, CBD, αξιοπιστίας ΑΔ και αξιοπιστίας ΜΔ (με τις μικρότερες τιμές MAFE και RMSFE) για την περίοδο $[2001, 2010]$ στην Ελλάδα. Οι προβλεπόμενες τιμές πιθανοτήτων θανάτου ηλικίας 65 ετών για τους άνδρες αναπαρίστανται στο αριστερό panel, ενώ οι αντίστοιχες τιμές για τις γυναίκες στο δεξί panel.



Σχήμα 3: Οι προβλεπόμενες τιμές πιθανοτήτων θανάτου ηλικίας 65 ετών για άνδρες (αριστερό panel) και γυναίκες (δεξί panel).

5. ΣΥΜΠΕΡΑΣΜΑΤΑ

Η εργασία αυτή χρησιμοποιώντας τεχνικές της θεωρίας αξιοπιστίας χαρτοφυλακίου προτείνει μια μέθοδο μοντελοποίησης των ποσοστών θνησιμότητας ανά ηλικία, ιδανική για χώρες με περιορισμένα δεδομένα ή/και ισχυρές τάσεις. Στην εφαρμογή μας στα ελληνικά δεδομένα, το υπόδειγμα αξιοπιστίας οδήγησε σε καλύτερες προβλέψεις και για τα δύο φύλα σε σύγκριση με τα υποδείγματα που χρησιμοποιούνται για αντίστοιχες μελέτες στη βιβλιογραφία.

Πίνακας 2: Μέσες τιμές MAFE και RMSFE στις ηλικίες $[15, 84]$ για την περίοδο $[2001, 2010]$.

$MAFE_{[15,84]}$		Υπόδειγμα LC		Υπόδειγμα Αξιοπιστίας		
Περίοδος Προσαρμογής	Φύλο	ΕΤ	ΜΠ	ΑΔ	ΜΔ	ΕΔ
$[1981, 2000]$	Άνδρες	0.1513	0.1569	0.1352	0.1256	0.1361
	Γυναίκες	0.0831	0.0861	0.0691	**0.0657**	0.0690
$[1986, 2000]$	Άνδρες	0.1684	0.1514	0.1203	0.1221	0.1206
	Γυναίκες	0.0625	0.0799	**0.0608**	0.0651	0.0613
$[1991, 2000]$	Άνδρες	0.1468	0.1681	0.1288	0.1289	0.1289
	Γυναίκες	0.0763	0.0959	**0.0622**	0.0663	0.0669
Μέσος		0.1147(4)	0.1231(5)	0.0961(2)	0.0956(1)	0.0971(3)

$RMSFE_{[15,84]}$		Υπόδειγμα LC		Υπόδειγμα Αξιοπιστίας		
Περίοδος Προσαρμογής	Φύλο	ΕΤ	ΜΠ	ΑΔ	ΜΔ	ΕΔ
$[1981, 2000]$	Άνδρες	0.3165	0.3220	0.2716	0.2511	0.2745
	Γυναίκες	0.1791	0.1825	0.1376	**0.1365**	0.1374
$[1986, 2000]$	Άνδρες	0.3543	0.3200	0.2362	0.2364	0.2375
	Γυναίκες	0.1307	0.1742	**0.1264**	0.1385	0.1288
$[1991, 2000]$	Άνδρες	0.3180	0.4010	0.2570	0.2551	0.2516
	Γυναίκες	0.1694	0.2415	**0.1302**	0.1438	0.1476
Μέσος		0.2447(4)	0.2735(5)	0.1932(1)	0.1936(2)	0.1962(3)

Πίνακας 3: Μέση τιμή MAFE και RMSFE στις ηλικίες $[55, 84]$ για την περίοδο $[2001, 2010]$.

$MAFE_{[55,84]}$		Υπόδειγμα LC		Υπόδειγμα CBD		Υπόδειγμα Αξιοπιστίας		
Περίοδος Προσαρμογής	Φύλο	ΕΤ	ΜΠ	ΕΤ	ΜΠ	ΑΔ	ΜΔ	ΕΔ
$[1981, 2000]$	Άνδρες	0.3389	0.3322	0.2924	0.3247	0.2871	**0.2673**	0.2870
	Γυναίκες	0.1976	0.1933	0.1694	0.1884	0.1629	**0.1448**	0.1627
$[1986, 2000]$	Άνδρες	0.3108	0.3186	0.2682	0.2988	0.2544	0.2581	**0.2541**
	Γυναίκες	0.1652	0.1769	0.1618	0.1708	0.1289	0.1351	**0.1288**
$[1991, 2000]$	Άνδρες	0.3286	0.3622	0.2790	0.3348	0.2538	**0.2493**	0.2525
	Γυναίκες	0.1805	0.2126	0.1659	0.1868	0.1363	0.1382	**0.1361**
Μέσος		0.2536(6)	0.2660(7)	0.2228(4)	0.2507(5)	0.2039(3)	0.1988(1)	0.2035(2)

$RMSFE_{[55,84]}$		Υπόδειγμα LC		Υπόδειγμα CBD		Υπόδειγμα Αξιοπιστίας		
Περίοδος Προσαρμογής	Φύλο	ΕΤ	ΜΠ	ΕΤ	ΜΠ	ΑΔ	ΜΔ	ΕΔ
$[1981, 2000]$	Άνδρες	0.4976	0.4848	0.3904	0.4467	0.4065	**0.3786**	0.4061
	Γυναίκες	0.2963	0.2842	0.2221	0.2512	0.2304	**0.2010**	0.2299
$[1986, 2000]$	Άνδρες	0.4648	0.4872	0.3551	0.4073	0.3631	0.3653	**0.3618**
	Γυναίκες	0.2481	0.2699	0.2165	0.2244	0.1803	0.1897	**0.1800**
$[1991, 2000]$	Άνδρες	0.5036	0.6129	0.3698	0.4625	0.3660	**0.3501**	0.3616
	Γυναίκες	0.2813	0.3721	0.2202	0.2510	0.1961	**0.1930**	0.1954
Μέσος		0.3820(6)	0.4185(7)	0.2957(4)	0.3405(5)	0.2904(3)	0.2796(1)	0.2891(2)

Στην αρχή παρουσιάστηκαν συνοπτικά τα υποδείγματα θνησιμότητας των Lee-Carter και CBD με τις δύο εναλλακτικές μεθόδους εκτίμησης των παραμέτρων τους. Στη συνέχεια περιγράψαμε μια μεθοδολογία πρόβλεψης των ποσοστών θνησιμότητας ενός πληθυσμού, αξιοποιώντας τις υποθέσεις του υποδείγματος παλινδρόμησης στην αξιοπιστία. Αφού δόθηκαν οι εκτιμήσεις των παραμέτρων του υποδείγματος αξιοπιστίας προτείναμε τρεις μεθόδους προβολής των ποσοστών θνησιμότητας.

Η εφαρμογή του προτεινόμενου υποδείγματος πάνω στα ελληνικά δεδομένα ανδρών και γυναικών αξιολογήθηκε ως προς την αποτελεσματικότητα των προβλέψεων σε σύγκριση με τις μεθόδους LC και CBD, όπου το υπόδειγμα αξιοπιστίας είχε καλύτερη απόδοση κατά μέσο όρο βάσει της αξιολόγησης των σφαλμάτων πρόβλεψης. Πιο συγκεκριμένα, η μέθοδος πρόβλεψης ΜΔ ξεχώρισε στο σύνολο των επιλεγμένων περιόδων και ηλικιών προσαρμογής, οδηγώντας μας στο συμπέρασμα ότι οι προβλέψεις για τον ελληνικό πληθυσμό θα πρέπει να βασίζονται στα πιο πρόσφατα δεδομένα θνησιμότητας.

Εν κατακλείδι, η χρήση του υποδείγματος παλινδρόμησης αξιοπιστίας χαρτοφυλακίου φαίνεται να παρουσιάζει ιδιαίτερο ενδιαφέρον, ειδικότερα για εφαρμογή σε δεδομένα χωρών περιορισμένης περιόδου παρατηρήσεων, καθώς μπορεί να συλλάβει αποτελεσματικά την υποκείμενη τάση θνησιμότητας για μια δεδομένη ηλικία, λαμβάνοντας υπόψιν την πληροφορία που έχει αποκτηθεί και από πληθυσμούς άλλων ηλικιών. Τέλος, αξίζει να αναφέρουμε ότι η εργασία αυτή μπορεί να αποτελέσει και έναν οδηγό εφαρμογής της μεθόδου μας, ενθαρρύνοντας τη χρήση τεχνικών αξιοπιστίας για μελλοντικές μελέτες πρόβλεψης θνησιμότητας και σε πληθυσμούς άλλων χωρών.

ABSTRACT

In this paper we propose the credibility regression model as a method of modeling and forecasting the mortality rates for a given population. After fitting this model to Greek data, we evaluate its forecasting performance using error-accuracy measures, leading to the conclusion that credibility forecasts are more accurate than the corresponding results of the Lee-Carter and the CBD mortality models.

Keywords: Lee-Carter, CBD, Credibility Theory, Regression, Mortality Forecasting, Greece.

Ευχαριστίες: Ο πρώτος συγγραφές αναγνωρίζει την πλήρη υποστήριξη από τη Γενική Γραμματεία Έρευνας και Τεχνολογίας (ΓΓΕΤ) και το Ελληνικό Ίδρυμα Έρευνας και Καινοτομίας (ΕΛΙΔΕΚ) (Κωδικός Υποτροφίας: 1286). Ο δεύτερος συγγραφέας αναγνωρίζει τη μερική υποστήριξη από το Κέντρο Ερευνών του Πανεπιστημίου Πειραιώς.

ΑΝΑΦΟΡΕΣ

Bozikas, A. and Pitselis, G. (2018). An Empirical Study on Stochastic Mortality Modelling under the Age-Period-Cohort Framework: The Case of Greece with Applications to Insurance Pricing. *Risks*, 6(2):1–34.

Brouhns, N., Denuit, M. and Vermunt, J. K. (2002). A Poisson log-bilinear regression approach to the construction of projected lifetables. *Insurance: Mathematics and Economics*, 31(3):373–393.

Bühlmann, H. and Gisler, A. (2005). *A course in credibility theory and its applications*. The Netherlands: Springer.

Cairns, A. J. G., Blake, D. and Dowd, K. (2006). A two-factor model for stochastic mortality with parameter uncertainty: Theory and calibration. *Journal of Risk and Insurance*, 73(4):687–718.

Cairns, A. J. G., Blake, D., Dowd, K., Coughlan, G. D., Epstein, D., Ong, A. and Balevich, I. (2009). A quantitative comparison of stochastic mortality models using data from England and Wales and the United States. *North American Actuarial Journal*, 13(1):1–35.

De Vylder, F. (1978). Parameter estimation in credibility theory. *ASTIN Bulletin*, 10(1):99–112.

Goovaerts, M. J., Kaas, R., Van Heerwaarden, A. E. and Bauwelinckx, T. (1990). *Effective actuarial methods*. North-Holland.

Hachemeister, C. (1975). Credibility for Regression Models with Application to Trend (Reprint). In Kahn, P., editor, *CREDIBILITY: Theory and Applications*, pages 307–348. Academic Press, Inc.

Hardy, M. R. and Panjer, H. H. (1998). A credibility approach to mortality risk. *Astin Bulletin*, 28(2):269–283.

Human Mortality Database (2017). University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Available at www.mortality.org.

Laliotis, I., Ioannidis, J. P. and Stavropoulou, C. (2016). Total and cause-specific mortality before and after the onset of the Greek economic crisis: an interrupted time-series analysis. *The Lancet Public Health*, 1(2):e56–e65.

Lee, R. D. and Carter, L. R. (1992). Modeling and Forecasting U.S. Mortality. *Journal of the American Statistical Association*, 87(419):659–671.

Oeppen, J. and Vaupel, J. W. (2002). Broken Limits to Life Expectancy. *Science*, 296:1029 – 1031.

R Core Team (2017). R: A language and environment for Statistical Computing.

Salhi, Y., Thérond, P. E. and Tomas, J. (2016). A credibility approach of the Makeham mortality law. *European Actuarial Journal*, 6(1):61–96.

Schinzinger, E., Denuit, M. M. and Christiansen, M. C. (2016). A multivariate evolutionary credibility model for mortality improvement rates. *Insurance: Mathematics and Economics*, 69:70–81.

Tsai, C. C. L. and Lin, T. (2017a). A Bühlmann credibility approach to modeling mortality rates. *North American Actuarial Journal*, 21(2):204–227.

Tsai, C. C. L. and Lin, T. (2017b). Incorporating the Bühlmann credibility into mortality models to improve forecasting performances. *Scandinavian Actuarial Journal*, 2017(5):419–440.

Tsai, C. C. L. and Yang, S. (2015). A Linear Regression Approach to Modeling Mortality Rates of Different Forms. *North American Actuarial Journal*, 19(1):1–23.

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 31ου Πανελληνίου Συνεδρίου Στατιστικής (2018), σελ.161-178

ΕΜΠΕΙΡΙΚΗ ΔΙΕΡΕΥΝΗΣΗ ΤΗΣ ΑΣΥΜΜΕΤΡΙΚΗΣ ΠΡΟΣΑΡΜΟΓΗΣ ΤΩΝ ΛΙΑΝΙΚΩΝ ΤΙΜΩΝ ΣΤΗΝ ΕΛΛΗΝΙΚΗ ΑΓΟΡΑ ΤΡΟΦΙΜΩΝ.

Ζ. Μπραγουδάκης
Τράπεζα της Ελλάδος
ZBragoudakis@bankofgreece.gr

ΠΕΡΙΛΗΨΗ

Ο βασικός σκοπός της εργασίας είναι να διερευνήσει εμπειρικά την ύπαρξη ασυμμετρίας στην μακροχρόνια προσαρμογή των λιανικών τιμών σε σχέση με τις μεταβολές των τιμών του παραγωγού στις βασικές κατηγορίες διατροφής: τα δημητριακά, τα γαλακτοκομικά , τα κρέατα, τα φρούτα και τα οπωροκηπευτικά. Η ύπαρξη ασυμμετρίας είναι ένδειξη έλλειψης ανταγωνισμού και στρεβλώσεων στην αγορά των τροφίμων. Για το σκοπό αυτό εφαρμόζεται η τεχνική του υποδείγματος διόρθωσης σφάλματος με συνολοκλήρωση αυτοπαλίνδρομου κατωφλιού. Τα δεδομένα έχουν μηνιαία συχνότητα και καλύπτουν την περίοδο από τον Ιανουάριο του 2002 έως τον Ιούνιο του 2016. Τα αποτελέσματα παρέχουν ενδείξεις για την ύπαρξη ασυμμετρικής προσαρμογής στις λιανικές τιμές των δημητριακών και των κρεάτων, δηλαδή οι τιμές αυξάνονται γρηγορότερα μετά από μια αύξηση των τιμών του παραγωγού, ενώ αργούν να μειωθούν ύστερα από μια πτώση των τιμών του παραγωγού. Αντίθετα, για τις περιπτώσεις των γαλακτοκομικών, των φρούτων και των οπωροκηπευτικών, η υπόθεση της συμμετρίας δεν μπορεί να απορριφθεί.

Λέξεις κλειδιά : Αγορά τροφίμων**,** Συνολοκλήρωση Αυτοπαλίνδρομου Κατωφλιού.

1. ΕΙΣΑΓΩΓΗ

Όταν οι τιμές καταναλωτή αργούν να μειωθούν ύστερα από μια πτώση των τιμών του παραγωγού, ενώ αυξάνονται γρηγορότερα μετά από μια αύξηση των τιμών αυτών, τότε τα περιθώρια κέρδους του παραγωγού αυξάνονται και καταγράφεται αύξηση του πλεονάσματος των προμηθευτών εις βάρος των καταναλωτών. Όσο πιο ανταγωνιστικά λειτουργεί ένας κλάδος τόσο μεγαλύτερες είναι οι πιθανότητες η μετάδοση των τιμών από τον παραγωγό στον τελικό καταναλωτή να είναι άμεση και συμμετρική.

Η πλειονότητα των εμπειρικών μελετών αναφέρουν ότι οι ασυμμετρίες στην προσαρμογή των τιμών προκύπτουν εξαιτίας του μεγάλου βαθμού μονοπωλιακής δύναμης των επιχειρήσεων που συμμετέχουν στην αλυσίδα εμπορίας των τροφίμων, δηλαδή της έντονα ολιγοπωλιακής δομής της χονδρικής και λιανικής αγοράς (Peltzman (2000) και Meyer και Taubadel (2004)).

Ο βασικός σκοπός της εργασίας είναι η εμπειρική διερεύνηση της ύπαρξης ασυμμετρικής ταχύτητας προσαρμογής των λιανικών τιμών προς το μακροχρόνιο επίπεδο ισορροπίας τους σε μια ομάδα βασικών τροφίμων. Οι κατηγορίες τροφίμων που εξετάζονται είναι τα δημητριακά προϊόντα, το κρέας και τα προϊόντα κρέατος, τα γαλακτοκομικά προϊόντα, τα φρούτα και τα οπωροκηπευτικά. Η συγκεκριμένη ομάδα τροφίμων επιλέχθηκε για τους εξής λόγους: α) υπάρχουν ενδείξεις ότι στους συγκεκριμένους κλάδους τροφίμων δεν λειτουργεί αποτελεσματικά ο ανταγωνισμός και υπάρχει υψηλή συγκέντρωση που οδηγεί σε ολιγοπωλιακή χειραγώγηση των τιμών και β) έχουν μεγάλη βαρύτητα στο καλάθι του μέσου καταναλωτή.

Για το σκοπό αυτό εφαρμόζεται η τεχνική του υποδείγματος διόρθωσης σφάλματος με συνολοκλήρωση αυτοπαλίνδρομου κατωφλιού (ECM with TAR cointegration) που αναπτύχθηκε από τους Enders και Granger (1998) και Enders και Siklos (2001). Η συνολοκλήρωση αυτοπαλίνδρομου κατωφλιού βασίζεται στην ιδέα της ασυμμετρικής σύγκλισης των ανισορροπιών προς τη μακροχρόνια σχέση συνολοκλήρωσης. Αυτή η παραδοχή στατιστικά σημαίνει ότι τα κατάλοιπα της μακροχρόνιας σχέσης έχουν ασυμμετρική αυτοπαλίνδρομη συμπεριφορά. Τα δεδομένα καλύπτουν την περίοδο από τον Ιανουάριο του 2002 έως τον Ιούνιο του 2016. Αξιοποιείται το πληρέστερο και πιο πρόσφατο δείγμα των διαθέσιμων μηνιαίων παρατηρήσεων των υπό εξέταση μεταβλητών συγκριτικά με οποιαδήποτε άλλη μελέτη που αφορά την ελληνική αγορά τροφίμων.

Το υπόλοιπο τμήμα του κεφαλαίου οργανώνεται ως εξής: Στην Ενότητα 2 παρουσιάζεται συνοπτικά η επισκόπηση της βιβλιογραφίας. Στην Ενότητα 3 περιγράφονται οι βασικές αιτίες της ασυμμετρικής προσαρμογής στις τιμές των τροφίμων. Στην Ενότητα 4 αναλύεται η σπουδαιότητα της ομάδας των βασικών τροφίμων για το μέσο νοικοκυριό. Τα δεδομένα που χρησιμοποιήθηκαν παρουσιάζονται στην Ενότητα 5. Στην Ενότητα 6 παρουσιάζεται η οικονομετρική μεθοδολογία. Στην Ενότητα 7 παρουσιάζονται τα εμπειρικά ευρήματα. Τέλος, στην Ενότητα 8 συνοψίζονται τα συμπεράσματα.

2. ΒΙΒΛΙΟΓΡΑΦΙΚΗ ΕΠΙΣΚΟΠΙΣΗ

Η βιβλιογραφία που αναφέρεται στην ελληνική αγορά τροφίμων είναι σχετικά περιορισμένη. Σε αυτές τις μελέτες γίνεται μια συνοπτική περιγραφή του κλάδου των τροφίμων ή μεμονωμένων κατηγοριών διατροφής, διερευνάται ο μηχανισμός μετάδοσης των τιμών και επισημαίνεται ο ελλιπής ανταγωνισμός. Πιο συγκεκριμένα:

Η Reziti (2004) εξέτασε την αγορά για τα μαρούλια, τα πορτοκάλια, τις πατάτες και το κρέας στην Ελλάδα για το διάστημα 2000-2003 χρησιμοποιώντας μηνιαία στοιχεία. Τα αποτελέσματα κατά τη βραχυχρόνια περίοδο υποστηρίζουν την ύπαρξη ασυμμετρίας (εκτός από την αγορά πορτοκαλιών) στην μετάδοση των τιμών. Στην περίπτωση που μειώνονται οι τιμές παραγωγού οι λιανέμποροι καθυστερούν να μειώσουν τις τιμές τους ενώ, αντίθετα όταν οι τιμές παραγωγού αυξάνουν οι λιανέμποροι σπεύδουν να μετακυλήσουν άμεσα τις αυξήσεις στους καταναλωτές.

Οι Reziti και Panagopoulos (2008,2006) μελέτησαν την αγορά τροφίμων, λαχανικών και φρούτων χρησιμοποιώντας μηνιαία στοιχεία για το διάστημα 1994-2005. Τα ευρήματα υποστηρίζουν την ύπαρξη ασυμμετρικής ταχύτητας στη μετάδοση των τιμών από τον παραγωγό στον καταναλωτή στα τρόφιμα και τα λαχανικά αλλά όχι στα φρούτα.

Η Επιτροπή Ανταγωνισμού (2011) έκανε κλαδική έρευνα για την περίοδο 2005-2011, στα οπωροκηπευτικά και επικεντρώθηκε σε δείγμα επτά οπωροκηπευτικών προϊόντων (Μήλα, Πορτοκάλια, Ροδάκινα, Πατάτες, Τομάτες, Αγγούρια και Μαρούλια) με κύριο κριτήριο τη συμμετοχή τους στο καλάθι της νοικοκυράς. Τα αποτελέσματα καταδεικνύουν σημαντικές ασυμμετρίες όσον αφορά στις ανταποκρίσεις σε θετικές και αρνητικές μεταβολές των τιμών στην εφοδιαστική αλυσίδα οπωροκηπευτικών.

Οι Rezitis και Reziti (2011) μελέτησαν την αγορά γάλακτος χρησιμοποιώντας μηνιαία στοιχεία για το διάστημα 1989-2009. Τα ευρήματα υποστηρίζουν την ύπαρξη ασυμμετρικής ταχύτητας στη μετάδοση των τιμών από τον παραγωγό στον καταναλωτή. Οι τιμές καταναλωτή αυξάνουν γρηγορότερα από την μεταβολή των τιμών του παραγωγού. Η Tsakistara και αλ.(2014) χρησιμοποιώντας ετήσια στοιχεία για το διάστημα 1990-2008 κατέληξαν ότι η αγορά γάλακτος λειτουργεί με ολιγοπωλιακές συνθήκες, οι οποίες έμμεσα ενδυναμώνουν τον ισχυρισμό για την ύπαρξη καρτέλ στον κλάδο. Η Reziti (2014) εξέτασε την αγορά γάλακτος χρησιμοποιώντας μηνιαία στοιχεία για το διάστημα 1998-2014. Τα ευρήματα καταδεικνύουν ότι ο μηχανισμός μετάδοσης των τιμών είναι ασυμμετρικός και την βραχυχρόνια και την μακροχρόνια περίοδο, υποννοώντας ότι οι έμποροι της λιανικής έχουν δύναμη αγοράς στην διαμόρφωση των τιμών σε σχέση με τους παραγωγωγούς.

3. ΑΙΤΙΕΣ ΑΣΥΜΕΤΡΙΩΝ ΣΤΗΝ ΑΓΟΡΑ ΤΡΟΦΙΜΩΝ

Η βιβλιογραφία έχει αναδείξει εμπειρικά τους παράγοντες που πιθανά ευθύνονται για την ασυμμετρική προσαρμογή των λιανικών τιμών στα τρόφιμα:

Πρώτον**, ο βαθμός μονοπωλιακής δύναμης** επιτρέπει στις ολιγοπωλιακές επιχειρήσεις που συμμετέχουν σε συμπαιγνία (μυστικές συνεννοήσεις) να πετυχαίνουν να διατηρούν συστηματικά υψηλότερα περιθώρια κέρδους (Miller και Hayenga (2001), Peltzman (2000), Meyer και Taubadel (2004)).

Δεύτερον, η ασυμμετρική μετάδοση των τιμών οφείλεται επίσης στην **υψηλή φθαρτότητα** κάποιων προϊόντων στην αγορά των τροφίμων (π.χ. γάλακτος, φρούτων). Ο Ward (1982) υποστηρίζει ότι οι παραγωγοί-αγρότες και οι μεταπωλητές (χονδρέμποροι και λιανέμποροι) που κατέχουν αναλώσιμα προϊόντα ενδεχομένως να μην αυξήσουν τις τιμές έπειτα από αύξηση των τιμών των εισροών, ώστε να αποφύγουν τον κίνδυνο να μείνουν με χαλασμένα προϊόντα.

Τρίτον, **τα κόστη ανατιμολόγησης** που προκύπτουν για τις επιχειρήσεις όταν αλλάζουν οι τιμές των εισροών. Οι αυξήσεις των τιμών των εισροών είναι πιο πιθανό να οδηγήσουν σε αναπροσαρμογή των λιανικών τιμών του τελικού προϊόντος από ό,τι οι μειώσεις των τιμών των εισροών. Εξαιτίας του πληθωρισμού κάποιες αυξήσεις

καθίστανται αναγκαίες λόγω της μείωσης στην πραγματική αξία του περιθωρίου κέρδους (Buckle και Carlson (2000).

Τέταρτον, η **κρατική παρέμβαση** με την ρύθμιση κατώτατων τιμών μπορεί να οδηγήσει σε ασυμμετρική μετάδοση των τιμών, εάν οδηγεί χονδρεμπόρους ή λιανοπωλητές να πιστεύουν ότι μια μείωση των γεωργικών τιμών θα είναι μόνο προσωρινή. Αυτό συμβαίνει διότι αναμένουν μια κρατική παρέμβαση στήριξης των τιμών του παραγωγού (Kinnucan και Forker, 1987).

Πέμπτον, **η ασύμμετρη πληροφόρηση** μεταξύ των εταιρειών. Οι μεγαλύτερες επιχειρήσεις συνήθως έχουν καλύτερη πληροφόρηση για την εξέλιξη των τιμών των παραγωγικών εισροών, όπως για παράδειγμα των διεθνών τιμών των καυσίμων ή των σιτηρών. Προμηθεύονται γρηγορότερα και φθηνότερα σε σχέση με τους ανταγωνιστές τους και έτσι έχουν περισσότερους βαθμούς ελευθερίας στην προσαρμογή των τιμών που πωλούν (Bailey και Brorsen,1989).

4. Η ΣΠΟΥΔΑΙΟΤΟΤΗΤΑ ΤΩΝ ΒΑΣΙΚΩΝ ΤΡΟΦΙΜΩΝ

Στην ενότητα αυτή καταδεικνύεται ο σημαντικός βαθμός βαρύτητας που έχει η συγκεκριμένη ομάδα τροφίμων (γαλακτοκομικά, κρέας, δημητριακά, φρούτα και οπωροκηπευτικά) ως προς το καταναλωτικό πρότυπο του μέσου νοικοκυριού.

Σύμφωνα με την Έρευνα Οικογενειακών Προυολογισμών (ΕΟΠ,2015) που δημοσιεύθηκε τον Ιούλιο του 2016 από την Ελληνική Στατιστική Αρχή (ΕΛΣΤΑΤ), η μέση μηνιαία δαπάνη των νοικοκυριών για το 2015 ήταν 1419,57 €. Η μέση μηνιαία δαπάνη των νοικοκυριών για είδη διατροφής (τρόφιμα και μη οινοπνευματώδη ποτά) ήταν 293,7930 € ή 20,7% επί της μέσης μηνιαίας δαπάνης των νοικοκυριών και αποτελεί την κυριότερη δαπάνη του νοικοκυριού διαχρονικά (βλέπε Πίνακα 4.1). Ακολουθούν οι δαπάνες για στέγαση (189,21€ ή 13,3%) και οι δαπάνες για μεταφορές (181,64 € ή 12,8%). Η μέση μηνιαία δαπάνη των νοικοκυριών σύμφωνα με την ΕΟΠ 2010 ήταν 1956,42 €, το 2015 έπεσε στα 1419,57 € καταγράφοντας μείωση κατά 27,4% ή 536,85 €. Αντιθέτως, η μέση μηνιαία δαπάνη των νοικοκυριών για είδη διατροφής την ίδια περίοδο 2010-2015, μειώθηκε μόνο κατά 16,6% (58,37€), δείχνοντας μια ανελαστική ζήτηση των καταναλωτών. Επισημαίνεται, ότι το ποσοστό των ειδών διατροφής στην μέση μηνιαία δαπάνη αυξήθηκε διαχρονικά από 18% το 2010 σε 20,7% το 2015. Το 2015 (αλλά και την περίοδο 2010-2015) η βαρύτητα της ομάδας τροφίμων (γαλακτοκομικά, κρέας, δημητριακά, φρούτα και οπωροκηπευτικά) με βάση την μέση μηνιαία δαπάνη του νοικοκυριού για την διατροφή είναι πολύ μεγάλη (βλ. Πίνακα 4.2). Το ποσοστό της εξεταζόμενης ομάδας τροφίμων αποτελεί το 75,0% (218 €) της μέσης μηνιαίας δαπάνης του νοικοκυριού για διατροφή (299,30 €). Το κρέας με ποσοστό 22,4% (65,56 €) είναι η μεγαλύτερη δαπάνη διατροφής για το νοικοκυριό και ακολουθούν τα άλευρα, το ψωμί και τα δημητριακά με ποσοστό 15,3% (44,84 €), τα γαλακτοκομικά με ποσοστό 17,6%(51,52 €), τα λαχανικά με ποσοστό 12,1% (35,62 €) και τα φρούτα με ποσοστό 7,2% (21,26 €). Συνεπώς, ο βαθμός στον οποίο λειτουργούν αποτελεσματικά οι αγορές των συγκεκριμένων

τροφίμων επηρεάζει σημαντικά το βιοτικό επίπεδο των νοικοκυριών. **Πίνακας 4.1.** Μέση μηνιαία δαπάνη νοικοκυριών για αγαθά και υπηρεσίες.

Έρευνα Οικογενειακών Προϋπολογισμών (ΕΟΠ 2015)

Αγαθά και υπηρεσίες	2010		2011		2012		2013		2014		2015		Μεταβολές 2010-2015	
Ποσά σε € και % επί της μέσης μηνιαίας δαπάνης	Αξία	%	Αξία	%	Αξία	%	Αξία	%	Αξία	%	Αξία	%	Αξία	%
Μέση μηνιαία δαπάνη	**1956.42**	**100.0%**	**1824.02**	**100.0%**	**1637.1**	**100.0%**	**1509.39**	**100.0%**	**1460.52**	**100.0%**	**1419.57**	**100.0%**	**-536.85**	**-27.4**
Είδη διατροφής	351.67	18.0%	355.05	19.5%	328.57	20.1%	307.33	20.4%	299.79	20.5%	293.30	20.7%	-58.37	-16.6
Αλκοολούχα ποτά και καπνός	68.7	3.5%	66.52	3.6%	62.71	3.8%	62.8	4.2%	58.8	4.0%	57.27	4.0%	-11.43	-16.6
Ένδυση και υπόδηση	140.84	7.2%	112.51	6.2%	95.34	5.8%	87.38	5.8%	85.7	5.9%	83.06	5.9%	-57.78	-41.0
Στέγαση	228.82	11.7%	230.16	12.6%	227.07	13.9%	206.99	13.7%	195.29	13.4%	189.21	13.3%	-39.61	-17.3
Διαρκή Αγαθά	130.47	6.7%	110.05	6.0%	94.97	5.8%	83.94	5.6%	72.76	5.0%	66.49	4.7%	-63.98	-49.0
Υγεία	124.43	6.4%	114.58	6.3%	104.71	6.4%	104.44	6.9%	105.76	7.2%	107.06	7.5%	-17.37	-14.0
Μεταφορές	264.87	13.5%	240.05	13.2%	209.88	12.8%	189.19	12.5%	184.82	12.7%	181.64	12.8%	83.23	31.4
Επικοινωνίες	78.46	4.0%	73.69	4.0%	68.19	4.2%	61.91	4.1%	60.08	4.1%	58.46	4.1%	-20	-25.5
Αναψυχή και πολιτισμός	91.8	4.7%	85.72	4.7%	72.87	4.5%	68.82	4.6%	68.71	4.7%	67.95	4.8%	-23.65	-25.8
Εκπαίδευση	64.21	3.3%	63.71	3.5%	57.33	3.5%	50.83	3.4%	50.84	3.5%	46.7	3.3%	-17.51	-27.3
Ξενοδοχεία , καφενεία, και εστιατόρια	209.75	10.7%	189.11	10.4%	160.47	9.8%	145.55	9.6%	143.49	9.8%	141.05	9.9%	-68.7	-32.8
Διάφορα αγαθά και υπηρεσίες	202.61	10.4%	182.69	10.0%	154.98	9.5%	140.19	9.3%	134.49	9.2%	127.37	9.0%	-75.24	-37.1

Πηγή:ΕΛΣΤΑΤ

Πίνακας 4.2. Ποσοστιαία κατανομή μηνιαίων δαπανών για είδη διατροφής.

Αξία σε ευρώ	ΕΟΠ 2015		ΕΟΠ 2014		ΕΟΠ 2015/2014	ΕΟΠ 2015/2014
Είδη διατροφής και μη οινοπνευματώδη ποτά	Αξία	Κατανομή %	Αξία	Κατανομή %	Μεταβολή αξίας %	Διαφορά ποσοστιαίας κατανομής
Σύνολο	293,30	100,0	299,79	100,0	3,9	
Αλεύρι,ψωμί,δημητριακά	44,94	15,3	46,99	15,7	-4,6	-0,4
Κρέας	65,56	22,4	68,06	22,7	-3,7	-0,3
Ψάρια	21,16	7,2	21,6	7,2	-2,0	0,0
Γαλακτοκομικά προϊόντα και αυγά	51,52	17,6	54,24	18,1	-5,0	-0,5
Έλαια και λίπη	17,1	6,0	17,59	5,9	0,9	0,1
Φρούτα	21,26	7,2	21,39	7,1	-0,6	0,1
Λαχανικά	35,62	12,1	35,36	11,8	0,7	0,3
Ζάχαρη,μαρμελάδα,μέλι,σιρόπια, σοκολάτα και ζαχαρωτά	14,57	5,0	13,76	4,6	5,9	0,4
Λοιπά είδη διατροφής	6,10	2,1	4,92	1,6	24,0	0,5
Καφές,τσάι και κακάο	6,21	2,1	6,68	2,2	-7,0	-0,1
Μεταλλικά νερά,χυμοί,αναψυκτικά,χυμοί φρούτων και λαχανικών	8,72	3,0	9,21	3,1	-5,3	-0,1

Πηγή:ΕΛΣΤΑΤ.

Ο δείκτης συγκέντρωσης της αγοράς HHI κατασκευάστηκε από τους Herfindahl(1950)-Hirschman(1964) και μετρά το βαθμό στον οποίο ένας μικρός αριθμός επιχειρήσεων αντιπροσωπεύει μεγάλο τμήμα της αγοράς.

Ο δείκτης HHI υπολογίζεται ως εξής: $HHI = \sum_{i=1}^{n} (s_i)^2 = s_1^2 + s_2^2 + s_3^2 + ... + s_n^2$ όπου

s_i είναι το μερίδιο αγοράς της επιχείρησης i, με συνολικό αριθμό των επιχειρήσεων n στην αγορά. Όταν ο HHI είναι κατώτερος του 1.000, η συγκέντρωση της αγοράς χαρακτηρίζεται ως χαμηλή, όταν κυμαίνεται μεταξύ 1.000 και 1.800 η συγκέντρωση της αγοράς θεωρείται μέτρια, ενώ με τον HHI άνω του 1.800 χαρακτηρίζεται ως υψηλή. Στον Πίνακα 4.3 παρουσιάζονται ο αριθμός των επιχειρήσεων κατά κλάδο τροφίμων, τα αντίστοιχα μερίδια αγοράς και ο δείκτης συγκέντρωσης (HHI) για το 2014.

Πίνακας 4.3. Ο δείκτης συγκέντρωσης Herfindahl-Hirschmann (HHI)

Κλάδοι Τροφίμων/ Επιχειρήσεις	Πολύ Μεγάλες		Μεγάλες		Μικρές και Μεσαίες		Σύνολο επιχειρήσεων του δείγματος		Δείκτης συγκέντρωσης αγοράς Herfindahl-Hirschmann (HHI)	Συμπέρασμα
	πάνω από € 250 εκατ. πωλήσεις		από € 50 εκατ. έως € 250 εκατ. πωλήσεις		από € 2 εκατ. έως €50 εκατ. πωλήσεις					
	Αριθμός επιχειρήσεων	Μερίδιο αγοράς	Αριθμός επιχειρήσεων	Μερίδιο αγοράς	Αριθμός επιχειρήσεων	Μερίδιο αγοράς	Αριθμός επιχειρήσεων	Μερίδιο αγοράς		
Δημητριακά και παρασκευάσματα [1]										
Αρτοποία και Ζαχαροπλαστική	**2**	**24%**	**11**	**42%**	**80**	**19%**	**93**	**85%**	**2926**	**Υψηλή Συγκέντρωση**
Κλάδος των Αλευρόμυλων			**1**		**17**		**18**	**90%**	**8200**	**Υψηλή Συγκέντρωση**
Προϊόντα κρέατος	**4**		**17**		**1.514** [2]					**Υψηλή Συγκέντρωση**
Κλάδος των αλλαντικών			**3**	**74%**	**5**	**11%**	**8**	**85%**	**5822**	**Υψηλή Συγκέντρωση**
Γαλακτοκομικά και αυγά	**4**	**53%**	**5**		**56**		**65**	**86%**	**4094**	**Υψηλή Συγκέντρωση**
Φρούτα και λαχανικά	**2** [3]	**20%**	**59**	**25%**		**55%**		**100%**	**4050**	**Υψηλή Συγκέντρωση**

[1] 4 αρτοποιητικές /εμπορικές επιχειρήσεις κατέχουν μαζί περί το 60% του συνόλου του μεριδίου της αγοράς και το 75% του κλάδου των αλευρόμυλων για το 2013 αντιπροσωπεύονται από 17 εταιρείες.

[2] 256 βιομηχανικές και 927 εμπορικές εταιρίες , ενώ μεσαίου μεγέθους εταιρίες ήταν 152 βιομηχανικές και 179 εμπορικές.

[3] Κεντρικές Αγορές:Πρόκειται για τα οργανωμένα σημεία χονδρεμπορικής πώλησης, τα σημαντικότερα από τα οποία είναι: α) ο Οργανισμός Κεντρικής Αγοράς Αθηνών Α.Ε. (ΟΚΑΑ), με 530 δροσοπηγοποιούμενες χονδρεμπορικές επιχειρήσεις νωπών οπωρολαχανικών και β)η Κεντρική Αγορά Θεσσαλονίκης Α.Ε.(ΚΑΘ), η οποία περιλαμβάνει 280 καταστήματα. Σύμφωνα με την ICAP(2015) οι 2 Κεντρικές Αγορές έχουν το 30% του συνόλου της αγοράς. οι χονδρεμπορικές επιχειρήσεις εκτός κεντρικών αγορών κατέχουν το 35% τα super markets το 25% και οι λαϊκές αγορές το υπόλοιπο 20%

Στον κλάδο των δημητριακών και συγκεκριμένα στον κλάδο των αρτοπαρασκευασμάτων και της ζαχαροπλαστικής δύο επιχειρήσεις κατέχουν μερίδιο αγοράς 24% σε όρους κύκλου εργασιών (πωλήσεων), ενώ 11 επιχειρήσεις αντιπροσωπεύουν το 42% των συνολικών πωλήσεων. Ο HHI

υπολογίστηκε στις 2.926 (>1.800) μονάδες, υποδηλώνοντας ότι ο κλάδος χαρακτηρίζεται από υψηλή συγκέντρωση. Επίσης, στον κλάδο των αλευρόμυλων το 75% του κλάδου σε όρους πωλήσεων για το 2013 αντιπροσωπεύεται από 17 επιχειρήσεις, με τον ΗΗΙ να είναι πολύ υψηλός αφού υπολογίζεται στις 8.200 (>1.800) μονάδες. Στον κλάδο του κρέατος και των παρασκευασμάτων του, επίσης υπάρχει έντονη συγκέντρωση της αγοράς, αφού τέσσερις είναι οι πολύ μεγάλες επιχειρήσεις του κλάδου και 17 οι μεγάλες, ενώ στον κλάδο των αλλαντικών ο δείκτης συγκέντρωσης είναι πολύ υψηλός και φθάνει τις 5.822 (>1.800) μονάδες. Στον κλάδο των γαλακτοκομικών, μόνο τέσσερις επιχειρήσεις αντιπροσωπεύουν πάνω από το 53% των συνολικών πωλήσεων για το 2013. Ο ΗΗΙ υπολογίστηκε στις 4.094 μονάδες (>1.800), συνεπώς και αυτός ο κλάδος χαρακτηρίζεται από υψηλή συγκέντρωση. Στον κλάδο των φρούτων και των λαχανικών, οι δύο κεντρικές αγορές της χώρας (ΟΚΑΑ και ΚΑΘ) κατέχουν το 20% του συνόλου της αγοράς, οι χονδρεμπορικές επιχειρήσεις εκτός κεντρικών αγορών κατέχουν το 35%, τα σούπερ μάρκετ το 25% και οι λαϊκές αγορές το υπόλοιπο 20%. Ο ΗΗΙ υπολογίστηκε στις 2.650 (>1.800) μονάδες, υποδηλώνοντας ότι ο κλάδος χαρακτηρίζεται επίσης από υψηλή συγκέντρωση.

5. ΤΑ ΔΕΔΟΜΕΝΑ

Τα δεδομένα που χρησιμοποιούνται είναι οι τιμές του παραγωγού και οι λιανικές τιμές για τις βασικές κατηγορίες τροφίμων: τα δημητριακά και τα παρασκευάσματα, τα κρέατα και προϊόντα κρέατος, τα γαλακτοκομικά και τα αυγά, τα φρούτα και τα οπωροκηπευτικά και καλύπτουν την περίοδο από τον Ιανουάριο του 2002 έως τον Ιούνιο του 2016. Οι πηγές των δεδομένων είναι η στατιστική βάση της Eurostat (https://ec.europa.eu/eurostat/web/experimental-statistics/food-price-monitoring) και η ΕΛΣΤΑΤ [Δείκτης Τιμών Παραγωγού (ΔΤΠ), https://www.statistics.gr/statistics/-/publication/DKT15/]. Οι τιμές λιανικής των τροφίμων εκφράζονται από τις αντίστοιχες κατηγορίες του Εναρμονισμένου Δείκτη Τιμών Καταναλωτή (ΕνΔΤΚ). Οι τιμές παραγωγού των τροφίμων εκφράζονται από τις αντίστοιχες κατηγορίες του (ΔΤΠ). Σημειώνεται ότι ενώ στην μέτρηση των τιμών που γίνεται με βάση τον ΕνΔΤΚ περιλαμβάνεται ο φόρος προστιθέμενης αξίας (ΦΠΑ), στον ΔΤΠ συλλέγονται οι βασικές τιμές πώλησης, δηλαδή οι τιμές χωρίς το ΦΠΑ και παρόμοιους εκπεστέους φόρους που συνδέονται άμεσα με τον κύκλο εργασιών, καθώς και χωρίς άλλους δασμούς ή φόρους στα προϊόντα, περιλαμβάνοντας όμως τις επιδοτήσεις στο προϊόν. Για το λόγο αυτό, από τον ΕνΔΤΚ αφαιρείται ο ΦΠΑ, προκειμένου να υπάρχει συνέπεια με τον ΔΤΠ.

6. Η ΟΙΚΟΝΟΜΕΤΡΙΚΗ ΜΕΘΟΔΟΛΟΓΙΑ

Για την εμπειρική διερεύνηση της ύπαρξης ασυμμετρικής προσαρμογής των λιανικών τιμών της βενζίνης προς το μακροχρόνιο επίπεδο ισορροπίας τους, χρησιμοποιείται η τεχνική του υποδείγματος διόρθωσης σφάλματος με συνολοκλήρωση

αυτοπαλίνδρομου κατωφλιού (ECM with TAR cointegration) όπως έχει προταθεί από τους Enders και Siklos (2001). Στην περίπτωση όπου η λιανική τιμή R_t^r και η τιμή του παραγωγού R_t^w ακολουθούν μια μακροχρόνια σχέση, αυτή μπορεί να περιγραφεί ως εξής :

$$R_t^r = a_0 + \beta_1 R_t^w + u_t \qquad \text{όπου } t=1,..., T \tag{1}$$

Με R_t^r συμβολίζουμε την λιανική τιμή του τροφίμου R την περίοδο t, ενώ με R_t^w συμβολίζουμε την τιμή παραγωγού ή χονδρικής του ίδιου αγαθού, αντίστοιχα την περίοδο t. Η παράμετρος α_0 αντιπροσωπεύει ένα σταθερό κόστος ενώ η παράμετρος β_1 αντανακλά το βαθμό προσαρμογής τις λιανικής τιμής σε μεταβολές τις τιμής του παραγωγού και αντανακλά έμμεσα την ένταση ανταγωνισμού στην αγορά. Όσο πιο έντονος είναι ο ανταγωνισμός τόσο μεγαλύτερος αναμένεται να είναι ο βαθμός προσαρμογής και τόσο λιγότερο αναμένεται να είναι το περιθώριο κέρδους. Με u_t συμβολίζεται ο όρος σφάλματος με σταθερό μέσο και διακύμανση ο οποίος αντανακλά τις βραχυχρόνιες αποκλίσεις από την μακροχρόνια σχέση ισορροπίας. Όταν οι τιμές R_t^r και R_t^w βρίσκονται στο μακροχρόνιο επίπεδο ισορροπίας τους, τότε τα κατάλοιπα u_t αναμένεται να είναι μηδέν. Αντίθετα στην περίπτωση όπου οι τιμές αποκλίνουν από τη μακροχρόνια σχέση, τότε ο όρος σφάλματος u_t μπορεί να είναι είτε θετικός, όταν η τιμή R_t^r είναι πάνω από το μακροχρόνιο επίπεδο ισορροπίας της, είτε αρνητικός στην αντίθετη περίπτωση.

Αν η προσαρμογή τη μακροχρόνια ισορροπία είναι ασυμμετρική, υποδηλώνοντας ότι η αυτοπαλίνδρομη συμπεριφορά των καταλοίπων τις μακροχρόνιας σχέσης $\hat{u}_t$, εξαρτάται από το πρόσημο των εξωγενών διαταραχών, τότε οι τυπικοί έλεγχοι Dickey-Fuller (1979) και οι επεκτάσεις τους για συνολοκλήρωση με τη μέθοδο των δυο σταδίων Engle και Granger(1987) έχουν λανθασμένη εξειδίκευση γιατί υποθέτουν συμμετρική προσαρμογή των καταλοίπων $\hat{u}_t$ την μακροχρόνια ισορροπία (Enders and Granger (1998) και Enders και Siklos (2001). Σύμφωνα με αυτή τη προσέγγιση η τυπική εξίσωση ελέγχου για την ύπαρξη συνολοκλήρωσης η οποία περιγράφεται από τη παρακάτω σχέση :

$$\Delta \hat{u}_t = \rho \hat{u}_{t-1} + v_t \tag{2}$$

επεκτείνεται, συμπεριλαμβάνοντας ένα συναρτησιακό δείκτη βαρύτητας I_t, σύμφωνα με τον οποίο διαχωρίζονται τα κατάλοιπα τις μακροχρόνιας σχέσης (1) σε δυο συνιστώσες με χρονική υστέρηση και περιγράφεται στην ακόλουθη σχέση :

$$\Delta \hat{u}_t = I_t \rho_1^{up} \hat{u}_{t-1} + (1 - I_t) \rho_2^{down} \hat{u}_{t-1} + v_t \quad (3)$$

Σε περίπτωση που παρατηρηθεί μετά την εκτίμηση της σχέσης (3) κάποια μορφή αυτοσυσχέτισης τότε προστίθεται στην σχέση ο κατάλληλος αριθμός χρονικών υστερήσεων $\Sigma\Delta\hat{u}_{t-j}$. Οι παράμετροι ρ_1^{up} και ρ_2^{down} μετρούν τις θετικές και τις αρνητικές αποκλίσεις από το κατώφλι $\hat{\tau}$.

Ο συναρτησιακός δείκτης βαρύτητας I_t, δίνεται στη σχέση (4) όπου :

$$I_t = \begin{cases} 1 & \text{εάν} \quad \hat{u}_{t-1} \geq \hat{\tau} \\ 0 & \text{εάν} \quad \hat{u}_{t-1} < \hat{\tau} \end{cases} \quad (4)$$

όπου $\hat{\tau}$ είναι η τιμή που λαμβάνει το κατώφλι TAR. Το κατώφλι TAR είναι στοχαστικό. Στη πραγματικότητα, η αληθινή τιμή του κατωφλίου είναι άγνωστη και πρέπει να εκτιμηθεί. Για να προσδιοριστεί η τιμή του κατωφλίου $\hat{\tau}$ όπου πιθανώς να εμφανίζεται στατιστικά σημαντική ασυμμετρία οι Enders και Siklos (2001) προτείνουν την εφαρμογή μιας διερευνητικής μεθόδου που είναι γνωστή στη βιβλιογραφίας ως "προσέγγιση του Chan". Ο Chan (1993) έδειξε ότι κάτω από κάποιες τυπικές συνθήκες η μέθοδος ελαχίστων τετραγώνων OLS αποδίδει ένα υπέρ-συνεπή εκτιμητή της αληθινής τιμής του κατωφλίου. Πιο συγκεκριμένα, σύμφωνα με αυτή τη διαδικασία, ανιχνεύονται (grid search procedure) ως πιθανά κατώφλια τις οι τιμές του διανύσματος των καταλοίπων, $\hat{u}_t$, που προκύπτουν από τη σχέση συνολοκλήρωσης (1). Το διάνυσμα των κατάλοιπων $\{\hat{u}_t\}$ ταξινομείται σε αύξουσα τάξη ως εξής $\{\hat{u}_1^{\tau} < \hat{u}_2^{\tau} < \ldots < \hat{u}_T^{\tau}\}$ όπου με Τ συμβολίζεται ο αριθμός των χρησιμοποιούμενων παρατηρήσεων. Από την κατανομή των ταξινομημένων τιμών των καταλοίπων κατά αύξουσα τάξη, αφαιρείται το μεγαλύτερο και το μικρότερο 15% των (ακραίων) τιμών και το υπόλοιπο (κεντρικό) 70% των καταλοίπων θεωρούνται πιθανά κατώφλια. Για κάθε ένα από αυτά τα πιθανά κατώφλια που ανήκουν στο 70% του δείγματος εκτιμάται διαδοχικά η σχέση (3) με βάση την εξειδίκευση του δείκτη βαρύτητας της σχέσης (4). Ως καταλληλότερο κατώφλι $\hat{\tau}$ επιλέγεται εκείνη η τιμή του καταλοίπου $\hat{u}_i^{\tau}$ με βάση την οποία ελαχιστοποιείται το άθροισμα των τετραγώνων των καταλοίπων (RSS,) που προκύπτει από την εκτίμηση της σχέσης (4). Με $\{v_t\}$ συμβολίζεται η σειρά των καταλοίπων της σχέσης (3) με μέσο μηδέν και σταθερή διακύμανση , έτσι ώστε το διάνυσμα $\{v_t\}$ να είναι ανεξάρτητο του $\hat{u}_t$. Ο συνδυασμός των σχέσεων (3) και (4) συνιστούν το υπόδειγμα αυτοπαλίνδρομου κατωφλίου TAR. Επιτρέποντας στις παραμέτρους ρ_1^{up} και ρ_2^{down} να λαμβάνουν διαφορετικές τιμές, το υπόδειγμα αναγνωρίζει ότι οι θετικές και οι αρνητικές αποκλίσεις από την ισορροπία μπορεί να διορθώνονται με διαφορετικές ταχύτητες προσαρμογής.

Ο έλεγχος συνολοκλήρωσης μπορεί να γίνει με βάση τις κρίσιμες τιμές των κατανομών t-Max και Φ* που προτείνονται από τους Enders και Siklos (2001). Στην περίπτωση που διαπιστωθεί η ύπαρξη συνολοκλήρωσης αυτοπαλίνδρομου κατωφλιού, δηλαδή δεν γίνει δεκτή η υπόθεση Ho: $\rho_1^{up} = \rho_2^{down} = 0$, οι εκτιμώμενοι με την μέθοδο ελαχίστων τετραγώνων παράμετροι ρ_1^{up} και ρ_2^{down} της σχέσης (3) ακολουθούν μια ασυμπτωτικά πολύ-μεταβλητή κανονική κατανομή και έτσι είναι δυνατόν να ελεγχθεί η υπόθεση Ho: $\rho_1^{up} = \rho_2^{down}$ με βάση έναν τυπικό έλεγχο F (πχ. Wald test), για να ελεγχθεί αν η προσαρμογή είναι συμμετρική.

Στην περίπτωση όπου επαληθευτεί η ύπαρξη συνολοκλήρωσης αυτοπαλίνδρομου κατωφλιού (TAR cointegration) τότε μπορεί να εκτιμηθεί ένα υπόδειγμα διόρθωσης σφάλματος με μακροχρόνια ασυμμετρική ισορροπία (ECM with long-run asymmetric equilibrium) το οποίο στη γενική του μορφή περιγράφεται από την ακόλουθη σχέση (5).

$$\Delta r_t^r = \mu_0 + \sum_{i=1}^{k_1} \beta_{1,i} \Delta r_{t-i}^r + \sum_{i=0}^{k_2} \beta_{2,i} \Delta r_{t-i}^w + a_1 \hat{u}_{t-1}^{up} + a_1 \hat{u}_{t-1}^{down} + e_t \qquad (5)$$

όπου $\hat{u}_{t-1}^{up} = \mathrm{I_t}\, \hat{u}_{t-1}$ και $\hat{u}_{t-1}^{down} = (1 - \mathrm{I_t})\, \hat{u}_{t-1}$.

7. ΕΜΠΕΙΡΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

Για την εκτίμηση των υποδειγμάτων έχει χρησιμοποιηθεί το οικονομετρικό πρόγραμμα EViews. Πριν από την εφαρμογή των ελέγχων για συνολοκλήρωση, εξετάζεται η ύπαρξη μοναδιαίας ρίζας για τις μεταβλητές των τιμών των τροφίμων εκτελώντας τους ελέγχους Augmented Dickey-Fuller (ADF) (1979,1981), Phillips-Perron (PP) (1988,1990) και DF-GLS (Elliot,1996). Σε όλες τις μεταβλητές γίνεται λογαριθμικός μετασχηματισμός των σειρών. Οι έλεγχοι μοναδιαίας ρίζας δείχνουν ότι ενώ η υπόθεση για την ύπαρξη μοναδιαίας ρίζας δεν μπορεί να απορριφθεί στα επίπεδα των σειρών, στις πρώτες διαφορές τους οι σειρές βρέθηκαν στάσιμες. Συνεπώς, τόσο οι σειρές των λιανικών τιμών των τροφίμων όσο και των τιμών του παραγωγού είναι ολοκληρωμένες I(1). Η εκτίμηση της σχέσης (1) γίνεται με τη μέθοδο FMOLS (Fully Modified OLS) η οποία παράγει υπέρ-συνεπείς εκτιμητές και τα κατάλοιπα που προκύπτουν αντιπροσωπεύουν στοχαστικές (μη-προβλέψιμες) διαταραχές. Η μέθοδος FMOLS υπερέχει έναντι της μεθόδου OLS, διότι αυξάνει την αποτελεσματικότητα και μειώνει την μεροληψία μικρού δείγματος συγκριτικά με την εκτιμήτρια OLS (Phillips και Hansen,1990). Επιπλέον, η FMOLS δίνει ασυμπτωτικά αμερόληπτο και αποτελεσματικό εκτιμητή και συνεπώς μας επιτρέπει να εκτελεστούν αξιόπιστοι στατιστικοί έλεγχοι. Οι έλεγχοι συνολοκλήρωσης απορρίπτουν την μηδενική υπόθεση ότι οι σειρές δεν είναι συνολοκληρωμένες (βλ. Πίνακα 7.1).

Οι τιμές που λαμβάνει ο βαθμός προσαρμογής σε όλες τις κατηγορίες των εξεταζόμενων τροφίμων είναι μικρότερες από την μονάδα. Αν οι επιμέρους κλάδοι των τροφίμων λειτουργούσαν πράγματι σε περιβάλλον τέλειου ανταγωνισμού τότε θα περιμέναμε ότι ο βαθμός προσαρμογής θα ήταν 100%, δηλαδή η ελαστικότητα θα ήταν β=1, δηλαδή θα υπήρχε πλήρης μετάδοση των μεταβολών που συντελούνται στην τιμή του παραγωγού στην λιανική τιμή του τροφίμου.

Πίνακας 7.1. Εκτίμηση μακροχρόνιας σχέσης για τις εξισώσεις τιμών των τροφίμων και έλεγχος για συνολοκλήρωση

	Εξισώσεις τιμών				
Μακροχρόνιος σχέσης	**Δημητριακά**	**Κρέατα**	**Γαλακτοκομικά**	**Φρούτα**	**Λαχανικά**
Προσαρμοσμένο εύρος δείγματος	2002M01-2016M06	2002M01-2016M06	2004M01-2016M06	2005M01-2016M06	2006M01-2016M06
Σύνολο παρατηρήσεων	174	174	150	138	126
α_0	**1,098*** (5,123)**	**0,386*** (2,681)**	**0,159 (1,288)**	**1,647*** (3,570)**	**2,251*** (4,079)**
β_1	**0,734** (13,971)**	**0,917** (28,800)**	**0,965** (28,936)**	**0,650** (19,328)**	**0,505** (4,158)**
Προσαρμοσμένος συντελεστής προσδιορισμού R^2	0,987	0,988	0,985	0,633	0,285
Τυπικό σφάλμα της παλινδρόμησης	0,012	0,012	0,006	0,043	0,041
Μακροχρόνια διακύμανση	0,000	0,000	0,000	0,006	0,006
Έλεγχοι συνολοκλήρωσης κατά Engle-Granger					
Engle-Granger tau-statistic	*-4,607** [0,018]*	*-5,604** [0,000]*	*-4,956** [0,001]*	*-5,300** [0,000]*	*-4,878** [0,000]*
Engle-Granger z-statistic	-37,657** [0,015]	-44,948** [0,000]	-43,313** [0,000]	-44,365** [0,000]	-40,276** [0,000]

Τα νούμερα στις παρενθέσεις αφορούν τις τιμές της στατιστικής t.

Τα νούμερα στις αγκύλες αφορούν τις τιμές P-value .

Η στατιστική σημαντικότητα σε επίπεδο 5% συμβολίζεται με διπλούς αστερίσκους (**).

Οι κρίσιμες τιμές για τον έλεγχο της συνολοκλήρωσης παρέχονται από τον MacKinnon(1996).

Οι χρονικές υστερήσεις των ελέγχων συνολοκλήρωσης κατά Engle-Granger βασίστηκαν στο πληροφοριακό κριτήριο Schwarz.

Τα αποτελέσματα της εκτίμησης του υποδείγματος TAR για συνολοκλήρωση αυτοπαλίνδρομου κατωφλιού παρουσιάζονται στον Πίνακα 7.2. Τα αποτελέσματα της εκτίμησης του υποδείγματος TAR, υποδεικνύουν ότι η συνεπής τιμή του κατωφλιού $\hat{\tau}$ για κάθε επιμέρους κατηγορία τροφίμων είναι διαφορετική. Δηλαδή το κρίσιμο σημείο πέραν του οποίου τα κατάλοιπα αντιδρούν και αλλάζουν την ταχύτητα προσαρμογής τους είναι διαφορετικό σε κάθε κατηγορία τροφίμου. Σύμφωνα με τις εκτιμήσεις του υποδείγματος TAR οι συντελεστές σύγκλισης των καταλοίπων της μακροχρόνιας σχέσης ρ_1^{up} και ρ_2^{down}, είναι στατιστικά σημαντικοί με βάση τις κρίσιμες τιμές της στατιστικής t-Max* και έχουν το αναμενόμενο αρνητικό πρόσημο σε όλες τις κατηγορίες των τροφίμων. Σε όλες τις κατηγορίες τροφίμων η παράμετρος ρ_2^{down} είναι μεγαλύτερη σε απολυτές τιμές από την ρ_1^{up}, υποδηλώνοντας ότι η ταχύτητα προσαρμογής των αποκλίσεων όταν τα κατάλοιπα βρίσκονται κάτω από το κατώφλι είναι μεγαλύτερη από ότι όταν αυτά βρίσκονται πάνω από τον μέσο τους.

Πίνακας 7.2. Έλεγχοι TAR για συνολοκλήρωση αυτοπαλίνδρομου κατωφλιού

Προσαρμοσμένο εύρος δείγματος			2002M02-2016M06 (n=173)			2002M02-2016M06 (n=173)			2004M02-2016M06 (n=149)			2005M02-2016M06 (n=137)			2006M02-2016M06 (n=128)		
Μέθοδος εκτίμησης: OLS			Δημητριακά			Κρέατα			Γαλακτοκομικά			Φρούτα			Λαχανικά		
			$\hat{\tau} = 0,012288$			$\hat{\tau} = 0,008340$			$\hat{\tau} = 0,006592$			$\hat{\tau} = -0,013076$			$\hat{\tau} = 0,023808$		
					P-value			P-value			P-value			P-value			P-value
	Enders-Siklos, t-MAX*	ρ_1^{up}	-0,148**	(-2,496)	[0,013]	-0,166**	(-2,176)	[0,030]	-0,231**	(-2,120)	[0,035]	-0,292**	(-4,299)	[0,000]	-0,265**	(-3,101)	[0,002]
	Enders-Siklos, t-MAX*	ρ_2^{down}	-0,238**	(-3,657)	[0,000]	-0,402**	(-5,678)	[0,000]	-0,341**	(-4,638)	[0,000]	-0,351**	(-3,162)	[0,001]	-0,397**	(-3,776)	[0,000]
Έλεγχος συνολοκλήρωσης αυτοπαλίνδρομου κατωφλιού	Ho: $\rho_1^{up} = \rho_2^{down} = 0$	Στατιστική Enders-Siklos Φ*		9,805**	[0,000]		17,874**	[0,000]		13,008**	[0,000]		14,244**	[0,000]		11,942**	[0,000]
Έλεγχος συμμετρίας	Ho: $\rho_1^{up} = \rho_1^{down}$	Wald F-test		1,048**	[0,307]		5,441	[0,020]		0,693**	[0,406]		0,227**	[0,634]		1,122**	[0,291]
	Συμπεράσματα		Συνολοκλήρωση Συμμετρική προσαρμογή των καταλοίπων			Συνολοκλήρωση Ασυμμετρική προσαρμογή των καταλοίπων			Συνολοκλήρωση Συμμετρική προσαρμογή των καταλοίπων			Συνολοκλήρωση Συμμετρική προσαρμογή των καταλοίπων			Συνολοκλήρωση Συμμετρική προσαρμογή των καταλοίπων		

Οι αριθμοί στις παρενθέσεις αφορούν τις τιμές της στατιστικής t και στις αγκύλες τις τιμές P-value

Η στατιστική σημαντικότητα σε επίπεδο 5% συμβολίζεται με διπλούς αστερίσκους (**).

Ο μέγιστος αριθμός των χρονικών υστερήσεων βασίστηκε στα πληροφοριακά κριτήρια AIC (1973) και SIC (1978)

Για τον έλεγχο συνολοκληρωτικής καταφυγής (TAR cointegration) οι κρίσιμες τιμές αναφέρονται στους Enders - Siklos (2001) και Wane (2001). Η κρίσιμη τιμή της στατιστικής Φ* (με κατωφλι που εκτιμάται με συνεπή τρόπο) χωρίς χρονικές υστερήσεις είναι η τιμή 8,85 (με 100 παρατηρήσεις) και 8,65 (με 250 παρατηρήσεις) σε επίπεδο στατιστικής σημαντικότητα 5%. Η κρίσιμη τιμή της στατιστικής tMax* χωρίς χρονικές υστερήσεις είναι η τιμή -1,85 (με 100 παρατηρήσεις) και -1,84 (με 250 παρατηρήσεις) σε επίπεδο στατιστικής σημαντικότητα 5% και 2,50.

Για τον έλεγχο της υπόθεσης ύπαρξης συμμετρίας, χρησιμοποιούνται οι τιμές της F με την προυπόθεση ότι υπάρχει συνολοκλήρωση αυτοπαλίνδρομου κατωφλιού.

Η ύπαρξη συνολοκλήρωσης αυτοπαλίνδρομου κατωφλιού επιτρέπει την εκτίμηση του υποδείγματος διόρθωσης σφάλματος με αυτοπαλίνδρομη συνολοκλήρωση κατωφλιού (ECM with TAR cointegration) το οποίο στη γενική του μορφή περιγράφεται από την σχέση (5). Με αυτό τον τρόπο διερευνάται η φύση της μακροχρόνιας προσαρμογής, δηλαδή αν υπάρχει ασύμμετρη συμπεριφορά των λιανικών τιμών των τροφίμων ως προς το μακροχρόνιο επίπεδο ισορροπίας τους. Τα αποτελέσματα των εκτιμήσεων παρουσιάζονται στον Πίνακα 7.3.

Σε όλες τις κατηγορίες τροφίμων οι εκτιμήσεις των συντελεστών που μετρούν την ασύμμετρη προσαρμογή προς την μακροχρόνια ισορροπία φαίνεται να διαφέρουν μεταξύ τους. Οι συντελεστές που μετρούν την ταχύτητα προσαρμογής όταν οι τιμές της λιανικής βρίσκονται κάτω από το επίπεδο ισορροπίας (καθεστώς συμπιεσμένου περιθωρίου κερδών), είναι μεγαλύτερες σε όλες τις κατηγορίες τροφίμων συγκριτικά με τους συντελεστές που μετρούν την ταχύτητα προσαρμογής όταν οι τιμές της λιανικής βρίσκονται πάνω από το επίπεδο ισορροπίας (καθεστώς διογκωμένου περιθωρίου κερδών). Συνεπώς, όταν τα κέρδη είναι συμπιεσμένα, η διάθεση των εμπόρων της λιανικής αγοράς να αυξήσουν τις λιανικές τιμές των τροφίμων είναι εντονότερη. Αντίθετα, όταν τα κέρδη είναι διογκωμένα η διάθεση των εμπόρων να μειώσουν τις τιμές είναι μικρότερη.

Πίνακας 7.3. Ασυμμετρικό υπόδειγμα διόρθωσης σφάλματος με συνολοκλήρωση αυτοπαλίνδρομου συνεπούς κατωφλιού (ECM with TAR cointegration)

Μέθοδος εκτίμησης: OLS (HAC standard errors & covariance)		ECM with TAR cointegration									
Προσαρμοσμένο εύρος δείγματος (n= παρατηρήσεις)		2002M2-2016N06 (n=173)		2002M2-2016N06 (n=173)		2004M2-2016N06 (n=149)		2005M2-2016N06 (n=137)		2006M2-2016N06 (n=128)	
		Δημητριακά		**Κρέατα**		**Γαλακτοκομικά**		**Φρούτα**		**Λαχανικά**	
	Συντελεστές		**p-value**		**p-value**		**p-value**		**p-value**		**p-value**
Σταθερός όρος	μ_0	0,000 (0,800)	[0,424]	0,000 (1,448)	[0,149]	0,001 (0,291)	[0,771]	-0,004 (-1,182)	[0,239]	-0,001 (-0,245)	[0,806]
$\Delta(r^{i})_{t-1}$	$\beta_{1,1}$	0,266*** (7,615)	[0,000]			0,224*** (2,439)	[0,016]			0,084 (1,043)	[0,298]
$\Delta(r^{i})_{t-2}$	$\beta_{1,2}$					0,098** (1,414)	[0,159]				
$\Delta(r^{i})_{t-3}$	$\beta_{1,3}$										
$\Delta(r^{i})_{t-5}$	$\beta_{1,5}$					0,137** (1,906)	[0,058]	0,215*** (2,772)	[0,006]		
$\Delta(r^{m})_{t}$	$\beta_{2,0}$					0,328*** (2,264)	[0,025]	0,651* (1,881)	[0,062]		
$\Delta(r^{m})_{t-1}$	$\beta_{2,1}$	0,153** (3,271)	[0,001]	0,203*** (5,253)	[0,000]						
$\Delta(r^{m})_{t-2}$	$\beta_{2,2}$			0,035** (1,781)	[0,076]					0,600** (1,993)	[0,048]
$\hat{u}_{t-1}^{up}$	a_1^{up}	**-0,017 (-0,546)**	**[0,585]**	**-0,047 (-1,058)**	**[0,291]**	**-0,151*** (-2,111)**	**[0,036]**	**-0,268*** (-4,175)**	**[0,000]**	**-0,225** (-2,207)**	**[0,029]**
$\hat{u}_{t-1}^{down}$	a_2^{down}	**-0,143*** (-4,392)**	**[0,004]**	**-0,133**(-2,808)**	**[0,005]**	**-0,260*** (-3,025)**	**[0,003]**	**-0,320*** (-2,902)**	**[0,004]**	**-0,424** (-3,362)**	**[0,001]**
Wald F-statistic	$a_1^{up} = a_2^{down}$	**7,206 df(1,165)**	**[0,008]**	**5,472 df(1,164)**	**[0,026]**	**0,934*** df(1,139)**	**[0,335]**	**0,125*** df(1,130)**	**[0,723]**	**1,136** df(1,118)**	**[0,288]**
Βασικό Συμπέρασμα		Απόρριψη συμμετρίας		Απόρριψη συμμετρίας		Αποδοχή συμμετρίας		Αποδοχή συμμετρίας		Αποδοχή συμμετρίας	
Διορθωμένος βαθμός προσαρμογής R^2		0,810		0,360		0,421		0,330		0,244	
Στατιστική DW		2,16		2,000		2,010		2,075		1,976	
LM(6) , Breusch-Godfrey Serial Correlation Test		1,966 Prob.F-stat(6,159)	[0,073]	0,928 Prob.F-stat(6,158)	[0,476]	1,002 Prob.F-stat(6,133)	[0,426]	1,220 Prob.F-stat(6,124)	[0,300]	0,615 Prob.F-stat(6,112)	[0,717]

Οι αριθμοί στις παρενθέσεις αφορούν τις τιμές της στατιστικής t και στις αγκύλες τις τιμές P-value .

Η στατιστική σημαντικότητα σε επίπεδο 5% συμβολίζεται με διπλούς αστερίσκους (**)

Η στατιστική σημαντικότητα σε επίπεδο 10% συμβολίζεται με αστερίσκο (*)

Ο μέγιστος αριθμός των χρονικών υστερήσεων βασίστηκε στα πληροφοριακά κριτήρια AIC(1973) και SBC(1978).

Στις περιπτώσεις των δημητριακών και των κρεάτων οι συντελεστές που μετρούν την προσαρμογή όταν οι λιανικές τιμές των δημητριακών είναι πάνω από το επίπεδο ισορροπίας τους (οπότε υπάρχει τάση να μειωθούν) είναι στατιστικά ασήμαντοι, δηλαδή επαγωγικά μπορεί να θεωρηθεί ότι είτε δεν υπάρχει διόρθωση των διογκωμένων περιθωρίων κέρδους είτε ότι η διόρθωση γίνεται με πολύ αργή ταχύτητα. Τα παραπάνω ευρήματα μας οδηγούν στην υποστήριξη της υπόθεσης της ασυμμετρικής προσαρμογής των λιανικών τιμών. Στις περιπτώσεις των γαλακτοκομικών, των φρούτων και των λαχανικών οι συντελεστές που μετρούν τις ταχύτητες προσαρμογής είναι στατιστικά σημαντικοί και φαίνεται να διαφέρουν μεταξύ τους, εντούτοις οι επαγωγικοί έλεγχοι υποστηρίζουν την ισότητα των συντελεστών προσαρμογής.

8. ΣΥΜΠΕΡΑΣΜΑΤΑ

Η εμπειρική διερεύνηση για την ύπαρξη ασύμμετρης συμπεριφοράς των λιανικών τιμών των εξεταζόμενων τροφίμων ως προς το μακροχρόνιο επίπεδο ισορροπίας τους παρέχει ισχυρές ενδείξεις ότι στην κατηγορία των δημητριακών και των κρεάτων υπάρχει ασυμμετρία. Τα εμπειρικά ευρήματα της ασυμμετρικής προσαρμογής των λιανικών τιμών πιθανά συνδέονται με φαινόμενα ολιγοπωλιακής διάρθρωσης των κλάδων όπου οι εταιρίες έχουν την τάση να κρατούν τις τιμές σχετικά αμετάβλητες σε περιόδους πτώσης των τιμών του παραγωγού, ενώ σπεύδουν να τις αυξήσουν σε φάσεις ανόδου των τιμών του παραγωγού. Αντίθετα, στις κατηγορίες των γαλακτοκομικών, των φρούτων και των λαχανικών οι αποκλίσεις που υπάρχουν στους συντελεστές μακροχρόνιας προσαρμογής δεν είναι τέτοιες που να οδηγούν στην απόρριψη της συμμετρικής προσαρμογής.

ABSTRACT

The main purpose of the paper is to investigate the existence of asymmetry in the long-term adjustment of retail prices in relation to changes in producer prices in the main categories of food: cereals, dairy, meat, fruits and vegetables. The existence of asymmetry implies a lack of competition and distortions in the food market. The technique of the error correction model with TAR cointegration is applied. The data are monthly and covers the period from January 2000 to June 2016. The results provide evidence of an asymmetric adjustment in cereal and meat retail prices, where prices rise faster following a producer price increase, while they adjust with a slower rate after a fall in producer prices. On the contrary, for dairy, fruits and vegetables, the hypothesis of symmetric adjustment cannot be rejected.

ΑΝΑΦΟΡΕΣ

Bailey D.V and Wade Brorsen B. (1989). "Price Asymmetry In Spatial Fed Cattle Markets." *Western Journal of Agricultural Economics*, 14, 246-252.

Buckle R.A. and Carlson J.A. (2000). "Inflation and asymmetric Price Adjustment." *Review of Economics and Statistics*, 82, 157-160.

Chan K.S. (1993), "Consistency and limiting distribution of the least squares estimation of a threshold autoregressive model." *The Annals of Statistics*, 21, 520-33.

Dickey D.A. and Fuller W.A. (1979). "Distribution of the estimators for autoregressive time series with a unit root." *Journal of the American Statistical Association*, 74, 427-431.

Dickey D.A and Fuller W.A. (1981). "Likelihood ratio statistics for autoregressive time series with a unit root." *Econometrica*, 49, 1057-1072.

Enders W. and Granger C.W.J. (1998). "Unit root tests and symmetric adjustment with an example using the term structure of interest rates." *Journal of Business and Economic Statistics*, 16, 304-11.

Elliott G., Rothenberg T. and Stock J. (1996). "Efficient tests for an autoregressive unit root." *Econometrica*, 64, 813-836.

Enders W. and Siklos P. (2001). "Cointegration and threshold adjustment." *Journal of Business and Economic Statistics*, 19, 166-76.

Engle R. and Granger C. (1987). "Cointegration and error-correction: representation, estimation and testing." *Econometrica*, 55, 251-76.

Επιτροπή Ανταγωνισμού (2011). "Κλαδική Έρευνα στα Νωπά Οπωροκηπευτικά. Συνοπτική Παρουσίαση Βασικών Διαπιστώσεων και Πορισμάτων", άρθρο 40 ν.3959/2011.

Herfindahl O.C. (1950). *Concertation in the U.S. Steel Industry*, Columbia University, Doctoral Dissertation.

Hirschman O. A. (1964). "The Paternity of an Index." *The American Economic Review*, 54, 761-762.

Kinnucan H.W. and Forker O.D. (1987). “Asymmetry in Farm-Retail Price Transmission for Major Dairy Products.” *American Journal of Agricultural Economics*, 69, 285-292.

Miller D.J. and Hayenga M.L. (2001). “Price Cycles and Asymmetric Transmission in the U.S. Pork Market.” *American Journal of Agricultural Economics*, 83, 551-562.

Meyer J. and Cramon-Taubadel S. (2004), “Asymmetric Price Transmission: A Survey.” *Journal of Agricultural Economics*, 55, 581-611.

Phillips P.C. and Hansen B. (1990). “Statistical Inference in Instrumental Variables Regression with I(1) Processes”, *The Review of Economic Studies*, 57, 99-125.

Phillips P.C. and Perron P. (1988). “Testing of a Unit Root in Time Series Regression.” *Biometrika*, 75, 335-346.

Perron P. (1990). “Tests of joint hypotheses in time series regression with a unit root.” In G.F. Rhodes and T.B. Fomby (eds), *Advances in Econometrics: Co-Integration, Spurious Regression, and Unit Roots*,. Greenwhich CT: JAO Press, Vol.8, 10-20.

Peltzman S. (2000). “Price Rise Faster than They Fall.” *Journal of Political Economy*, 108, 466-502.

Ρεζίτη Ι. (2004). “Διαμόρφωση τιμών βασικών αγροτικών προϊόντων.” ΚΕΠΕ *Οικονομικές εξελίξεις*, 6, 58-64.

Reziti I. and Panagopoulos Y. (2006). “The Price Transmission Mechanism in the Greek Agri-food Sector: An Empirical Approach.” Athens: *Central of Planning and Economic Research*.

Reziti I. and Panagopoulos, Y. (2008). “Asymmetric price transmission in the greek agri 3food sector: some tests.” *Agribusiness*, 24, 16-30.

Rezitis A.N. and Reziti, I. (2011). "Threshold Cointegration in the Greek Milk Market." *Journal of International Food, Agribusiness Marketing,* 23, 231-246.

Reziti I. (2014). “Price transmission analysis in the Greek milk market.” SPOUDAI *Journal of Economics and Business*, 64, 75-86.

Tsakistara D., Sdrolias L., Polyzos S., and Kakkos N. (2008). 'Analysis of Greek milk market. Quo vadis?', *MIBES-Oral*, 457-471.

Ward R.W. (1982). "Asymmetry in Retail, Wholesale, and Shipping Point Pricing for Fresh Vegetables." *American Journal of Agricultural Economics*, 64, 205-212.

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 31ου Πανελληνίου Συνεδρίου Στατιστικής (2018), σελ.179-190

ΜΕΛΕΤΗ ΜΗ ΠΑΡΑΜΕΤΡΙΚΩΝ ΔΙΑΓΡΑΜΜΑΤΩΝ ΕΛΕΓΧΟΥ ΜΕ ΧΡΗΣΗ ΜΕΘΟΔΩΝ ΠΡΟΣΟΜΟΙΩΣΗΣ

Ν. Ι. Παναγιώτου, Ι.Σ. Τριανταφύλλου
Τμήμα Πληροφορικής με εφαρμογές στη Βιοϊατρική, Σχολή Θετικών Επιστημών, Πανεπιστήμιο Θεσσαλίας
nikopanag@gmail.com, itriantafyllou@uth.gr

ΠΕΡΙΛΗΨΗ

Στα συνήθη διαγράμματα ελέγχου, βασική προϋπόθεση για την εφαρμογή τους είναι η κανονικότητα του υπό μελέτη χαρακτηριστικού. Αντίθετα, τα μη-παραμετρικά διαγράμματα ελέγχου δεν απαιτούν την γνώση της κατανομής που διέπει την υπό μελέτη διεργασία. Στην παρούσα εργασία, μελετώνται μη-παραμετρικά διαγράμματα ελέγχου στα οποία απεικονίζεται η στατιστική συνάρτηση *Wilcoxon* και τα οποία χρησιμοποιούν επιπροσθέτως γνωστούς κανόνες ροών. Στο πλαίσιο της παρούσας μελέτης, προτείνεται και περιγράφεται αλγόριθμος προσομοίωσης για τον υπολογισμό των βασικών χαρακτηριστικών απόδοσης των προαναφερθέντων μη-παραμετρικών διαγραμμάτων ελέγχου, τόσο σε εντός ελέγχου όσο και σε εκτός ελέγχου καταστάσεις της διεργασίας. Με εφαρμογή του προτεινόμενου αλγορίθμου και χρήση κατάλληλου λογισμικού, παρουσιάζονται αριθμητικά αποτελέσματα για την αποτύπωση και τη συγκριτική αξιολόγηση της ικανότητας των υπό μελέτη διαγραμμάτων ελέγχου να εντοπίζουν μετατοπίσεις στην κατανομή που διέπει την παραγωγική διεργασία.

Λέξεις Κλειδιά: Μη-παραμετρικά διαγράμματα ελέγχου, πιθανότητα λανθασμένου συναγερμού (*FAR*), στατιστική συνάρτηση *Wilcoxon*, κανόνες ροών, προσομοίωση *Monte Carlo*.

1. ΕΙΣΑΓΩΓΗ

Σε μια παραγωγική διαδικασία ενδιαφερόμαστε να μελετήσουμε τη συμπεριφορά ενός (μετρήσιμου) χαρακτηριστικού X των προϊόντων που παράγονται. Για το σκοπό αυτό παίρνουμε τυχαία δείγματα του προϊόντος σε διαφορετικές χρονικές στιγμές και κάνοντας μετρήσεις, υπολογίζουμε την τιμή μιας κατάλληλα επιλεγμένης στατιστικής συνάρτησης. Ένα τυπικό διάγραμμα

ελέγχου είναι μια γραφική παράσταση της στατιστικής συνάρτησης, τις τιμές της οποίας έχουμε υπολογίσει με τη βοήθεια ενός τυχαίου δείγματος, έναντι του αύξοντα αριθμού του δείγματος αυτού. Σκοπός των διαγραμμάτων ελέγχου είναι να δείξουν πότε μια διαδικασία είναι εντός ή εκτός ελέγχου με τη μεγαλύτερη πιθανότητα ορθής εκτίμησης και ταυτόχρονα να εντοπίσουν το δυνατόν συντομότερα τυχόν μετατόπιση της κατανομής της διεργασίας.

Ο Shewhart (1927) παρουσίασε το παραμετρικό διάγραμμα ελέγχου $\overline{X}$ για την παρακολούθηση της μέσης τιμής της κατανομής ενός χαρακτηριστικού X των προϊόντων που παράγονται από μια παραγωγική διεργασία (υπό την υπόθεση της κανονικότητας). Για την κατασκευή του παραπάνω διαγράμματος, πιο συχνά χρησιμοποιείται το μοντέλο ορίων L σίγμα που τοποθετεί τα όρια ελέγχου σε απόσταση L μονάδων *τυπικής απόκλισης* $\sigma_{\overline{X}}$ *από* την κεντρική γραμμή $\mu_{\overline{X}}$. Στα παραμετρικά διαγράμματα ελέγχου, η δημοφιλέστερη επιλογή για την απόσταση των ορίων από την κεντρική γραμμή είναι τρεις μονάδες τυπικής απόκλισης και το συγκεκριμένο διάγραμμα είναι γνωστό ως μοντέλο τριών σίγμα (*three sigma control chart*).

Έκτοτε έχουν παρουσιαστεί ποικίλες παραλλαγές του διαγράμματος ελέγχου *Shewhart*, οι περισσότερες εκ των οποίων αφορούν παραμετρικά διαγράμματα, δηλαδή εφαρμόζονται υπό την υπόθεση ότι οι μετρήσεις που λαμβάνονται ακολουθούν κάποια γνωστή κατανομή, συνήθως την κανονική. Ωστόσο αυτή η υπόθεση συχνά δεν ικανοποιείται στην πράξη και συνεπώς τα προκύπτοντα διαγράμματα ελέγχου ενδέχεται να μην είναι αξιόπιστα. Για να ξεπεραστεί αυτό το ζήτημα στη βιβλιογραφία έχουν προταθεί μη-παραμετρικά ή ελεύθερα κατανομής διαγράμματα ελέγχου (*non-parametric or distribution-free control charts*), τα οποία αφενός διατηρούν την παραδοσιακή δομή ενός τυπικού διαγράμματος ελέγχου και αφετέρου τα στατιστικά χαρακτηριστικά του διαγράμματος δεν εξαρτώνται από το είδος της (συνεχούς) κατανομής την οποία ακολουθούν τα δεδομένα που συλλέγονται.

Για παράδειγμα, ο Bakir (2006) δημιούργησε ένα διάγραμμα ελέγχου τύπου *Shewhart* με τη χρήση της προσημικής-βαθμολογικής στατιστικής συνάρτησης ελέγχου (*signed–rank statistic*), ενώ οι Chakraborti & Eryilmaz (2007) εναλλακτικά προτείνουν για την παρακολούθηση της διάμεσου, δύο νέα μη-παραμετρικά διαγράμματα ελέγχου με βάση την ίδια στατιστική συνάρτηση. Επίσης, οι Chakraborti *et al*. (2004) μελέτησαν μια οικογένεια μη-παραμετρικών διαγραμμάτων ελέγχου συγκρίνοντας την *j*-οστή διατεταγμένη παρατήρηση από ένα τυχαίο δείγμα ελέγχου (*test sample*), με όρια ελέγχου που υπολογίστηκαν χρησιμοποιώντας κατάλληλα επιλεγμένες διατεταγμένες παρατηρήσεις από ένα δείγμα αναφοράς (*reference sample*). Οι Balakrishnan *et al*. (2010) για να βελτιώσουν το διάγραμμα των Chakraborti *et al*. (2004), πρότειναν ένα νέο μη-παραμετρικό διάγραμμα ελέγχου το οποίο λαμβάνει υπόψη του όχι μόνο τη θέση της *j*-οστής διατεταγμένης παρατήρησης (όπως η διάμεσος) του εξεταζόμενου δείγματος ελέγχου αλλά και το πλήθος των παρατηρήσεών του που βρίσκονται μεταξύ των ορίων ελέγχου. Επιπλέον, οι ίδιοι συγγραφείς το 2009, εφάρμοσαν την αθροιστική βαθμολογική στατιστική

συνάρτηση του *Wilcoxon* για να ανιχνεύσουν πιθανές μετατοπίσεις μιας παρακολουθούμενης διεργασίας. Ορισμένα πρόσφατα ερευνητικά αποτελέσματα σε αυτό το θέμα μπορούν επίσης να βρεθούν στις δημοσιεύσεις των Mukherjee *et al*. (2013), Malela-Majika *et al*. (2016a, 2016b) ή Triantafyllou (2017, 2018a, 2018b). Για μια εκτεταμένη επισκόπηση των ελεύθερων κατανομής διαγραμμάτων ελέγχου για συνεχείς μεταβλητές, οι ενδιαφερόμενοι αναγνώστες μπορούν να ανατρέξουν στις εργασίες των Chakraborti & Graham (2007) και του Chakraborti (2011).

Τα διαγράμματα ελέγχου τύπου *Shewhart* δεν είναι ιδιαίτερα ευαίσθητα στην ανίχνευση μικρών μετατοπίσεων του μέσου μιας διεργασίας. Η ευαισθητοποίησή τους συνήθως βελτιώνεται με τη χρήση κανόνων ροών, που σχετίζονται με την εμφάνιση ειδικών σχηματισμών (*patterns*) των σημείων του διαγράμματος ελέγχου. Η εμφάνιση ενός τέτοιου ειδικού σχηματισμού σημείων μας δίνει σήμα για εκτός ελέγχου διεργασία. Ο Klein (2000) πρότεινε δύο διαγράμματα ελέγχου ως εναλλακτικά του $\overline{X}$ διαγράμματος ελέγχου *Shewhart*, τα οποία βασίζονται στη χρήση των κανόνων ροών δύο από τα δύο (2/2) και δύο από τα τρία (2/3), χρησιμοποιώντας συμμετρικά όρια ελέγχου και εφαρμόζοντας την μεθοδολογία εμφύτευσης σε Μαρκοβιανή αλυσίδα. Τα δύο προαναφερθέντα διαγράμματα υλοποιούνται εύκολα και έχουν καλύτερη απόδοση ως προς το μέσο μήκος ροής (*ARL*) από το τυπικό διάγραμμα *Shewhart* για μετατοπίσεις του μέσου της διεργασίας έως και 2.6 τυπικές αποκλίσεις. Ο Khoo (2004) επέκτεινε το έργο του Klein (2000) δημοσιεύοντας μια μελέτη προσομοίωσης της απόδοσης του μέσου μήκους ροής (*ARL*) των διαγραμμάτων ελέγχου χρησιμοποιώντας τους κανόνες ροής: 2/2, 2/3, 2/4, 3/3 και 3/4. Επιπλέον, πρότεινε δύο νέα μη-παραμετρικά διαγράμματα τύπου *Shewhart* για την παρακολούθηση της άγνωστης παραμέτρου θέσης ενός συνεχούς πληθυσμού. Στα διαγράμματα αυτά τα όρια ελέγχου προκύπτουν από δύο κατάλληλα επιλεγμένες διατεταγμένες παρατηρήσεις από το δείγμα αναφοράς και γίνεται χρήση κανόνων ροών.

Στην παρούσα εργασία εφαρμόζουμε τον κανόνα 2-από-3 και 2-από-4 στο μη-παραμετρικό διάγραμμα ελέγχου φάσης II που προτάθηκε από τους Balakrishnan *et al*. (2009) οι οποίοι χρησιμοποίησαν την αθροιστική βαθμολογική συνάρτηση τύπου *Wilcoxon (Wilcoxon rank-sum statistic)* για να εξετάσουν αν μια διεργασία παραμένει εντός ελέγχου ή έχει μετατοπιστεί σε εκτός ελέγχου κατάσταση. Στην Ενότητα 2 περιγράφεται σύντομα το παραπάνω μη-παραμετρικό διάγραμμα, στη συνέχεια εφαρμόζονται οι κανόνες ροής 2/3, 2/4 στο προαναφερθέν διάγραμμα και τέλος παρουσιάζεται λεπτομερώς ο προτεινόμενος αλγόριθμος προσομοίωσης. Στην Ενότητα 3, υπολογίζονται οι τιμές της πιθανότητας λανθασμένου συναγερμού (*FAR*) για διαφορετικούς σχεδιασμούς. Η απόδοση του προτεινόμενου διαγράμματος ελέγχου διερευνάται χρησιμοποιώντας δύο διαφορετικές υποκείμενες κατανομές.

2. ΤΟ ΝΕΟ ΜΗ ΠΑΡΑΜΕΤΡΙΚΟ ΔΙΑΓΡΑΜΜΑ ΕΛΕΓΧΟΥ

2.1 Το μη-παραμετρικό διάγραμμα των Balakrishnan *et al.* (2009)

Έστω $X_1, X_2, \ldots, X_m$ είναι ένα τυχαίο δείγμα μεγέθους m που προέρχεται από την εντός ελέγχου αθροιστική κατανομή $F_X = F$ και ως όρια ελέγχου του διαγράμματος

χρησιμοποιούνται δύο συγκεκριμένες διατεταγμένες παρατηρήσεις $X_{\alpha:m}$, $X_{b:m}$, δηλαδή

$$LCL=X_{\alpha:m}, UCL=X_{b:m} \text{ με } 1 \le \alpha \le b \le m.$$

Οι παράμετροι *α, b* είναι παράμετροι σχεδιασμού του διαγράμματος και είναι κατάλληλα επιλεγμένοι έτσι ώστε να επιτυγχάνεται ένα συγκεκριμένο επίπεδο λανθασμένων συναγερμών (*FAR*) ή μια προκαθορισμένη τιμή του εντός ελέγχου μέσου μήκους ροής (ARL_{in}).

Ας υποθέσουμε ότι εν συνεχεία συλλέγονται ανεξάρτητα τυχαία δείγματα από τη διεργασία και θέλουμε να εξετάσουμε αν παραμένει εντός ελέγχου ή αν έχει μετατοπιστεί σε εκτός ελέγχου κατάσταση. Αν $Y_1, Y_2, \ldots, Y_n$ είναι ένα εξεταζόμενο δείγμα ελέγχου μεγέθους *n* που προέρχεται από την κατανομή $F_Y=G$, ο στόχος μας είναι να ανιχνεύσουμε μια πιθανή μετατόπιση της εντός ελέγχου κατανομής από *F* σε *G*, δηλαδή να ελέγξουμε τη μηδενική υπόθεση H_0*: F=G* έναντι της μονόπλευρης εναλλακτικής H_1*: F<G*. Με άλλα λόγια, η μηδενική υπόθεση εκφράζει την ισότητα των δύο κατανομών, ενώ η εναλλακτική υπόθεση εκφράζει ότι η μία κατανομή είναι στοχαστικά μεγαλύτερη από την άλλη.

Οι Balakrishnan *et al*. (2009) για να αποφασίσουν αν η διεργασία παραμένει εντός ελέγχου ή όχι, χρησιμοποιούν το άθροισμα των βαθμών (στο από κοινού δείγμα των *X* και *Y* παρατηρήσεων) των *Y* παρατηρήσεων που βρίσκονται ανάμεσα στα όρια ελέγχου. Η συγκεκριμένη επιλογή αντιστοιχεί στη γνωστή αθροιστική βαθμολογική συνάρτηση τύπου *Wilcoxon*. Πιο συγκεκριμένα, έστω ότι το M_i, $i = 1,2,\ldots,m$ εκφράζει το πλήθος των *Y* παρατηρήσεων του εξεταζόμενου δείγματος ελέγχου που βρίσκονται μεταξύ της (*i*-1)-οστής και *i*-οστής διατεταγμένης παρατήρησης του *X* δείγματος αναφοράς (ισχύει η σύμβαση: $X_{(0)} = -\infty$). Με άλλα λόγια, τα M_i είναι τα μήκη των ροών των *Y* παρατηρήσεων μεταξύ διαδοχικών *X* παρατηρήσεων. Αθροίζοντας τους βαθμούς (στο από κοινού δείγμα των *X* και *Y*) των *Y* παρατηρήσεων που βρίσκονται μεταξύ των ορίων ελέγχου, υπολογίζεται η στατιστική βαθμολογική συνάρτηση τύπου *Wilcoxon* :

$$W = \sum_{i=a+1}^{b} W_i$$

όπου το W_i εκφράζει το άθροισμα των βαθμών των *Y* παρατηρήσεων που βρίσκονται μεταξύ των $X_{(i-1)}$ και $X_{(i)}$. Στη συνέχεια, προσδιορίζεται η αθροιστική βαθμολογική συνάρτηση τύπου *Wilcoxon* με τη βοήθεια των M_i , οπότε προκύπτει

$$W_i = \sum_{j=1}^{M_i}((i-1) + \sum_{r=1}^{i-1} M_r + j) = M_i((i-1) + \sum_{r=1}^{i-1} M_r) + \frac{M_i(M_i+1)}{2}$$

και υλοποιώντας αλγεβρικές πράξεις, αναδεικνύεται ο ακόλυθος τύπος

$$W = \frac{1}{2}\left(\sum_{i=a+1}^{b} M_i\right)^2 + \sum_{i=a+1}^{b} iM_i + (M_0 + a - \frac{3}{2})\sum_{i=a+1}^{b} M_i ,$$

όπου $M_0 = \sum_{i=1}^{a} M_i$ εκφράζει το πλήθος των παρατηρήσεων του *Y* δείγματος πριν το $LCL=X_{\alpha:m}$.

2.2 Το νέο μη-παραμετρικό διάγραμμα

Η κατηγορία των μη-παραμετρικών διαγραμμάτων ελέγχου που εισάγεται στην παρούσα εργασία κάνει χρήση των κανόνων ροής 2-από-3 και 2-από-4. Ειδικότερα, η διεργασία κηρύσσεται εκτός ελέγχου εάν τουλάχιστον δύο από τα τρία ή δύο από τα τέσσερα διαδοχικά δείγματα ελέγχου αντιστοίχως δεν πληρούν τουλάχιστον μία από τις ακόλουθες δύο συνθήκες:

***Συνθήκη* 1.** Η στατιστική βαθμολογική συνάρτηση τύπου *Wilcoxon W* δεν υπερβαίνει ένα προκαθορισμένο όριο, δηλαδή $W \leq c_1$, όπου c_1 είναι μια θετική παράμετρος.

***Συνθήκη* 2.** Ο αριθμός των παρατηρήσεων του δείγματος ελέγχου *Y* πριν από την παρατήρηση $X_{\alpha:m}$ είναι μικρότερος ή ίσος του c_2, δηλαδή $M_0 = \sum_{i=1}^{a} M_i \leq c_2$, όπου c_2 είναι μια θετική παράμετρος.

Κάθε μια από τις παραπάνω συνθήκες απαιτεί την παράλληλη κατασκευή δύο ξεχωριστών μονόπλευρων διαγραμμάτων ελέγχου, από τα οποία δημιουργείται το προτεινόμενο μη-παραμετρικό διάγραμμα ελέγχου. Οι σταθερές c_1, c_2 είναι παράμετροι σχεδιασμού, έχοντας το ρόλο των αντίστοιχων άνω ορίων ελέγχου των προαναφερθέντων διαγραμμάτων καθώς και τη χρησιμότητα, κατά την εφαρμογή του διαγράμματος, ο χρήστης να μπορεί να επιλέξει κατάλληλες τιμές ώστε να επιτύχει το επιθυμητό επίπεδο εντός ή εκτός ελέγχου απόδοσης του διαγράμματος. Είναι προφανές ότι, αν τουλάχιστον μία από τις συνθήκες παραβιάζεται, το αντίστοιχο δείγμα χαρακτηρίζεται ως εκτός ελέγχου.

Αξίζει να σημειωθεί ότι οι κανόνες 2-από-3 και 2-από-4 διαφέρουν ως προς την «αυστηρότητα» που παρουσιάζουν ως προς την απόφαση για το πότε η υπό μελέτη διεργασία είναι εκτός ελέγχου. Πιο συγκεκριμένα, όταν εφαρμόζεται ο κανόνας 2-από-3, η διεργασία χαρακτηρίζεται εκτός ελέγχου αν τουλάχιστον 2 από 3 συνεχόμενα δείγματα βρεθούν εκτός των ορίων ελέγχου του προτεινόμενου διαγράμματος. Αντίθετα, όταν εφαρμόζεται ο κανόνας 2-από-4, η διεργασία χαρακτηρίζεται εκτός ελέγχου αν τουλάχιστον 2-από-4 συνεχόμενα δείγματα βρεθούν εκτός των ορίων ελέγχου του προτεινόμενου διαγράμματος. Συνεπώς, είναι άμεσο ότι ο κανόνας 2-από-3 είναι «αυστηρότερος» ως προς την αποδοχή μίας παραγωγικής διεργασίας ως εντός ελέγχου έναντι του κανόνα 2-από-4.

Η απόδοση ενός διαγράμματος ελέγχου αξιολογείται από την κατανομή του μήκους ροής ή / και τη δυνατότητα του να παράγει (να μην παράγει) ένα σήμα συναγερμού όταν η διαδικασία είναι εκτός ελέγχου (εντός ελέγχου, αντίστοιχα). Ο όρος "μήκος ροής" αναφέρεται στην τυχαία μεταβλητή του χρόνου αναμονής, η οποία αντιπροσωπεύει τον αριθμό των δειγμάτων έως ότου έχουμε το πρώτο σήμα εκτός ελέγχου από το γράφημα. Ένα άλλο μέτρο απόδοσης ενός διαγράμματος είναι η πιθανότητα λανθασμένου συναγερμού (*FAR*), η οποία εκφράζει την πιθανότητα να παράγει (λανθασμένα) το διάγραμμα σήμα ότι η διεργασία μετατοπίστηκε σε εκτός ελέγχου κατάσταση, ενώ στην πραγματικότητα παραμένει εντός ελέγχου. Στο

πλαίσιο της παρούσας εργασίας, για τη μέτρηση της απόδοσης του προτεινόμενου διαγράμματος ελέγχου θα υπολογισθεί η πιθανότητα λανθασμένου συναγερμού.

2.3 Ο αλγόριθμος προσομοίωσης

Για τη μέτρηση της απόδοσης του προτεινόμενου διαγράμματος, εφαρμόζουμε κατάλληλο υπολογιστικό πρόγραμμα προσομοίωσης *Monte Carlo*, προκειμένου να προσδιοριστεί αριθμητικά η σταθερότητα και η αποτελεσματικότητά του στην ανίχνευση πιθανών μετατοπίσεων της υποκείμενης κατανομής. Ο αλγόριθμος προσομοίωσης βασίζεται στην υπόθεση ότι κάθε φορά που η διεργασία βρίσκεται σε εντός ελέγχου κατάσταση, τόσο τα δείγματα αναφοράς όσο και τα δείγματα ελέγχου προέρχονται από την ίδια κατανομή (ή εναλλακτικά λέμε ότι δεν υπάρχει αλλαγή στην κατανομή της διεργασίας, δηλαδή $\delta = 0$). Από την άλλη πλευρά, όταν η διεργασία μετατοπιστεί σε κατάσταση εκτός ελέγχου, τα δείγματα ελέγχου αντλούνται από μια κατανομή που έχει την ίδια μορφή με την αρχική κατανομή, αλλά με μετατόπιση της παραμέτρου θέσης κατά δ μονάδες. Μια σύντομη περιγραφή του αλγορίθμου για τον προσδιορισμό της απόδοσης του προτεινόμενου μη-παραμετρικού διαγράμματος ελέγχου δίνεται ακολούθως.

Βήμα 1. Ανάλογα με την κατάσταση που θα προσομοιωθεί, δηλαδή διεργασία εντός ή εκτός ελέγχου, καθορίζεται η κατανομή και οι παράμετροί της για τα δείγματα αναφοράς και ελέγχου όπως έχει περιγραφεί παραπάνω. Προσδιορίζεται το μέγεθος του δείγματος αναφοράς και δείγματος ελέγχου (m, n αντίστοιχα), καθώς και το πλήθος προσομοιώσεων k.

Βήμα 2. Επιλέγονται κατάλληλα οι παράμετροι σχεδιασμού α , b, c_1, c_2 .

Βήμα 3. Παράγεται τυχαίο δείγμα αναφοράς Φάσης Ι από καθορισμένη κατανομή (βλ. Βήμα 1) και προσδιορίζονται τα όρια ελέγχου $LCL=X_{\alpha:m}$, $UCL=X_{b:m}$ με $1 \le \alpha \le b \le m$.

Βήμα 4. Παράγεται τυχαίο δείγμα ελέγχου Φάσης ΙΙ από την ίδια κατανομή με την Φάση Ι (για διεργασία που βρίσκεται σε εντός ελέγχου κατάσταση). Υπολογίζεται το πλήθος M_0 των παρατηρήσεων του δείγματος ελέγχου που είναι μικρότερα του $LCL=X_{\alpha:m}$ και το πλήθος M_i, $i = \alpha+1, \alpha+2,\ldots,b$ των παρατηρήσεων του δείγματος ελέγχου που βρίσκονται ανάμεσα σε κάθε υποδιάστημα που σχηματίζεται από δύο διαδοχικές διατεταγμένες παρατηρήσεις του δείγματος αναφοράς, του διαστήματος ($X_{\alpha:m}$, $X_{b:m}$). Στη συνέχεια υπολογίζεται η βαθμολογική συνάρτηση W και καταγράφεται σε πίνακα T η τιμή 1 ή 0, αντίστοιχα με το αν το δείγμα ικανοποιεί (εντός ελέγχου) ή όχι τις συνθήκες 1 ή 2.

Βήμα 5. Επανάληψη του Βήματος 4 για k φορές.

Βήμα 6. Για κάθε μία εκ των συνεχόμενων (χωρίς διακοπή) τριάδων $T_{(i)}$, $T_{(i+1)}$, $T_{(i+2)}$ με $i = 1, 2,\ldots, k$-2 ή τετράδων $T_{(i)}$, $T_{(i+1)}$, $T_{(i+2)}$, $T_{(i+3)}$, με $i = 1, 2,\ldots, k$-3, αποτελούμενων από στοιχεία του πίνακα T ελέγχουμε αν ικανοποιείται ή όχι ο κανόνας 2-από-τα-3 ή 2-από-τα-4 αντιστοίχως. Πιο συγκεκριμένα, ελέγχεται για κάθε τριάδα (ή τετράδα) αν υπάρχουν τουλάχιστον δύο δείγματα που δεν ικανοποιούν τις συνθήκες 1 ή 2. Βάσει του προαναφερθέντος ελέγχου σε κάθε τριάδα (ή τετράδα), καταγράφεται ο δείκτης RL_i για κάθε τιμή $i = 1, 2,\ldots, k$-2 (ή $i = 1, 2,\ldots, k$-3). Ο δείκτης RL_i λαμβάνει την τιμή 0, αν η ελεγχόμενη τριάδα (ή τετράδα) περιλαμβάνει

τουλάχιστον δύο δείγματα εκτός ελέγχου, δηλαδή περιλαμβάνει τουλάχιστον δύο τιμές ίσες με 0 από τον πίνακα *T*. Πρόσθετα, ο δείκτης RL_i λαμβάνει την τιμή 1, αν η ελεγχόμενη τριάδα (ή τετράδα) περιλαμβάνει λιγότερα από δύο δείγματα εκτός ελέγχου, δηλαδή περιλαμβάνει λιγότερες από δύο τιμές ίσες με 0 από τον πίνακα *T*.

Βήμα 7. Υπολογίζεται η πιθανότητα λανθασμένου συναγερμού

$$FAR = 1 - (\Sigma RL_i) / (k\text{-}2) \quad \text{ή} \quad FAR = 1 - (\Sigma RL_i) / (k\text{-}3)$$

για τον κανόνα 2-από-3 ή 2-από-4 αντίστοιχα.

Για την υλοποίηση του αλγόριθμου χρησιμοποιήθηκε το υπολογιστικό πακέτο MATLAB. Αρχικά επιβεβαιώθηκε αριθμητικά η ορθότητα του αλγορίθμου προσομοίωσης, συγκρίνοντας τα αποτελέσματά για την πιθανότητα λανθασμένου συναγερμού (*FAR*) που προκύπτουν με την εφαρμογή του, για το διάγραμμα των Balakrishnan *et al*. (2009) με τα αντίστοιχα θεωρητικά αποτελέσματα που υπάρχουν στην εργασία τους.

3. ΑΠΟΔΟΣΗ ΤΟΥ ΝΕΟΥ ΔΙΑΓΡΑΜΜΑΤΟΣ ΕΛΕΓΧΟΥ

3.1 Εντός ελέγχου επίδοση του νέου διαγράμματος

Η απόδοση ενός διαγράμματος ελέγχου κρίνεται συνήθως με τη βοήθεια της κατανομής του μήκους ροής και την ικανότητα ανίχνευσης μιας πιθανής μετατόπισης της διεργασίας. Υλοποιήσαμε αριθμητικούς υπολογισμούς για τα δυο προτεινόμενα διαγράμματα που αποδεικνύουν την αποτελεσματικότητα και την αντοχή τους, τόσο σε εντός όσο και σε εκτός ελέγχου καταστάσεις. Η χρήση μεγάλου πλήθους παραμέτρων σχεδιασμού για τα δύο διαγράμματα ελέγχου μας προσφέρει την ευελιξία να σταθεροποιήσουμε ορισμένες από αυτές και στη συνέχεια να αναζητήσουμε τη βέλτιστη επιλογή για τις υπόλοιπες.

Στον επόμενο Πίνακα 1, έχουμε επιλέξει να σταθεροποιήσουμε το μέγεθος m του δείγματος αναφοράς και το μέγεθος n του δείγματος ελέγχου και παρουσιάζουμε διάφορες επιλογές των άλλων τεσσάρων παραμέτρων σχεδιασμού έτσι ώστε τα τρία υπό μελέτη διαγράμματα (το ήδη υπάρχον διάγραμμα των Balakrishnan *et al*.(2009) που συμβολίζεται ως *Διάγραμμα B-T-K* και τα δύο νέα διαγράμματα που εφαρμόζουν τους κανόνες ροής 2-από-τα-3 και 2-από-τα-4 που συμβολίζονται *Νέο Διάγραμμα B-T-K-*$RR_{2\text{-}of\text{-}3}$ και *Νέο Διάγραμμα B-T-K-*$RR_{2\text{-}of\text{-}4}$ αντίστοιχα) να επιτυγχάνουν μια προκαθορισμένη πιθανότητα λανθασμένου συναγερμού (*FAR*) περίπου ίση με 5%. Τα αποτελέσματα που συνοψίζονται στον Πίνακα 1 παρήχθησαν με εφαρμογή του νέου αλγορίθμου προσομοίωσης που παρουσιάστηκε νωρίτερα.

Για παράδειγμα, αν έχουμε συλλέξει δείγμα αναφοράς μεγέθους m=100 και επιθυμούμε να εξετάσουμε τυχαία δείγματα ελέγχου μεγέθους n=25 που εξάγονται από τη διεργασία, τότε για να επιτύχουμε πιθανότητα λανθασμένου συναγερμού ίση με 5% χρησιμοποιούμε:

α) για το διάγραμμα των Balakrishnan *et al*.(2009) , την 15η και 18η παρατήρηση του δείγματος αναφοράς με c_1 = 105 και c_2 = 8 (ακριβές *FAR* = 0.0498)

β) για το προτεινόμενο διάγραμμα ***B-T-K-***$\boldsymbol{RR_{2\text{-}of\text{-}3}}$ που αξιοποιεί την αθροιστική στατιστική συνάρτηση *Wilcoxon* και ταυτόχρονα ενσωματώνει τον κανόνα 2-από-3,

την 14η και 18η παρατήρηση του δείγματος αναφοράς με $c_1 = 83$ και $c_2 = 8$ (ακριβές *FAR* = 0.0499)

γ) για το προτεινόμενο διάγραμμα ***B-T-K-RR*$_{2\text{-}of\text{-}4}$** που αξιοποιεί την αθροιστική στατιστική συνάρτηση *Wilcoxon* και ταυτόχρονα ενσωματώνει τον κανόνα 2-από-4, την 10η και 15η παρατήρηση του δείγματος αναφοράς με $c_1 = 80$ και $c_2 = 9$ (ακριβές *FAR* = 0.0495).

Πίνακας 1. *Διαγράμματα ελέγχου με πιθανότητα FAR≈0.05*

Διάγραμμα B-T-K						
m	***n***	***a***	***b***	c_1	c_2	***FAR***
40	5	13	17	90	3	0.0499
	11	7	12	100	4	0.0506
	25	5	8	85	8	0.0496
60	5	20	23	150	3	0.0492
	11	11	15	58	6	0.0486
	25	10	13	130	8	0.0481
100	5	15	19	50	2	0.0464
	11	15	19	68	4	0.0499
	25	15	18	105	8	0.0498
***Νέο διάγραμμα B-T-K-RR*$_{2\text{-}of\text{-}3}$**						
m	***n***	***a***	***b***	c_1	c_2	***FAR***
40	5	17	22	115	3	0.0487
	11	15	19	100	8	0.0497
	25	14	17	185	14	0.0492
60	5	24	28	100	3	0.0506
	11	10	15	50	7	0.0508
	25	9	12	76	8	0.0509
100	5	10	17	40	1	0.0505
	11	13	18	40	4	0.0498
	25	14	18	83	8	0.0499
***Νέο διάγραμμα B-T-K-RR*$_{2\text{-}of\text{-}4}$**						
m	***n***	***a***	***b***	c_1	c_2	***FAR***
40	5	15	19	90	3	0.0502
	11	11	16	95	6	0.0493
	25	6	10	129	8	0.0505
60	5	22	26	94	3	0.0506
	11	10	15	55	7	0.0496
	25	8	12	96	8	0.0508
100	5	10	18	60	1	0.0499
	11	12	16	49	4	0.0496
	25	10	15	80	9	0.0495

3.2 Εκτός ελέγχου επίδοση του νέου διαγράμματος

Για να μελετήσουμε την απόδοση των προτεινόμενων διαγραμμάτων, τα δοκιμάζουμε ως προς την ικανότητα να ανιχνεύουν μικρές μετατοπίσεις του μέσου της διεργασίας. Για τα δείγματα αναφοράς χρησιμοποιούμε δύο διαφορετικές κατανομές: (α) την τυπική κανονική κατανομή $N(0,1)$ και (β) την εκθετική κατανομή με παράμετρο $\lambda=1$. Παίρνουμε δείγματα ελέγχου από κατανομές τις ίδιας μορφής με μετατόπιση είτε του μέσου κατά -1.5 και -1 μονάδες (για την κανονική κατανομή) είτε της παραμέτρου ρυθμού (για την εκθετική κατανομή) κατά -0.7 και -0.4 μονάδες και υπολογίζουμε τις πιθανότητες ορθής ανίχνευσης της μετατόπισης που είναι γνωστές ως πιθανότητες συναγερμού (*alarm rates, AR*) των τριών διαγραμμάτων ελέγχου για τις παραμέτρους σχεδιασμού που έχουν ήδη προσδιορισθεί στον Πίνακα 1. Στον Πίνακα 2 παρουσιάζονται ορισμένα εκ των αποτελεσμάτων που προέκυψαν.

Πίνακας 2. *Πιθανότητες ορθού συναγερμού (AR) των διαγραμμάτων με FAR≈0.05*

Διάγραμμα B-T-K									
						AR με δείγμα N(0,1)		***AR με δείγμα Exp(λ=1)***	
m	***n***	***a***	***b***	c_1	c_2	***N(-1.5,1)***	***N(-1,1)***	***Exp(λ=0.3)***	***Exp(λ=0.6)***
40	5	13	17	90	3	0.8135	0.5231	0.5638	0.1751
60	25	10	13	130	8	0.9980	0.9167	0.8207	0.2659
100	11	15	19	68	4	0.9702	0.7330	0.6323	0.1826
Νέο διάγραμμα B-T-K-RR$_{2\text{-}of\text{-}3}$									
40	5	17	22	115	3	0.9690	0.7688	0.8508	0.2877
60	25	9	12	76	8	0.9997	0.9616	0.8923	0.3033
100	11	13	18	40	4	0.9923	0.7896	0.6833	0.1925
Νέο διάγραμμα B-T-K-RR$_{2\text{-}of\text{-}4}$									
40	5	15	19	90	3	0.9841	0.8041	0.8603	0.2933
60	25	8	12	96	8	0.9999	0.9747	0.8962	0.3128
100	11	12	16	49	4	0.9970	0.8422	0.7025	0.1939

Για παράδειγμα, για δείγματα αναφοράς και ελέγχου μεγέθους $m=40$ και $n=5$ αντίστοιχα, υποθέτουμε ότι η εντός ελέγχου διεργασία ακολουθεί τυπική κανονική κατανομή. Αν υποθέσουμε ότι η διεργασία μετατοπίζεται κατά 1 (1.5) μονάδες, τότε το διάγραμμα των *Balakrishnan et al.*(2009) έχει πιθανότητα ορθού συναγερμού 0.5231 (0.8135), το προτεινόμενο διάγραμμα που ενσωματώνει τον κανόνα 2-από-τα-3 (***Νέο διάγραμμα B-T-K-RR$_{2\text{-}of\text{-}3}$***) έχει πιθανότητα ορθού συναγερμού 0.7688 (0.9690), ενώ το προτεινόμενο διάγραμμα που ενσωματώνει τον κανόνα 2-από-τα-4 (***Νέο διάγραμμα B-T-K-RR$_{2\text{-}of\text{-}4}$***) έχει πιθανότητα ορθού συναγερμού 0.8041 (0.9841).

Αξίζει να σημειωθεί ότι η συνήθης τακτική που εφαρμόζεται στη διεθνή βιβλιογραφία για σύγκριση μεταξύ μη-παραμετρικών διαγραμμάτων ελέγχου είναι αρχικά να γίνει η επιλογή των μεγεθών δειγμάτων που θα χρησιμοποιηθούν ως δείγματα αναφοράς και δείγματα ελέγχου, δηλαδή ο αρχικός προσδιορισμός των

παραμέτρων *m*, *n* και εν συνεχεία να προσδιορισθούν οι υπόλοιπες παράμετροι σχεδιασμού που οδηγούν στο επιθυμητό επίπεδο *FAR* (βλ. Balakrishnan, Triantafyllou and Koutras (2009, 2010), Chakraborti, van der Laan and van de Wiel (2004), Malela-Majika and Rapoo (2016), Chakraborti and van de Wiel (2008), Chowdhury, Mukherjee and Chakraborti (2014)). Συνεπώς, οι συνδυασμοί που χρησιμοποιούνται στα εξαγόμενα αριθμητικά αποτελέσματα της παρούσας εργασίας δεν αποτελούν απαραίτητα τους βέλτιστους συνδυασμούς παραμέτρων που επιτυγχάνουν από τη μία μεριά την επιθυμητή εντός ελέγχου απόδοση του προτεινόμενου διαγράμματος ελέγχου και από την άλλη τη μέγιστη δυνατή εκτός ελέγχου απόδοση. Ωστόσο, οι συνδυασμοί που έχουν χρησιμοποιηθεί επαληθεύουν αριθμητικά την υπεροχή των προτεινόμενων διαγραμμάτων ελέγχου έναντι ήδη υπαρχόντων και ανταγωνιστικών μη-παραμετρικών διαγραμμάτων ελέγχου.

Συμπερασματικά, με την εφαρμογή των κανόνων ροής που χρησιμοποιήθηκαν, το διάγραμμα αυξάνει την πιθανότητα να ανιχνεύσει ορθώς τη (σχετικώς μεγάλη) μετατόπιση της κατανομής της διεργασίας. Με άλλα λόγια, με χρήση πρωτίστως του 2-από τα-4 αλλά και του 2-από τα-3 κανόνα, το διάγραμμα καθίσταται ικανότερο στο να αξιολογήσει ορθά την πραγματική κατάσταση της υπό μελέτη διεργασίας. Αξίζει να επισημανθεί ότι αριθμητικά ο κανόνας 2-4 φαίνεται να δίνει καλύτερα αποτελέσματα έναντι των αντίστοιχων αποτελεσμάτων του κανόνα 2-3. Το συγκεκριμένο συμπέρασμα επαληθεύεται αριθμητικά και σε αντίστοιχα αποτελέσματα στη διεθνή βιβλιογραφία για χρήση παραμετρικών διαγραμμάτων με εφαρμογή των προαναφερθέντων κανόνων ροής (βλ. Khoo & Quah (2004)).

Από τον Πίνακα 2, είναι σαφές ότι τα μη-παραμετρικά διαγράμματα ελέγχου με τους κανόνες ροής 2-από τα-3 και 2-από τα-4 που εφαρμόζονται στην παρούσα εργασία, παρουσιάζουν καλύτερη απόδοση, αφού, για την ίδια πιθανότητα λανθασμένου συναγερμού (σχεδόν), πετυχαίνουν σαφώς καλύτερες τιμές για την πιθανότητα συναγερμού (*AR*), δηλαδή παρουσιάζουν συγκριτικά μεγαλύτερη πιθανότητα να ανιχνεύσουν τη μετατόπιση που πραγματοποιήθηκε στην κατανομή της διεργασίας.

4. ΣΥΜΠΕΡΑΣΜΑΤΑ

Στην παρούσα εργασία, μελετήθηκε το ελεύθερο κατανομής διάγραμμα ελέγχου που αρχικά εισήγαγαν οι *Balakrishnan et al*. (2009) και προτάθηκαν τρόποι βελτίωσης της απόδοσης του. Πιο συγκεκριμένα, προτείνεται η ενσωμάτωση κανόνων ροής, ενώ παρουσιάζεται νέος αλγόριθμος προσομοίωσης για τον προσδιορισμό βασικών χαρακτηριστικών της απόδοσης των τροποποιημένων διαγραμμάτων ελέγχου. Τα αριθμητικά αποτελέσματα που παρήχθησαν με την εφαρμογή του προτεινόμενου αλγορίθμου, αναδεικνύουν την υπεροχή των νέων μη-παραμετρικών διαγραμμάτων ελέγχου έναντι του διαγράμματος ελέγχου των *Balakrishnan et al*. (2009). Δεδομένου ότι στη διεθνή βιβλιογραφία, έχουν προταθεί πιο σύνθετες και ενδεχομένως πιο αποτελεσματικές μορφές κανόνων ροών, είναι αναμενόμενο ότι η εφαρμογή τους στο υπό μελέτη μη-παραμετρικό διάγραμμα θα αποφέρει ακόμη μεγαλύτερη βελτίωση στην απόδοση του τροποποιημένου διαγράμματος. Στο πλαίσιο μελλοντικής ερευνητικής δραστηριότητας και σε

συνέχεια των όσων αποτελεσμάτων έχουν έως τώρα προκύψει, παρουσιάζει ερευνητικό ενδιαφέρον η εφαρμογή νέων κανόνων ροών στο διάγραμμα έλεγχου που χρησιμοποιεί ως στατιστική συνάρτηση την *Wilcoxon* προκειμένου να διερευνηθεί το επίπεδο βελτίωσης της απόδοσης που συνεπάγεται η εφαρμογή τους.

ABSTRACT

In the present article, we implement runs-rules for enhancing the performance of the nonparametric Phase II Shewhart-type chart based on the Wilcoxon-type rank-sum statistic, which was first introduced by Balakrishnan *et al*. (2009). The false alarm rate and several design aspects of the proposed control charts are studied through extensive simulations. Finally, we investigate the respective alarm rate under certain out-of-control alternatives.

ΑΝΑΦΟΡΕΣ

Bakir, S.T. (2006). Distribution–free quality control charts based on signed–rank–like statistics. *Communications in Statistics: Theory & Methods*, **35**, 743-757.

Balakrishnan, N., Triantafyllou, I.S. and Koutras, M.V. (2009). Nonparametric control charts based on runs and Wilcoxon–type rank–sum statistics. *Journal of Statistical Planning and Inference*, **139**, 3177-3192.

Balakrishnan, N., Triantafyllou, I.S. and Koutras, M.V. (2010). A distribution-free control chart based on order statistics. *Communication in Statistics: Theory & Methods*, **39**, 3652-3677.

Chakraborti, S. (2011). Nonparametric (Distribution-free) Quality Control Charts. *Encyclopedia of Statistical Sciences*, 1-27.

Chakraborti, S. & van de Wiel, M.A. (2008). A nonparametric control chart based on the Mann-Whitney statistic. *IMS Collections* [*Beyond parametrics in interdisciplinary research*: Festschrift in Honour of Professor Pranab K. Sen], **1**, 156-172.

Chakraborti, S. and Eryilmaz, S. (2007). A nonparametric Shewhart–type signed–rank control chart based on runs. *Communications in Statistics: Simulation & Computation*, **36**, 335-356.

Chakraborti, S. and Graham, M.A. (2007). Nonparametric control charts. *Encyclopedia of Quality and Reliability*, New York: John Wiley.

Chakraborti, S., van der Laan, P. and van der Wiel, M.A. (2004). A class of distribution-free control charts. *Journal of the Royal Statistical Society, Series C-Applied Statistics*, **53**, 443-462.

Chowdhury S., Mukherjee A. & Chakraborti S. (2014). A new distribution-free control chart for joint monitoring of unknown location and scale parameters of continuous distributions. *Quality and Reliability Engineering International*, **30**, 191-204.

Derman,C. and Ross, S. M. (1997). *Statistical Aspects of Quality Control*. San Diego: Academic Press.

Khoo, M.B.C. (2004). Design of runs rules schemes. *Quality Engineering*, **16**, 27-43.

Khoo, M.B.C. & Quah, S.H. (2003). Multivariate control chart for process dispersion based on individual observations, *Quality Engineering*, **15**, 639-642.

Klein, M. (2000). Two alternatives to the Shewhart $\overline{X}$ control chart. *Journal of Quality Technology*, **32**, 427-431.

Malela-Majika, J.C. & Rapoo, E. (2016). Distribution-free cumulative sum and exponentially weighted moving average control charts based on the Wilcoxon rank-sum statistic using ranked set sampling for monitoring mean shifts, *Journal of Statistical Computation and Simulation*, **86**, 3715-3734.

Malela-Majika, J.C., Chakraborti, S. and Graham, M.A. (2016a). Distribution-free Phase II Mann–Whitney control charts with runs-rules. *International Journal Advanced Manufacturing Technology*, **86**, 723-735.

Malela-Majika, J.C., Chakraborti, S. and Graham, M.A. (2016b). Distribution-free precedence control charts with improved runs rules. *Applied Stochastic Models in Business and Industry*, **32**, 423-439.

Mukherjee, A., Graham, M.A., and Chakraborti, S. (2013). Distribution-Free Exceedance CUSUM Control Charts for Location. *Communication in Statistics*: *Simulation & Computation*, **42**, 1153-1187.

Triantafyllou, I. S. (2017). Nonparametric control charts based on order statistics: some advances. *Communication in Statistics*: *Simulation and Computation*, DOI: 10.180/03610918.2017.1359283.

Triantafyllou, I. S. (2018a). A new distribution-free control scheme based on order statistics. *Journal of Nonparametric Statistics*, DOI: 10.1080/10485252.2018. 1518524.

Triantafyllou, I. S. (2018b). A new distribution-free reliability monitoring scheme and applications in Engineering. In *Modeling and Simulation Based Analysis in Reliability Engineering* (ed. M. Ram), CRC Press: Taylor & Francis Group, 215-230.

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 31ου Πανελληνίου Συνεδρίου Στατιστικής (2018), σελ.191-203

ΓΡΑΦΗΜΑΤΑ ΚΑΙ ΔΕΙΓΜΑΤΟΛΗΨΙΑ

Παπαγεωργίου Μυρτώ[1], Φαρμάκης Νικόλαος[2]
[1,2]Τμήμα Μαθηματικών Α.Π.Θ.
myrtomp@hotmail.gr, farmakis@math.auth.gr

ΠΕΡΙΛΗΨΗ

Στην παρούσα εργασία αναπτύσσουμε την κατασκευή πεπερασμένων γεωμετριών-γραφημάτων με αξιωματική θεμελίωση με σκοπό τη μελέτη αποτελεσμάτων δειγματοληψίας στα γραφήματα αυτά. Γίνεται προσέγγιση του προβλήματος με την παράθεση σειράς προτάσεων που αποδεικνύονται με βάση τα αρχικά αξιώματα. Επίσης, επιδιώκεται το σύστημα αξιωμάτων να έχει τρεις βασικές ιδιότητες: Συνέπεια, Πληρότητα, Ανεξαρτησία. Όλα τα παραπάνω επιτυγχάνονται μέσα από τα γραφήματα (graphos) που απεικονίζονται κυρίως με τη μορφή πίνακα (Matrix) και είναι μια μορφή δικτύων. Διαπιστώνεται ότι σε κάποια δίκτυα (που παριστάνονται από γραφήματα) τα δείγματα που προέρχονται κατά βάση από Απλή Τυχαία Δειγματοληψία δίνουν αμερόληπτες εκτιμήτριες της μέσης τιμής τυχαίας μεταβλητής X. Αυτό κάνει ευκολότερη τη διαχείριση θεμάτων δειγματοληψίας σε δίκτυα που είναι ένα σημαντικό αντικείμενο και αρκετά πολύπλοκο. Δίνονται και παραδείγματα που κάνουν σαφέστερα τα ευρήματα αυτά.

Λέξεις κλειδιά: Γράφημα (Graphos), Πεπερασμένη Γεωμετρία, Πίνακας (Matrix).
MCS2010: 97K20, 05A19, 05B20

1. ΕΙΣΑΓΩΓΗ

Δίνεται Πληθυσμός ***Π*** μεγέθους *N*. Στον ***Π*** ορίζονται δύο τυχαίες μεταβλητές (τ.μ.), οι *X* και *Y*, που παίρνουν τις τιμές {1,2,3,…,*N*} με την ίδια πιθανότητα, δηλ.

$$\mathrm{P}\left(X=i\right)=\mathrm{P}\left(Y=i\right)=1/N,\ \ i=1,2,..,N\ .$$

Είναι γνωστά από την βιβλιογραφία τα κάτωθι, για τις τ.μ.*X, Y*

$$T=X,Y:\ ET=\mu_T=\left(N+1\right)/2 \tag{1}$$

$$S_T^2=\frac{1}{N-1}\sum_{i=1}^{N}\left(t_i-\overline{t}\right)^2=...=\frac{N\cdot\left(N+1\right)}{12}=\frac{N}{N-1}\cdot E\left(T-\frac{N+1}{2}\right)^2,\ T=X,Y. \tag{2}$$

Ορίζουμε

$$U=X+Y$$

και είναι προφανές, λόγω της (1), το

$$EU=EX+EY=N+1. \tag{3}$$

Θεωρούμε πίνακα **Α** διάστασης $N \times N$ με στοιχεία

$$\alpha_{ij} = i + j,\ i, j = 1,2,3,...,N\ ,$$

δηλαδή:

$$\boldsymbol{A} = \begin{bmatrix} 2 & 3 & 4 & .... & N & N+1 \\ 3 & 4 & 5 & .... & N+1 & N+2 \\ 4 & 5 & 6 & .... & N+2 & N+3 \\ ... & ... & ... & .... & ... & ... \\ N & N+1 & N+2 & .... & 2N-2 & 2N-1 \\ N+1 & N+2 & N+3 & .... & 2N-1 & 2N \end{bmatrix}$$

Από όλα τα παραπάνω προκύπτει ότι αν οι τ.μ. *X,Y* είναι ανεξάρτητες, όλες οι N^2τιμές της *U=X+Y* δίνονται στον πίνακα ***A*** και έχουν την ίδια πιθανότητα $1/N^2$. Η τιμή της Uπου αντιστοιχεί στο στοιχείο (i, j) είναι το άθροισμα των δύο συντεταγμένων της θέσης του στοιχείου. Δηλαδή είναι η

$$U = i + j = \alpha_{ij}, i, j = 1,2,3,...,N\ .$$

Θεωρούμε στη συνέχεια τον περιορισμό της τ.μ. Uπάνω στην κύρια διαγώνιο του ***A***, την τ.μ.

$$Z = \alpha_{ii} = 2i,\ i = 1,2,3,\ldots,N.$$

Προφανώς ισχύει:

$$EZ = 2 \cdot \frac{N+1}{2} = N+1\ . \tag{4}$$

Επίσης θεωρούμε τους περιορισμούς *W* και *R* της τ.μ. *U* πάνω στην άνω δεξιά (ΑΔ) τριγωνική περιοχή του ***A***, κατά τις σχέσεις

$$W = \alpha_{ij},\ i < j = 2,3,4...,N \text{ και } R = \alpha_{ij},\ j < i = 2,3,4...,N \tag{5}$$

αμφότεροι με πλήθος τιμών

$$\binom{N}{2}.$$

Η μεταθετική ιδιότητα της πρόσθεσης επιτρέπει να έχουμε σε πλήρη αντιστοίχιση τις τιμές των δύο τ.μ. W και R, (1 - 1) και επί(αμφιμονοσήμαντη αντιστοίχιση). Αυτό επειδή στον πίνακα ***A*** ισχύει η σχέση

$$\alpha_{ij} = \alpha_{ji}, \forall i \neq j,\ \ i, j = 1,2,3,...,N \text{ και } ER = EW. \tag{6}$$

$$\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_{ij} = \sum_{i=1}^{N}\sum_{j=1}^{N}\left(a_i + a_j\right) = N\left(\sum_{i=1}^{N}a_i + \sum_{j=1}^{N}a_j\right) = 2\frac{N^2(N+1)}{2} = N^2(N+1) \tag{7}$$

$$\sum_{i=1}^{N}\sum_{j=1}^{N}\left(a_i+a_j\right)=\sum_{i=j}^{N}\left(a_i+a_j\right)+\sum_{i<j}^{N}\left(a_i+a_j\right)+\sum_{i>j}^{N}\left(a_i+a_j\right)=N\left(N+1\right)+2C \qquad (8)$$

όπου τέθηκε

$$\sum_{i<j}^{N}\left(a_i+a_j\right)=\sum_{i>j}^{N}\left(a_i+a_j\right)=C$$

και άρα

$$N\left(N+1\right)+2C=N^2\left(N+1\right) \;\; \acute{\eta} \;\; C=N\left(N+1\right)\left(N-1\right)/2 \qquad (9)$$

και τελικά έχουμε

$$E(W)=E(R)=\frac{2}{N(N-1)}C=N+1 \qquad (10)$$

Για τη διασπορά των τ.μ. U, X, Y, W, *Z* και *R* με αξιοποίηση και του πίνακα ***A*** έχουμε τα παρακάτω:

1.1 Τυχαία μεταβλητή U

Παίρνουμε το άθροισμα των τετραγώνων των διαφορών των τιμών της τ.μ. U από τη μέση τιμή της την *EU*=*N*+1

$$\sum_{i=1}^{N^2}\left(U-EU\right)^2=\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\left(i+j\right)-\left(N+1\right)\right)^2=\ldots=\frac{N^2\left(N+1\right)\left(2N+1\right)}{3}-\frac{N^2\left(N+1\right)^2}{2}$$

Άρα,

$$VarU=\sigma_U^2=\frac{\sum_{i=1}^{N^2}\left(U-EU\right)^2}{N^2}=\frac{\left(N-1\right)\cdot\left(N+1\right)}{6} \qquad (11)$$

και

$$S_U^2=\frac{N^2\cdot\sigma_U^2}{N^2-1}=\frac{N^2}{6}\,. \qquad (12)$$

1.2 Τυχαίες μεταβλητές *X* και *Y*

(βλέπε σχέσεις (1) και (2))

1.3 Τυχαία μεταβλητή *Z*

$$VarZ=\sigma_Z^2=Var\left(2X\right)=4VarX=\frac{\left(N-1\right)\cdot\left(N+1\right)}{3} \qquad (13)$$

και από τη σχέση (2)

$$S_Z^2 = \frac{N\cdot\sigma_Z^2}{N-1} = \frac{N\cdot(N-1)\cdot(N+1)}{3(N-1)} = \frac{N\cdot(N+1)}{3}. \qquad (14)$$

1.4 Τυχαίες μεταβλητές W και R

Από την (4) έχουμε σχετικά με την τ.μ. Z

$$\sum_{i=1}^{N}(Z-EZ)^2 = \sum_{i=1}^{N}\{Z-(N+1)\}^2 = \sum_{i=1}^{N}\{2X-(N+1)\}^2 = \ldots = \frac{(N-1)\cdot N\cdot(N+1)}{3}. \qquad (15)$$

Από (11) και (15) προκύπτει

$$\sum_{i=1}^{N(N-1)/2}(W-EW)^2 = \frac{1}{2}\left(\sum_{i=1}^{N^2}(U-EU)^2 - \sum_{i=1}^{N}(Z-EZ)^2\right) = \ldots = \frac{(N-2)\cdot(N-1)\cdot N\cdot(N+1)}{12}, \qquad (16)$$

οπότε διαιρώντας με το πλήθος των τιμών της τ.μ. W, το $\frac{N\cdot(N-1)}{2}$ έχουμε

$$VarW = \sigma_W^2 = \frac{(N-2)\cdot(N+1)}{6} \text{ και } S_W^2 = \frac{\binom{N}{2}\sigma_W^2}{\binom{N}{2}-1} = \frac{N\cdot(N-1)}{6}. \qquad (17)$$

Όμοια έχουμε και τις τιμέγια την τ.μ. *R* στην κάτω αριστερή περιοχή του ***A*** (ΚΑ)

$$\sigma_R^2 = \sigma_W^2,\ S_R^2 = S_W^2. \qquad (18)$$

2. ΔΟΜΕΣ ΚΑΙ ΓΡΑΦΗΜΑΤΑ

Θεωρούμε το σύνολο V με N στοιχεία που τα ονομάζουμε **κορυφές** (Vertices). Συνηθίζεται να τα συμβολίζουμε με τους N αρχικούς φυσικούς αριθμούς, 1,2,3,…,N ή με τα πρώτα Ν γράμματα και συνδυασμούς γραμμάτων μιας αλφαβήτου, όπως Α,Β,Γ,..,Ω,..,ΒΔ,…κλπ. Άρα γράφουμε V={1,2,3,…,N}. Θεωρούμε επίσης και το καρτεσιανό γινόμενο

$$V\times V = \{(x,y),\ x,y\in V\}$$

και ένα μη κενό υποσύνολό του $E\subseteq V\times V$, όχι κατ' ανάγκη διάφορο του $V\times V$. Τα στοιχεία του E ονομάζονται συνήθως **ακμές** (edges). Το σύνολο $\{V,E\}$ ονομάζεται συνήθως **δομή** των V και E και είναι ένα βασικό εργαλείο για τη μελέτη διαφόρων καταστάσεων που μπορούν να περιγραφούν με βάση τα V και E, καθώς αυτά και η δομή εφοδιάζονται συνήθως με κάποιες ιδιότητες και συναρτήσεις που συνδέουν αυτά ή υποσύνολά τους. Ειδικότερα:

Ορισμός 1: Η δομή Γ=(V, E) ονομάζεται **γράφημα** (Graphos), όταν ισχύει

$$\varnothing \neq E \subseteq V\times V.$$

Σημείωση 1η:Το γράφημα Γ μπορεί να παρασταθεί με πολλούς τρόπους, όπως με σχήμα πάνω σε επίπεδο με κορυφές γεωμετρικά σημεία (της Ευκλείδειας

Γεωμετρίας) που τα συνδέουν ακμές ή παριστάνεται με χρήση πινάκων (matrices) ***M***, διάστασης *NxN*, με στοιχεία ίσα με 0 και 1. Όταν το στοιχείο

$$(i, j),\ i, j = 1,2,3,...,N$$

του πίνακα είναι 1, σημαίνει ότι το αντίστοιχο στοιχείο υπάρχει στο σύνολο *E* της δομής (είναι ακμή της). Όταν είναι 0 τότε το αντίστοιχο στοιχείο δεν υπάρχει στο Ε.
Πολλές φορές για την εξυπηρέτηση σχετικά πολύπλοκων καταστάσεων έρευνας, οι τιμές των στοιχείων των πινάκων ***M*** παίρνουν και άλλες τιμές εκτός από τις 0 και 1.
Σημείωση 2η: Τη μορφή του γραφήματος καθορίζει και αναδεικνύει μία σειρά από προτάσεις. Μερικές από αυτές θεωρείται ότι ισχύουν χωρίς να χρειάζεται να υπάρξει απόδειξη. Ονομάζονται **αξιώματα**. Έπονται και συνυπάρχουν συνήθως και άλλες προτάσεις που η απόδειξή τους είναι αναγκαία και στηρίζεται στα αξιώματα. Αυτές ονομάζονται **θεωρήματα**. Σχηματικά εδώ έχουμε αρχικά τη δομή-γράφημα Γ(*V, E*), εφοδιασμένη με ένα σύνολο αξιωμάτων Α.
Ορισμός 2: Η δομή Γ(*V, E*), εφοδιασμένη με το σύνολο αξιωμάτων Α αποτελεί μια νέα δομή που ονομάζεται **αξιωματικό γράφημα**. Συμβολίζεται Γ(*V, E, A*).
Ενδιαφέρον παρουσιάζουν τα αξιωματικά γραφήματα που η μορφή τους επιτρέπει τη χρήση απλών κλασσικών τεχνικών δειγματοληψίας σε αυτά, ώστε να προκύπτουν αμερόληπτες εκτιμήτριες βασικών παραμέτρων των τ.μ. που υπεισέρχονται στο πρόβλημα, π.χ. της μέσης τιμής μιας τ.μ. Χ. ***Αυτός είναι και ο βασικός στόχος της παρούσας εργασίας.***

3. ΔΕΙΓΜΑΤΙΚΕΣ ΕΚΤΙΜΗΤΡΙΕΣ ΣΕ ΓΡΑΦΗΜΑΤΑ

Έστω το αξιωματικό γράφημα Γ(*V*,*E*,*A*) με *V*={1,2,3,…,*N*}, το καρτεσιανό γινόμενο $E = V \times V$ και το σύνολο *A* που περιέχει τα εξής 5 αξιώματα:
Αξίωμα 1: Κάθε ζεύγος από στοιχεία του *V* έχει τουλάχιστο ένα κοινό στοιχείο του *E*.
Αξίωμα 2: Κάθε ζεύγος από στοιχεία του *V* δεν έχει περισσότερα από ένα κοινό στοιχείοτου *E*.
Αξίωμα 3: Κάθε στοιχείο του *E* ανήκει σε δύο τουλάχιστο στοιχεία του *V*.
Αξίωμα 4: Κάθε στοιχείο του *E* δεν ανήκει σε περισσότερα από δύο στοιχεία του *V*.
Αξίωμα 5: Το *V* περιέχει συνολικά *N* στοιχεία. (βλ. και Αρτεμιάδης Ν., 1977).

Το σύνολο *V* της δομής θα αποτελεί τον Πληθυσμό-στόχο από τον οποίο και θα παίρνονται τα δείγματα με κάποια γνωστή τεχνική δειγματοληψίας, π.χ. απλή τυχαία δειγματοληψία (ΑΤΔ), Φαρμάκης (2016, 2015), Cochran (1977). Μπορεί να θεωρηθεί (σπανιότερα σε εφαρμογές) ότι είναι ένα δείγμα μεγέθους *N* που προέκυψε από Πληθυσμό πολύ μεγαλύτερου μεγέθους από το *N*.
Θεωρούμε την τ.μ. *Y* που απεικονίζει (κατά τυχαίο τρόπο) το *V* στο σύνολο των πραγματικών ***R*** σύμφωνα με τη σχέση

$$Y(i) = Y_i,\ \ i = 1,2,3,...,N\ .$$

Όμοια στη συνέχεια θεωρούμε τις τ.μ. *U, Z, X, R* που απεικονίζουν το καρτεσιανό γινόμενο $E = V \times V$ στο διδιάστατο επίπεδο των πραγματικών αριθμών$\boldsymbol{R}^2$,
σύμφωνα με τις σχέσεις:

$$U(i,j)=Y_i+Y_j,\quad i,j=1,2,3,...,N$$
$$Z(i,i)=Y_i+Y_i=2Y_i,\ i=1,2,3,...,N$$
$$W(i,j)=Y_i+Y_j,\ i<j=2,3,4,...,N$$
$$R(i,j)=Y_i+Y_j,\ j<i=2,3,4,...,N\ .$$

Στον παρακάτω τώρα πίνακα **A** το εκάστοτε ζεύγος (*i,j*) παριστάνει την τιμή της συνάρτησης $\Theta\left(Y_i,Y_j\right)$,ανάλογα με τις εκάστοτε ανάγκες της ανάλυσης και σε σχέση με τις τιμές*i*, *j*, όπου

$$i,j\in\{1,2,3,...,N\}\text{, π.χ.}\Theta\left(Y_i,Y_j\right)=Y_i+Y_j\ .$$

$$\boldsymbol{A}=\begin{bmatrix}(1,1) & (1,2) & (1,3) & ..... & (1,N-2) & (1,N-1) & (1,N)\\ (2,1) & (2,2) & (2,3) & ..... & (2,N-2) & (2,N-1) & (2,N)\\ (3,1) & (3,2) & (3,3) & ..... & (3,N-2) & (3,N-1) & (3,N)\\ ... & ... & ... & ..... & ... & ... & ...\\ (N-2,1) & (N-2,2) & (N-2,3) & ..... & (N-2,N-2) & (N-2,N-1) & (N-2,N)\\ (N-1,1) & (N-1,2) & (N-1,3) & ..... & (N-1,N-2) & (N-1,N-1) & (N-1,N)\\ (N,1) & (N,2) & (N,3) & ..... & (N,N-2) & (N,N-1) & (N,N)\end{bmatrix}$$

Ο πίνακας ***A*** κανονικά αποτελεί το πεδίο ορισμού της τ.μ. *U*. Προφανώς μέρη του αποτελούν το αντίστοιχο πεδίο ορισμού των τ.μ. Z, *W*, R, π.χ. η άνω δεξιά περιοχή του ***A*** αποτελεί το πεδίο ορισμού της τ.μ. *W*. Μετά όμως από τη σύμβαση που προηγήθηκε του ***A***, το στοιχείο (*i, j*) του ***A*** αντιστοιχεί στο στοιχείο $\left(Y_i,Y_j\right)$ και στη συνέχεια μπορεί να οριστεί (παρακάτω) ότι αντιστοιχεί και με την ποσότητα Y_i+Y_j, σύμφωνα με τις ανάγκες της ανάλυσης. Επίσης ο πίνακας ***A*** με τις εκάστοτε μορφές του που μόλις αναφέραμε παριστάνει ένα γράφημα χρήσιμο για τη μελέτη των τ.μ. *U, Z, R,W*(και άλλων) που μας ενδιαφέρουν. Θα προσπαθήσουμε να προσεγγίσουμε το πρόβλημα της μελέτης τωντ.μ. *U, Z, R, W* με δειγματοληψία. Πληθυσμός στόχος το *V*μεγέθους*N* και μέγεθος δείγματος *n*<*N*. Μέθοδος δειγματοληψίας η απλή τυχαία δειγματοληψία (ΑΤΔ), Φαρμάκης (2016). Κατά την ΑΤΔ υπάρχουν $\binom{N}{n}$ δείγματα μεγέθους *n*, καθώς έχουμε τυχαία δειγματοληψία χωρίς επανάθεση. Στο δείγμα που θα προκύψει θα γίνουν οι υπολογισμοί και γενικά η ανάλυση των δεδομένων του δείγματος και θα προκύψει η δειγματική μέση τιμή της τ.μ. που μελετούμε, π.χ. η $\overline{u}$ για την τ.μ. *U*, η $\overline{w}$ για την τ.μ. *W* κλπ. Η δειγματική μέση τιμή είναι η εκτιμήτρια της μέσης τιμής του Πληθυσμού για την εκάστοτε τ.μ. που μελετούμε. Έτσι έχουμε τις εκτιμήτριες $\overline{u}=\hat{\mu}_U$, $\overline{w}=\hat{\mu}_W$ κλπ. Χωρίς περιορισμό της γενικότητας μελετώντας

την τ.μ. U και από το i-δείγμα, $i=1,2,3,...,\binom{N}{n}$, προκύπτει η μέση δειγματική τιμή $\overline{u}_i$ ως εκτιμήτρια της πληθυσμιακής μέσης τιμής EU. Έχουμε ήδη μία νέα τ.μ. την τ.μ. $\overline{u}$ που παίρνει τις τιμές $\overline{u}_i$, $i=1,2,3,...,\binom{N}{n}$.

Ορισμός 3:Η εκτιμήτρια $\overline{u}$ ονομάζεται αμερόληπτη (unbiased) αν ισχύει: $E\overline{u}=\mu_U$ και γενικότερα αν ισχύει $E\hat{M}=M$, για οποιοδήποτε μέγεθος M και την εκτιμήτριά του $\hat{M}$, Φαρμάκης (2016), Cochran (1977).

Υπενθυμίζεται ότι η τ.μ. U ορίζεται σε όλον τον πίνακα $\boldsymbol{A}$. Οι τ.μ. Wκαι R ορίζονται στις τριγωνικές του περιοχές, άνω δεξιά (ΑΔ) και κάτω αριστερά (ΚΑ) αντίστοιχα. Τέλος η τ.μ. Z πάνω στην κύρια διαγώνιο του πίνακα $\boldsymbol{A}$.

Η δειγματοληψία που θα χρησιμοποιηθεί εδώ είναι η εξής:

Παίρνουμε δείγμα μεγέθους $n<N$ από τον πληθυσμό-στόχο με ΑΤΔ και έστω ότι είναι το $\delta=\{i_1,i_2,i_3,...,i_n\}$. Οπότε, αν η ΑΤΔ γίνεται στο σύνολο $V=\{1,2,...,N\}$ όπου ορίζεται η τ.μ.X, υπάρχουν $m=\binom{N}{n}$ δείγματα. Κάθε στοιχείο το $V=\{1,2,...,N\}$ περιέχεται σε $\binom{N-1}{n-1}$ δείγματα, το άθροισμα των στοιχείων του V είναι $N(N+1)/2$ και η πιθανότητα να πάρουμε ένα συγκεκριμένο δείγμα είναι $1/\binom{N}{n}$. Άρα η μέση τιμή του δείγματος είναι

$$E(\overline{x})=\frac{\binom{N-1}{n-1}N(N+1)}{2n\binom{N}{n}}=\frac{N(N+1)}{2N}=\frac{N+1}{2}=E(X).$$

Θεωρούμε την τ.μ. W. Η μέση τιμή της W για τον πληθυσμό (ΑΔ περιοχή του **Α**) είναι από τη σχέση (10):

$$EW=N+1$$

Μία εκτιμήτρια θα προκύψει με την παρακάτω διαδικασία:

Οι δείκτες στο ανωτέρω δείγμα δ θεωρείται ότι βαίνουν αυξανόμενοι (χωρίς βλάβη της όποιας γενικότητας). Από το δείγμα δ προκύπτει ένα δείγμα με στοιχεία του $\boldsymbol{A}$ που θα αποτελέσει το «υλοποιημένο» δείγμα $\delta_υ$, με βάση τα στοιχεία του οποίου θα

υπολογιστεί η τιμή $\overline{w}$, της εκτιμήτριας της *EW*. Όλες τις τιμές της τ.μ. *W* που αντιστοιχούν στο δ_υ αντιστοιχούν ουσιαστικά σε στοιχεία της ΑΔ περιοχής του ***A*** με συντεταγμένες (i,j) που πληρούν τον κανόνα:

$$\forall i \in \delta, (i,j) \acute{\eta} (j,i) \in \text{ΑΔ}, i \neq j = 1,2,3,...,N. \tag{19}$$

Δηλ. αν *i*=3 δείγμα δ_υ θα υπάρχουν τα στοιχεία της ΑΔ περιοχής που βρίσκονται στην τρίτη γραμμή και στην τρίτη στήλη. Εξαιρείται το (3,3) στοιχείο του ***A***. Άρα στο δ_υ ανήκουν τελικά $n \cdot (N-1)$ στοιχεία του **A** και άρα τόσες αντίστοιχες τιμές της τ.μ. $W=Y_i+Y_j$ στο «υλοποιημένο» δείγμα δ_υ. ***Έχουμε μια δισταδιακή δειγματοληψία που στηρίζεται και ξεκινάει από ΑΤΔ στο V και συνεχίζει στον A.*** Με τις τιμές αυτές της τ.μ. *W υπολογίζουμε* τη δειγματική μέση τιμή (εκτιμήτρια) που πρακτικά είναι η:

$$\overline{w} = \frac{\sum_{i=i_1}^{i_n} \sum_{i \neq j=1}^{N} (Y_i + Y_j)}{n \cdot (N-1)} \tag{20}$$

με

$$n \cdot (N-1)$$

όρους $Y_i + Y_j = W_{ij}$

$$W_{ij} = Y_i + Y_j, i \in \{i_1, i_2, i_3, ..., i_n\} = \delta,\ i \neq j = 1,2,3,...,N$$

στον αριθμητή του κλάσματος. Αυτό επειδή για κάθε δείκτη του δείγματος δ (πρώτο στάδιο), έχουμε *N*-1στοιχεία στο «υλοποιημένο» δείγμα (δεύτερο στάδιο) που μας δίνουν την τιμή του αριθμητή στην σχέση (20). Σχηματικά τα στοιχεία του «υλοποιημένου» δείγματος είναι όλα τα στοιχεία της *i*-γραμμής και της *j*-στήλης της ΑΔ περιοχής του ***A***,και άρα εκτός από το διαγώνιο στοιχείο που δεν ανήκει στην ΑΔ.

Θεώρημα 1:Η μέση τιμή $\overline{w}$ είναι αμερόληπτη εκτιμήτρια της πληθυσμιακής μέσης τιμής *EW*. Δηλαδή ισχύει $E\overline{w} = EW$.

Απόδειξη: Συμβολίζουμε με $W_{ij} = Y_i + Y_j$ την τιμή της *W* για

$$(i,j),\ i < j = 2,3,4,...N\,.$$

Έστω $\delta = \{i_1, i_2, i_3, ..., i_n\}$ τυχαίο δείγμα εκλεγμένο από το σύνολο $V=\{1,2,3,\ldots,N\}$.Λόγω της συμμετρικής μορφής των W_{ij} και του πίνακα ***A*** η (19) μπορεί να γραφεί ως:

$$\forall i \in \delta, (i,j) \in \boldsymbol{A}, i \neq j = 1,2,3,...,N. \tag{21}$$

Από (21) έχουμε: Στο διπλό άθροισμα του αριθμητή της $\overline{w}$ η κάθε μια από τις τιμές $Y_{i_\lambda}, \lambda = 1,2,3,...,n$ συμμετέχει *N*-1φορές ως πρώτος προσθετέος

Y_{i_λ}, $\lambda = 1,2,3,...,n$ και n-1 φορές ως δεύτερος προσθετέος. Δηλαδή στον αριθμητή της $\overline{w}$ θα έχουμε τον όρο

$$\left(N+n-2\right)\cdot Y_{i_\lambda},\ \forall \lambda = 1,2,3,...,n$$

όπου το Y έχει δείκτη από το δείγμα δ και συμμετέχει ως προσθετέος στην τιμή $W_{ij} = Y_i + Y_j$ ή $W_{ji} = Y_j + Y_i$. Ανάλογα, για όλους τους δείκτες του συνόλου {1,2,3,…,N}-δ που δεν κληρώθηκαν στο δείγμα, οι αντίστοιχες τιμές Y_j εμφανίζονται n φορές. Δηλαδή στον αριθμητή της εκτιμήτριας $\overline{w}$ υπάρχει ο όρος

$$n\cdot Y_j,\ j \in \left\{j_1, j_2, j_3, ..., j_{N-n}\right\} = \left\{1,2,3,...,N\right\} - \delta .$$

Ενώ ο αριθμητής της $\overline{w}$ στην (20) έχει στοιχεία της Wτελικά οι πράξεις γίνονται με τιμές της Y=1,2,3,..,N.
Άρα το δείγμα δ παρέχει ως αριθμητή της $\overline{w}$ το άθροισμα

$$\left(N+n-2\right)\cdot \sum_{\lambda=1}^{n} Y_{i_\lambda} + n\cdot \sum_{m=1}^{N-n} Y_{j_m} .$$

Πρέπει να υπολογιστεί η μέση τιμή όλων των δειγματικών μέσων τιμών $\overline{w}$. Θα υπολογιστεί η ποσότητα

$$E\overline{w} = \sum_{m=1}^{\binom{N}{n}} \overline{w}_m \Big/ \binom{N}{n}.$$

Στον αριθμητή της ποσότητας αυτής υπάρχουν όλοι οι αριθμητές της $\overline{w}$. Αυτό σημαίνει ότι όλες οι τιμές $Y_{i_\lambda}, \lambda = 1,2,3,...,n$ συμμετέχουν εκεί $\left(N+n-2\right)$ φορές, επί το πλήθος των δειγμάτων που περιέχουν τον δείκτη i_λ, λ=1,2,3,…,n.Αυτά τα δείγματα είναι σε πλήθος $\binom{N-1}{n-1}$, Φαρμάκης (2016). Αντίστοιχα κάθε τιμή Y_j, με $j \in \left\{1,2,3,...,N\right\} - \delta$, συμμετέχει στον παραπάνω αριθμητή nφορές επί τον αριθμό των δειγμάτων που δεν περιέχουν το δείκτη j. Τα δείγματα αυτά είναι σε πλήθος $\binom{N}{n} - \binom{N-1}{n-1}$,δηλαδή είναι όλα τα n-μελή δείγματα μείον όσα περιέχουν το δείκτηj.
Όλες οι τιμές Y_m, $m = 1,2,3,...,N$ συμμετέχουν στον αριθμητή της ποσότητας

$$\sum_{k=1}^{\binom{N}{n}} \overline{w}_k$$

με την ίδια πολλαπλότητα, λόγω ορισμού του τυχαίου δείγματος, Φαρμάκης (2016). Θα συμμετέχουν δηλαδή όλοι οι όροι

$$Y_m,\ m = 1,2,3,...,N\ ,\ \left(N+n-2\right)\cdot \binom{N-1}{n-1}$$

φορές ως όροι που έχουν δείκτη από το δείγμα και

$$n\cdot\left\{\binom{N}{n}-\binom{N-1}{n-1}\right\}$$

φορές ως όροι με δείκτη εκτός του εκάστοτε εξεταζόμενου δείγματος. Άρα ο αριθμητής της ποσότητας $\sum_{k=1}^{\binom{N}{n}}\overline{w}_k$ θα είναι τελικά

$$\left\{(N+n-2)\cdot\binom{N-1}{n-1}+n\cdot\left\{\binom{N}{n}-\binom{N-1}{n-1}\right\}\right\}\cdot\sum_{m=1}^{N}Y_m=S_{o\lambda}$$

και άρα

$$E\overline{w}=\frac{S_{o\lambda}}{n\cdot(N-1)\cdot\binom{N}{n}}=\frac{2\cdot\binom{N-1}{n-1}}{n\cdot\binom{N}{n}}\cdot\sum_{m=1}^{N}Y_m=2\cdot EY=EW\text{, ό.έ.δ.}$$

Μια άλλη εκτιμήτρια της *EW* μπορεί να προκύψει από το δείγμα δ, που πήραμε με ΑΤΔ, με την εξής διαδικασία:

Από την ΑΔ περιοχή του **A** κρατάμε όλα τα σημεία $(i,j), i,j\in\delta, i\neq j$. Η τιμή $\overline{w}$, της εκτιμήτριας της *EW* προκύπτει τώρα ως:

$$\overline{w}=\frac{\sum_{i=i_1}^{i_{n-1}}\sum_{i<j=i_2}^{i_n}(Y_i+Y_j)}{\binom{n}{2}}. \qquad (22)$$

Θεώρημα 2: Η μέση τιμή $\overline{w}$ της σχέσης (22) είναι αμερόληπτη εκτιμήτρια της πληθυσμιακής μέσης τιμής *EW*. Δηλαδή ισχύει $E\overline{w}=EW$.

Απόδειξη: Έστω $\delta=\{i_1,i_2,i_3,...,i_n\}$ δείγμα εκλεγμένο με ΑΤΔ από πληθυσμό-στόχο $V=\{1,2,3,...,N\}$. Βάζουμε σε αύξουσα σειρά τους δείκτες-στοιχεία του δείγματος, αν δεν είναι ήδη. Συμβολίζουμε με $W_{ij}=Y_i+Y_j$ την τιμή της W για $(i,j), i,j\in\delta$. Εξετάζουμε την ΑΔ περιοχή του $n\times n$ πίνακα $\boldsymbol{\alpha}$ με στοιχεία τα $(i,j), i,j\in\delta$.

Παρατηρούμε ότι το στοιχείο i_1εμφανίζεται στην ΑΔ περιοχή του $\boldsymbol{\alpha}$,n-1 φορές. Το δεύτερο στοιχείο του δείγματος εμφανίζεται n-2φορές ως πρώτο στοιχείο ζεύγους και 1 φορά ως δεύτερο στοιχείο. Συνολικά n-2+1=n-1 φορές. Γενικότερα το στοιχείο i_λ εμφανίζεται συνολικά n-λ+λ-1=n-1 φορές. Υπολογίζοντας από το δ τη μέση τιμή της τ.μ. W, τη $\overline{w}$, έχουμε στον αριθμητή n-1 φορές τον κάθε όρο $Y_{i_m}, m=1,2,3,...,n$. Άρα

ο αριθμητής έχει την τιμή $(n-1)\cdot\sum_{m=1}^{n} Y_{i_m} = (n-1)\cdot n\cdot \overline{y}$. Αντικαθιστούμε αυτήν την ποσότητα στην τιμή της $\overline{w}$ στην (22). Κάνουμε τις πράξεις και είναι $\overline{w} = 2\cdot\overline{y}$. Μετά παίρνουμε τη μέση τιμή των δύο μελών της ισότητας ως προς όλα τα δείγματα δ. Έχουμε υπόψη και την ιδιότητα της μέσης δειγματικής τιμής με δείγμα από την ΑΤΔ, την: $E\overline{y} = EY$, Φαρμάκης (2016). Τελικά $E\overline{w} = E2\cdot\overline{y} = 2\cdot EY = EW$, ό.έ.δ.

Είναι ενδιαφέρουσα περίπτωση η ΑΤΔ να γίνει στον πληθυσμό «ΑΔ περιοχή του ***A***» που έχει $m=N(N-1)/2$ στοιχεία. Η μελέτη γίνεται για την τ.μ. Xμε τιμές (στην ΑΔ) τις

$$x_{ij} = i + j,\quad i < j = 2,3,4,...,N,$$

όπου, λόγω της (10), είναι

$$EX = N + 1.$$

Θεώρημα 3: Αν $\overline{x}$ είναι η μέση τιμή του δείγματος τότε $E\overline{x} = EX = N + 1$.

Απόδειξη: Η πιθανότητα να πάρουμε ένα συγκεκριμένο δείγμα $\delta_i = \{i_1, i_2, i_3, ..., i_n\}$ είναι

$$\frac{1}{m}\frac{1}{m-1}...\frac{1}{m-n+1} = \frac{(m-n)!}{m!}$$

και το πλήθος αυτών των δειγμάτων μεγέθους $n<m$ είναι

$$\binom{m}{n} = \frac{m(m-1)...((m-n+1)}{n!}.$$

Η μέση τιμή του δείγματος είναι

$$\overline{x}_i = \sum_{j=1}^{n} x_{ij} / n,\ i = 1,2,3,..., \binom{m}{n}$$

και το κάθε στοιχείο του δείγματος περιέχεται σε

$$\binom{m-1}{n-1}$$

δείγματα: είναι όλα τα δείγματα μεγέθους n-1 που δεν περιέχουν το εν λόγω στοιχείο παρμένα από τον πληθυσμό χωρίς το στοιχείο αυτό. Επομένως πριν πάρουμε το δείγμα με ΑΤΔ αναμένεται η μέση τιμή του δείγματος να είναι (expected mean value)

$$E\overline{x} = \frac{1}{n}\cdot\binom{m-1}{n-1}\cdot C / \binom{m}{n} = \frac{1}{n}\cdot\binom{m-1}{n-1}\cdot\frac{N(N+1)(N-1)}{2} / \binom{m}{n} = N+1 = EX,$$

όπου από την (9) έχουμε για το άθροισμα των στοιχείων της ΑΔ περιοχής του ***A***

$$C = N(N+1)(N-1)/2.$$

ABSTRACT

In this paper, a family of finite geometries is proposed and studied. It is based on the idea of graphs (graphos). We try an axiomatic foundation in order to study the sampling results in

these graphs. We assume the graphs and some parts of them as the Population for the sampling trials. The problem is approximated by the quoting of a series of propositions that are based on the original axioms. It is also intended that the system of axioms has three basic properties: Consistency, Completeness and Independence. The entire above are achieved through the graphs (graphs) that are mainly represented in the form of a matrix (Matrix) and are a form of networks. It can be seen that in some networks (represented by graphs) samples which are (basically)derived from Simple Random Sampling give impartial estimators of the mean value of random variable. This makes easier to manage sampling issues on networks that are an important subject and quite complex. Illustrating examples are given, making clear the above ideas.

Ευχαριστίες: Οι συγγραφείς ευχαριστούν τους κριτές για την εποικοδομητική κριτική και τις υποδείξεις που συντέλεσαν στην αρτιότερη μορφή της εργασίας.

ΑΝΑΦΟΡΕΣ

Αρτεμιάδης, Ν. (1977) «Στοιχειώδης Γεωμετρία από ανώτερη σκοπιά», Εκδόσεις ΕΜΕ, Αθήνα.

Φαρμάκης, Ν. (2016) «Εισαγωγή στη Δειγματοληψία», Αφοί Κυριακίδη Εκδόσεις Α.Ε., Θεσσαλονίκη.

Φαρμάκης, Ν. (2015), «Δειγματοληψία και εφαρμογές», Ελληνικά Ακαδημαϊκά Ηλεκτρονικά Συγγράμματα & Βοηθήματα, Αθήνα ISBN:978-960-603-093-2.

Cochran W. (1977) "Sampling Techniques", John Wiley & Sons, Inc, New York, Toronto.

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 31ου Πανελληνίου Συνεδρίου Στατιστικής (2018), σελ.204-219

ΠΑΡΑΓΟΝΤΕΣ ΚΑΤΑ ΤΗ ΔΙΑΡΚΕΙΑ ΤΗΣ ΖΩΗΣ ΠΟΥ ΚΑΘΟΡΙΖΟΥΝ ΤΗΝ ΚΑΤΑΘΛΙΨΗ ΣΕ ΑΤΟΜΑ ΗΛΙΚΙΑΣ 50 ΚΑΙ ΑΝΩ: ΔΙΑΦΟΡΟΠΟΙΗΣΕΙΣ ΜΕΤΑΞΥ ΑΝΔΡΩΝ ΚΑΙ ΓΥΝΑΙΚΩΝ ΣΤΗΝ ΕΥΡΩΠΗ

Ελένη Σεραφετινίδου[1] και Γεωργία Βερροπούλου[1]
[1] Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης, Πανεπιστήμιο Πειραιώς
e-mails: elen_sta@outlook.com.gr, gverrop@unipi.gr

ΠΕΡΙΛΗΨΗ

Στόχος της παρούσας μελέτης είναι η διερεύνηση παραγόντων που προβλέπουν κατάθλιψη ανάμεσα σε μεταβλητές της παιδικής ηλικίας, της ενήλικης ζωής καθώς και τις συγχρονικές συνθήκες σε άτομα ηλικίας 50 ετών και άνω και ο εντοπισμός των διαφορών κατά φύλο μεταξύ γεωγραφικών περιοχών της Ευρώπης. Τα στοιχεία προέρχονται από την έρευνα για την Υγεία, τη Γήρανση και τη Συνταξιοδότηση στην Ευρώπη (SHARE). Για την ανάλυση χρησιμοποιήθηκαν μοντέλα λογιστικής παλινδρόμησης. Τα ευρήματα δείχνουν ότι: α. οι παιδικοί προγνωστικοί παράγοντες είναι οι λιγότερο σημαντικοί, με εξαίρεση τους κατοίκους της Βόρειας, Κεντρικής και Νότιας Ευρώπης β. από την ενήλικη ζωή η εμπειρία μιας περιόδου άγχους αποτελεί τον πιο σημαντικό παράγοντα για τις γυναίκες σε όλες τις ευρωπαϊκές περιοχές ενώ για τους άνδρες μόνο σε χώρες της Κεντρικής και Νότιας Ευρώπης και γ. οι παράγοντες πρόβλεψης που αφορούν στην μετέπειτα ζωή, συμπεριλαμβανομένης της σωματικής υγείας και του κοινωνικοοικονομικού επιπέδου έχουν την ισχυρότερη σχέση με την κατάθλιψη για όλες τις ομάδες.

Λέξεις κλειδιά: κατάθλιψη; υγεία; κοινωνικοοικονομική κατάσταση; δυσμενείς εμπειρίες

1. ΕΙΣΑΓΩΓΗ

Η κατάθλιψη είναι ένα σημαντικό θέμα που αφορά στην δημόσια υγεία για πολλές χώρες και επιβαρύνει ιδιαίτερα τα άτομα, τις οικογένειες τους και την κοινωνία στο σύνολο της (WHO, 2002). Έχει αποδειχθεί ότι σχετίζεται με αυξημένη νοσηρότητα και θνησιμότητα από χρόνιες παθήσεις (Farrokhi et al., 2014) καθώς και με έκπτωση στην ποιότητα ζωής (D'Alisa et al., 2006).

Προηγούμενες μελέτες έχουν ομαδοποιήσει τους παράγοντες κινδύνου για την κατάθλιψη στην μεταγενέστερη ζωή σε διάφορες κατηγορίες, συμπεριλαμβανομένων των ψυχο-κοινωνικών, βιολογικών και κοινωνικο-δημογραφικών παραγόντων καθώς και των χαρακτηριστικών της προσωπικότητας. Οι ψυχο-κοινωνικοί παράγοντες

περιλαμβάνουν άγχη προερχόμενα από οικονομικές αντιξοότητες και άλλα δυσμενή βιώματα (Bruce, 2001) ενώ οι βιολογικοί παράγοντες περιλαμβάνουν κακή υγεία που μπορεί να προέρχεται από σοβαρά προβλήματα υγείας ή χρόνιο σωματικό πόνο. Στους κοινωνικο-δημογραφικούς παράγοντες κινδύνου ανήκουν η χαμηλή κοινωνικοοικονομική κατάσταση που εκφράζεται μέσα από το χαμηλό εισόδημα καθώς και το μορφωτικό επίπεδο (Mills & Henretta, 2001).

Άλλες έρευνες έχουν εξετάσει παράγοντες που σχετίζονται με την κατάθλιψη σε συγκεκριμένες χώρες ή για πολλές ευρωπαϊκές χώρες μαζί. Οι Alonso et al. (2004) μελέτησαν τις ψυχικές διαταραχές σε ερωτώμενους άνω των 18 ετών σε 6 χώρες, μεταξύ των οποίων το Βέλγιο, η Γαλλία, η Γερμανία, η Ιταλία, η Ολλανδία και η Ισπανία. Τα αποτελέσματα έδειξαν ότι οι γυναίκες είχαν διπλάσιες πιθανότητες να υποφέρουν από διαταραχές διάθεσης και άγχους σε σύγκριση με τους άνδρες. Επιπρόσθετα, το επίπεδο της εκπαίδευσης θεωρείται ένας άλλος ισχυρός προγνωστικός παράγοντας για την κατάθλιψη στην μεταγενέστερη ζωή. Η Ladin (2008) στην μελέτη που έκανε σε άτομα άνω των 50 ετών που προέρχονταν από 10 ευρωπαϊκές χώρες βρήκε ότι μεταξύ των ενηλίκων με χαμηλό μορφωτικό επίπεδο, η πιθανότητα να ασθενήσει κάποιος από κατάθλιψη ήταν περίπου διπλάσια της πιθανότητας να μην ασθενήσει. Συνακόλουθα, η χαμηλή κοινωνικοοικονομική κατάσταση των ατόμων όταν αξιολογείται από δείκτες υλικών προτύπων διαβίωσης συνδέεται με υψηλότερα επίπεδα της κατάθλιψης (Lorant et al., 2003). Τέλος, η διαμονή χωρίς τον σύντροφο αυξάνει τον αριθμό των καταθλιπτικών συμπτωμάτων ενώ η βελτίωση των σχέσεων μεταξύ των συντρόφων επιφέρει το αντίθετο αποτέλεσμα σύμφωνα με την μελέτη των Lorant et al. (2007) σε ερωτηθέντες άνω των 16 ετών.

Ιδιαίτερο ενδιαφέρον έχουν έρευνες που επικεντρώνονται στις διαφορές στην κατάθλιψη μεταξύ των φύλων. Οι Van de Velde et al. (2010) εξέτασαν το χάσμα που υπάρχει στην κατάθλιψη για άνδρες και γυναίκες ηλικίας 15 ετών και άνω σε 23 ευρωπαϊκές χώρες. Οι γυναίκες ανέφεραν υψηλότερα επίπεδα κατάθλιψης σε όλες τις χώρες και οι διαφορές ήταν πιο έντονες στην Ανατολική και Νότια Ευρώπη. Στην πλειονότητα των χωρών, οι κοινωνικοοικονομικοί παράγοντες είχαν τη μεγαλύτερη συσχέτιση με την κατάθλιψη τόσο για άνδρες όσο και για γυναίκες.

Συγκριτικά με τις προηγούμενες έρευνες, η παρούσα εργασία έχει σαν σκοπό να καλύψει το κενό της εκτίμησης προγνωστικών παραγόντων για την κατάθλιψη από διάφορους τομείς της ζωής (υγείας, κοινωνικοοικονομικής κατάστασης, δυσάρεστων εμπειριών) και σε όλο το χρονικό φάσμα ζωής των ατόμων ηλικίας 50 ετών και άνω που διαμένουν στην Ευρώπη. Σαν κύριοι στόχοι της μελέτης τίθενται οι διαφοροποιήσεις μεταξύ των ευρωπαϊκών ομάδων αλλά και μεταξύ των φύλων αυτών.

2. ΔΕΔΟΜΕΝΑ ΚΑΙ ΜΕΘΟΔΟΙ

2.1 Δεδομένα

Για την στατιστική ανάλυση χρησιμοποιήθηκαν δεδομένα από την ηλεκτρονική βάση για την Έρευνα, την Υγεία και την Συνταξιοδότηση στην Ευρώπη (SHARE). Το SHARE είναι μία διαχρονική (longitudinal) και διακρατική (cross national) βάση δεδομένων που περιέχει μικροδεδομένα (micro data) που αφορούν σε δημογραφικά χαρακτηριστικά καθώς και χαρακτηριστικά για την υγεία και την κοινωνικοοικονομική κατάσταση ερωτώμενων ηλικίας 50 ετών και άνω με προέλευση από διάφορες ευρωπαϊκές χώρες (Börsch–Supan et al., 2013). Για τους σκοπούς της παρούσας μελέτης το δείγμα που επιλέχθηκε αναφέρεται σε διακρατικά δεδομένα από το δεύτερο κύμα της έρευνας που πραγματοποιήθηκε το 2006-2007 σε συνδυασμό με αναδρομικές πληροφορίες από το τρίτο κύμα της έρευνας (SHARELIFE) που πραγματοποιήθηκε το 2008-2009. Το δείγμα περιλαμβάνει 22667 ερωτώμενους, και συγκεκριμένα 9997 άνδρες και 12670 γυναίκες από 13 διαφορετικές ευρωπαϊκές χώρες που συμμετέχουν αμφότερα στα δύο κύματα. Ειδικότερα, τα άτομα προέρχονται από την Βόρεια Ευρώπη (Σουηδία και Δανία), την Νότια Ευρώπη (Ισπανία, Ιταλία και Ελλάδα), την Κεντρική Ευρώπη (Αυστρία, Γερμανία, Ολλανδία, Γαλλία, Ελβετία και Βέλγιο) και την Ανατολική Ευρώπη (Τσεχία και Πολωνία).

Η επιλογή των ερωτώμενων από τις ευρωπαϊκές χώρες έγινε με διαφορετικά δειγματοληπτικά σχέδια που να είναι αντιπροσωπευτικά για κάθε χώρα. Πιο συγκεκριμένα, η βάση δεδομένων του SHARE παρέχει βάρη για τον σχεδιασμό δειγματοληψίας προκειμένου οι εκτιμητές των παραμέτρων του πληθυσμού να είναι αμερόληπτοι. Στην παρούσα εργασία όπου το μέγεθος του δείγματος συνολικά αλλά και στις επιμέρους ομάδες χωρών είναι μεγάλο, το σφάλμα μεροληψίας των εκτιμητών τείνει στο μηδέν και για αυτό δεν κρίθηκε απαραίτητη η στάθμιση λόγω δειγματοληπτικού σχεδίου.

Συνακόλουθα, το αρχικό δείγμα των ερωτώμενων ήταν 24431 άτομα εκ των οποίων τα 1764 (7.2%) αποτελούν ελλείπουσες τιμές. Για αυτές δεν χρησιμοποιήθηκε κάποια μέθοδος υποκατάστασης τους (imputation method). Οι χώρες με τα υψηλότερα ποσοστά ελλειπουσών τιμών στην παρούσα εργασία ήταν η Γαλλία (14.9%), η Ισπανία (11.5%) και η Ελλάδα (9.3%). Επιπρόσθετα, με βάση τους παράγοντες που χρησιμοποιήθηκαν στην ανάλυση και που θα περιγραφούν παρακάτω, το βίωμα μιας περιόδου άγχους έχει τις περισσότερες ελλείπουσες τιμές (29.8%) ενώ παράγοντες που σχετίζονται με προσωπικές εκτιμήσεις των ερωτώμενων όπως «προσδιοριζόμενη υγεία» είχε τις λιγότερες ελλείπουσες τιμές (2.5%).

2.2 Εξαρτημένη μεταβλητή

Η μεταβλητή απόκρισης είναι η κατάθλιψη σε δυαδική μορφή και συγκρίνει τους ερωτώμενους που δήλωσαν ότι πάσχουν από 4 και άνω συμπτώματα και συνεπώς υποφέρουν από κατάθλιψη με τους ερωτώμενους που δήλωσαν από 0 μέχρι 3 συμπτώματα και οι οποίοι θεωρούνται υγιείς, χωρίς ενδείξεις καταθλιπτικών

συμπτωμάτων (ως κατηγορία αναφοράς έχει οριστεί αυτή με τα 0-3 συμπτώματα). Η μέτρηση της κατάθλιψης στο SHARE γίνεται βάσει της κλίμακας EURO-D που περιλαμβάνει τα εξής 12 στοιχεία: καταθλιπτική διάθεση, πεσιμιστική διάθεση, τάσεις αυτοκτονίας, αίσθημα ενοχής, απώλεια ύπνου, έλλειψη ενδιαφέροντος, εριστική διάθεση, απώλεια όρεξης, κόπωση, έλλειψη συγκέντρωσης, έλλειψη ευχαρίστησης και δακρύρροια και έχει χρησιμοποιηθεί ευρέως από κοινωνικούς επιστήμονες διαφόρων χωρών ως δείκτης κατάθλιψης (Börch-Supan et al., 2005).

2.3 Ανεξάρτητες μεταβλητές

Οι ανεξάρτητες μεταβλητές περιλαμβάνουν επεξηγηματικούς παράγοντες από την παιδική ηλικία, την ενηλικίωση και την μεταγενέστερη ζωή.

i) *Παράγοντες στην παιδική ηλικία.* Η προκείμενη κατηγορία μεταβλητών περιλαμβάνει χαρακτηριστικά για την υγεία των ερωτώμενων στα πρώτα χρόνια της ζωής τους (μέχρι την ηλικία των 15 ετών), την εκπαίδευση και γνωστική λειτουργία που ανέπτυξαν, ενδεχόμενες δυσμενείς εμπειρίες που βίωσαν καθώς και το κοινωνικοοικονομικό επίπεδο των γονέων. Η υγεία των ατόμων μετριέται με την μεταβλητή της αυτό-προσδιοριζόμενης υγείας στην παιδική ηλικία που είναι κατηγορική διατάξιμη με πέντε επίπεδα: άριστη που είναι η κατηγορία αναφοράς, πολύ καλή, καλή, μέτρια και φτωχή. Ωστόσο, στα πλαίσια της ανάλυσης, τα επίπεδα που αναφέρονται στην πολύ καλή και καλή υγεία συγχωνεύθηκαν εξαιτίας μικρού πλήθους συχνοτήτων με αποτέλεσμα να χρησιμοποιηθούν τέσσερα επίπεδα για την μεταβλητή. Η διαδικασία της ενοποίησης των κατηγοριών με μικρές συχνότητες είναι συνήθης διαδικασία που συχνά βελτιώνει την αξιοπιστία της ανάλυσης (Linacre, 2002). Η εκπαίδευση/γνωστική λειτουργία των ερωτώμενων μετριέται με το πλήθος των βιβλίων στα οποία είχε πρόσβαση το άτομο στην ηλικία των 10 ετών (ταξινομημένη σε πέντε κατηγορίες: κανένα ή πολύ λίγα: 0-10 βιβλία, αρκετά για να γεμίσει ένα ράφι: 11-25 βιβλία, αρκετά για να γεμίσει μία βιβλιοθήκη: 26-100 βιβλία και αρκετά για να γεμίσουν δύο ή περισσότερες βιβλιοθήκες: 200 βιβλία ή περισσότερα, ωστόσο η μεταβλητή στην ανάλυση των δεδομένων θα αντιμετωπιστεί ως συνεχής. Παρόμοιο εγχείρημα με την χρήση αυτής της μεταβλητής σε συνεχή μορφή έχει γίνει και από άλλους ερευνητές (Van Bergen et al., 2017). Επίσης η εκπαίδευση αντιπροσωπεύεται και από την μεταβλητή που συγκρίνει τις επιδόσεις που είχε ο ερωτώμενος στα Μαθηματικά σε σχέση με τους συμμαθητές του (ταξινομημένη σε τέσσερις κατηγορίες: πολύ καλύτερη που είναι η κατηγορία αναφοράς, καλύτερη, σχεδόν ίδια, χειρότερη και πολύ χειρότερη). Η τελευταία κατηγορία έχει προκύψει και πάλι από την συγχώνευση των δύο επιμέρους κατηγοριών: χειρότερη και πολύ χειρότερη, εξαιτίας του μικρού αριθμού συχνοτήτων. Οι δυσάρεστες εμπειρίες αντιπροσωπεύονται από το αν οι γονείς των ερωτώμενων κατανάλωναν μεγάλες ποσότητες αλκοόλ. Είναι δίτιμη μεταβλητή με κατηγορία αναφοράς το αρνητικό ενδεχόμενο. Τέλος, η κοινωνικοοικονομική κατάσταση των ατόμων στην παιδική ηλικία εκφράζεται από το επάγγελμα του κυρίου οικογενειάρχη του νοικοκυριού όταν ο ερωτώμενος ήταν 10 ετών, μία δίτιμη μεταβλητή που διακρίνει τα επαγγέλματα σε χειρωνακτικά, που είναι η κατηγορία αναφοράς, και μη χειρωνακτικά.

ii) *Παράγοντες στην ενηλικίωση*. Σε αυτήν την κατηγορία μεταβλητών εντάσσονται τυχόν δυσάρεστες εμπειρίες που βίωσαν οι ερωτώμενοι στην ενήλικη τους ζωή εκπεφρασμένες με την μορφή δυαδικών μεταβλητών. Χαρακτηριστικά, εδώ ανήκουν η περίοδος του άγχους που μπορεί να έζησε ο ερωτώμενος για κάποιο διάστημα και αναφορικά με την υγεία του κάποια περίοδος που παρατηρήθηκε ότι είχε φτωχή υγεία. Επιπλέον, σαν δείκτες κοινωνικοοικονομικής κατάστασης στην ενηλικίωση χρησιμοποιήθηκαν το μορφωτικό επίπεδο (ταξινομημένο σε τρεις κατηγορίες: 0-6 έτη σχολικής φοίτησης, 7-12 έτη και περισσότερα από 13 έτη φοίτησης σε σχολείο και άλλα ιδρύματα με κατηγορία αναφοράς την τελευταία) καθώς και οι περίοδοι οικονομικής δυσπραγίας και πείνας. Όλες οι μεταβλητές των δυσμενών εμπειριών θεωρούν σαν κατηγορία αναφοράς το αρνητικό ενδεχόμενο.

iii) *Παράγοντες στην μεταγενέστερη ζωή*. Οι προβλεπτικοί παράγοντες που ανήκουν στην όψιμη ενήλικη ζωή περιλαμβάνουν κοινωνικοδημογραφικά χαρακτηριστικά των ερωτώμενων και στοιχεία για την πρόσφατη υγεία τους και την κατάσταση της γνωστικής λειτουργίας τους. Πιο συγκεκριμένα, η σωματική υγεία αντιπροσωπεύεται από έναν δείκτη που δηλώνει αν ο ερωτώμενος ανέφερε δύο κατ' ελάχιστο φυσικά συμπτώματα με χρονική διάρκεια τουλάχιστον έξι μηνών από έναν κατάλογο με δώδεκα συμπτώματα που περιλαμβάνει 1) πόνο στην μέση, στα γόνατα, στα ισχία ή σε οποιαδήποτε άλλη άρθρωση 2) καρδιακό πρόβλημα ή στηθάγχη, θωρακικό πόνο ή πόνο κατά την άσκηση 3) δύσπνοια, δυσκολία στην αναπνοή 4) επίμονο βήχα 5) πρησμένα πόδια 6) προβλήματα ύπνου 7) πτώσεις ή πεσίματα 8) φόβο για πτώσεις 9) ζάλη, λιποθυμίες ή σκοτοδίνη 10) προβλήματα στομάχου ή εντέρου, συμπεριλαμβανομένων της δυσκοιλιότητας, του μετεωρισμού, της διάρροιας 11) ακράτεια ούρων ή ακούσια απώλεια ούρων και 12) κόπωση. Επιπρόσθετα, συμπεριλαμβάνεται και η αυτοπροσδιοριζόμενη υγεία που εκφράζει την υποκειμενική υγεία των ατόμων σε πέντε διαφορετικά επίπεδα: εξαιρετική (1), πολύ καλή (2), καλή (3), μέτρια (4) και φτωχή (5). Παρόλα αυτά η μεταβλητή χρησιμοποιείται στην ανάλυση ως συνεχής. Η επιλογή χρήσης της μεταβλητής σε συνεχή μορφή έχει υιοθετηθεί και από άλλους ερευνητές (Paul et al., 2016). Συνακόλουθα, η γνωστική λειτουργία είναι συνεχής και μετράται με μία βαθμολογία που αντανακλά την ικανότητα του ερωτώμενου να προσδιορίζει σωστά τον χρόνο, κυμαινόμενη από 0 (κακή) μέχρι 4 (καλή). Επίσης, η τρέχουσα κοινωνικοοικονομική κατάσταση των ερωτώμενων δηλώνεται με μία κατηγορική μεταβλητή που εξετάζει αν το νοικοκυριό μπορούσε να ανταποκριθεί στις οικονομικές απαιτήσεις τον προηγούμενο χρόνο της έρευνας με μεγάλη δυσκολία, με κάποια δυσκολία, αρκετά εύκολα ή εύκολα που είναι και η κατηγορία αναφοράς. Τέλος, χρησιμοποιείται και η δυαδική δημογραφική μεταβλητή που δηλώνει αν το άτομο διαμένει στο νοικοκυριό με σύντροφο που αποτελεί και την κατηγορία αναφοράς ή μόνο του.

2.4 Στατιστική ανάλυση

Η ανάλυση των δεδομένων περιλαμβάνει την εφαρμογή μοντέλων λογιστικής παλινδρόμησης με εξίσωση

$$\log(p_i/(1-p_i)) = \log(odds) = b_0 + b_1 X_{i1} + \ldots + b_{16} X_{i16},$$

για $i = 1, 2, \ldots N$ όπου N είναι το μέγεθος του δείγματος και X_{ij}, $j = 1, 2, \ldots, 16$ είναι οι επεξηγηματικές μεταβλητές του μοντέλου όπως έχουν περιγραφεί παραπάνω και οι οποίες διευκρινίζονται στους πίνακες. Το odds ορίζεται ως ο λόγος (ή η αναλογία) των συμπληρωματικών πιθανοτήτων. Αντίστοιχα, η έννοια Odds Ratio (OR) που χρησιμοποιείται παρακάτω στην ανάλυση των αποτελεσμάτων είναι ο λόγος των odds ή διαφορετικά, λόγος των σχετικών πιθανοτήτων. Τα αποτελέσματα που παρουσιάζονται στους πίνακες που ακολουθούν προκύπτουν από το μοντέλο λαμβάνοντας υπόψη την πληροφορία από τους παράγοντες από όλες τις περιόδους ζωής των ατόμων, από την παιδική ηλικία μέχρι και την μεταγενέστερη ζωή και εκτιμά τους παράγοντες που μπορούν να επηρεάσουν την κατάθλιψη σε πολύ μεγάλες ηλικίες. Η ανάλυση διεξήχθη χωριστά για τέσσερις κύριες ομάδες χωρών, που αντιπροσωπεύουν τη Βόρεια, Νότια, Κεντρική και Ανατολική Ευρώπη. Επιπλέον, τα μοντέλα διεξήχθησαν ξεχωριστά ανά φύλο για τον προσδιορισμό των διαφορών μεταξύ ανδρών και γυναικών. Οι υπολογισμοί πραγματοποιήθηκαν χρησιμοποιώντας το SPSS, έκδοση 20. Δεν υπήρχε ένδειξη πολυσυγγραμμικότητας, καθώς οι συντελεστές VIF (Variance Inflation Factors) των ανεξάρτητων μεταβλητών ανήκαν στο διάστημα (1.023 – 1.461) και ως εκ τούτου κανένας δεν ξεπερνούσε την τιμή 2. Οι p-τιμές των ελέγχων που διεξήχθησαν συγκρίθηκαν με τρία διαφορετικά επίπεδα σημαντικότητας (τιμές άλφα ή αλλιώς πιθανότητες να έχουμε σφάλμα τύπου Ι), τις εξής: 0.01, 0.05 και 0.10. Αν για μία μεταβλητή του μοντέλου ισχύει «p-τιμή του ελέγχου της μεταβλητής<0.01 (ή <0.05 ή <0.10)» τότε συμπεραίνουμε ότι «η μεταβλητή είναι στατιστικά σημαντική σε επίπεδο σημαντικότητας 1% (ή 5% ή 10%)». Η σημαντικότητα και η αξιοπιστία της προσαρμογής των μοντέλων ελέγχθηκε με βάση τους πίνακες ταξινόμησης που δείχνουν το ποσοστό των περιπτώσεων της εξαρτώμενης μεταβλητής που είχαν προβλεφθεί σωστά από το μοντέλο καθώς και από την ευαισθησία και ειδικότητα του μοντέλου. Για όλα τα μοντέλα, η στατιστική ανάλυση έδειξε καλή προσαρμογή σύμφωνα με τις τιμές του στατιστικού Hosmer and Lemeshow και του ψευδο-συντελεστή προσδιορισμού Nagelkerke R^2 που αποτελεί μία προσαρμοσμένη έκδοση του ψευδο-συντελεστή προσδιορισμού Cox and Snell R^2, λαμβάνει τιμές ώστε να καλύπτει ολόκληρο το φάσμα από 0 έως 1 και δείχνει πόσο καλά εξηγείται η εξαρτημένη μεταβλητή του μοντέλου από τις ανεξάρτητες.

3. ΑΠΟΤΕΛΕΣΜΑΤΑ

3.1 ΠΕΡΙΓΡΑΦΙΚΗ ΑΝΑΛΥΣΗ

Ο πίνακας 1 δείχνει τα μεγέθη των δειγμάτων για τις ομάδες των χωρών που χρησιμοποιήθηκαν στην ανάλυση. Το συνολικό δείγμα περιλαμβάνει 22667

Πίνακας 1. *Μεγέθη δειγμάτων και ποσοστά κατάθλιψης για τις ομάδες των χωρών της Ευρώπης.*

Group of countries	Sample Size (N)	Sample based on total (%)	Sample based on groups (%)	Depression (%)
Europe	22667	100.0		23.6
Males	9997	44.1		14.9
Females	12670	55.9		30.4
North Europe	3417	15.1		15.1
Males	1535	15.4=1535/9997	44.9=1535/3417	9.0
Females	1882	14.9=1882/12670	55.1=1882/3417	20.1
Central Europe	9583	42.3		21.4
Males	4237	42.4=4237/9997	44.2=4237/9583	13.9
Females	5346	42.2=5346/12670	55.8=5346/9583	27.3
South Europe	6453	28.5		26.1
Males	2871	28.7=2871/9997	44.5=2871/6453	15.5
Females	3582	28.3=3582/12670	55.5=3582/6453	34.6
East Europe	3214	14.2		34.0
Males	1354	13.5=1354/9997	42.1=1354/3214	23.3
Females	1860	14.7=1860/12670	57.9=1860/3214	41.7

ερωτηθέντες και το 44.1% αυτών είναι άνδρες. Οι χώρες που συμμετέχουν στο SHARE έχουν ομαδοποιηθεί ώστε να καλύπτουν την Βόρεια, Κεντρική, Νότια και Ανατολική Ευρώπη όπως περιγράφονται παραπάνω στα δεδομένα. Είναι εμφανές ότι η Κεντρική και η Νότια Ευρώπη αντιστοιχούν στα μεγαλύτερα δείγματα (9583 και 6453 άτομα αντίστοιχα) ενώ η Βόρεια και η Ανατολική Ευρώπη αντιστοιχούν στα μικρότερα (3417 και 3214 αντίστοιχα). Σε όλες τις ομάδες χωρών τα ποσοστά των γυναικών είναι στα ίδια επίπεδα, γύρω στο 55%, με εξαίρεση την Ανατολική Ευρώπη που υπάρχει ένα προβάδισμα κοντά στο 58%. Η τελευταία στήλη του πίνακα 1 δείχνει το ποσοστό των ατόμων με κατάθλιψη, έχοντας δηλώσει τουλάχιστον τέσσερα σχετικά συμπτώματα, ανά φύλο, για κάθε ομάδα χωρών.

Στο συνολικό δείγμα, μεγαλύτερο ποσοστό γυναικών (30.4%), σχεδόν διπλάσιο από των ανδρών (14.9%) αναφέρει ότι υποφέρει από κατάθλιψη σε μεγάλες ηλικίες και η διαφορά είναι σημαντική για όλες τις ομάδες χωρών της Ευρώπης. Επιπλέον, οι αναλογίες των ατόμων με κατάθλιψη παρουσιάζουν μία αύξουσα πορεία από τον Βορρά προς την Ανατολή. Μόνο το 15.1% των Ευρωπαίων των Βόρειων Χωρών δήλωσε ότι έχει τουλάχιστον τέσσερα καταθλιπτικά συμπτώματα ενώ για την Κεντρική Ευρώπη το αντίστοιχο ποσοστό φθάνει το 21.4%. Σε ερωτώμενους από την Νότια Ευρώπη το ποσοστό της κατάθλιψης στην μεταγενέστερη ζωή ανέρχεται στο 26.1% ενώ στον πληθυσμό της Ανατολικής Ευρώπης φθάνει το 34%.

Οι επόμενοι πίνακες παρουσιάζουν περιγραφικά αποτελέσματα για τους προγνωστικούς παράγοντες της κατάθλιψης στην παιδική ηλικία, κατά την διάρκεια

της ζωής και σε μεταγενέστερο στάδιο ηλικίας για άνδρες και γυναίκες και για τις διάφορες ομάδες των χωρών.

Σχετικά με την παιδική ηλικία, οι γυναίκες φαίνεται να έχουν χειρότερη υγεία από ότι οι άνδρες και ταυτόχρονα, σε μεγαλύτερο βαθμό λιγότερο καλές επιδόσεις στα Μαθηματικά σε σύγκριση με συμμαθητές τους. Αναφορικά με τις υπόλοιπες μεταβλητές η εικόνα που λαμβάνουμε είναι ίδια και για τα δύο φύλα. Οι ερωτώμενοι συνολικά είχαν στην βιβλιοθήκη τους γύρω στα 11-25 βιβλία στην ηλικία των 10 ετών ενώ ένα ποσοστό της τάξεως του 8.3% δήλωσε ότι οι γονείς του κατανάλωναν μεγάλες ποσότητες αλκοόλ. Επιπρόσθετα, περίπου το 27% των ερωτώμενων είχαν γονείς που απασχολούνταν σε μη χειρωνακτικό επάγγελμα. Οι διαφοροποιήσεις που λαμβάνουμε εξετάζοντας τις ομάδες των χωρών δείχνουν ότι άτομα από την Βόρεια και Νότια Ευρώπη είχαν άριστη υγεία ως παιδιά ενώ πολύ υψηλά είναι και ποσοστά της πολύ καλής και καλής υγείας των ατόμων και στις τέσσερις ομάδες. Σχετικά με τις σχολικές επιδόσεις στα Μαθηματικά, τις χειρότερες και με μεγάλη διαφορά τις είχαν οι Ευρωπαίοι των Νότιων χωρών και τις καλύτερες οι Ευρωπαίοι από τις

Πίνακας 2. *Περιγραφικά στατιστικά στοιχεία (ποσοστά και διάμεσες τιμές) για τους παράγοντες στην παιδική ηλικία για τις διάφορες ομάδες χωρών*

	Ευρώπη	ΒΕ	ΚΕ	ΝΕ	ΑΕ
Φυσική και Ψυχική υγεία					
Προσδιοριζόμενη υγεία (sph) (X_1)					
Άριστη (κατ. αναφ.)	35.9	48.8	28.5	45.2	25.6
Πολύ καλή/Καλή	55.9	43.5	60.7	49.7	67.6
Αρκετά καλή	5.6	4.7	7.6	3.6	4.8
Φτωχή	2.5	3.0	3.2	1.5	2.0
Εκπαίδευση/Γνωστική λειτουργία					
Αριθμός βιβλίων στην ηλικία των 10 ετών (X_2)	2* 2**	3* 2**	2* 2**	1* 1**	2* 2**
Σχετικές επιδόσεις με τα άλλα παιδιά στα Μαθηματικά (X_3)					
Πολύ καλύτερη (κατ. αναφ.)	10.3	15.9	11.3	6.1	9.5
Καλύτερη	23.8	28.1	24.4	21.6	21.6
Σχεδόν ίδια	49.5	47.3	50.9	45.7	55.5
Χειρότερη και πολύ χειρότερη	16.5	8.6	13.4	26.7	13.3
Δυσμενείς εμπειρίες					
Βαριά κατανάλωση αλκοόλ από τους γονείς (X_4)					
Όχι (κατ. αναφ.)	91.7	91.0	91.5	91.9	93.0
Ναι	8.3	9.0	8.5	8.1	7.0
Παιδική Κοινωνικοοικονομική Κατάσταση					
Επάγγελμα οικογενειάρχη (X_5)					
Χειρωνακτικό (κατ. αναφ.)	72.7	67.1	67.5	80.1	79.4
Μη χειρωνακτικό	27.3	32.9	32.5	19.9	20.6
Μέγεθος δείγματος (N)	22667	3417	9583	6453	3214

* Διάμεση τιμή ** Ενδοτεταρτημοριακό εύρος

Βόρειες Χώρες. Ερωτώμενοι από την Ανατολική Ευρώπη έζησαν σε μικρότερο ποσοστό την δυσάρεστη εμπειρία των αλκοολικών γονέων (7%) ενώ άτομα από την Κεντρική και Βόρεια Ευρώπη είχαν τα μεγαλύτερα ποσοστά γονέων που ασκούσαν μη χειρωνακτικό επάγγελμα (περίπου 33%).

Αναφορικά με τα γεγονότα κατά την διάρκεια της ζωής οι γυναίκες συνολικά σε μεγαλύτερα ποσοστά έχουν βιώσει στο παρελθόν περιόδους άγχους (54.5%) και κυρίως αυτές που προέρχονται από την Κεντρική Ευρώπη (62.3%), φτωχής υγείας (42.9%) ιδιαίτερα όσες ανήκουν σε χώρες της Βόρειας Ευρώπης (47.2%) και οικονομικών δυσκολιών (34.6%) σε σχέση με τους άνδρες (ποσοστά ίσα με 47%, 39.5% και 30.7% αντίστοιχα). Η δυσμενής εμπειρία της περιόδου πείνας επικρατεί περισσότερο στους Ευρωπαίους άνδρες (8.1%) ιδίως σε όσους προέρχονται από την Νότια Ευρώπη (12.2%) συγκριτικά με τις Ευρωπαίες γυναίκες (6.9%), με το πιο χαμηλό ποσοστό να αναφέρεται σε αυτές που διαμένουν στην Βόρεια Ευρώπη (2.2%). Επιπλέον, το μορφωτικό επίπεδο με πάνω από 13 έτη εκπαίδευσης είναι υψηλό κυρίως για τους άνδρες (35.2%) συγκριτικά με τις γυναίκες (26.5%) και τα ποσοστά είναι μεγαλύτερα σε όσους προέρχονται από τις Βόρειες Χώρες (58.2% για τους άνδρες και 50.7% για τις γυναίκες). Τέλος, το μέσο μορφωτικό επίπεδο, που καλύπτει 7-12 έτη εκπαίδευσης, είναι πιο υψηλό για τις γυναίκες (53.3%) και ιδιαίτερα για αυτές που διαμένουν σε χώρες της Ανατολικής Ευρώπης (69.6%). Συνολικά αποτελέσματα για τις ομάδες των χωρών φαίνονται στον πίνακα 3.

Πίνακας 3. *Περιγραφικά στατιστικά στοιχεία (ποσοστά) για τους παράγοντες κατά την διάρκεια της ζωής για τις διάφορες ομάδες χωρών*

	Ευρώπη	ΒΕ	ΚΕ	ΝΕ	ΑΕ
Δυσάρεστες εμπειρίες					
Περίοδος στρες (X_6)					
Όχι (κατ. αναφ.)	48.8	46.2	40.2	59.5	55.8
Ναι	51.2	53.8	59.8	40.5	44.2
Φυσική υγεία					
Περίοδος φτωχής υγείας (X_7)					
Όχι (κατ. αναφ.)	58.6	55.9	54.9	65.6	58.4
Ναι	41.4	44.1	45.1	34.4	41.6
Κοινωνικοοικονομική Κατάσταση					
Εκπαιδευτικό επίπεδο (X_8)					
13+ έτη (κατ. αναφ.)	30.3	54.1	34.0	15.9	23.2
0-6 έτη	19.0	2.6	7.6	50.0	8.0
7-12 έτη	50.7	43.3	58.4	34.1	68.8
Περίοδος οικονομικών δυσκολιών (X_9)					
Όχι (κατ. αναφ.)	67.1	65.0	63.9	71.4	70.3
Ναι	32.9	35.0	36.1	28.6	29.7
Περίοδος πείνας (X_{10})					
Όχι (κατ. αναφ.)	92.2	97.9	90.9	90.0	94.0
Ναι	7.8	2.1	9.1	10.0	6.0
Μέγεθος δείγματος (N)	22667	3417	9583	6453	3214

Ολοκληρώνοντας την περιγραφική ανάλυση με τους παράγοντες στην μεταγενέστερη ζωή οι γυναίκες συνολικά δηλώνουν χειρότερη υγεία μόνο αναφορικά με το πλήθος των πάνω από δύο συμπτωμάτων φυσικής υγείας που μπορεί να αντιμετωπίζουν (ποσοστό ίσο με 48.3%) σε σχέση με τους άνδρες (ποσοστό ίσο με 33.5%) ενώ και τα δύο φύλα παρουσιάζουν την ίδια εικόνα ως προς την αυτοπροσδιοριζόμενη υγεία την οποία θεωρούν καλή καθώς και ως προς τον προσδιορισμό της ημερομηνίας και του χρόνου (επίπεδο 4) γεγονός που υποδηλώνει πολύ καλή γνωστική λειτουργία. Ειδικότερα, το μεγαλύτερο πλήθος συμπτωμάτων έχουν οι άνδρες και οι γυναίκες από την Ανατολική Ευρώπη με ποσοστά ίσα με 47.9% και 61.7%, αντίστοιχα. Παρόμοια αποτελέσματα δίνει και για τα δύο φύλα ο παράγοντας που αναφέρεται στην δυνατότητα του νοικοκυριού να ανταπεξέλθει στις οικονομικές απαιτήσεις της οικογένειας. Οι γυναίκες και οι άνδρες που προέρχονται από την Βόρεια Ευρώπη έχουν τα υψηλότερα ποσοστά που δηλώνουν εύκολη οικονομική διαχείριση, ίσα με 50.9% και 55.3%, και τα χαμηλότερα ποσοστά σε οικονομικές καταστάσεις μεγάλης δυσκολίας, ίσα με 2.4% και 1.5% αντίστοιχα. Συνακόλουθα, μεγάλες οικονομικές δυσκολίες αντιμετωπίζουν οι ερωτώμενοι από χώρες της Νότιας Ευρώπης με ποσοστά ίσα με 19.3% για τους άνδρες και 22.5% για τις γυναίκες. Η παρουσία συντρόφου παρατηρείται σε μεγαλύτερα ποσοστά στους άνδρες Ευρωπαίους σε σχέση με τις γυναίκες αφού μόνο το 16.4% αυτών μένουν μόνοι τους έναντι του 34.2% των γυναικών. Τα υψηλότερα ποσοστά μοναξιάς στις γυναίκες εμφανίζονται σε αυτές που προέρχονται από χώρες της Ανατολικής (37%) και Κεντρικής (35.3%) Ευρώπης ενώ για τους άνδρες σε όσους διαμένουν σε χώρες της Βόρειας (17.8%) και Κεντρικής (17.7%) Ευρώπης και είναι κατά πολύ μικρότερα από αυτά των γυναικών. Τέλος, η διάμεση ηλικία για τις γυναίκες συνολικά είναι 64 έτη και για τους άνδρες τα 65 έτη. Την χαμηλότερη διάμεση ηλικία έχουν οι ερωτώμενοι από χώρες της Ανατολικής Ευρώπης και συγκεκριμένα είναι τα 63 έτη για τις γυναίκες και τα 64 για τους άνδρες. Στον πίνακα 4 που ακολουθεί βλέπουμε πως διαφοροποιούνται τα αποτελέσματα για τις διάφορες ομάδες των χωρών ανεξάρτητα από το φύλο.

3.2 Αποτελέσματα μοντέλων λογιστικής παλινδρόμησης

Το συνολικό μοντέλο λογιστικής παλινδρόμησης εντόπισε τους στατιστικά σημαντικούς παράγοντες που δύνανται να αυξήσουν ή να μειώσουν τα επίπεδα της κατάθλιψης στην μεταγενέστερη ζωή.

Σχετικά με την παιδική ηλικία το μοντέλο δείχνει ως σημαντικό παράγοντα για όλο το δείγμα των Ευρωπαίων το πλήθος βιβλίων που είχε ο ερωτώμενος στην ηλικία των 10 ετών. Χαρακτηριστικά, τα άτομα με έφεση στην μελέτη στην παιδική ηλικία έχουν κατά 8.3% μειωμένη την πιθανότητα να εμφανίσουν κατάθλιψη σε μεγάλες ηλικίες από ότι να μην την εμφανίσουν (OR=0.92, 95%ΔΕ:0.9-1.0, p<0.001) και ιδιαίτερα ο πληθυσμός της Ανατολικής Ευρώπης στον οποίο επιτυγχάνεται η μεγαλύτερη μείωση τόσο συνολικά (OR=0.82, 95%ΔΕ:0.7-0.9, p<0.001) όσο και για τα δύο φύλα ξεχωριστά ((OR=0.76, 95%ΔΕ:0.6-0.9, p=0.001) για τους άνδρες και (OR=0.84, 95%ΔΕ:0.8-0.9, p=0.003) για τις γυναίκες). Συνακόλουθα, οι επιδόσεις των ερωτώμενων στα Μαθηματικά συγκριτικά με τους συμμαθητές τους είναι σημαντικός παράγοντας παρουσίας της κατάθλιψης στην μεταγενέστερη ζωή. Στους

Πίνακας 4. *Περιγραφικά στατιστικά στοιχεία (ποσοστά και διάμεσες τιμές) για τους παράγοντες στην μεταγενέστερη ζωή για τις διάφορες ομάδες χωρών*

	Ευρώπη	ΒΕ	ΚΕ	ΝΕ	ΑΕ
Τρέχουσα υγεία					
Συμπτώματα (X_{11})					
Λιγότερα από δύο (κατ. αναφ.)	58.2	62.0	60.0	60.6	44.1
2+ συμπτώματα	41.8	38.0	40.0	39.4	55.9
Αυτοπροσδιοριζόμενη υγεία (sph) (X_{12})	3* 2**	2* 1**	3* 2**	3* 2**	3* 1**
Γνωστική Λειτουργία					
Προσδιορισμός του χρόνου (X_{13})	4*	4*	4*	4*	4*
0 (κακή) – 4 (καλή)					
Κοινωνικοδημογραφικά Χαρακτηριστικά					
Το νοικοκυριό ανταποκρίνεται (X_{14})					
Εύκολα (κατ. αναφ.)	27.3	52.9	37.4	8.9	6.7
Με μεγάλη δυσκολία	11.6	2.0	6.0	21.1	19.6
Με μερική δυσκολία	28.2	10.9	18.9	42.5	45.5
Αρκετά εύκολα	33.0	34.2	37.8	27.6	28.2
Σύντροφος στο νοικοκυριό (X_{15})					
Διαμονή με σύντροφο (κατ. αναφ.)	73.7	74.3	72.5	76.0	71.8
Διαμονή χωρίς σύντροφο	26.3	25.7	27.5	24.0	28.2
Ηλικία του ερευνώμενου (X_{16})	65*	65*	65*	65*	63*
Μέγεθος δείγματος (N)	22667	3417	9583	6453	3214

* Διάμεση τιμή ** Ενδοτεταρτημοριακό εύρος

ερωτώμενους από την Βόρεια Ευρώπη με τις χειρότερες επιδόσεις στα Μαθηματικά, η πιθανότητα εμφάνισης της νόσου είναι 2.18 φορές μεγαλύτερη της πιθανότητας μη εμφάνισης (OR=2.18, 95%ΔΕ:1.4-3.4, p<0.001) σε σχέση με ερωτώμενους που είχαν καλύτερες επιδόσεις. Συνακόλουθα, η πιθανότητα να εμφανιστεί η κατάθλιψη παραμένει εξίσου υψηλή και για τους κατοίκους της Νότιας Ευρώπης, ίση με 1.80 φορές μεγαλύτερη της πιθανότητας να μην εμφανιστεί (OR=1.80, 95%ΔΕ:1.3-2.5, p=0.001). Η εικόνα ανατρέπεται στην περίπτωση των ανδρών καθώς αυτός ο παράγοντας παρουσιάζεται στατιστικά μη σημαντικός ενώ φαίνεται να επηρεάζει μόνο τις γυναίκες και ιδιαίτερα αυτές από την Βόρεια και Νότια Ευρώπη ((OR=2.78, 95%ΔΕ:1.6-4.8, p<0.001) και (OR=2.19, 95%ΔΕ:1.4-3.4, p=0.001) αντίστοιχα). Η κατανάλωση αλκοόλ από τους γονείς ως δυσμενής εμπειρία στην παιδική ηλικία είναι καθοριστικός παράγοντας για τους ερωτώμενους από την Βόρεια, Κεντρική και Νότια Ευρώπη και ειδικότερα τις γυναίκες. Αυτές που διαμένουν σε χώρες της Βόρειας Ευρώπης έχουν 1.80 φορές μεγαλύτερη πιθανότητα εμφάνισης του χαρακτηριστικού της κατάθλιψης σε σχέση με την πιθανότητα μη εμφάνισης σε πολύ μεγάλες ηλικίες (OR=1.80, 95%ΔΕ:1.2-2.6, p=0.003) ενώ για τις γυναίκες από την Κεντρική Ευρώπη η πιθανότητα εμφάνισης της νόσου είναι 1.42 φορές μεγαλύτερη της πιθανότητας μη εμφάνισης (OR=1.42, 95%ΔΕ:1.1-1.8, p=0.002). Η υγεία όπως προσδιορίζεται από τους ερωτώμενους καθώς και το επάγγελμα των γονέων των

ατόμων δεν φαίνεται να είναι στατιστικά σημαντικοί παράγοντες για τις διάφορες ομάδες των ευρωπαϊκών χωρών.

Αναφορικά με τα γεγονότα κατά την διάρκεια της ζωής το βίωμα του άγχους για μία περίοδο είναι σημαντικός παράγοντας για όλους τους Ευρωπαίους από κάθε ομάδα χωρών και για τα δύο φύλα. Την μικρότερη επίδραση από αυτόν τον παράγοντα υφίστανται οι γυναίκες από την Ανατολική Ευρώπη (OR=1.42, 95%ΔΕ:1.1-1.8, p=0.004) ενώ οι γυναίκες από την Κεντρική και Νότια Ευρώπη επηρεάζονται κατά τον ίδιο τρόπο (αφού οι πιθανότητες επικράτησης της κατάθλιψης σε μεγάλες ηλικίες είναι 1.58 (OR=1.58, 95%ΔΕ:1.4-1.8, p<0.001) και 1.57 φορές αντίστοιχα (OR=1.57, 95%ΔΕ:1.3-1.9, p<0.001) μεγαλύτερες της συμπληρωματικής πιθανότητας). Η περίοδος φτωχής υγείας προκαλεί μικρή επίδραση με εξαίρεση τους κατοίκους της Ανατολικής Ευρώπης, καθώς η πιθανότητα μελλοντικής κατάθλιψης είναι 1.38 φορές αυξημένη της πιθανότητας να μην συμβεί (OR=1.38, 95%ΔΕ:1.1-1.7, p=0.001). Ιδιαίτερα πλήττονται οι άνδρες της Κεντρικής Ευρώπης (OR=1.33, 95%ΔΕ:1.1-1.6, p=0.008) και οι γυναίκες της Ανατολικής Ευρώπης για τις οποίες η πιθανότητα να εμφανιστεί κατάθλιψη σε μεγάλες ηλικίες είναι 1.42 φορές μεγαλύτερη της πιθανότητας να μην εμφανιστεί (OR=1.42, 95%ΔΕ:1.1-1.8, p=0.003). Η περίοδος οικονομικών δυσκολιών επιβαρύνει τους άνδρες Ευρωπαίους, κυρίως αυτούς από την Ανατολική Ευρώπη οι οποίοι έχουν κατά 62.8% αυξημένη την πιθανότητα να παρουσιάσουν την νόσο από ότι να μην την παρουσιάσουν (OR=1.63, 95%ΔΕ:1.2-2.3, p=0.006) ενώ δεν δείχνει να είναι σημαντικός παράγοντας για τις γυναίκες. Τέλος, η εμπειρία της περιόδου πείνας αποδεικνύεται σημαντική μόνο για τις γυναίκες της Ανατολικής Ευρώπης αφού έχουν 2.10 φορές μεγαλύτερη πιθανότητα να νοσήσουν μελλοντικά ως προς το να μην νοσήσουν (OR=2.10, 95%ΔΕ:1.3-3.5, p=0.004) σε σχέση με τις γυναίκες που δεν βίωσαν περίοδο πείνας στην ζωή τους.

Οι παράγοντες στην μεταγενέστερη ζωή είναι οι πιο σημαντικοί για την επικράτηση της κατάθλιψης σε όλες τις ομάδες χωρών και για τα δύο φύλα. Πιο συγκεκριμένα, η παρουσία δύο κατ' ελάχιστο συμπτωμάτων δίνει τα μεγαλύτερα OR εμφάνισης της κατάθλιψης σε μεγάλες ηλικίες. Ιδιαίτερα για τον πληθυσμό της Βόρειας Ευρώπης η τιμή αυτή ανέρχεται στο 3.34, 95%ΔΕ:2.6-4.3, p<0.001. Αντίστοιχη είναι η εικόνα και για τα δύο φύλα με τους άνδρες της Βόρειας Ευρώπης και τις γυναίκες της Κεντρικής Ευρώπης να έχουν τα μεγαλύτερα OR, ίσα με 4.37, 95%ΔΕ:2.8-6.8, p<0.001 και 3.21, 95%ΔΕ:2.8-3.7, p<0.001. Επιπλέον, μεγάλες τιμές για την αυτοπροσδιοριζόμενη υγεία (sph) έχουν σαν αποτέλεσμα να αυξάνει κατά πολύ η πιθανότητα εμφάνισης της κατάθλιψης σε όλες τις χώρες της Ευρώπης, ιδίως δε στις Νότιες όπου η εν λόγω πιθανότητα είναι 1.96 φορές μεγαλύτερη της πιθανότητας να μην συμβεί (OR=1.96, 95%ΔΕ:1.8-2.1, p<0.001) αλλά και ανάμεσα στα φύλα, κυρίως σε άνδρες από την Ανατολική Ευρώπη (OR=2.28, 95%ΔΕ:1.9-2.9, p<0.001) και γυναίκες από την Νότια Ευρώπη (OR=1.97, 95%ΔΕ:1.8-2.2, p<0.001). Η ικανότητα προσδιορισμού του χρόνου είναι παράγοντας μείωσης των σχετικών πιθανοτήτων της κατάθλιψης σε όλες τις ομάδες χωρών. Μάλιστα, σε χώρες της Βόρειας Ευρώπης η μείωση της πιθανότητας να συμβεί μελλοντικά η νόσος ως προς το να μην συμβεί ανέρχεται στο 19.4% (OR=0.81, 95%ΔΕ:0.6-1.0, p=0.079) ενώ σε

χώρες της Νότιας Ευρώπης αγγίζει το 30.4% (OR=0.70, 95%ΔΕ:0.6-0.8, p<0.001). Σχετικά με την οικονομική κατάσταση του νοικοκυριού όσοι ερωτώμενοι αντιμετωπίζουν μεγάλες δυσκολίες έχουν και τις μεγαλύτερες σχετικές πιθανότητες εμφάνισης κατάθλιψης. Ξεχωρίζουν τα άτομα από την Βόρεια Ευρώπη με OR=3.57, 95%ΔΕ:2.0-6.4, p<0.001 και ιδιαίτερα οι άνδρες με OR=10.56, 95%ΔΕ:3.7-30.1, p<0.001. Τέλος, η απουσία του συντρόφου δεν φαίνεται να επηρεάζει καθόλου τις γυναίκες ως προς την εμφάνιση της κατάθλιψης σε μεγάλη ηλικία ενώ από τους άνδρες, μόνο αυτούς που διαμένουν στην Κεντρική και Νότια Ευρώπη, καθώς η πιθανότητα εμφάνισης της νόσου είναι 1.48 (OR=1.48, 95%ΔΕ:1.2-1.9, p=0.001) και 1.44 φορές αντίστοιχα μεγαλύτερη της πιθανότητας μη εμφάνισης (OR=1.44, 95%ΔΕ:1.1-1.9, p=0.017).

Η τιμή του ψευδο-συντελεστή προσδιορισμού Nagelkerke R^2 κυμαίνεται από 0.239 μέχρι 0.360 για τους άνδρες και από 0.259 μέχρι 0.326 για τις γυναίκες. Το ποσοστό των τιμών της εξαρτημένης μεταβλητής που προβλέφθηκε σωστά από το μοντέλο κυμαίνεται από 76.2% για τους ερωτώμενους της Ανατολικής Ευρώπης μέχρι 86.2% για τους ερωτώμενους της Βόρειας Ευρώπης. Στον πίνακα 5 βλέπουμε τις τιμές των OR για τους προγνωστικούς παράγοντες της κατάθλιψης στην μεταγενέστερη ζωή για τις διάφορες ευρωπαϊκές χώρες.

Πίνακας 5. *OR και 95% διαστήματα εμπιστοσύνης με βάση το μοντέλο λογιστικής παλινδρόμησης για τους στατιστικά σημαντικούς παράγοντες πρόβλεψης της κατάθλιψης από την παιδική ηλικία, την ενηλικίωση και την μεταγενέστερη ζωή για τις διάφορες ομάδες χωρών*

	Ευρώπη	BE	KE	NE	AE
Αριθμός βιβλίων στα 10 έτη	0.917a	1.068d	0.925a	0.874a	0.816a
	(0.883 0.952)	(0.974 1.171)	(0.877 0.976)	(0.795 0.962)	(0.743 0.895)
Σχετικές Μαθηματικές επιδόσεις					
Χειρότερη και πολύ χειρότερη	1.467a	2.184a	1.382a	1.803a	1.255d
	(1.261 1.706)	(1.423 3.351)	(1.100 1.736)	(1.282 2.535)	(0.856 1.839)
Βαριά κατανάλωση αλκοόλ από τους γονείς	1.339a	1.424b	1.426a	1.287b	1.233d
	(1.188 1.509)	(1.029 1.972)	(1.189 1.709)	(1.028 1.611)	(0.892 1.704)
Περίοδος στρες	1.542a	1.783a	1.550a	1.508a	1.356a
	(1.428 1.665)	(1.419 2.240)	(1.370 1.754)	(1.313 1.731)	(1.121 1.639)
Περίοδος φτωχής υγείας	1.151a	1.012c	1.164b	1.077d	1.380a
	(1.067 1.242)	(0.806 1.269)	(1.036 1.309)	(0.935 1.240)	(1.144 1.666)
Περίοδος οικονομικών δυσκολιών	1.149a	1.062d	1.102d	1.159c	1.367a
	(1.063 1.242)	(0.850 1.327)	(0.979 1.241)	(0.998 1.346)	(1.122 1.667)
Περίοδος πείνας	1.213a	1.148d	1.203b	1.198c	1.168d
	(1.070 1.374)	(0.629 2.095)	(1.000 1.447)	(0.971 1.480)	(0.816 1.672)
2+ Συμπτώματα	2.770a	3.337a	3.118a	2.387a	2.435a
	(2.562 2.995)	(2.618 4.254)	(2.761 3.522)	(2.078 2.742)	(1.992 2.975)
Αυτοπροσδιοριζόμενη υγεία (sph)	1.825a	1.603a	1.788a	1.959a	1.848a
	(1.747 1.905)	(1.433 1.794)	(1.670 1.914)	(1.804 2.127)	(1.653 2.067)

Προσδιορισμός του χρόνου	0.715a	0.806d	0.729a	0.696a	0.714a
	(0.667 0.767)	(0.633 1.025)	(0.652 0.816)	(0.620 0.782)	(0.591 0.862)
Το νοικοκυριό ανταποκρίνεται					
Με μεγάλη δυσκολία	1.968a	3.572a	1.715a	1.712a	2.391a
	(1.720 2.252)	(1.983 6.433)	(1.370 2.147)	(1.281 2.288)	(1.549 3.691)
Μέγεθος δείγματος (N)	22667	3417	9583	6453	3214

a p-τιμή<0.01 b p-τιμή<0.05 c p-τιμή<0.10 d μη σημαντικό

Σημείωση: Επιπλέον χαρακτηριστικά του μοντέλου των Ευρωπαίων (N=2667): τιμή του ψευδο-συντελεστή προσδιορισμού Nagelkerke R^2 = 0.316, p-τιμή με βάση το τεστ των Hosmer and Lemeshow = 0.108, ευαισθησία του μοντέλου = 38%, ειδικότητα του μοντέλου = 93.6%, ποσοστό ορθής ταξινόμησης = 80.5%

4. ΣΥΖΗΤΗΣΗ

Η παρούσα εργασία είχε σαν στόχο τον προσδιορισμό προγνωστικών παραγόντων της κατάθλιψης σε μελλοντικές ηλικίες για τα δύο φύλα και για διάφορες ομάδες ευρωπαϊκών χωρών λαμβάνοντας υπόψη όλες τις περιόδους ζωής των ενηλίκων, από τα παιδικά χρόνια μέχρι την παρούσα ζωή. Τα αποτελέσματα έδειξαν ότι οι παράγοντες από τα παιδικά χρόνια δεν είναι σημαντικοί για τους άνδρες όλων των ομάδων χωρών ενώ για τις γυναίκες έχουν την μικρότερη ισχύ συγκριτικά με άλλους προβλεπτικούς παράγοντες της μεταγενέστερης ζωής. Πιο σημαντικές κρίνονται: η δυσμενής εμπειρία της κατανάλωσης μεγάλης ποσότητας αλκοόλ από τους γονείς για τις γυναίκες της Βόρειας και της Κεντρικής Ευρώπης και ο αριθμός των βιβλίων στην ηλικία των 10 ετών για τις γυναίκες της Κεντρικής, Νότιας και Ανατολικής Ευρώπης. Αποτελέσματα άλλων ερευνών έχουν δείξει ότι οι αντιξοότητες της παιδικής ηλικίας σχετικά με δυσλειτουργικά χαρακτηριστικά των γονέων αποτελούν τους ισχυρότερους προγνωστικούς παράγοντες ψυχικών διαταραχών (Kessler et al., 2010). Σχετικά με τα γεγονότα στη διάρκεια της ζωής η περίοδος άγχους συνδέεται με την μελλοντική κατάθλιψη για τις γυναίκες όλων των ευρωπαϊκών ομάδων ενώ για τους άνδρες μόνο σε όσους διαμένουν στην Κεντρική και Νότια Ευρώπη. Συνακόλουθα, για τους άνδρες είναι σημαντική και η εμπειρία οικονομικών δυσκολιών με εξαίρεση αυτούς που ζουν στην Κεντρική Ευρώπη. Προηγούμενες μελέτες υποστηρίζουν τα συμπεράσματα της παρούσας έρευνας (Gallagher et al., 2013). Τέλος, οι παράγοντες που αναφέρονται στην παρούσα ζωή και κυρίως οι μεταβλητές υγείας καθορίζουν σε μεγάλο βαθμό την ύπαρξη κατάθλιψης για όλες τις ομάδες των χωρών ανεξάρτητα από τα φύλα, συμπέρασμα που είναι συνεπές με την προηγούμενη βιβλιογραφία (Kamiya et al., 2012). Σαν αντιστάθμισμα, η καλή γνωστική λειτουργία των ερωτώμενων λειτουργεί ανασταλτικά στην επικράτηση της νόσου σε μεγάλες ηλικίες. Τέλος, η παρουσία οικονομικών δυσκολιών προκαλεί αύξηση στα επίπεδα της κατάθλιψης και για τα δύο φύλα ανεξάρτητα από την ομάδα χωρών στην οποία ανήκουν, συμπέρασμα που επιβεβαιώνεται και από άλλους ερευνητές (Van de Velde et al., 2010).

Η παραπάνω μελέτη υπόκειται σε κάποιους περιορισμούς που καλό θα ήταν να συζητηθούν. Αρχικά, αναφορικά με το τεχνικό μέρος της εργασίας, ο σχεδιασμός της έρευνας στηρίχθηκε στην εφαρμογή του μοντέλου λογιστικής παλινδρόμησης.

Επιπρόσθετα, ιδιαίτερο ενδιαφέρον θα είχε και η εφαρμογή ενός μοντέλου τυχαίων συντελεστών (random coefficient model) σύμφωνα με το οποίο η κλίση του μοντέλου ποικίλλει με βάση κάποια κατανομή, ιδιαίτερα όταν μελετώνται πληθυσμιακές ομάδες με ετερογενή χαρακτηριστικά όπως πρόκειται στην παρούσα ανάλυση όπου οι χώρες ανά ομάδα παρουσιάζουν διαφορές μεταξύ τους. Με δεδομένο ότι το συγκεκριμένο μοντέλο μοντελοποιεί όχι μόνο τις μέσες τιμές αλλά και τις διασπορές των ομάδων η εφαρμογή του θα προσέδιδε βαρύτητα στις διαφοροποιήσεις των προς μελέτη παραγόντων στις ομαδοποιημένες χώρες της Ευρώπης οι οποίες υφίστανται εξαιτίας διακρατικών διαφορών και θεμάτων αντίληψης. Η εφαρμογή αυτού του μοντέλου θα ήταν εξαιρετικά ενδιαφέρουσα σε μελλοντική ανάλυση. Επιπλέον, ο αποκλεισμός των ελλειπουσών τιμών θα μπορούσε να αντιμετωπιστεί με μία μέθοδο υποκατάστασης αυτών (imputation method) ώστε να μην υπάρχει απώλεια πληροφορίας. Βέβαια, η τεχνική υποκατάστασης των παρατηρήσεων αν και στηρίζεται σε στατιστικές τεχνικές εκτίμησης, ωστόσο δεν αντιπροσωπεύει την πραγματική τιμή του παράγοντα που χρησιμοποιείται και για το λόγο αυτό δεν επιλέχθηκε να εφαρμοστεί στην συγκεκριμένη ανάλυση. Τέλος, κάποιοι άλλοι περιορισμοί συνδέονται με λάθη ανάκλησης της μνήμης και ενδεχόμενες εσφαλμένες απαντήσεις (recall errors) εξαιτίας της σύνδεσης του μεγάλου χρονικού διαστήματος που καλύπτει την ενηλικίωση την στιγμή συμμετοχής στην έρευνα και την παιδική ηλικία των ερωτώμενων.

ABSTRACT

The aim of this study is to investigate factors that predict depression among childhood circumstances, adulthood and later life conditions in people aged 50 or higher, and to identify gender differences across groups of countries in Europe. Data are derived from the Health, Aging and Retirement Survey in Europe (SHARE). The analysis was performed using logistic regression models. The findings show that: a. the children's predictors are the least important, with the exception of inhabitants of Northern, Central and Southern Europe b. having experienced a period of stress is the most important factor for women in all European regions while for men only in Central and Southern European countries and c. factors related to later life, including physical health and socio-economic level, have the strongest association with depression for all groups.

Ευχαριστίες: "Η εργασία αυτή υποστηρίχθηκε πλήρως από τη Γενική Γραμματεία Έρευνας και Τεχνολογίας (ΓΓΕΤ) και το Ελληνικό Ίδρυμα Έρευνας και Καινοτομίας (ΕΛΙΔΕΚ) (κωδικός υποτροφίας: 991)".

ΑΝΑΦΟΡΕΣ

Alonso, J., et al. (2004). Prevalence of mental disorders in Europe: results from the European Study of the Epidemiology of Mental Disorders (ESEMeD) project. *Acta psychiatrica scandinavica,* **109** (420), 21-27.

Börsch-Supan, A. & Jurges, H. (2005). *The Survey of Health, Aging and Retirement in Europe. Methodology,* Manheim, Germany.

Börsch-Supan, A., Brandt, M., Hunkler, C., Kneip, T., Korbmacher, J., Malter, F., Schaan, B., Stuck, S. & Zuber, S. (2013). Data Resource Profile: The Survey of Health, Ageing and Retirement in Europe (SHARE). *Int J Epidemiol,* **42**(4), 992- 1001.

Bruce, M. L. (2001). Psychological risk factor for depressive disorder in late life. *Biological Psychiatry,* **52**,175-184.

D'Alisa ,S, Miscio ,G, Baudo, S, Simone, A, Tesio L and Mauro, A. (2006). Depression is the main determinant of quality of life in multiple sclerosis: a classification-regression (CART) study. *Disabil Rehabil*, **28**, 307–314.

Farrokhi, F., Abedi, N., Beyene J., Kurdyak, P., Jassal S.V. (2014). Association Between Depression and Mortality in Patients Receiving Long-term Dialysis: A Systematic Review and Meta-analysis. *Am J Kidney Dis*, **63**(4), 623-35

Gallagher, D., Sawa, G.M., Kenny, R. and Lawlor, B.A. (2013). What predicts persistent depression in older adults across Europe? Utility of clinical and neuropsychological predictors from the SHARE study, *Journal of Affective Disorders,* **147**, 192-197.

Kamiya Y., Doyle, M., Henretta, J.C. & Timonen, V. (2013). Depressive symptoms among older adults: The impact of early and later life circumstances and marital status, Aging & Mental Health, **17**(3), 349-357.

Kessler, R.C., McLaughlin KA, Green JG, et al. (2010) Childhood adversities and adult psychopathology in the WHO World Mental Health Surveys. The British Journal of Psychiatry, **197**(5), 378-385.

Ladin, K. (2008) Risk of late-life depression across 10 European Union countries: deconstructing the education effect. *Journal of aging and health,* **20(**6), 653-670.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness, *J Appl Meas,* **3**(1), 85-106.

Lorant, V., Deliege D., Eaton, W., Robert A., Philippot, P. and Ansseau, M. (2003). Socioeconomic inequalities in depression: a meta-analysis. *American journal of epidemiology,* **157**(2), 98-112.

Lorant, V., Croux C., Weich, S., Deliege D., Mackenbach ,J. and Ansseau M. (2007). Depression and socio-economic risk factors: 7-year longitudinal population study. *The British journal of psychiatry,* **190**(4), 293-298.

Mills, T. N. and Henretta, J. C. (2001). Racial, ethnic and sociodemographic differences in the level of psychosocial difference among older Americans. *Research on Ageing*, **23**, 131-152.

Paul, P., Hakobyan, M. and Valtonen, H. (2016). The association between self-perceived health status and satisfaction with healthcare services: Evidence from Armenia. *BMC health services research*, **16**(67).

Van, Bergen E., van Zuijen T., Bishop, D. and de Jong P.F. (2017). Why Are Home Literacy Environment and Children's Reading Skills Associated? What Parental Skills Reveal. *READING RESEARCH QUARTERLY*, **52**(2), 147-160.

Van, de Velde S., Bracke, P. and Levecque, K. (2010). Gender differences in depression in 23 European countries. Cross-national variation in the gender gap in depression. *Social science & medicine,* **71**(2), 305-313.

Van, de Velde, S., Bracke, P., Levecque, K. and Meuleman, B. (2010). Gender differences in depression in 25 European countries after eliminating measurement bias in the CES-D 8. *Social Science Research,* **39**(3), 396-404.

World Health Organization (2002) Prevention and promotion in mental health. Geneva 2002.

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 31ου Πανελληνίου Συνεδρίου Στατιστικής (2018), σελ.220-235

ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ ΤΩΝ ΤΟΥΡΙΣΤΙΚΩΝ ΠΡΟΟΡΙΣΜΩΝ ΤΗΣ ΕΛΛΑΔΑΣ ΜΕ ΒΑΣΗ ΤΟΥΣ ΚΑΛΛΙΚΡΑΤΙΚΟΥΣ ΔΗΜΟΥΣ: ΜΙΑ ΕΜΠΕΙΡΙΚΗ ΑΝΑΛΥΣΗ

Α. Σταμάτης[1], Ε. Ταμπούρης[1], Κ. Ταραμπάνης[2]
[1]Τμήμα Εφαρμοσμένης Πληροφορικής, Πανεπιστήμιο Μακεδονίας
{tambouris , a.stamatis}@uom.edu.gr
[2]Τμήμα Οργάνωσης & Διοίκησης Επιχειρήσεων, Πανεπιστήμιο Μακεδονίας
kat@uom.gr

ΠΕΡΙΛΗΨΗ

Ο τουρισμός είναι ο πιο σημαντικός κλάδος της ελληνικής οικονομίας συνεισφέροντας περισσότερο από το 20% του ΑΕΠ της χώρας (Ανδρεάδη, 2016). Ο κλάδος έχει ερευνηθεί συστηματικά (Plog, 2001) ενώ τελευταία οι αναλύσεις που βασίζονται στην ανάλυση δεδομένων κερδίζουν έδαφος (Dolnicar, 2002). Ο βασικός στόχος της εργασίας είναι η τμηματοποίηση των ελληνικών προορισμών με χρήση σχετικών τουριστικών δεδομένων, με απώτερο σκοπό την συμβολή στην λήψη σχετικών στρατηγικών αποφάσεων. Η μέθοδος που χρησιμοποιήθηκε είναι η ανάλυση κατά συστάδες (clustering analysis) και συγκεκριμένα ο αλγόριθμος κ-μέσων (k-means++) (Arthur, Arthur, & Vassilvitskii, 2007) . Για την τροφοδότηση του μοντέλου μας χρησιμοποιήσαμε ανοικτά δεδομένα από την ΕΛΣΤΑΤ, την τράπεζα της Ελλάδος και τον Σύνδεσμο Ελληνικών Τουριστικών Επιχειρήσεων (ΣΕΤΕ). Στην εργασία μας παρουσιάζουμε δύο μοντέλα: το πρώτο κατατάσσει 260 τουριστικούς προορισμούς της Ελλάδας σε 5 ομάδες λαμβάνοντας ως είσοδο τον συνολικό αριθμό αφίξεων και το ποσοστό πληρότητας του εκάστοτε προορισμού. Το δεύτερο μοντέλο τους κατατάσσει σε 6 ομάδες λαμβάνοντας ως επιπλέον είσοδο τη μέση τιμή διανυκτέρευσης του εκάστοτε προορισμού. Η εκτίμηση της μέσης τιμής διανυκτέρευσης έγινε με συλλογή δεδομένων από την ιστοσελίδα booking.com με χρήση ενός HTML scraper (“jsoup Java HTML Parser”) , που δημιουργήσαμε για το σκοπό αυτό. Το αποτέλεσμα της ομαδοποίησης μπορεί να αποτελέσει χρήσιμο εργαλείο για τους τουριστικούς προορισμούς και τις επιχειρήσεις που δραστηριοποιούνται σε αυτούς. Για τους τουριστικούς προορισμούς αποτελεί έναν δείκτη της θέσης τους σε σχέση με τους υπόλοιπους προορισμούς τόσο της ίδιας όσο και των υπολοίπων ομάδων. Η γνώση αυτής θα μπορούσε να συμβάλει επικουρικά στην λήψη τουριστικών στρατηγικών αποφάσεων. Για τις επιχειρήσεις αποτελεί μέτρο σύγκρισης με τις υπόλοιπες επιχειρήσεις του προορισμού στον οποίο ανήκει αλλά και με τον μέσο όρο των επιχειρήσεων του προορισμού αυτού. Με βάση αυτή την πληροφορία, κάθε επιχείρηση μπορεί να βοηθηθεί στη λήψη σχετικών επιχειρηματικών αποφάσεων.

Λέξεις Κλειδιά: Data Drive Tourism, Cluster Analysis, Web Scrapper.

`

1. ΕΙΣΑΓΩΓΗ

1.1 Περιγραφή του προβλήματος

Ο τουρισμός είναι ένας από τους μεγαλύτερους κλάδους στην ελληνική οικονομία καθώς συμμετέχει στο Εθνικό Ακαθάριστο Εγχώριο Προϊόν κατά 20% του συνόλου (Ανδρεάδη, 2016). Το παραπάνω γεγονός από μόνο του κάνει τον τουρισμό ιδιαίτερα σημαντικό. Για αυτό είναι σκόπιμο να μελετήσουμε και να διερευνήσουμε περαιτέρω τρόπους ανάπτυξής του τουρισμού σε όλη την ελληνική επικράτεια.

Ένα από τα προβλήματα του ελληνικού τουρισμού σύμφωνα με τον (Buhalis, 2001) (Buhalis, 2001b) είναι η ανεπάρκεια στρατηγικού σχεδιασμού. Οι τουριστικές επιχειρήσεις της χώρας δεν λαμβάνουν αποφάσεις με γνώμονα κάποιο αντικειμενικό σχέδιο που να είναι βασισμένο σε δεδομένα. Αυτό έχει πολλές φορές ως αποτέλεσμα οι επιχειρηματίες να βασίζονται στο λεγόμενο «επιχειρηματικό ένστικτο». Να πράττουν δηλαδή με απλές υποθέσεις οι οποίες προκύπτουν από προηγούμενες εμπειρίες. Η ελλιπής ανάπτυξη στρατηγικού σχεδίου μπορεί να οδηγήσει σε λανθασμένες αποφάσεις και σε μια μονόπλευρη - «οριζόντια» ανάπτυξη.

Δεύτερο πρόβλημα είναι η εικόνα της Ελλάδας ως ένας φτηνός, απλός, αδιαφοροποίητος, ηλιόλουστος προορισμός (Buhalis, 2001). Η αύξηση των αφίξεων τουριστών, αλλά με μείωση της κατά κεφαλήν δαπάνης, είναι ακόμα ένα γεγονός που μας δείχνει ότι ο ελληνικός τουρισμός δεν αναπτύσσεται με τον καλύτερο δυνατό τρόπο. Η μείωση της κατά κεφαλήν δαπάνης είναι ίσως ένα σημάδι πτώσης της ποιότητας των τουριστικών υπηρεσιών.

1.2 Αντικείμενο και στόχοι της μελέτης

Στα πλαίσια αυτής της εργασίας έχουμε ως στόχο να κατηγοριοποιήσουμε κατά συστάδες τους τουριστικούς προορισμούς της Ελλάδας. Ο βασικός στόχος της εργασίας είναι να παρουσιάσει ένα μοντέλο, το οποίο βασιζόμενο στα δεδομένα θα δείχνει την θέση κάθε ελληνικού δήμου – προορισμού στην ελληνική επικράτεια. Ουσιαστικά θα παρουσιάσουμε μια βασική τμηματοποίηση του ελληνικού τουρισμού με σκοπό να αποτελέσει ένα εργαλείο που να συμμετέχει επικουρικά στη διαδικασία λήψης αποφάσεων στον κλάδο του τουρισμού.

Επιμέρους στόχος της εργασίας είναι η συλλογή και η αξιολόγηση δεδομένων που αφορούν τον κλάδο του τουρισμού. Θα γίνει προσπάθεια να αξιολογηθούν οι πηγές των δεδομένων, η ποιότητα και το πλήθος τους. Τέλος, στοχεύουμε να αναδείξουμε αδυναμίες και δυνατότητες, σχετικά με τα ανοικτά και γενικότερα τα μεγάλα δεδομένα του κλάδου.

`

2. ΑΝΑΣΚΟΠΗΣΗ ΒΙΒΛΙΟΓΡΑΦΙΑΣ

Η πρώτη σημαντική απόπειρα δημιουργίας ενός γενικού μοντέλου ερμηνείας της τουριστικής αγοράς έγινε από τον Plog το 1972. Το άρθρο αυτό επικαιροποιήθηκε το 2001 για να συμπεριλάβει τις αλλαγές στις στατιστικές μεθόδους και τα προφίλ των τουριστικών προορισμών που έγιναν μετά την αρχική δημοσίευση (Plog, 2001). Ακολουθώντας το μεταπολεμικό ρεύμα του συμπεριφορισμού και την πρόσφατη ανάπτυξη της νευροεπιστήμης από τα μέσα της δεκαετίας του 1990, ο Plog χρησιμοποιεί την "ψυχογραφική μέθοδο" ανάλυσης για την κατάτμηση και ανάλυση της τουριστικής αγοράς. Αρχικά, το δείγμα της ζήτησης (ταξιδιώτες) κατανέμεται ανάλογα με την προτίμησή του σε κοντινούς, γνωστούς, οικείους προορισμούς ή σε μακρινούς, άγνωστους προορισμούς. Η κατανομή είναι προσεγγιστικά κανονική. Στην αριστερή ουρά της κατανομής βρίσκονται οι λεγόμενοι ψυχοκεντρικοί ή εξαρτώμενοι ταξιδιώτες, που δείχνουν πολύ μεγάλη ή αποκλειστική προτίμηση σε σχετικά κοντινούς προορισμούς μαζικού τουρισμού που εξασφαλίζουν προβλέψιμο πρόγραμμα διακοπών. Στη δεξιά ουρά της κατανομής βρίσκονται οι λεγόμενοι αλλοκεντρικοί ταξιδιώτες ή εξερευνητές.

Η έρευνα των Bigné, Sanchez και Sanchez (Bigné, Sánchez, & Sánchez, 2001) δείχνει ότι η στοχευμένη προσέγγιση των δυνητικών τουριστών που προτείνει ο Plog θα είναι αποτελεσματική αν έχει ως αντικείμενο την εικόνα που έχει ο δυνητικός τουρίστας για τον τόπο.

Εικόνα 1. ***Κατάταξη Προορισμών κατά Plog (πηγή: Plog, 2001).***

Οι Elbe κι Emmoth (Elbe & Emmoth, 2014) ασχολούνται με την ανάλυση συμμετόχων στον τουριστικό κλάδο από νεοθεσμική σκοπιά, με τη χρήση της έννοιας της νομιμοποίησης (legitimacy). Η βασική υπόθεση είναι ότι κάθε οργανισμός

`

παροχής τουριστικών υπηρεσιών καλείται να πείσει τους πελάτες του, τους προμηθευτές του, τους ανταγωνιστές τους, τις αρχές και την τοπική κοινωνία ότι καλώς υπάρχει και δραστηριοποιείται στον συγκεκριμένο κλάδο (ηθική νομιμοποίηση), τον οποίο γνωρίζει εκτενώς (γνωσιακή νομιμοποίηση) κι επομένως μπορεί να είναι ένας καλός συνεταίρος, συνεργάτης ή έστω συνομιλητής (πραγματική νομιμοποίηση). Σε αυτή την προσπάθειά του, κάθε οργανισμός επιστρατεύει τη ρητορική πειθώ των μελών του. Οι Elbe κι Emmoth (Elbe & Emmoth, 2014) κατηγοριοποιούν την πειθώ σε επίκληση στη λογική (logos), επίκληση στο συναίσθηση (pathos) κι επίκληση στο ήθος του συνομιλητή (ethos) και στη συνέχεια ισχυρίζονται ότι κάθε συμμέτοχος πείθει και πείθεται από διαφορετικά είδη ρητορικής, από τα οποία αντλεί ή δίνει διαφορετικά είδη νομιμοποίησης.

Οι Noni, Orsi και Zanderighi (De Noni, Orsi, & Zanderighi, 2014) εντοπίζουν ως σημαντικότερους παράγοντες μεγαλύτερης αυτοδέσμευσης ενός ταξιδιώτη με έναν τόπο τις επιλογές εκπαίδευσης και τον πολιτισμό που ο τόπος αυτός προσφέρει. Οι ευκαιρίες απασχόλησης είναι σημαντικός παράγοντας αλλά όχι ο πιο σημαντικός, δεδομένου ότι η σύγχρονη αγορά εργασίας απαιτεί ολοένα και μεγαλύτερη κινητικότητα από τους εργαζομένους. Είναι αξιοσημείωτο ότι οι δημόσιες παροχές ενός αστικού κέντρου δεν παίζουν μεγάλο ρόλο ούτε στην τουριστική του εικόνα ούτε στην προοπτική του ως πιθανό τόπο μετεγκατάστασης.

Σε αντίθεση με τα παραπάνω, ο Zemla (Żemła, 2008) υποστηρίζει ότι η συναίνεση και συνεργασία των συμμετόχων σε έναν τουριστικό τόπο δεν είναι αρκετή για να αυξήσει την ανταγωνιστικότητά του και να του δώσει συγκριτικό πλεονέκτημα. Συχνά οι τουρίστες θα χρησιμοποιήσουν έναν τουριστικό τόπο ως βάση για να εξερευνήσουν μία ευρύτερη περιοχή. Επίσης, ο αυξημένος ανταγωνισμός μεταξύ γειτονικών και ομοειδών τουριστικών προορισμών μειώνει την αποδοτικότητα της προώθησης

Η βασικότερη συνεισφορά στην ανάλυση του τουρισμού βάσει δεδομένων προέρχεται από τη Sara Dolnicar (Dolnicar, 2002; Dolnicar, Grün, Leisch, & Schmidt, 2014). Η μεταέρευνα αυτή ανέδειξε πως η ανάλυση του τουριστικού κλάδου βάσει δεδομένων με τη χρήση συστάδων ακολουθούσε άτυπα ορισμένες παραδοχές σε μεγάλο βαθμό: τα δεδομένα ήταν συνήθως τακτικά (η αριθμητική τιμή δεν έχει φυσική σημασία και κατά συνέπεια το ίδιο ισχύει για την ευκλείδεια απόστασή τους) και προεπεξεργασμένα (συνήθως παραγοντοποιημένα), ο διαχωρισμός σε συστάδες γινόταν με βάση τον αλγόριθμο του Ward, ο προσδιορισμός του πλήθους τους γινόταν με καθαρά υποκειμενικά κριτήρια του ερευνητή και η ευστάθεια της λύσης δεν εξεταζόταν.

Η Dolnicar (Dolnicar, 2002) παρατήρησε ότι η χρήση εκ των προτέρων χαρακτηριστικών ή τυπολογιών για τη δημιουργία συστάδων (π.χ. φύλο) τείνει να μειωθεί προς όφελος της χρήσης εκ των υστέρων χαρακτηριστικών ή ταξονομιών (π.χ. ικανοποίηση). Στην πράξη, τα εκ των υστέρων στοιχεία αποδείχτηκαν

`

χρησιμότερα στη δημιουργία συστάδων και τον εντοπισμό του συγκριτικού πλεονεκτήματος ή του σημείου επικέντρωσης. Επίσης, διαπίστωσε ότι το μέγεθος του δείγματος και το πλήθος των συστάδων δεν συσχετίζεται θετικά, αντίθετα απ' ό,τι αναμενόταν. Αυτό το αντιδιαισθητικό συμπέρασμα δείχνει μία τάση συμπερίληψης υπερβολικά πολλών μεταβλητών για το μέγεθος του εκάστοτε δείγματος, η οποία κατά κανόνα δυσκολεύει τον εντοπισμό συστάδων, ακόμα και με το απλό κριτήριο της ελάχιστης συνδιακύμανσης, υπονομεύοντας έτσι την ανθεκτικότητα (robustness) της "προτεινόμενης" δομής δεδομένων.

Την ίδια χρονιά, οι Johns και Gyimothy (Johns & Gyimóthy, 2002) δημοσιεύουν ένα άρθρο κατάτμησης τουριστικής αγοράς βάσει δεδομένων, με τη χρήση συστάδων που σχεδόν στερεοτυπικά ανταποκρίνεται στο προφίλ που σκιαγράφησε η Dolnicar (Dolnicar, 2002). Τα δεδομένα είναι τακτικά, χρησιμοποιείται παραγοντική ανάλυση και ιεραρχικός αλγόριθμος για τον προσδιορισμό των συστάδων, ενώ το τελικό αποτέλεσμα επαληθεύεται βάσει εξωτερικής πληροφορίας (αποτελέσματα παρόμοιων ερευνών).

Μία μεταγενέστερη έρευνα (Dolnicar et al., 2014) προσπαθεί να εντοπίσει έναν γενικό εμπειρικό κανόνα επιλογής μεγέθους δείγματος ως προς το πλήθος των μεταβλητών. Με τη χρήση του προσαρμοσμένου δείκτη Ραντ, που υπολογίζει κατά πόσο η δομή των δεδομένων είναι συμβατή με τον επιλεγμένο διαχωρισμό σε συστάδες, καταλήγει σε έναν εμπειρικό κανόνα μεγέθους δείγματος τουλάχιστον 70 φορές μεγαλύτερο από τον αριθμό των μεταβλητών. Δεδομένου ότι στα ίδια άρθρα συνιστάται η χρήση λίγων μεταβλητών και η εκ των προτέρων αποκοπή των μεταβλητών θορύβου, συμπεραίνεται ότι ένα καλό μέγεθος δείγματος για ένα μοντέλο τριών ερμηνευτικών μεταβλητών, όπως αυτό της παρούσας εργασίας, είναι τουλάχιστον 210.

Ο Akerkar (Fuchs et al., 2014) (2012) εστιάζει περισσότερο στις μελλοντικές προοπτικές χρήσης μεγάλων δεδομένων στον τουριστικό τομέα. Στην εργασία του, επισημαίνει ότι οι εμπλεκόμενοι στην προσφορά τουριστικών υπηρεσιών έχουν εδώ και χρόνια πρόσβαση σε πολλά δεδομένα, αλλά δεν τα αξιοποιούν κατάλληλα. Οι λόγοι είναι τόσο τεχνικοί (έλλειψη υπολογιστικής ισχύος της υπάρχουσας υποδομής τους) όσο και ουσιαστικοί: τα μεγάλα δεδομένα είναι συχνά αδόμητα και περιλαμβάνουν πολλή περιττή για τους σκοπούς του αναλυτή πληροφορία. Οι Fuchs, Hopken και Lexhagen (Fuchs, Höpken, & Lexhagen, 2014) εντοπίζουν το ίδιο πρόβλημα της ανεπαρκούς αξιοποίησης των μεγάλων δεδομένων και προσπαθούν να το λύσουν με τη χρήση πληροφοριακών συστημάτων διοίκησης προορισμών (Destination management information systems – DMIS).

Οι Hjalager και Nordin (Hjalager & Nordin, 2011) στη δική τους έρευνα - επισκόπηση βιβλιογραφίας εξετάζουν τις διάφορες μεθόδους ανάλυσης της πελατοκεντρικής ή βασισμένης στη χρήση καινοτομίας, η οποία τέμνεται με την

`

ανάλυση αγοράς βάσει δεδομένων στον βαθμό που πολλά από αυτά τα δεδομένα προέρχονται ακριβώς από τη δραστηριότητα των πελατών. Ορισμένες από αυτές τις μεθόδους απαιτούν την ενεργή συμμετοχή των χρηστών και κάποιες άλλες όχι, ενώ ορισμένες μέθοδοι επιλέγουν να παίρνουν "λίγα από πολλούς" (λίγες μεταβλητές – μεγάλο δείγμα), για περισσότερες δυνατότητες στατιστικού χειρισμού των δεδομένων, και άλλες "πολλά από λίγους" (πολλές μεταβλητές – μικρό δείγμα), ώστε να εξασφαλίσουν την καλύτερη δυνατή εξέταση συγκεκριμένων μόνο χρηστών.

3. ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΕΡΜΗΝΕΙΑ ΕΥΡΗΜΑΤΩΝ

3.1 Απόκτηση Δεδομένων & Μεταβλητές

Μετά από βιβλιογραφική έρευνα, με σκοπό να βρούμε τις μεταβλητές εκείνες που δείχνουν την τουριστική δυναμική του εκάστοτε προορισμού, επιλέξαμε τρεις βασικές μεταβλητές. Η κάθε μια από αυτές μετράει κάτι διαφορετικό. Η πρώτη είναι οι αφίξεις του εκάστοτε προορισμού σε ένα έτος. Με αυτήν την μεταβλητή μετράμε το μέγεθος της τουριστικής κίνησης σε έναν προορισμό. Η δεύτερη μεταβλητή είναι το ποσοστό πληρότητας ενός προορισμού σε ένα έτος. Είναι μια μεταβλητή που μας δείχνει έμμεσα αν ο προορισμός που αναλύουμε έχει υποδομές και άμεσα μας δείχνει, αν ο εκάστοτε προορισμός εκμεταλλεύεται τις υποδομές του. Ως τρίτη και τελευταία μεταβλητή θέλουμε μια μεταβλητή τέτοια που να μπορεί να δείξει την «προστιθέμενη αξία» ενός προορισμού. Έτσι επιλέξαμε μια μεταβλητή που μετριέται σε χρηματική αξία. Η μεταβλητή μας είναι η μέση τιμή δωματίου για μια διανυκτέρευση στον εκάστοτε προορισμό που είναι προς ανάλυση.

Τα δεδομένα αντλήθηκαν από τις ιστοσελίδες statistics.gr & bankofgreece.gr κατόπιν υποδείξεων από την ΕΛΣΤΑΤ σε συνεργασία με την Τράπεζα της Ελλάδος . Τα δεδομένα είναι σε μορφή λογιστικού φύλλου και έχουν ως περιγραφή γενικές αφίξεις ανά «Καλλικρατικό» δήμο. Τα δεδομένα έχουν τίτλο «Αφίξεις στα καταλύματα ξενοδοχειακού τύπου (πλην κάμπινγκ), κατά Περιφερειακή Ενότητα και Δήμο» και αναφέρονται στο έτος 2015. Υπάρχει επίσης μια πολύ σημαντική σημείωση: «Τα στοιχεία των Δήμων τα οποία εμπίπτουν στο στατιστικό απόρρητο αθροίζονται με τα στοιχεία όμορων Δήμων». Λόγω της παραπάνω σημείωσης έχουμε 260 εγγραφές με δεδομένα δήμων και όχι 325 που είναι το σύνολο των δήμων σύμφωνα με το σχέδιο Καλλικράτης. Ένα στοιχείο που κάνει τα δεδομένα αφίξεων ικανοποιητικά είναι ότι δεν προσμετρούνται μόνο αεροπορικές αφίξεις αλλά το σύνολο των αφίξεων, οι οποίες χωρίζονται σε ξένες και ημεδαπές. Εκτός από αφίξεις στο ίδιο αρχείο δεδομένων παίρνουμε το ετήσιο ποσοστό πληρότητας των διαθέσιμων δηλωμένων κλινών. Κάνουμε την υπόθεση, καθώς δεν διευκρινίζεται ρητά, πως το ποσοστό αυτό προκύπτει ως μέση τιμή από χαμηλές τιμές κατά τη διάρκεια των μη τουριστικών περιόδων του έτους και από υψηλές τιμές κατά τη διάρκεια των τουριστικών περιόδων του έτους. Ειδικά στους λεγόμενους θερινούς προορισμούς κάνουμε την υπόθεση ότι υπάρχει μεγάλη τυπική απόκλιση στην πληρότητα ανάμεσα στους μήνες

`

έντονης τουριστικής δραστηριότητας και αυτούς με περιορισμένη τουριστική δραστηριότητα.

Δεδομένα για την 3η και τελευταία μεταβλητή μας δυσκολευτήκαμε αρκετά να βρούμε στις παραπάνω θεσμικές πηγές. Στόχος μας ήταν να βρούμε δεδομένα που θα αποτελούν έναν δείκτη ποσοτικής συσχέτισης με βάση την αξία του τουριστικού προϊόντος. Επιλέξαμε να μετρήσουμε την αξία της διανυκτέρευσης με την θεώρηση ότι είναι ο πυρήνας του τουριστικού προϊόντος. Δεύτερος στόχος ήταν και πάλι η γεωγραφική ανάλυση. Δηλαδή, αναζητούμε τιμές δωματίων για κάθε Καλλικρατικό δήμο. Δεν βρήκαμε δεδομένα αυτής της ανάλυσης στις θεσμικές πηγές και για αυτό κατευθυνθήκαμε στα δεδομένα που παρέχει η αγορά. Επιλέξαμε την ιστοσελίδα Booking.com ως εργαλείο αναζήτησης τιμών δωματίων καθώς είναι ο μεγαλύτερος ηλεκτρονικός τουριστικός πράκτορας στην χώρα μας. Είναι γεγονός πως το Booking.com δεν περιέχει το 100% των ελληνικών καταλυμάτων, ωστόσο το επιλέξαμε για αντλήσουμε τα δεδομένα μας λόγω της απλότητάς του. Για την απόσπαση των δεδομένων από το Booking.com αναζητήσαμε κάποιο API που να μας δίνει την δυνατότητα να αντλήσουμε τα δεδομένα αυτοματοποιημένα και σε επεξεργάσιμη μορφή. Δεν βρέθηκε API ούτε αντίστοιχος μηχανισμός άντλησης δεδομένων. Έτσι κατασκευάσαμε ένα HTML web scraper σύστημα σε Java με σκοπό να πάρουμε τις τιμές δωματίων ανά δήμο της χώρας από το Booking.com.

Η εργασία είχε όμως και περιορισμούς σχετικά με τα δεδομένα. Βασικότερος περιορισμός είναι ότι έχουμε δεδομένα μόνο για τους 260 από τους συνολικά 325 Καλλικρατικούς δήμους της Ελλάδας. Το γεγονός αυτό είναι μια αδυναμία του ελληνικού τουριστικού κλάδου να παράγει αναλυτικά ανοικτά δεδομένα και να τα διαθέσει. Τα οφέλη που θα προκύψουν από την συλλογή και διάθεση προς έρευνα τέτοιων δεδομένων είναι αρκετά. Ένας δεύτερος περιορισμός της εργασίας μας σχετικά με τα δεδομένα αφορά στην περίπτωση της μέσης τιμής διανυκτέρευσης. Για αυτή τη μεταβλητή δε βρέθηκαν επίσημα δεδομένα και η συλλογή των δεδομένων που κάναμε ήταν περιορισμένη σε τιμές από μια συγκεκριμένη περίοδο του τουριστικού έτους και από μια πηγή (Booking.com). Επίσης κάθε φόρα στον υπολογισμό κάναμε χρήση 15 διαφορετικών τιμών από 15 διαφορετικά καταλύματα κάθε προορισμού. Είναι σίγουρα μια πρόκληση η τροφοδότηση του μοντέλου μας με δεδομένα στη μεταβλητή της μέσης τιμή διανυκτέρευσης από πολλές πηγές, από πολλές ημέρες του έτους και από πολλά διαφορετικά καταλύματα κάθε προορισμού.

3.2 Επιλογή και χρήση προγράμματος Rapid Miner

Για την επεξεργασία των δεδομένων χρησιμοποιήσαμε το Rapid Miner. Το Rapid Miner είναι ένα εργαλείο ανάλυσης δεδομένων. Έχει τρεις εκδόσεις ανάλογα με το μέγεθος των δεδομένων που θέλουμε να αναλύσουμε και τους χρήστες που θα εργαστούν. Η πρώτη βασική έκδοση που επιλέξαμε μπορεί να εγκατασταθεί σε προσωπικό υπολογιστή και να γίνει τοπικά η επεξεργασία των δεδομένων.

`

3.3 ΑΠΑΛΟΙΦΉ ΔΕΔΟΜΈΝΩΝ ΧΩΡΊΣ ΤΙΜΉ & ΚΑΝΟΝΙΚΟΠΟΊΗΣΗ ΔΕΔΟΜΈΝΩΝ

Στα δεδομένα που αντλήσαμε για τις δύο πρώτες μεταβλητές μας υπήρχαν σημεία χωρίς περιεχόμενο. Σε αυτές τις περιπτώσεις, με την βοήθεια του RapidMiner, εξαλείψαμε τις κενές τιμές με την τιμή 0.

Η δεύτερη επεξεργασία που κάναμε στα δεδομένα ήταν η κανονικοποίηση με σκοπό όλες οι τιμές διαφορετικών μεταβλητών να μην έχουν μεταξύ τους απόσταση λόγω της φύσης τους. Για παράδειγμα, έστω ότι έχουμε δυο μεταβλητές όπου η πρώτη παίρνει τιμές από το 1 έως και το 10 ενώ η δεύτερη έχει δεδομένα στην κλίμακα των χιλιάδων. Σε αυτήν την περίπτωση είναι εύκολα αντιληπτό ότι η μεταβλητή με τιμές στην κλίμακα των χιλιάδων θα καθορίσει και το αποτέλεσμα της ανάλυσης ενώ ή μεταβλητή με τιμές από 1 έως 10 δεν θα συμμετέχει σχεδόν καθόλου σε αυτό. Στο μοντέλο μας έχουμε τέτοιες μεταβλητές. Οι αφίξεις μετριούνται σε χιλιάδες, η πληρότητα σε ποσοστό και η μέση τιμή διανυκτέρευσης σε δεκάδες ή εκατοντάδες. Σε περίπτωση μη κανονικοποίησης των δεδομένων οι αφίξεις θα είχαν κυρίαρχο ρόλο στην ανάλυση σε σχέση με τις άλλες δύο μεταβλητές.

Για να κάνουμε τα παραπάνω χρησιμοποιήσαμε το λογισμικό RapidMiner. Η μέθοδος κανονικοποίησης που επιλέξαμε από το RapidMiner ήταν η range transformation η οποία ορίζει ένα μέγιστο και ένα ελάχιστο και τα θέτει αντίστοιχα στην μέγιστη και ελάχιστη τιμή κάθε μεταβλητής. Όλες οι υπόλοιπες τιμές παίρνουν μια τιμή ανάμεσα στο μέγιστο και ελάχιστο, διατηρώντας πάντοτε αναλογικά την αρχική απόσταση τους. Ως μέγιστο θέτουμε το 1 και ως ελάχιστο το 0. Με τον τρόπο αυτό καταφέρνουμε να κανονικοποιήσουμε τις τρείς μεταβλητές μας.

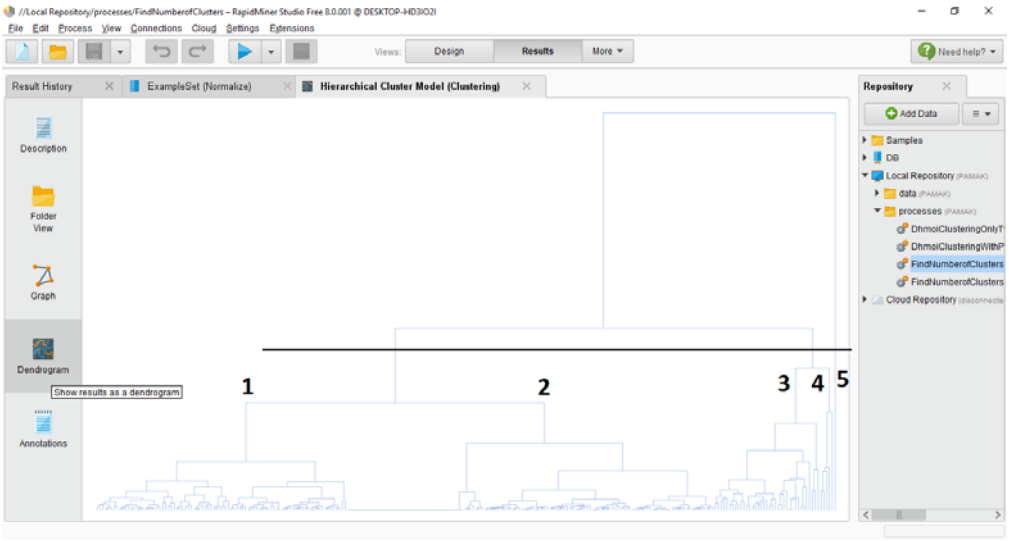
3.4 Αλγόριθμος k-means & εύρεση του k του πλήθους των ομάδων

Για την επιλογή της κατάλληλης μεθόδου αναζητήσαμε στη βιβλιογραφία μεθόδους εξόρυξης και ανάλυσης δεδομένων (Liao et al., 2012), που να μπορούν να ομαδοποιήσουν τα δεδομένα μας. Η μέθοδος k-means μας επιτρέπει να δημιουργήσουμε ομάδες προορισμών, σύμφωνα με την απόσταση των τιμών που έχουν οι μεταβλητές του κάθε προορισμού.

Για να επιλέξουμε σωστά το πλήθος των ομάδων που θα θέσουμε στον αλγόριθμο k-μέσων (k-means), πρέπει να ακολουθήσουμε μια μεθοδολογία και να μην το ορίσουμε τυχαία. Η μεθοδολογία που θα ακολουθήσουμε βασίζεται στην εκτέλεση μιας διαδικασίας ομαδοποίησης με μια ιεραρχική μέθοδο. Το αποτέλεσμα της ιεραρχικής μεθόδου ομαδοποίησης αποτελεί έναν δείκτη για το πλήθος των ομάδων που θα ζητήσουμε από τον αλγόριθμο k-means.

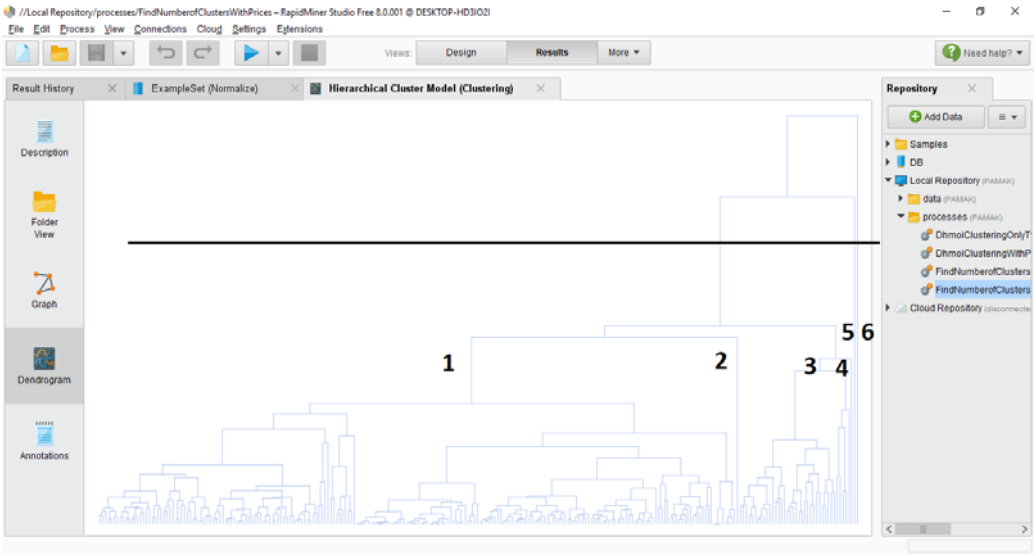
`

Εικόνα 2. *Αποτέλεσμα ιεραρχικής διαδικασίας μοντέλου δύο μεταβλητών*



Αν εφαρμόσουμε τη διαδικασία με τις δύο μεταβλητές παίρνουμε την γραφική απεικόνιση της διαδικασίας (εικόνα 2) και προκύπτουν ως προτεινόμενη τιμή οι 5 συστάδες. Αντίστοιχα, από την γραφική απεικόνιση που προέκυψε από την διαδικασία και με τις 3 μεταβλητές (εικόνα 3), παίρνουμε ως προτεινόμενη τιμή 6 συστάδες.

Εικόνα 3. *Αποτέλεσμα ιεραρχικής διαδικασίας μοντέλου τριών μεταβλητών*



`

3.5 Επιλογή αρχικής τιμής για τα κέντρα των συστάδων

Μια από τις πιο καθοριστικές ρυθμίσεις για τον αλγόριθμό k-means είναι ο τρόπος που ο αλγόριθμος θα θέσει τις αρχικές τιμές των κέντρων πριν ξεκινήσει τις συγκρίσεις με τα δείγματα και τον υπολογισμό εκ' νέου του εκάστοτε κέντρου. Η μέθοδος που θα ακολουθήσουμε είναι αυτή που πρότειναν οι David Arthur και Sergei Vassilvitskii (Arthur & Vassilvitskii, 2007). Ο αλγόριθμός δεν θέτει τυχαία τις αρχικές τιμές των κέντρων, αλλά λαμβάνει υπ' όψη τα δείγματα με σκοπό να θέσει ικανοποιητικές αρχικές τιμές. Με αυτόν τον τρόπο πετυχαίνουμε καλύτερο αποτέλεσμα στην ομαδοποίηση.

3.6 Ανάλυση κατά συστάδες: μοντέλο με δυο μεταβλητές

Είναι σημαντικό να αναφέρουμε τις ρυθμίσεις που ορίσαμε στον αλγόριθμο k-means στο λογισμικό RapidMiner. Η βασικότερη ρύθμιση είναι το πλήθος των ομάδων που αναζητούμε. Στην περίπτωση που τρέχουμε τον αλγόριθμο μόνο με δύο μεταβλητές (Σύνολο Αφίξεων, Ποσοστό Πληρότητας) είναι k= 5, όπως υπολογίστηκε σε προηγούμενη παράγραφο. Δεύτερη πολύ σημαντική ρύθμιση είναι να επιλέξουμε την επιλογή «determine good start values» που ενεργοποιεί τον αλγόριθμων k-means++ που περιγράψαμε στην προηγούμενη ενότητα. Επίσης, ρυθμίζουμε «measure type» σε Numerical καθώς έχουμε μόνο αριθμητικές τιμές, την ρύθμιση «numerical measure» σε «Euclidian Distance» και θέτουμε ως τον αριθμό των επαναλήψεων του αλγορίθμου την τιμή 100.

3.7 Αποτέλεσμα: τιμές των κέντρων

Στην εικόνα 4 βλέπουμε τις τιμές των κέντρων για κάθε μια από τις μεταβλητές μας που προκύπτει μετά την εκτέλεση του αλγορίθμου.

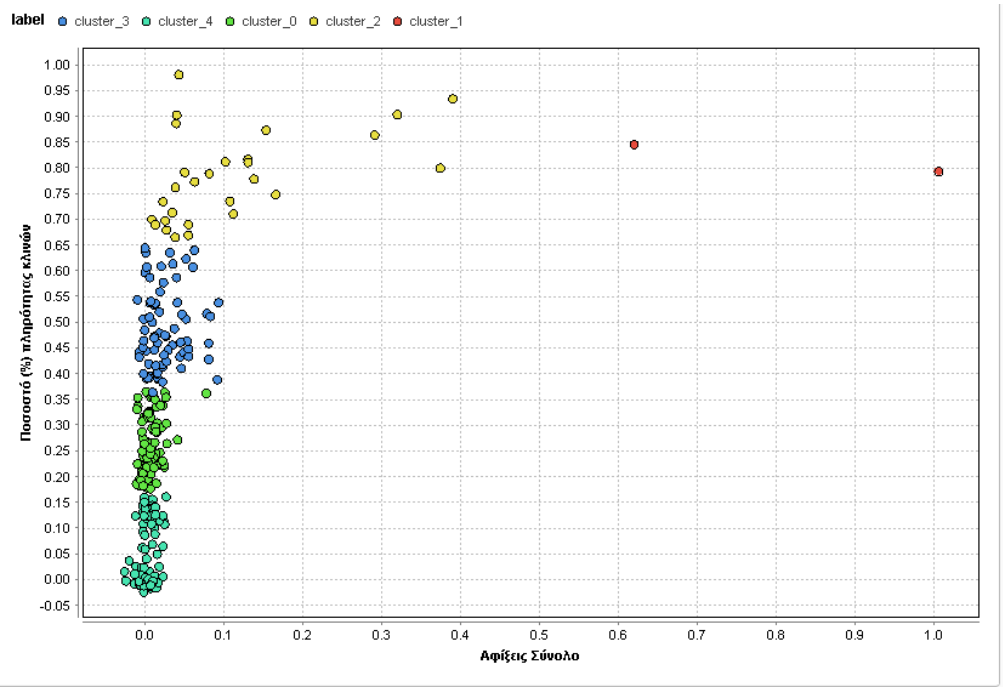
Εικόνα 4. *Τιμές των κέντρων μοντέλου δύο μεταβλητών*

Attribute	cluster_0	cluster_1	cluster_2	cluster_3	cluster_4
Αφίξεις Σύνολο	0.008	0.808	0.110	0.024	0.001
Ποσοστό (%) πληρότητας ...	0.267	0.826	0.783	0.488	0.051

3.8 Αποτέλεσμα: Συστάδες

Στην εικόνα 5 βλέπουμε πώς διαμορφώνονται σε γραφική παράσταση οι παρατηρήσεις μέσα στις ομάδες. Κάθε ομάδα έχει ένα χρώμα, ενώ οι άξονες είναι οι μεταβλητές μας. Από το παρακάτω σχήμα μπορούμε να δούμε βασικά χαρακτηριστικά όπως το μέγεθος των ομάδων, το πόσο κοντά βρίσκεται η μια ομάδα με την άλλη και το πόσο πυκνή είναι η κάθε ομάδα.

Εικόνα 5. *Διάγραμμα ομάδων πρώτου μοντέλου*



3.9 Ανάλυση κατά συστάδες: μοντέλο με τρεις μεταβλητές

Σε σχέση με το μοντέλο με δύο μεταβλητές, η μόνη αλλαγή στις ρυθμίσεις που ορίσαμε στον Operator k-means του λογισμικού RapidMiner είναι ο αριθμός των ομάδων που τώρα είναι k = 6, σύμφωνα με την ιεραρχική ανάλυση που προηγήθηκε.

3.10 Αποτέλεσμα: τιμές των κέντρων

Στην εικόνα 6 βλέπουμε τις τιμές των κέντρων για κάθε μια από τις μεταβλητές μας.

Εικόνα 6. *Τιμές των κέντρων μοντέλου τριών μεταβλητών*

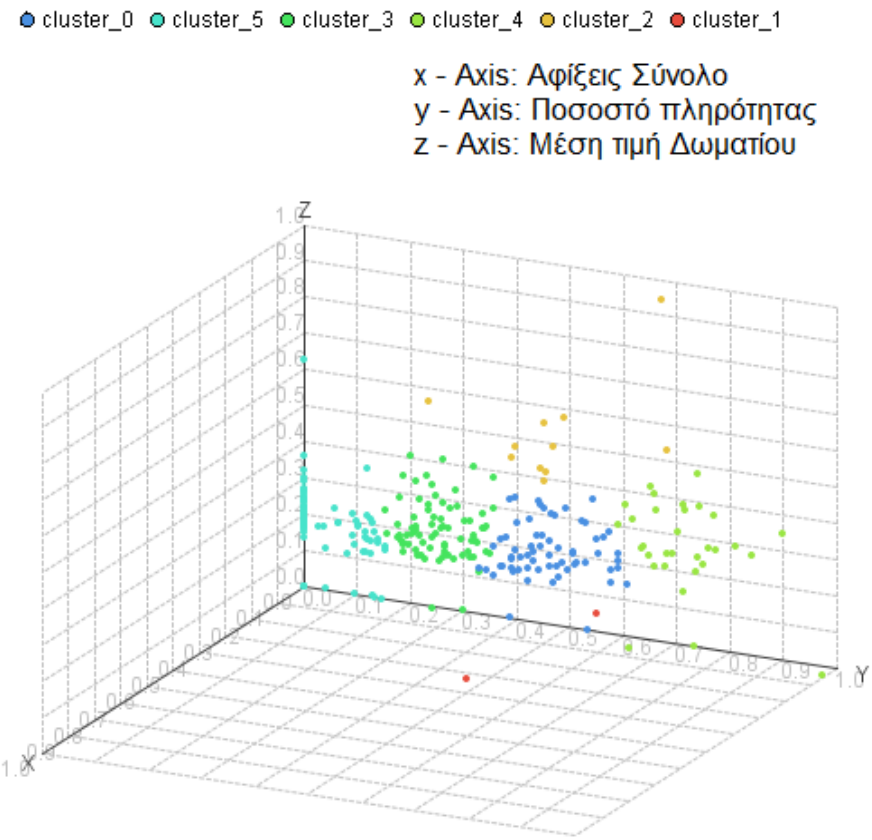
Attribute	cluster_0	cluster_1	cluster_2	cluster_3	cluster_4	cluster_5
Αφίξεις Σύνολο	0.024	0.808	0.026	0.007	0.100	0.001
Ποσοστό (%) πληρότητας κλινών	0.473	0.826	0.480	0.259	0.764	0.048
Μέση Τιμή Δωματίου 1- 10 Αυγούστου.	0.213	0.392	0.542	0.217	0.325	0.183

`

3.11 Αποτέλεσμα: Συστάδες

Στην εικόνα 7 βλέπουμε πώς διαμορφώνονται σε γραφική παράσταση οι παρατηρήσεις μέσα στις ομάδες.

Εικόνα 7. *Διάγραμμα ομάδων επαυξημένου μοντέλου*



3.12 Συμπεράσματα ανάλυσης

Το αποτέλεσμα της ανάλυσης μας, με είσοδο τις μεταβλητές αριθμός αφίξεων και πληρότητα κλινών, μας έδωσε ως αποτέλεσμα μια ομάδα με πολύ μεγάλες τιμές αφίξεων, όπου υπήρχαν και μεγάλες τιμές πληρότητας (Δήμος Αθηναίων και Δήμος Ρόδου). Επίσης, στα αποτελέσματα είχαμε μια ομάδα με όλους τους τουριστικά δημοφιλείς προορισμούς (Καλλικρατικούς Δήμους). Τέλος, δύο ομάδες με προορισμούς χαμηλής τουριστικής κίνησης και μια ομάδα με μηδενική τουριστική κίνηση ή έλλειψη στοιχείων. Το αποτέλεσμα δείχνει μια πρώτη εικόνα κατάταξης

`

των ελληνικών Καλλικρατικών δήμων σε πέντε ομάδες, σύμφωνα με τον αλγόριθμο k-μέσων και την ευκλείδεια απόσταση από το εκάστοτε κέντρο, τονίζοντας με αυτό τον τρόπο την αντικειμενικότητα της μεθόδου που επιλέξαμε.

Το αποτέλεσμα της ανάλυσης μας με είσοδο και τις τρεις μεταβλητές μας (δηλαδή το αρχικό μοντέλο επαυξημένο με ακόμα μια μεταβλητή αυτή της μέσης τιμής διανυκτέρευσης) μας έδωσε ως αποτέλεσμα έξι ομάδες. Ο σκοπός του επαυξημένου μοντέλου είναι να περάσουμε στο μοντέλο μας και την έννοια της προστιθέμενης αξίας του τουριστικού προϊόντος. Εάν συγκρίνουμε την κατανομή των τιμών και την κατανομή της πληρότητας θα δούμε ότι δεν υπάρχει συσχέτιση. Επίσης, προκύπτει από τα δεδομένα μας ότι σε διαφορετικούς προορισμούς υπάρχει διαφορετική τιμή στο ίδιο ποσοστό πληρότητας, άρα και διαφορετική αξία – μια προστιθέμενη αξία του ενός προορισμού σε σχέση με το άλλον.

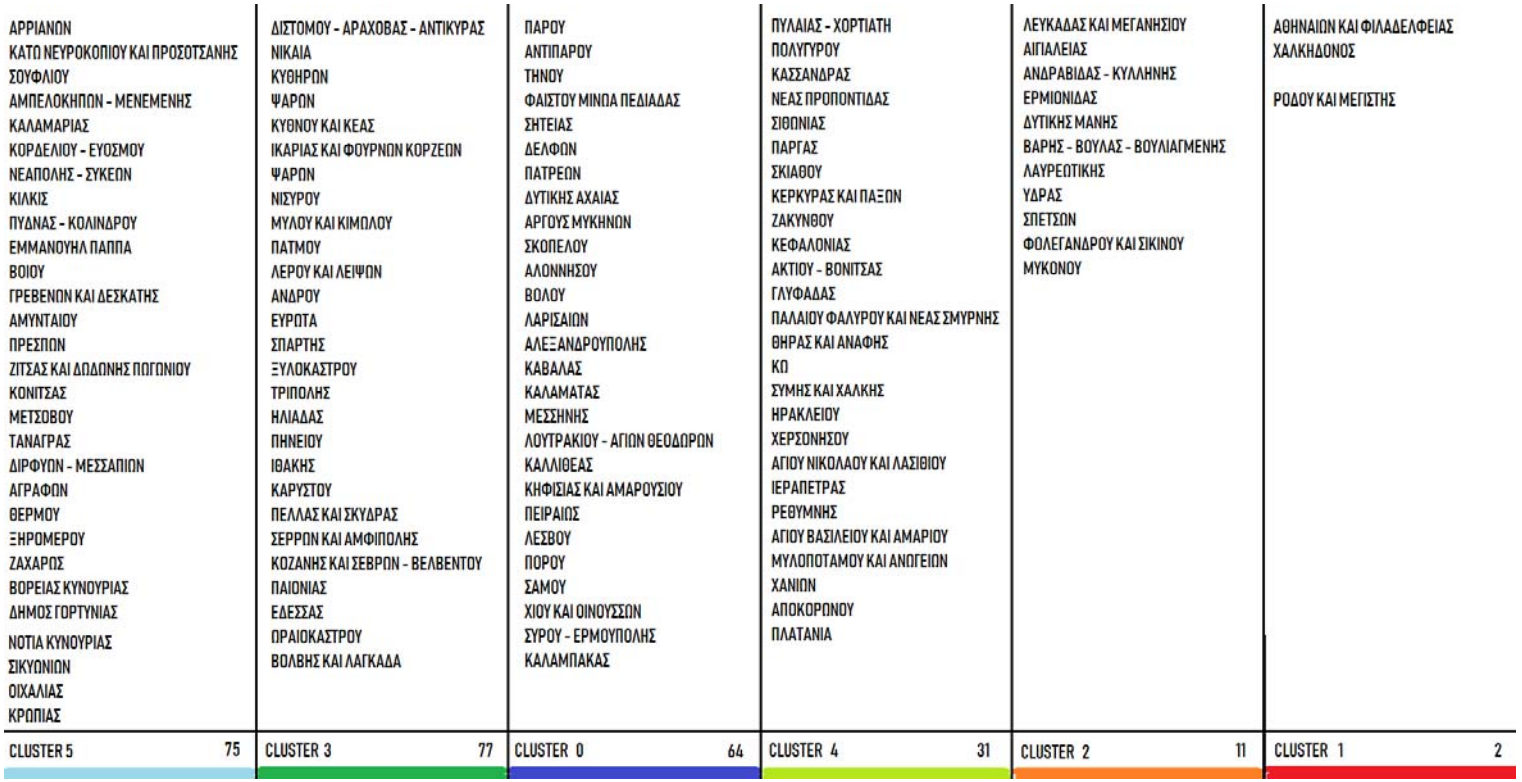
Η ανάλυση μας καταλήγει και ξεχωρίζει στο επαυξημένο μοντέλο μια ομάδα (Cluster 2) του δείγματος που συγκεντρώνει μεγάλες τιμές στην μεταβλητή μέση τιμή διανυκτέρευσης. Αυτό το τελευταίο εύρημα είναι και το μοναδικό που μπορεί να μας δώσει ένα στοιχείο για την προστιθέμενη αξία. Πρέπει να σημειωθεί ότι σε αυτήν την ομάδα (Cluster 2) δεν συναντάμε τα μεγαλύτερα ποσοστά πληρότητας. Δηλαδή, έχουμε μικρές πληρότητες και σχετικά μεγάλες τιμές γεγονός που και πάλι καταδεικνύει ότι εδώ οι τιμές δεν είναι μεγάλες λόγω περιορισμένης διαθεσιμότητας.

Οι υπόλοιπες πέντε ομάδες του επαυξημένου μοντέλου διακρίνονται στις εξής: Μια ομάδα (Cluster 1) με τις ακραίες τιμές αφίξεων (Δήμος Αθηναίων και Δήμος Ρόδου). Μια ομάδα που δεν παρουσιάζει τουριστικό ενδιαφέρον (Cluster 5). Μία ομάδα (Cluster 4) με έντονο τουριστικό ενδιαφέρον χωρίς όμως να έχουμε μεγάλες μέσες τιμές διανυκτέρευσης, ουσιαστικά εδώ έχουμε δημοφιλείς προορισμούς με μεγάλες πληρότητες. Τέλος έχουμε δύο ομάδες με μέτριο και ελάχιστο τουριστικό ενδιαφέρον αντίστοιχα (Cluster 0 , Cluster 3).

Ένα πρώτο συμπέρασμα που προκύπτει από το δεύτερο μοντέλο στην εργασία μας είναι ότι η ομάδα 2 (Cluster 2) που περιέχει προορισμούς όπου συναντάμε μεγαλύτερες μέσες τιμές διανυκτέρευσης (δηλαδή μεγαλύτερη «προστιθέμενη αξία»), αριθμεί μόνο 11 δήμους. Δηλαδή, μόνο το 4% του δείγματος, σύμφωνα με το μοντέλο μας, παρουσιάζει σημαντική «προστιθέμενη αξία» στις τουριστικές τους υπηρεσίες.

Στην εικόνα 8 έχουμε μια αναπαράσταση σαν αυτή του Plog που είδαμε στην εικόνα 1. Στην εικόνα 8 στα δεξιά έχουμε τους προορισμούς με πολύ μεγάλους αριθμούς αφίξεων, η δεύτερη από δεξιά στήλη είναι εκείνη με τους προορισμούς που εμφανίζουν μεγαλύτερες τιμές, στο κέντρο έχουμε τους δημοφιλείς τουριστικούς προορισμούς, ενώ στα αριστερά τους μη δημοφιλείς. Βλέπουμε ότι όσο πάμε προς τα δεξιά οι πληθυσμοί γίνονται μικρότεροι.

Εικόνα 8. *Απεικόνιση αποτελέσματος*



`

4. ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΠΡΟΤΑΣΕΙΣ

Τα αποτελέσματα από το μοντέλο μας θα μπορούσαν να αποτελούν έναν δείκτη της θέσης κάποιου προορισμού. Θα μπορούσαν συνεπώς να συμβάλουν επικουρικά στην λήψη στρατηγικών αποφάσεων. Είναι ένα μοντέλο οδηγούμενο από δεδομένα, με ποσοτικές μεταβλητές που μπορεί να παρουσιάσει μια εικόνα για τον εκάστοτε προορισμό. Δείχνει την θέση κάποιου προορισμού σε σχέση με κάποιον άλλο αλλά και με το σύνολο. Επίσης, το μοντέλο μας μπορεί να είναι χρήσιμο στην περίπτωση σύγκρισης της εκάστοτε τουριστικής επιχείρησης με την θέση του προορισμού στον οποίου ανήκει. Με αυτόν τον τρόπο κάθε επιχείρηση μπορεί να έχει ένα εργαλείο που να της δείχνει την θέση της σε σχέση με τον μέσο όρο του προορισμού.

Μια τελευταία παρατήρηση σχετικά με τα δεδομένα είναι ότι θα είχαν μεγάλο ενδιαφέρον τα αποτελέσματα με μεγαλύτερη γεωγραφική ανάλυση από αυτή του δήμου. Δηλαδή, να βρεθούν δεδομένα σε επίπεδο που να περιγράφουν καλύτερα την έννοια «τουριστικός προορισμός». Ακόμα, ένα σημείο περαιτέρω ανάπτυξης είναι να περιορίσουμε τους προορισμούς που μελετούμε μόνο σε όσους θεωρούνται τουριστικοί. Σε αυτή την περίπτωση θα είχε ενδιαφέρον να συγκριθούν τόσο τα αποτελέσματα όσο και οι ομάδες που προκύπτουν.

ABSTRACT

Tourism is the most important economic sector in the Greek economy. It has been growing since 1950 and its total contribution to the economy is more than 20% of the country's GDP. (Andreadis, 2016) The industry has been systematically researched by academics (Plog, 2001), while scientific papers based on data analysis have gained momentum last years (Dolnicar, 2002). The main objective of the current work is the segmentation of Greek destinations using relevant tourism data, with the final goal of making relevant strategic decisions. The method we used is clustering analysis and in particular the k-means algorithm (Arthur, Arthur, & Vassilvitskii, 2007). For our model, we used open data from ELSTAT, the bank of Greece and the Association of Greek Tourist Enterprises (SETE). In our document we present two models: the first classifies 260 tourist destinations of Greece in 5 groups taking as input the total number of arrivals and the occupancy rate of each destination. The second model classifies them in 6 groups, taking as an additional input the average night's price for each destination. The data for the third variable we collected from the booking.com website using an HTML scraper ("jep Java HTML Parser") that we created for this purpose. As a result of our model is a useful table of groups that could be a decion making tool for tourist destinations and businesses that are located in them. For tourist destinations our work could be an indicator of their position relative to the other destinations wich located in the same group. For businesses, it is a measure of comparison with the other companies and the average metrics of the whole destination.

`

ΑΝΑΦΟΡΕΣ

Arthur, D., & Vassilvitskii, S. (2007). K-Means++: The Advantages of Careful Seeding. In *Proc. of the Annu. ACM-SIAM Symp. on Discrete Algorithms* (Vol. 8, pp. 1027–1035). https://doi.org/10.1145/1283383.1283494

Bigné, J. E., Sánchez, M. I., & Sánchez, J. (2001). Tourism image, evaluation variables and after purchase behaviour: inter-relationship. *Tourism Management*, *22*(6), 607–616. https://doi.org/10.1016/S0261-5177(01)00035-8

Buhalis, D. (2001). Tourism in Greece: Strategic analysis and challenges. Retrieved from http://epubs.surrey.ac.uk/1116/

De Noni, I., Orsi, L., & Zanderighi, L. (2014). Attributes of Milan influencing city brand attractiveness. *Journal of Destination Marketing & Management*, *3*(4), 218–226. https://doi.org/10.1016/j.jdmm.2014.06.001

Dolnicar, S. (2002). A Review of Data-Driven Market Segmentation in Tourism. *Journal of Travel & Tourism Marketing*, *12*(1), 1–22. https://doi.org/10.1300/J073v12n01_01

Dolnicar, S., Grün, B., Leisch, F., & Schmidt, K. (2014). Required Sample Sizes for Data-Driven Market Segmentation Analyses in Tourism. *Journal of Travel Research*, *53*(3), 296–306. https://doi.org/10.1177/0047287513496475

Elbe, J., & Emmoth, A. (2014). The use of rhetoric in legitimation strategies when mobilizing destination stakeholders. *Journal of Destination Marketing & Management*, *3*(4), 210–217. https://doi.org/10.1016/J.JDMM.2014.08.001

Fuchs, M., Höpken, W., & Lexhagen, M. (2014). Big data analytics for knowledge generation in tourism destinations – A case from Sweden. *Journal of Destination Marketing & Management*, *3*(4), 198–209. https://doi.org/10.1016/j.jdmm.2014.08.002

Hjalager, A.-M., & Nordin, S. (2011). User-driven Innovation in Tourism—A Review of Methodologies. *Journal of Quality Assurance in Hospitality & Tourism*, *12*(4), 289–315. https://doi.org/10.1080/1528008X.2011.541837

Johns, N., & Gyimóthy, S. (2002). Market Segmentation and the Prediction of Tourist Behavior: The Case of Bornholm, Denmark. *Journal of Travel Research*, *40*(3), 316–327. https://doi.org/10.1177/0047287502040003009

jsoup Java HTML Parser, with best of DOM, CSS, and jquery. (n.d.). Retrieved September 14, 2018, from https://jsoup.org/

Plog, S. (2001). Why destination areas rise and fall in popularity: an update of a Cornell Quarterly classic. *The Cornell Hotel and Restaurant Administration Quarterly*, *42*(3), 13–24. https://doi.org/10.1016/S0010-8804(01)81020-X

Żemła, M. (2008). The product quality of Polish ski-resorts: A case study of Silesian skiers' requirements, satisfaction and complaints. *Turizam : Međunarodni Znanstveno-Stručni Časopis*, *56*(1), 41–58.

Ανδρεάδη, Α. (2016). *Ομιλία Α. Ανδρεάδη στην Κλειστή Συνεδρίαση. 24ης Τακτικής Γενικής Συνέλευσης του ΣΕΤΕ*. Retrieved from http://sete.gr/el/kentro-typou/omilies-parousiaseis/2016/omilia-a-andreadi-stin-kleisti-sunedriasi/

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 31ου Πανελληνίου Συνεδρίου Στατιστικής (2018), σελ.236-250

ΕΠΙΛΟΓΗ ΥΠΟΔΕΙΓΜΑΤΩΝ ΜΕ ΠΛΗΡΟΦΟΡΙΑΚΑ ΚΡΙΤΗΡΙΑ ΣΕ ΧΩΡΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ

Α. Τσιμπάνος[1] και Χ.Αγιακλόγλου[2]
[1]Τμήμα Στατιστικής και Αναλογιστικών-Χρηματοοικονομικών Μαθηματικών, Πανεπιστήμιο Αιγαίου
[2]Τμήμα Οικονομικής Επιστήμης, Πανεπιστήμιο Πειραιώς
atsibanos@aegean.gr, agiaklis@unipi.gr

ΠΕΡΙΛΗΨΗ

Στην ανάλυση χωρικών δεδομένων το εξεταζόμενο δείγμα θεωρείται ότι προέρχεται από την πραγματοποίηση κάποιας χωρικής διαδικασίας. Οι πιο γνωστές χωρικές διαδικασίες είναι η χωρική αυτοπαλίνδρομη διαδικασία πρώτης τάξης SAR(1), η χωρική διαδικασία κινητού μέσου πρώτης τάξης SMA(1) και η μικτή χωρική αυτοπαλίνδρομη κινητού μέσου διαδικασία SARMA(1,1) που αποτελούν επέκταση των αντίστοιχων διαδικασιών της ανάλυσης χρονοσειρών στο γεωγραφικό χώρο επιτρέποντας εξάρτηση προς πολλές κατευθύνσεις. Στην εργασία αυτή, με τη βοήθεια προσομοίωσης, διερευνάται η δυνατότητα των τριών πιο συχνά χρησιμοποιούμενων πληροφοριακών κριτηρίων, δηλαδή του κριτηρίου AIC του Akaike, του κριτηρίου BIC του Schwarz και του κριτηρίου HQC των Hannan και Quinn να επιλέγουν το καλύτερο υπόδειγμα. Τα αποτελέσματα δείχνουν ότι η συμπεριφορά των τριών κριτηρίων εξαρτάται από τις τιμές των συντελεστών και το μέγεθος του δείγματος και σε μερικές περιπτώσεις οδηγούν σε παραπλανητικές επιλογές.

Λέξεις Κλειδιά: Χωρικές Διαδικασίες, Χωρική Αυτοσυσχέτιση, Πληροφοριακά Κριτήρια

1. ΕΙΣΑΓΩΓΗ

Βασικό χαρακτηριστικό των χωρικών δεδομένων (*spatial data*), δηλαδή των διαστρωμματικών δεδομένων που έχουν συλλεχθεί από διαφορετικές γεωγραφικές περιοχές, είναι η παρουσία χωρικής αυτοσυσχέτισης στις παρατηρήσεις. Η χωρική αυτοσυσχέτιση (*spatial autocorrelation*) είναι αποτέλεσμα της χωρικής εξάρτησης (*spatial dependence*), η οποία μαζί με τη χωρική ετερογένεια (*spatial heterogeneity*) αποτελούν τις χωρικές επιδράσεις (*spatial effects*) που οφείλονται στο είδος των δεδομένων (Anselin, 1988). Η ύπαρξη χωρικής αυτοσυσχέτισης, δηλαδή συσχέτισης των παρατηρήσεων σύμφωνα με τις γεωγραφικές τους θέσεις, παραβιάζει την υπόθεση του τυχαίου δείγματος κάτι που θα πρέπει να λαμβάνεται υπ' όψιν όταν τα χωρικά δεδομένα χρησιμοποιούνται για την εκτίμηση ενός οικονομετρικού υποδείγματος.

Αναμφισβήτητα, όταν μελετάται ένα χωρικό δείγμα δεν θα πρέπει να θεωρείται ότι είναι τυχαίο αλλά ότι προέρχεται από την πραγματοποίηση κάποιας χωρικής διαδικασίας όπως συμβαίνει και στις χρονοσειρές. Κατά συνέπεια, η γνωστή από την ανάλυση χρονοσειρών μεθοδολογία των Box και Jenkins (1976) έχει επεκταθεί στην ανάλυση χωρικών δεδομένων και αντίστοιχα υποδείγματα έχουν προταθεί που εκφράζουν τη χωρική εξάρτηση μεταξύ των παρατηρήσεων (Cliff και Ord, 1981). Ωστόσο, στη χωρική οικονομετρία, η ταυτοποίηση και εκτίμηση αυτών των υποδειγμάτων δεν αποσκοπεί στην υλοποίηση προβλέψεων όπως στις χρονοσειρές αλλά στην εμβάθυνση του ερευνητή στο πρόβλημα που μελετάει αναγνωρίζοντας τη δομή των δεδομένων.

Η εφαρμογή πληροφοριακών κριτηρίων στην ανάλυση χρονοσειρών έχει καθιερωθεί ως βοήθημα για την επιλογή του καταλληλότερου υποδείγματος. Στην εργασία αυτή, με τη βοήθεια προσομοίωσης, διερευνάται η ικανότητα των τριών πιο συχνά χρησιμοποιούμενων πληροφοριακών κριτηρίων, δηλαδή του κριτηρίου AIC του Akaike, του κριτηρίου BIC του Schwarz και του κριτηρίου HQC των Hannan και Quinn να επιλέγουν χωρικές διαδικασίες. Τα αποτελέσματα δείχνουν ότι η επιτυχία επιλογής του σωστού υποδείγματος εξαρτάται από τις πραγματικές τιμές των παραμέτρων αλλά και από το μέγεθος του δείγματος.

2. ΧΩΡΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ ΚΑΙ ΠΛΗΡΟΦΟΡΙΑΚΑ ΚΡΙΤΗΡΙΑ

Στην ενότητα αυτή περιγράφονται συνοπτικά οι πιο συνηθισμένες χωρικές διαδικασίες, δηλαδή η χωρική αυτοπαλίνδρομη διαδικασία πρώτης τάξης (*Spatial Autoregressive Process*–SAR(1)), η χωρική διαδικασία κινητού μέσου πρώτης τάξης (*Spatial Moving Average Process*–SMA(1)) και η μικτή χωρική αυτοπαλίνδρομη κινητού μέσου διαδικασία (*Spatial Autoregressive Moving Average Process*–SARMA(1,1)) και ο τρόπος εκτίμησής τους. Επιπρόσθετα, παρουσιάζονται τα πληροφοριακά κριτήρια που θα χρησιμοποιηθούν στις προσομοιώσεις για τη σύγκριση μεταξύ των τριών διαδικασιών και την επιλογή αυτής που προσαρμόζεται καλύτερα σε κάθε δείγμα.

Η χωρική αυτοπαλίνδρομη διαδικασία πρώτης τάξης παρουσιάστηκε αρχικά από τους Whittle (1954) και Besag (1974) και αποτελεί επέκταση της γνωστής από την ανάλυση χρονοσειρών αυτοπαλίνδρομης διαδικασίας AR(1) στο γεωγραφικό χώρο. Χρησιμοποιώντας μήτρες (LeSage και Pace, 2009) και θεωρώντας μηδενική μέση τιμή για την εξεταζόμενη μεταβλητή η SAR(1) διαδικασία ορίζεται ως $\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \boldsymbol{\varepsilon}$ όπου $\mathbf{y}$ είναι το $(n\times 1)$ διάνυσμα των παρατηρήσεων της διαδικασίας στα *n* σημεία του χώρου, $\mathbf{W}$ είναι η $(n\times n)$ μήτρα χωρικών σταθμίσεων που καθορίζει την αλληλεπίδραση μεταξύ των *n* γειτονικών περιοχών, ρ είναι ο χωρικός αυτοπαλίνδρομος συντελεστής και $\boldsymbol{\varepsilon}$ ένα $(n\times 1)$ τυχαίο διάνυσμα με τιμές λευκού θορύβου. Υποθέτοντας ότι $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ η λογαριθμική συνάρτηση πιθανοφάνειας της SAR(1) διαδικασίας είναι:

$$\ln L\left(\rho,\sigma^2\right) = -\left(n/2\right)\ln\left(2\pi\right) - \left(n/2\right)\ln\sigma^2 - \left(1/2\sigma^2\right)\left(\mathbf{y}-\rho\mathbf{W}\mathbf{y}\right)'\left(\mathbf{y}-\rho\mathbf{W}\mathbf{y}\right) + \ln\left|\mathbf{I}-\rho\mathbf{W}\right|$$

όπου $\left|\mathbf{I}-\rho\mathbf{W}\right|$ είναι η ορίζουσα από το μετασχηματισμό του τυχαίου διανύσματος $\boldsymbol{\varepsilon}$ στο διάνυσμα $\mathbf{y}$ και επειδή για τον εκτιμητή μέγιστης πιθανοφάνειας της διακύμανσης

της διαδικασίας ισχύει $\sigma^2_{ML} = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}/n$ η λογαριθμική συνάρτηση πιθανοφάνειας γράφεται ισοδύναμα ως εξής:

$$\ln L(\rho) = -(n/2)\ln(2\pi) - (n/2) - (n/2)\ln\left((1/n)(\mathbf{y} - \rho\mathbf{W}\mathbf{y})'(\mathbf{y} - \rho\mathbf{W}\mathbf{y})\right) + \ln|\mathbf{I} - \rho\mathbf{W}|$$

Η χωρική διαδικασία κινητού μέσου πρώτης τάξης (Haining, 1978) ορίζεται ως $\mathbf{y} = \boldsymbol{\varepsilon} - \lambda\mathbf{W}\boldsymbol{\varepsilon}$ όπου $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ και λ είναι ο συντελεστής του χωρικού κινητού μέσου. Η λογαριθμική συνάρτηση πιθανοφάνειας της SMA(1) διαδικασίας είναι:

$\ln L(\lambda, \sigma^2) = -(n/2)\ln(2\pi) - (n/2)\ln\sigma^2 - (1/2\sigma^2)\left[(\mathbf{I} - \lambda\mathbf{W})^{-1}\mathbf{y}\right]'(\mathbf{I} - \lambda\mathbf{W})^{-1}\mathbf{y} - \ln|\mathbf{I} - \lambda\mathbf{W}|$ με $|\mathbf{I} - \lambda\mathbf{W}|$ την ορίζουσα από το μετασχηματισμό του τυχαίου διανύσματος $\boldsymbol{\varepsilon}$ στο διάνυσμα $\mathbf{y}$ και επειδή ισχύει πάλι $\sigma^2_{ML} = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}/n$ γράφεται ως συνάρτηση μόνο της παραμέτρου λ ως εξής:

$$\ln L(\lambda) = -(n/2)\ln(2\pi) - (n/2) - (n/2)\ln\left((1/n)\left[(\mathbf{I} - \lambda\mathbf{W})^{-1}\mathbf{y}\right]'(\mathbf{I} - \lambda\mathbf{W})^{-1}\mathbf{y}\right) - \ln|\mathbf{I} - \lambda\mathbf{W}|$$

Στην ανάλυση χρονοσειρών ορίζονται μικτές διαδικασίες που συνδυάζουν αυτοπαλίνδρομους όρους και όρους κινητών μέσων. Με τον ίδιο τρόπο ο Huang (1984) ορίζει μικτές χωρικές διαδικασίας. Η μικτή χωρική αυτοπαλίνδρομη κινητού μέσου διαδικασία SARMA(1,1) θα είναι $\mathbf{y} = \rho\mathbf{W}_1\mathbf{y} + \boldsymbol{\varepsilon} - \lambda\mathbf{W}_2\boldsymbol{\varepsilon}$ με $\mathbf{W}_1$ και $\mathbf{W}_2$ τις $(n \times n)$ μήτρες χωρικών σταθμίσεων για τον αυτοπαλίνδρομο όρο και τον όρο κινητού μέσου. Στο σημείο αυτό πρέπει να αναφερθεί ότι για να οριστεί η SARMA(1,1) διαδικασία είναι αναγκαίο να χρησιμοποιηθούν διαφορετικές μήτρες χωρικών σταθμίσεων για τα δύο μέρη της διαδικασίας (Mur και Angulo, 2007). Επειδή η SAR(1) διαδικασία εκφράζει συσχέτιση που διαχέεται σε ολόκληρη τη γεωγραφική περιοχή ενώ η SMA(1) διαδικασία συσχέτιση μεταξύ γειτονικών περιοχών στις εμπειρικές εφαρμογές το αυτοπαλίνδρομο μέρος στοχεύει στην περιγραφή των ολικών επιδράσεων και το μέρος του κινητού μέσου στην περιγραφή των τοπικών επιδράσεων. Επιπρόσθετα, εφόσον $\mathbf{W}_1 \neq \mathbf{W}_2$ η SARMA(1,1) διαδικασία ορίζεται και όταν ισχύει $\rho = \lambda$ σε αντίθεση με την ARMA(1,1) διαδικασία της ανάλυσης χρονοσειρών που σε αυτή την περίπτωση εμφανίζεται το πρόβλημα του κοινού παράγοντα και το υπόδειγμα δεν μπορεί να εκτιμηθεί. Η λογαριθμική συνάρτηση πιθανοφάνειας της SARMA(1,1) διαδικασίας είναι:

$$\ln L(\rho, \lambda, \sigma^2) = -(n/2)\ln(2\pi) - (n/2)\ln\sigma^2 - (1/2\sigma^2)\left[\mathbf{B}^{-1}\mathbf{A}\mathbf{y}\right]'\mathbf{B}^{-1}\mathbf{A}\mathbf{y} + \ln|\mathbf{A}| - \ln|\mathbf{B}|$$

με $\mathbf{A} = \mathbf{I} - \rho\mathbf{W}_1$ και $\mathbf{B} = \mathbf{I} - \lambda\mathbf{W}_2$ και που γράφεται ισοδύναμα ως:

$$\ln L(\rho, \lambda) = -(n/2)\ln(2\pi) - (n/2) - (n/2)\ln\left((1/n)\left[\mathbf{B}^{-1}\mathbf{A}\mathbf{y}\right]'\mathbf{B}^{-1}\mathbf{A}\mathbf{y}\right) + \ln|\mathbf{A}| - \ln|\mathbf{B}|$$

Στις χρονοσειρές είναι γνωστό ότι για να είναι στάσιμη η AR(1) διαδικασία και αντιστρέψιμη η MA(1) διαδικασία οι συντελεστές ρ και λ θα πρέπει να λαμβάνουν τιμές αυστηρά στο διάστημα (-1, +1). Στις αντίστοιχες χωρικές διαδικασίες το εφικτό εύρος των χωρικών συντελεστών δεν περιορίζεται στο προηγούμενο διάστημα αλλά,

όπως έχει δείξει ο Ord (1975), στο διάστημα $\left(1/\omega_{\min},1/\omega_{\max}\right)$ όπου $\omega_{\min}$ και $\omega_{\max}$ είναι τα πραγματικά μέρη της ελάχιστης και μέγιστης ιδιοτιμής της μήτρας χωρικών σταθμίσεων. Ωστόσο, σε μια τυποποιημένη κατά γραμμή μήτρα χωρικών σταθμίσεων η μεγαλύτερη ιδιοτιμή της είναι πάντα +1 κάτι που εξασφαλίζει ότι το επάνω όριο του διαστήματος θα είναι +1 ενώ για το κάτω όριο δεν υπάρχει κάποια συγκεκριμένη τιμή που συνήθως είναι μικρότερο του -1. Εφόσον τηρείται το προηγούμενο διάστημα εξασφαλίζεται ότι η ορίζουσα του Ιακωβιανού μετασχηματισμού θα είναι θετική και κατά συνέπεια θα ορίζονται ο λογάριθμός της και οι συναρτήσεις πιθανοφάνειας των διαδικασιών. Οι συναρτήσεις πιθανοφάνειας θα πρέπει χρησιμοποιώντας το διαθέσιμο χωρικό δείγμα να μεγιστοποιηθούν με κάποια αριθμητική μέθοδο για να εκτιμηθούν οι τιμές των συντελεστών ρ και λ**.**

Στη διαδικασία της μεγιστοποίησης διευκολύνει η διάσπαση της Ιακωβιανής ορίζουσας στις ιδιοτιμές της όπως έχει προτείνει ο Ord (1975) δηλαδή, παραδείγματος χάριν για το συντελεστή ρ, ως $\ln\left|\mathbf{I}-\rho\mathbf{W}\right|=\sum_{j=1}^{n}\ln\left(1-\rho\,\omega_j\right)$ όπου με ω_i συμβολίζονται οι ιδιοτιμές της μήτρας χωρικών σταθμίσεων **W**. Ένα πρόβλημα που προκύπτει αφορά την ασυμμετρία της μήτρας χωρικών σταθμίσεων που μπορεί να προκύψει λόγω της τυποποίησης ανά γραμμή ή λόγω της κατασκευής της με μη συμμετρικό κριτήριο χωρικής γειτονίας όπως το κριτήριο των κοντινών γειτόνων. Σε αυτές τις περιπτώσεις, η μήτρα **W** θα έχει και μιγαδικές ιδιοτιμές που θα πρέπει να ληφθούν υπ' όψιν στη διαδικασία της μεγιστοποίησης. Στηριζόμενοι στο ότι οι μιγαδικές ιδιοτιμές θα εμφανίζονται πάντα σε συζυγή ζεύγη οι Bivand *et al.* (2013) τροποποιούν τη μέθοδο του Ord (1975) ώστε να μπορεί να χρησιμοποιηθεί και σε μη συμμετρικές μήτρες χωρικών σταθμίσεων. Ειδικότερα, εάν $\omega_j=\alpha_j+ib_j$ και $\bar{\omega}_j=\alpha_j-ib_j$ είναι οι $2k$ μιγαδικές ιδιοτιμές και ζ_j οι *n-k* πραγματικές ιδιοτιμές τότε για τον λογάριθμο της Ιακωβιανής ορίζουσας ισχύει $\ln\left|\mathbf{I}-\rho\mathbf{W}\right|=L_1+L_2$, όπου

$$L_1=\sum_{j=1}^{k}\ln\left[\left(1-\rho\,\omega_j\right)\left(1-\rho\,\bar{\omega}_j\right)\right]=\sum_{j=1}^{k}\ln\left[\left(1-\rho\alpha_j\right)^2+\left(\rho b_j\right)^2\right]\ \text{και}\ L_2=\sum_{j=k+1}^{n-k}\ln\left(1-\rho\,\zeta_j\right).$$

Τα υποδείγματα των τριών χωρικών διαδικασιών εκτιμώνται με την εφαρμογή της μεθόδου της μέγιστης πιθανοφάνειας και στη συνέχεια ο ερευνητής θα πρέπει να επιλέξει αυτό που είναι πιο αντιπροσωπευτικό στο εξεταζόμενο δείγμα. Τα πληροφοριακά κριτήρια που χρησιμοποιούνται ευρέως για την επιλογή υποδειγμάτων τόσο στην ανάλυση χρονοσειρών όσο και στο πλαίσιο της γραμμικής παλινδρόμησης θα μπορούσαν να βοηθήσουν και στην επιλογή της σωστής χωρικής διαδικασίας. Η φιλοσοφία αυτών των κριτηρίων βασίζεται στην ποσοτικοποίηση της καλής προσαρμογής ενός υποδείγματος στις διαθέσιμες παρατηρήσεις περιλαμβάνοντας κάποιο τύπο ποινής για τον αριθμό των παραμέτρων του και το μέγεθος του δείγματος. Χρονολογικά το πρώτο κριτήριο αυτής της μορφής είναι το πληροφοριακό κριτήριο AIC (*Akaike Information Criterion*) του Akaike (1973) που ορίζεται ως $\mathrm{AIC}=-2\ln\hat{L}+2p$. Δυο άλλα γνωστά κριτήρια είναι το BIC (*Bayesian Information Criterion*) του Schwarz (1978) που ορίζεται ως $\mathrm{BIC}=-2\ln\hat{L}+p\ln n$ και το

πληροφοριακό κριτήριο HQC (*Hannan Quinn Information Criterion*) των Hannan και Quinn (1978) που ορίζεται ως $\mathrm{HQC} = -2\ln\hat{L} + 2p\ln\ln n$. Στα κριτήρια αυτά, $\ln\hat{L}$ είναι η μέγιστη τιμή της λογαριθμικής συνάρτησης πιθανοφάνειας, p ο αριθμός των παραμέτρων κάθε υποδείγματος και n το μέγεθος του δείγματος με το οποίο έγινε η εκτίμηση. Για όλα τα υποψήφια υποδείγματα υπολογίζονται οι τιμές του πληροφοριακού κριτηρίου και επιλέγεται ως καταλληλότερο αυτό για το οποίο η τιμή του κριτηρίου είναι ελάχιστη.

3. ΠΡΟΣΟΜΟΙΩΣΗ

Στην ενότητα αυτή εξετάζεται με τη βοήθεια προσομοίωσης η συμπεριφορά και η επίδοση των τριών πληροφοριακών κριτηρίων, δηλαδή του AIC, του BIC και του HQC να εντοπίζουν τη σωστή χωρική διαδικασία μεταξύ των εναλλακτικών SAR(1), SMA(1) και SARMA(1,1). Για τις ανάγκες της προσομοίωσης, η χωρική εξάρτηση ορίστηκε με μία ($n \times n$) τυποποιημένη κατά γραμμή μήτρα χωρικών σταθμίσεων που κατασκευάστηκε με βάση τα κριτήρια του πύργου (*rook contiguity*) για το χωρικό αυτοπαλίνδρομο όρο (μήτρα $\mathbf{W}_1$**)** και της βασίλισσας (*queen contiguity*) για τον όρο του χωρικού κινητού μέσου (μήτρα $\mathbf{W}_2$**)** επάνω σε ένα ομαλό τετραγωνικό πλέγμα διαστάσεων 10x10, και 20x20 παρέχοντας δείγματα 100 και 400 παρατηρήσεων αντίστοιχα. Επιπρόσθετα, για να εξεταστεί η περίπτωση μη συμμετρικών μητρών χωρικών σταθμίσεων χρησιμοποιήθηκε η γεωγραφική δομή της Ελλάδας σε επίπεδο δήμων του προγράμματος Καλλικράτης (325 δήμοι) και με γεωγραφικές συντεταγμένες τα γεωγραφικά κεντροειδή κάθε πολυγώνου κατασκευάστηκαν δύο τυποποιημένες κατά γραμμή μήτρες χωρικών σταθμίσεων εφαρμόζοντας τα κριτήρια των τεσσάρων κοντινότερων γειτόνων για τον αυτοπαλίνδρομο όρο (μήτρα $\mathbf{W}_1$**)** και των οκτώ κοντινότερων γειτόνων για τον κινητό μέσο (μήτρα $\mathbf{W}_2$**).** Όπως φαίνεται στον Πίνακα 1 που περιέχει το εφικτό εύρος των χωρικών συντελεστών σύμφωνα με τις μήτρες χωρικών σταθμίσεων που αναφέρθηκαν, το κάτω όριο των συντελεστών σε όλες τις μήτρες εκτός από αυτές που κατασκευάστηκαν με το κριτήριο του πύργου είναι σημαντικά μικρότερο του -1. Επιπρόσθετα, στις μήτρες χωρικών σταθμίσεων που δημιουργήθηκαν με τα κριτήρια του πύργου και των κοντινότερων γειτόνων προκύπτουν και μιγαδικές ιδιοτιμές που πρέπει να ληφθούν υπ' όψιν στη διαδικασία της μεγιστοποίησης.

Η SAR(1) διαδικασία παράγεται από τη σχέση $\mathbf{y} = \left(\mathbf{I} - \rho\mathbf{W}_1\right)^{-1}\boldsymbol{\varepsilon}$ ενώ η SMA(1) διαδικασία από τη σχέση $\mathbf{y} = \left(\mathbf{I} - \lambda\mathbf{W}_2\right)\boldsymbol{\varepsilon}$ με τους συντελεστές ρ και λ να λαμβάνουν τις τιμές -0,9, -0,8, -0,6, -0,4, -0,2, 0,2, 0,4, 0,6, 0,8 και 0,9. Η SARMA(1,1) διαδικασία παράγεται ως $\mathbf{y} = \left(\mathbf{I} - \rho\mathbf{W}_1\right)^{-1}\left(\mathbf{I} - \lambda\mathbf{W}_2\right)\boldsymbol{\varepsilon}$ και θεωρήθηκαν για τους συντελεστές ρ και λ συνδυασμοί τιμών με $\rho \in$(0,2, 0,5, 0,9), $\lambda \in$ (-0,9, -0,8, -0,6, -0,4, -0,2, 0,2, 0,4, 0,6, 0,8, 0,9) και $\lambda \in$(0,2, 0,5, 0,9), $\rho \in$ (-0,9, -0,8, -0,6, -0,4, -0,2, 0,2, 0,4, 0,6, 0,8, 0,9). Το διάνυσμα των τυχαίων σφαλμάτων $\boldsymbol{\varepsilon}$ προέρχεται από $N\left(\mathbf{0}, \mathbf{I}\right)$. Η προσομοίωση πραγματοποιήθηκε στο στατιστικό πρόγραμμα R με τη βοήθεια του πακέτου SPDEP (Bivand, 2015) στη δημιουργία των μητρών χωρικών σταθμίσεων.

Εκτελέστηκαν 1.000 επαναλήψεις και εκτιμήθηκαν με τη μέθοδο της μέγιστης πιθανοφάνειας οι συντελεστές των τριών χωρικών διαδικασιών χρησιμοποιώντας τη συνάρτηση nlminb, υπολογίστηκαν οι τιμές των κριτηρίων AIC, BIC και HQC και κατεγράφησαν τα ποσοστά επιλογής κάθε διαδικασίας.

Πίνακας 1. *Πάνω και κάτω όρια χωρικών συντελεστών*

N	Σταθμίσεις	Μιγαδικές Ιδιοτιμές	Ελάχιστη Ιδιοτιμή Πραγματικό Μέρος	Μέγιστη Ιδιοτιμή	Κάτω Όριο ρ, λ	Πάνω Όριο ρ, λ
100	Πύργος	Ναι	-1	1	-1	1
100	Βασίλισσα	Όχι	-0,508	1	-1,970	1
400	Πύργος	Ναι	-1	1	-1	1
400	Βασίλισσα	Όχι	-0,520	1	-1,921	1
325	4 Γείτονες	Ναι	-0,676	1	-1,478	1
325	8 Γείτονες	Ναι	-0,368	1	-2,720	1

Πίνακας 2. *Ποσοστά επιλογής των τριών υποδειγμάτων για όταν ισχύει η χωρική αυτοπαλίνδρομη διαδικασία πρώτης τάξης SAR(1) για διάφορες τιμές του ρ*

N	***ρ***	**-0,9**	**-0,8**	**-0,6**	**-0,4**	**-0,2**	**0,2**	**0,4**	**0,6**	**0,8**	**0,9**
100	AIC-SAR(1)	**80,7**	**81,1**	**81,9**	**76,2**	**57,4**	**58,8**	**76,4**	**83,4**	**83,2**	**83,8**
	AIC-SMA(1)	0	0	1,2	11,5	34,1	31,3	13,4	3	0	0
	AIC-SARMA(1,1)	19,3	18,9	16,9	12,3	8,5	9,9	10,2	13,6	16,8	16,2
	BIC-SAR(1)	**96,8**	**96,5**	**95,5**	**85,1**	**62,4**	**64,9**	**83,4**	**93,8**	**96,7**	**96,9**
	BIC-SMA(1)	0	0	1,7	13	36,5	32,8	13,7	3,8	0,4	0
	BIC-SARMA(1,1)	3,2	3,5	2,8	1,9	1,1	2,3	2,9	2,4	2,9	3,1
	HQC-SAR(1)	**89,3**	**89,3**	**89,1**	**81,9**	**60,7**	**63,5**	**81**	**89,7**	**92,6**	**91,7**
	HQC-SMA(1)	0	0	1,4	12,6	35,9	32,2	13,6	3,6	0,1	0
	HQC-SARMA(1,1)	10,7	10,7	9,5	5,5	3,4	4,3	5,4	6,7	7,3	8,3
400	AIC-SAR(1)	**83,1**	**83,2**	**84,5**	**84,7**	**76,8**	**75,8**	**82,5**	**83,2**	**82,8**	**82,1**
	AIC-SMA(1)	0	0	0	0,1	13,1	14,5	0,7	0	0	0
	AIC-SARMA(1,1)	16,9	16,8	15,5	15,2	10,1	9,7	16,8	16,8	17,2	17,9
	BIC-SAR(1)	**98,3**	**98,3**	**98,6**	**98,6**	**85,5**	**84,4**	**97,5**	**98,6**	**98,7**	**98,6**
	BIC-SMA(1)	0	0	0	0,4	13,9	15,2	1,8	0	0	0
	BIC-SARMA(1,1)	1,7	1,7	1,4	1	0,6	0,4	0,7	1,4	1,3	1,4
	HQC-SAR(1)	**94,1**	**94,7**	**94,3**	**94,2**	**83,4**	**82,4**	**94,3**	**93,8**	**93,8**	**93**
	HQC-SMA(1)	0	0	0	0,3	13,8	15,2	1,4	0	0	0
	HQC-SARMA(1,1)	5,9	5,3	5,7	5,5	2,8	2,4	4,3	6,2	6,2	7
325	AIC-SAR(1)	**83,8**	**84,2**	**84,6**	**83,7**	**69,5**	**69,4**	**82,7**	**84,7**	**84,3**	**83,9**
	AIC-SMA(1)	0	0	0	2,8	22,2	21,3	2,9	0	0	0
	AIC-SARMA(1,1)	16,2	15,8	15,4	13,5	8,3	9,3	14,4	15,3	15,7	16,1
	BIC-SAR(1)	**98,1**	**98**	**97,9**	**93,8**	**76,3**	**77,3**	**93,8**	**97,9**	**98,1**	**97,8**
	BIC-SMA(1)	0	0	0,3	4,7	22,6	21,6	4,9	0,2	0	0
	BIC-SARMA(1,1)	1,9	2	1,8	1,5	1,1	1,1	1,3	1,9	1,9	2,2
	HQC-SAR(1)	**94,7**	**94,6**	**94,6**	**91,5**	**74**	**75,3**	**91,7**	**93**	**94,6**	**93,8**
	HQC-SMA(1)	0	0	0	4,6	22,5	21,4	4,6	0,1	0	0
	HQC-SARMA(1,1)	5,3	5,4	5,4	3,9	3,5	3,3	3,7	6,9	5,4	6,2

Πίνακας 3. *Ποσοστά επιλογής των τριών υποδειγμάτων για όταν ισχύει η χωρική διαδικασία κινητού μέσου πρώτης τάξης SMA(1) για διάφορες τιμές του λ*

N	*λ*	**-0,9**	**-0,8**	**-0,6**	**-0,4**	**-0,2**	**0,2**	**0,4**	**0,6**	**0,8**	**0,9**
100	AIC-SAR(1)	4,5	8,2	16,7	28	38,8	32,2	15,1	3,1	0,1	0
	AIC-SMA(1)	**83,4**	**81,9**	**75,5**	**65,2**	**53,9**	**58,9**	**75,4**	**84,1**	**84,1**	**83**
	AIC-SARMA(1,1)	12,1	9,9	7,8	6,8	7,3	8,9	9,5	12,8	15,8	17
	BIC-SAR(1)	6,8	9,7	17,3	28,9	40,9	35	15,2	4	0,2	0
	BIC-SMA(1)	**92,3**	**89,4**	**82**	**70,2**	**57,8**	**63**	**82**	**92,7**	**96,2**	**93,7**
	BIC-SARMA(1,1)	0,9	0,9	0,7	0,9	1,3	2	2,8	3,3	3,6	6,3
	HQC-SAR(1)	6	9,5	17,2	28,6	40	33,8	15,2	3,8	0,1	0
	HQC-SMA(1)	**89,5**	**87,3**	**79,8**	**68,4**	**56,5**	**61,3**	**79,8**	**89,9**	**92,3**	**89,8**
	HQC-SARMA(1,1)	4,5	3,2	3	3	3,5	4,9	5	6,3	7,6	10,2
400	AIC-SAR(1)	0	0	0,5	6,9	25,4	18,9	2,2	0	0	0
	AIC-SMA(1)	**82,6**	**82,7**	**83**	**81,5**	**66,1**	**72,4**	**83,2**	**84,6**	**83,8**	**83,5**
	AIC-SARMA(1,1)	17,4	17,3	16,5	11,6	8,5	8,7	14,6	15,4	16,2	16,5
	BIC-SAR(1)	0	0,1	2	8,1	26,1	20	3,4	0,3	0	0
	BIC-SMA(1)	**98,7**	**98,7**	**97,4**	**91,4**	**73,5**	**79,3**	**95,5**	**98,4**	**98,5**	**98,7**
	BIC-SARMA(1,1)	1,3	1,2	0,6	0,5	0,4	0,7	1,1	1,3	1,5	1,3
	HQC-SAR(1)	0	0	1,4	7,9	25,8	19,3	3	0,2	0	0
	HQC-SMA(1)	**94,5**	**94,3**	**93,3**	**89,2**	**71,3**	**77,4**	**92,5**	**93,9**	**94,4**	**93,7**
	HQC-SARMA(1,1)	5,5	5,7	5,3	2,9	2,9	3,3	4,5	5,9	5,6	6,3
325	AIC-SAR(1)	0,7	1,5	7,2	16,7	32,7	26,8	6,7	0	0	0
	AIC-SMA(1)	**83,3**	**83,2**	**81,7**	**74,6**	**59,1**	**63,8**	**81,8**	**83,8**	**84,6**	**83,5**
	AIC-SARMA(1,1)	16	15,3	11,1	8,7	8,2	9,4	11,5	16,2	15,4	16,5
	BIC-SAR(1)	1,7	3,5	8,3	16,9	34,1	28,1	7,8	0,6	0	0
	BIC-SMA(1)	**97,7**	**96,1**	**91,2**	**82,6**	**65,4**	**71,2**	**91,3**	**98,2**	**97,8**	**98**
	BIC-SARMA(1,1)	0,6	0,4	0,5	0,5	0,5	0,7	0,9	1,2	2,2	2
	HQC-SAR(1)	1,2	2,9	8,3	16,9	33,8	28	7,7	0,3	0	0
	HQC-SMA(1)	**93,4**	**92,2**	**88,8**	**80,5**	**63,9**	**69,7**	**88,8**	**93,5**	**94,1**	**94,3**
	HQC-SARMA(1,1)	5,4	4,9	2,9	2,6	2,3	2,3	3,5	6,2	5,9	5,7

Τα ποσοστά που τα πληροφοριακά κριτήρια επιλέγουν τις τρεις υποψήφιες χωρικές διαδικασίες παρουσιάζονται στον Πίνακα 2 για την περίπτωση που η πραγματική διαδικασία είναι η SAR(1). Όπως φαίνεται σε αυτόν τον πίνακα, η συμπεριφορά των κριτηρίων εξαρτάται από τις τιμές του συντελεστή ρ αλλά και το μέγεθος του δείγματος. Ειδικότερα, στις πολύ μικρές κατ' απόλυτη τιμή τιμές του ρ η ικανότητα των κριτηρίων να υποδείξουν τη σωστή διαδικασία είναι περιορισμένη και αρκετές φορές επιλέγεται εσφαλμένα η SMA(1) διαδικασία. Τα ποσοστά σωστής επιλογής είναι σαφώς πιο ικανοποιητικά για μέτριες και μεγαλύτερες απόλυτες τιμές του ρ. Οι επιδόσεις τους βελτιώνονται στα δείγματα των 325 και των 400 παρατηρήσεων. Γενικά, την καλύτερη επίδοση στον εντοπισμό της SAR(1) διαδικασίας την επιτυγχάνει το κριτήριο BIC με ποσοστά επιλογής που πλησιάζουν στο 98% όταν ο συντελεστής ρ σε απόλυτη τιμή είναι μεγαλύτερος ή ίσος του 0,4. Το κριτήριο που ακολουθεί με ποσοστά σωστής επιλογής αρκετά κοντά σε αυτά του BIC είναι το κριτήριο HQC. Τα χαμηλότερα ποσοστά σωστής επιλογής διαπιστώνονται για το κριτήριο AIC το όποιο επιλέγει σε αξιοσημείωτα ποσοστά την SARMA(1,1) διαδικασία κάτι σε κάποιο βαθμό αναμενόμενο καθώς είναι γνωστή η τάση αυτού του κριτηρίου να επιλέγει υποδείγματα με μεγάλο αριθμό παραμέτρων.

Πίνακας 4. *Ποσοστά επιλογής των τριών υποδειγμάτων για όταν ισχύει η μικτή χωρική διαδικασία SARMA(1,1) με ρ =0,2 και διάφορες τιμές του λ*

N	*λ*	**-0,9**	**-0,8**	**-0,6**	**-0,4**	**-0,2**	**0,2**	**0,4**	**0,6**	**0,8**	**0,9**
100	AIC-SAR(1)	10,5	16,8	32,5	45,8	57,6	43,9	16,7	2,4	0,1	0
	AIC-SMA(1)	58,3	58,2	55,1	49,9	39,4	34,1	51,9	62,8	62,1	56,1
	AIC-SARMA(1,1)	**31,2**	**25**	**12,4**	**4,3**	**3**	**22**	**31,4**	**34,8**	**37,8**	**43,9**
	BIC-SAR(1)	20,3	25,8	37,3	48,2	59	52,8	21,6	3	0,2	0
	BIC-SMA(1)	75,5	72,7	62,3	51,6	40	41,6	68,2	83,9	84,9	81
	BIC-SARMA(1,1)	**4,2**	**1,5**	**0,4**	**0,2**	**1**	**5,6**	**10,2**	**13,1**	**14,9**	**19**
	HQC-SAR(1)	16	22	35,8	47,5	58,3	48,3	19,1	2,7	0,2	0
	HQC-SMA(1)	68	67,5	61,3	51,6	39,8	38,9	61,7	74,8	73	69
	HQC-SARMA(1,1)	**16**	**10,5**	**2,9**	**0,9**	**1,9**	**12,8**	**19,2**	**22,5**	**26,8**	**31**
400	AIC-SAR(1)	0	0	1,4	17,5	58	37,4	1,5	0	0	0
	AIC-SMA(1)	17,5	18,7	21,5	24,3	24,2	15,5	31,9	31	26,2	19,8
	AIC-SARMA(1,1)	**82,5**	**81,3**	**77,1**	**58,2**	**17,8**	**47,1**	**66,6**	**69**	**73,8**	**80,2**
	BIC-SAR(1)	0,1	0,9	13,5	43,2	68,2	62,1	3,8	0	0	0
	BIC-SMA(1)	55	57,3	61,4	53,4	31,8	25,2	71,1	71,9	67,6	59,1
	BIC-SARMA(1,1)	**44,9**	**41,8**	**25,1**	**3,4**	**0**	**12,7**	**25,1**	**28,1**	**32,4**	**40,9**
	HQC-SAR(1)	0	0,1	6,4	29,8	66,2	53,4	2,3	0	0	0
	HQC-SMA(1)	32,5	34	38,5	40,6	30,5	21	50,6	51,4	44	36,6
	HQC-SARMA(1,1)	**67,5**	**65,9**	**55,1**	**29,6**	**3,3**	**25,6**	**47,1**	**48,6**	**56**	**63,4**
325	AIC-SAR(1)	1,1	2,6	13,5	36,4	58,5	43,1	5	0,1	0	0
	AIC-SMA(1)	37,1	39	41,7	42	36,3	22,1	40,1	38,8	21,8	6,1
	AIC-SARMA(1,1)	**61,8**	**58,4**	**44,8**	**21,6**	**5,2**	**34,8**	**54,9**	**61,1**	**78,2**	**93,9**
	BIC-SAR(1)	7,8	14,6	30,6	46,1	62,1	63,5	9,3	0,2	0	0
	BIC-SMA(1)	74,1	73,2	66,2	53,8	37,9	28,6	70,6	73,1	56,9	27,7
	BIC-SARMA(1,1)	**18,1**	**12,2**	**3,2**	**0,1**	**0**	**7,9**	**20,1**	**26,7**	**43,1**	**72,3**
	HQC-SAR(1)	2,5	7,4	23,5	44,4	61,6	55,4	6,3	0,2	0	0
	HQC-SMA(1)	53,8	55,7	57,7	51,8	37,7	25,6	54,9	54,2	38	14,1
	HQC-SARMA(1,1)	**43,7**	**36,9**	**18,8**	**3,8**	**0,7**	**19**	**38,8**	**45,6**	**62**	**85,9**

Στον Πίνακα 3 συνοψίζονται τα ποσοστά επιλογής των τριών εναλλακτικών χωρικών διαδικασιών από τα πληροφοριακά κριτήρια όταν η πραγματική διαδικασία είναι η SMA(1). Είναι φανερό ότι η συμπεριφορά των κριτηρίων παρουσιάζει αρκετές ομοιότητες με την περίπτωση της SAR(1) διαδικασίας. Δηλαδή, στο δείγμα των 100 παρατηρήσεων για SMA(1) διαδικασίες με μικρές απόλυτες τιμές του λ τα ποσοστά επιλογής εσφαλμένων διαδικασιών είναι αρκετά υψηλά. Αντίθετα, για υψηλές τιμές του λ τα ποσοστά σωστής επιλογής ξεπερνάνε το 80%. Όπως παρατηρήθηκε στη SAR(1) διαδικασία έτσι και για όταν ισχύει η SMA(1) διαδικασία για το κριτήριο BIC καταγράφηκαν τα πιο μεγάλα ποσοστά εντοπισμού της πραγματικής διαδικασίας και ακολουθούν τα κριτήρια HQC και AIC με το τελευταίο να έχει τα χειρότερα ποσοστά. Στα δείγματα των 325 και των 400 παρατηρήσεων και για μεγάλες απόλυτες τιμές του λ τα ποσοστά σωστής επιλογής είναι εμφανώς υψηλότερα και για τα τρία κριτήρια. Παραδείγματος χάριν, στο δείγμα των 400 παρατηρήσεων για τιμές του λ ίσες με -0,4, -0,6, -0,8, -0,9 και 0,4, 0,6, 0,8, 0,9 τα ποσοστά επιλογής της SMA(1) διαδικασία από το κριτήριο BIC κυμαίνονται από 91,4% μέχρι 98,7%.

Πίνακας 5. *Ποσοστά επιλογής των τριών υποδειγμάτων για όταν ισχύει η μικτή χωρική διαδικασία SARMA(1,1) με ρ =0,5 και διάφορες τιμές του λ*

N	*λ*	**-0,9**	**-0,8**	**-0,6**	**-0,4**	**-0,2**	**0,2**	**0,4**	**0,6**	**0,8**	**0,9**
100	AIC-SAR(1)	11,1	20,8	43	63,6	77,6	69,2	48,4	15,5	0,6	0,1
	AIC-SMA(1)	4,3	5,1	6,8	8,8	8,9	5,5	5,6	13,9	16,3	10,7
	AIC-SARMA(1,1)	**84,6**	**74,1**	**50,2**	**27,6**	**13,5**	**25,3**	**46**	**70,6**	**83,1**	**89,2**
	BIC-SAR(1)	32,9	46	67,7	80,1	85,6	85,9	68,6	30,5	1,7	0,1
	BIC-SMA(1)	16,8	17,3	18,4	15,8	13,1	5,8	7,7	25,9	40,5	29,6
	BIC-SARMA(1,1)	**50,3**	**36,7**	**13,9**	**4,1**	**1,3**	**8,3**	**23,7**	**43,6**	**57,8**	**70,3**
	HQC-SAR(1)	21	31,3	54,5	74,6	82,8	78,3	58,6	22,7	1	0,1
	HQC-SMA(1)	8,4	9,8	11,7	12,1	11,7	5,7	6,8	19,2	24,9	17,8
	HQC-SARMA(1,1)	**70,6**	**58,9**	**33,8**	**13,3**	**5,5**	**16**	**34,6**	**58,1**	**74,1**	**82,1**
400	AIC-SAR(1)	0	0	4,2	26,4	65,9	59,2	15	0,4	0	0
	AIC-SMA(1)	0	0	0	0	0	0	0	0	0	0
	AIC-SARMA(1,1)	**100**	**100**	**95,8**	**73,6**	**34,1**	**40,8**	**85**	**99,6**	**100**	**100**
	BIC-SAR(1)	0,4	2,6	25,1	66,9	91,6	89,4	47,5	4,3	0	0
	BIC-SMA(1)	0	0	0,1	0,2	0,4	0,1	0	2,2	2	0,3
	BIC-SARMA(1,1)	**99,6**	**97,4**	**74,8**	**32,9**	**8**	**10,5**	**52,5**	**93,5**	**98**	**99,7**
	HQC-SAR(1)	0	0,4	10,5	45,7	80,7	76,8	28,6	1,5	0	0
	HQC-SMA(1)	0	0	0	0	0	0	0	0,4	0,1	0
	HQC-SARMA(1,1)	**100**	**99,6**	**89,5**	**54,3**	**19,3**	**23,2**	**71,4**	**98,1**	**99,9**	**100**
325	AIC-SAR(1)	1,2	3,1	15,8	42,2	70,1	65,8	23,8	1,9	0	0
	AIC-SMA(1)	0	0	0	0,1	0,2	0,7	0,1	0,2	0,2	0
	AIC-SARMA(1,1)	**98,8**	**96,9**	**84,2**	**57,7**	**29,7**	**33,5**	**76,1**	**97,9**	**99,8**	**100**
	BIC-SAR(1)	10,7	20,8	50	76,8	93,9	90,8	61	9,2	0	0
	BIC-SMA(1)	0,3	0,7	1,3	1,8	1,9	0,7	0,1	1,7	0,8	0,1
	BIC-SARMA(1,1)	**89**	**78,5**	**48,7**	**21,4**	**4,2**	**8,5**	**38,9**	**89,1**	**99,2**	**99,9**
	HQC-SAR(1)	3,4	8,8	30,3	59,4	83,6	80,7	41,4	3,8	0	0
	HQC-SMA(1)	0	0	0,2	0,4	0,9	0,7	0,1	0,3	0,3	0
	HQC-SARMA(1,1)	**96,6**	**91,2**	**69,5**	**40,2**	**15,5**	**18,6**	**58,5**	**95,9**	**99,7**	**100**

Στον Πίνακα 4 περιλαμβάνονται τα ποσοστά επιλογής κάθε διαδικασίας από τα πληροφοριακά κριτήρια για την περίπτωση της SARMA(1,1) διαδικασίας με ρ ίσο με 0,2 και διάφορες τιμές για το λ. Το βασικό συμπέρασμα από αυτόν τον πίνακα είναι ότι στο δείγμα των 100 παρατηρήσεων και τα τρία κριτήρια αποτυγχάνουν να επιλέξουν τη SARMA(1,1) διαδικασία. Για τιμές του λ ίσες με -0,2 και 0,2 επιλέγονται σχεδόν αποκλειστικά οι SAR(1) και SMA(1) διαδικασίες. Επειδή ο συντελεστής ρ έχει πολύ χαμηλή τιμή αγνοείται το αυτοπαλίνδρομο μέρος της διαδικασίας με αποτέλεσμα για τιμές του λ μεγαλύτερες του 0,2 και μικρότερες του -0,2 να υπερισχύει το μέρος του κινητού μέσου και τα κριτήρια να υποδεικνύουν σε μεγάλα ποσοστά τη SMA(1) διαδικασία. Τα αποτελέσματα βελτιώνονται στα δείγματα των 325 και 400 παρατηρήσεων αλλά και πάλι τα ποσοστά εσφαλμένων επιλογών είναι πολύ υψηλά. Τα πιο μεγάλα ποσοστά εντοπισμού της πραγματικής διαδικασίας προκύπτουν για το κριτήριο AIC κάτι πάλι αναμενόμενο λόγω της συμπεριφορά αυτού του κριτηρίου να επιλέγει υποδείγματα με πολλές παραμέτρους και ακολουθούν τα κριτήρια HQC και BIC.

Πίνακας 6. *Ποσοστά επιλογής των τριών υποδειγμάτων για όταν ισχύει η μικτή χωρική διαδικασία SARMA(1,1) με ρ =0,9 και διάφορες τιμές του λ*

N	*λ*	**-0,9**	**-0,8**	**-0,6**	**-0,4**	**-0,2**	**0,2**	**0,4**	**0,6**	**0,8**	**0,9**
100	AIC-SAR(1)	2,5	5,1	20,8	47,4	72,1	79,4	63,7	47,3	24,3	9,7
	AIC-SMA(1)	0	0	0	0	0	0	0	0,1	0,1	0,1
	AIC-SARMA(1,1)	**97,5**	**94,9**	**79,2**	**52,6**	**27,9**	**20,6**	**36,3**	**52,6**	**75,6**	**90,2**
	BIC-SAR(1)	10	19,4	46,7	74,7	91,2	94,9	87,4	73,3	48,1	22,6
	BIC-SMA(1)	0	0	0	0	0	0	0,1	0,1	0,2	0,4
	BIC-SARMA(1,1)	**90**	**80,6**	**53,3**	**25,3**	**8,8**	**5,1**	**12,5**	**26,6**	**51,7**	**77**
	HQC-SAR(1)	4,6	10,5	30,4	59,3	83,8	89,4	75,8	58,8	33,2	14,3
	HQC-SMA(1)	0	0	0	0	0	0	0	0,1	0,2	0,3
	HQC-SARMA(1,1)	**95,4**	**89,5**	**69,6**	**40,7**	**16,2**	**10,6**	**24,2**	**41,1**	**66,6**	**85,4**
400	AIC-SAR(1)	0	0	0,1	10,8	53,5	63,1	29,6	9,5	0,6	0,1
	AIC-SMA(1)	0	0	0	0	0	0	0	0	0	0
	AIC-SARMA(1,1)	**100**	**100**	**99,9**	**89,2**	**46,5**	**36,9**	**70,4**	**90,5**	**99,4**	**99,9**
	BIC-SAR(1)	0	0	5,1	41,7	86,8	91,3	65,7	35,2	7,8	0,6
	BIC-SMA(1)	0	0	0	0	0	0	0	0	0	0
	BIC-SARMA(1,1)	**100**	**100**	**94,9**	**58,3**	**13,2**	**8,7**	**34,3**	**64,8**	**92,2**	**99,4**
	HQC-SAR(1)	0	0	1,2	22,3	71,4	79,4	46,9	19,3	2,1	0,3
	HQC-SMA(1)	0	0	0	0	0	0	0	0	0	0
	HQC-SARMA(1,1)	**100**	**100**	**98,8**	**77,7**	**28,6**	**20,6**	**53,1**	**80,7**	**97,9**	**99,7**
325	AIC-SAR(1)	0	0,2	5,1	24,1	63,5	66,3	27	3,4	0,3	0
	AIC-SMA(1)	0	0	0	0	0	0	0	0	0	0
	AIC-SARMA(1,1)	**100**	**99,8**	**94,9**	**75,9**	**36,5**	**33,7**	**73**	**96,6**	**99,7**	**100**
	BIC-SAR(1)	1	3,7	21,7	62,3	89,7	92,5	63,3	17,4	1	0,1
	BIC-SMA(1)	0	0	0	0	0	0	0	0	0	0
	BIC-SARMA(1,1)	**99**	**96,3**	**78,3**	**37,7**	**10,3**	**7,5**	**36,7**	**82,6**	**99**	**99,9**
	HQC-SAR(1)	0,1	0,9	10,6	41,2	79,4	82,2	43,9	6,8	0,5	0
	HQC-SMA(1)	0	0	0	0	0	0	0	0	0	0
	HQC-SARMA(1,1)	**99,9**	**99,1**	**89,4**	**58,8**	**20,6**	**17,8**	**56,1**	**93,2**	**99,5**	**100**

Τα αποτελέσματα των προσομοιώσεων της SARMA(1,1) διαδικασίας για διάφορες τιμές του λ και για τιμές του ρ ίσες με 0,5 και 0,9 παρουσιάζονται, αντίστοιχα, στους Πίνακες 5 και 6. Η βασική διαπίστωση από αυτούς τους πίνακες είναι ότι σε αντίθεση με την περίπτωση που το ρ ισούταν με 0,2 και που παρουσιάστηκε στον Πίνακα 4, τώρα για τιμές του λ ίσες με -0,6, -0,4, -0,2, 0,2, 0,4, και 0,6 και τα τρία κριτήρια επιλέγουν εσφαλμένα σε μεγάλα ποσοστά τη SAR(1) διαδικασία. Μεγαλύτερη επιτυχία σωστής επιλογής σημειώνεται και εδώ για το κριτήριο AIC και ακολουθούν τα HQC και BIC. Ωστόσο, για πολύ υψηλές ή πολύ μικρές τιμές του λ ιδιαίτερα στα μεγαλύτερα δείγματα τα τρία κριτήρια επιλέγουν την πραγματική SARMA(1,1) διαδικασία σε ποσοστά σχεδόν 100%. Αξιοσημείωτο είναι το γεγονός ότι η SMA(1) διαδικασία δεν επιλέγεται ποτέ όταν το ρ ισούται με 0,9 αλλά και όταν ισούται με 0,5 και ταυτόχρονα τα δείγματα είναι μεγάλα. Η επιλογή των κριτηρίων γίνεται μεταξύ των διαδικασιών SARMA(1,1) και SAR(1). Επομένως, θα μπορούσε να εξαχθεί το συμπέρασμα ότι για να επιλέξουν με επιτυχία τα τρία κριτήρια τη SARMA(1,1) διαδικασία θα πρέπει οι τιμές του χωρικού αυτοπαλίνδρομου συντελεστή ρ να είναι μεγάλες, οι τιμές του συντελεστή του χωρικού μέσου λ να είναι και αυτές σε απόλυτες τιμές αρκετά υψηλές και τα

Πίνακας 7. *Ποσοστά επιλογής των τριών υποδειγμάτων για όταν ισχύει η μικτή χωρική διαδικασία SARMA(1,1) με λ =0,2 και διάφορες τιμές του ρ*

N	***ρ***	**-0,9**	**-0,8**	**-0,6**	**-0,4**	**-0,2**	**0,2**	**0,4**	**0,6**	**0,8**	**0,9**
100	AIC-SAR(1)	59,2	59,6	60,2	60,3	48,2	43,9	63,4	73	79,9	79,4
	AIC-SMA(1)	0	0	1,4	16,5	45,4	34,1	11,2	2,9	0,1	0
	AIC-SARMA(1,1)	**40,8**	**40,4**	**38,4**	**23,2**	**6,4**	**22**	**25,4**	**24,1**	**20**	**20,6**
	BIC-SAR(1)	84	84,3	83,2	72,4	50,8	52,8	79,3	88,9	94,4	94,9
	BIC-SMA(1)	0	0,1	3,5	24,1	48,9	41,6	12,3	3,2	0,2	0
	BIC-SARMA(1,1)	**16**	**15,6**	**13,3**	**3,5**	**0,3**	**5,6**	**8,4**	**7,9**	**5,4**	**5,1**
	HQC-SAR(1)	70,7	70,7	72,2	68,3	50,3	48,3	72,3	82,2	88	89,4
	HQC-SMA(1)	0	0	2,5	21,2	48,1	38,9	11,4	3,2	0,2	0
	HQC-SARMA(1,1)	**29,3**	**29,3**	**25,3**	**10,5**	**1,6**	**12,8**	**16,3**	**14,6**	**11,8**	**10,6**
400	AIC-SAR(1)	22	23,8	27,2	31,2	33,9	37,4	54,7	63,3	66,9	63,1
	AIC-SMA(1)	0	0	0	0,1	23,4	15,5	0,3	0	0	0
	AIC-SARMA(1,1)	**78**	**76,2**	**72,8**	**68,7**	**42,7**	**47,1**	**45**	**36,7**	**33,1**	**36,9**
	BIC-SAR(1)	62,3	64,6	69	73,1	56,9	62,1	87,6	91,5	92,8	91,3
	BIC-SMA(1)	0	0	0	1,7	40,3	25,2	0,4	0	0	0
	BIC-SARMA(1,1)	**37,7**	**35,4**	**31**	**25,2**	**2,8**	**12,7**	**12**	**8,5**	**7,2**	**8,7**
	HQC-SAR(1)	39,5	42,1	45,5	50,2	47,9	53,4	74,1	78,7	82	79,4
	HQC-SMA(1)	0	0	0	0,7	34,4	21	0,4	0	0	0
	HQC-SARMA(1,1)	**60,5**	**57,9**	**54,5**	**49,1**	**17,7**	**25,6**	**25,5**	**21,3**	**18**	**20,6**
325	AIC-SAR(1)	44,3	45,4	48,2	50,7	47	43,1	63,4	68	69,9	66,3
	AIC-SMA(1)	0	0	0	3,3	39,1	22,1	1,5	0	0	0
	AIC-SARMA(1,1)	**55,7**	**54,6**	**51,8**	**46**	**13,9**	**34,8**	**35,1**	**32**	**30,1**	**33,7**
	BIC-SAR(1)	81,4	82,5	84,3	80,4	55,5	63,5	89,7	92,5	93,2	92,5
	BIC-SMA(1)	0	0	0,5	13,8	44,2	28,6	1,9	0,3	0	0
	BIC-SARMA(1,1)	**18,6**	**17,5**	**15,2**	**5,8**	**0,3**	**7,9**	**8,4**	**7,2**	**6,8**	**7,5**
	HQC-SAR(1)	63,2	65	66,6	69,2	54,2	55,4	78,8	82,8	84,8	82,2
	HQC-SMA(1)	0	0	0	8	43,5	25,6	1,9	0,2	0	0
	HQC-SARMA(1,1)	**36,8**	**35**	**33,4**	**22,8**	**2,3**	**19**	**19,3**	**17**	**15,2**	**17,8**

δείγματα μεγάλα. Σε διαφορετικές περιπτώσεις, φαίνεται ότι υπερισχύει ο συντελεστής ρ και επιλέγονται SAR(1) διαδικασίες.

Ο Πίνακας 7 περιλαμβάνει τα αποτελέσματα των προσομοιώσεων για την περίπτωση της SARMA(1,1) διαδικασίας με λ ίσο με 0,2 και διάφορες τιμές για το ρ. Ο πίνακας αυτός περιέχει για λόγους ευκολίας στις συγκρίσεις και τις περιπτώσεις που το ρ ισούται με 0,2 και 0,9 οι οποίες έχουν παρουσιαστεί και σε προηγούμενους πίνακες. Όπως διαπιστώνεται, τα κριτήρια δυσκολεύονται να εντοπίσουν τη SARMA(1,1) διαδικασία και παρουσιάζουν συμπεριφορά παρόμοια με την περίπτωση του Πίνακα 4 που το ρ διατηρούταν σταθερό ίσο με 0,2. Ωστόσο, τώρα όταν ο συντελεστής ρ παίρνει τιμές μεγαλύτερες του 0,2 ή μικρότερες του -0,2 επικρατεί το αυτοπαλίνδρομο μέρος της διαδικασίας και τα κριτήρια έχουν την τάση να επιλέγουν τη SAR(1) διαδικασία. Επιπρόσθετα, και εδώ τα ποσοστά σωστής επιλογής βελτιώνονται λίγο στα δείγματα των 325 και 400 παρατηρήσεων αλλά γενικά η συμπεριφορά των κριτηρίων είναι απογοητευτική. Το κριτήριο AIC έχει τις περισσότερες επιτυχίες και ακολουθούν τα κριτήρια HQC και BIC.

Πίνακας 8. *Ποσοστά επιλογής των τριών υποδειγμάτων για όταν ισχύει η μικτή χωρική διαδικασία SARMA(1,1) με λ =0,5 και διάφορες τιμές του ρ*

N	*ρ*	**-0,9**	**-0,8**	**-0,6**	**-0,4**	**-0,2**	**0,2**	**0,4**	**0,6**	**0,8**	**0,9**
100	AIC-SAR(1)	6,4	7,3	8,3	10,7	12	6,4	21,6	42,2	54,3	56
	AIC-SMA(1)	0	0	1	18,2	63	60,8	21,7	2,2	0,3	0
	AIC-SARMA(1,1)	**93,6**	**92,7**	**90,7**	**71,1**	**25**	**32,8**	**56,7**	**55,6**	**45,4**	**44**
	BIC-SAR(1)	19,7	20,4	25,5	28,6	19,1	7,7	33,8	63,3	77,7	81,1
	BIC-SMA(1)	0	0	4,1	39,2	77,6	80,7	36,4	4,3	0,3	0,1
	BIC-SARMA(1,1)	**80,3**	**79,6**	**70,4**	**32,2**	**3,3**	**11,6**	**29,8**	**32,4**	**22**	**18,8**
	HQC-SAR(1)	10,5	11,8	13,8	17,1	15,4	7,3	28,1	52,3	64,6	68,2
	HQC-SMA(1)	0	0	2,7	27,4	71,5	71,5	28,1	3	0,3	0,1
	HQC-SARMA(1,1)	**89,5**	**88,2**	**83,5**	**55,5**	**13,1**	**21,2**	**43,8**	**44,7**	**35,1**	**31,7**
400	AIC-SAR(1)	0	0	0	0	0,3	0,3	2,2	7,2	17,9	17,4
	AIC-SMA(1)	0	0	0	0,1	24,7	31,9	0,8	0	0	0
	AIC-SARMA(1,1)	**100**	**100**	**100**	**99,9**	**75**	**67,8**	**97**	**92,8**	**82,1**	**82,6**
	BIC-SAR(1)	0	0,1	0,4	0,5	1,2	0,4	11,5	30,8	50,9	51,2
	BIC-SMA(1)	0	0	0	1,8	61,1	72,4	8,5	0	0	0
	BIC-SARMA(1,1)	**100**	**99,9**	**99,6**	**97,7**	**37,7**	**27,2**	**80**	**69,2**	**49,1**	**48,8**
	HQC-SAR(1)	0	0	0,1	0,4	0,4	0,4	5,3	15,3	31,1	31,1
	HQC-SMA(1)	0	0	0	0,6	39,6	51,7	3,4	0	0	0
	HQC-SARMA(1,1)	**100**	**100**	**99,9**	**99**	**60**	**47,9**	**91,3**	**84,7**	**68,9**	**68,9**
325	AIC-SAR(1)	0,2	0,3	0,8	1,2	1,9	0,8	4,8	10,5	14,4	10,4
	AIC-SMA(1)	0	0	0	1,7	37,7	41,5	2,4	0	0	0
	AIC-SARMA(1,1)	**99,8**	**99,7**	**99,2**	**97,1**	**60,4**	**57,7**	**92,8**	**89,5**	**85,6**	**89,6**
	BIC-SAR(1)	3	3,6	5,7	7,9	8,2	0,9	21,8	39,7	45,3	41
	BIC-SMA(1)	0	0	0	11,9	72,7	75,6	11,8	0	0	0
	BIC-SARMA(1,1)	**97**	**96,4**	**94,3**	**80,2**	**19,1**	**23,5**	**66,4**	**60,3**	**54,7**	**59**
	HQC-SAR(1)	1,3	1,3	1,9	2,4	3,7	0,8	11,1	20,2	28,5	23
	HQC-SMA(1)	0	0	0	4	53,6	56,7	5,7	0	0	0
	HQC-SARMA(1,1)	**98,7**	**98,7**	**98,1**	**93,6**	**42,7**	**42,5**	**83,2**	**79,8**	**71,5**	**77**

Οι Πίνακες 8 και 9 συνοψίζουν τα αποτελέσματα των προσομοιώσεων όταν η πραγματική διαδικασία είναι η SARMA(1,1) με τιμές του λ ίσες με 0,5 και 0,9 και για διάφορες τιμές του ρ. Όπως προκύπτει από τους δυο πίνακες, στα μεγάλα δείγματα και όταν ο συντελεστής ρ παίρνει πολύ μεγάλες θετικές ή πολύ μικρές αρνητικές τιμές τα κριτήρια επιλέγουν τη SARMA(1,1) με ποσοστά επιτυχίας που σε ορισμένες περιπτώσεις προσεγγίζουν το 100%. Εντούτοις, για μικρές και μέτριες απόλυτες τιμές του ρ και για λ ίσο με 0,9 στον Πίνακα 9 παρατηρούνται αξιόλογα ποσοστά επιλογής της SMA(1) διαδικασίας ενώ η SAR(1) διαδικασία δεν επιλέγεται σχεδόν ποτέ ανεξάρτητα της τιμής του συντελεστή ρ. Δηλαδή, ουσιαστικά η επιλογή γίνεται ανάμεσα στις SARMA(1,1) και SMA(1) διαδικασίες. Αντίθετα, στον Πίνακα 8 όπου ο συντελεστής λ λαμβάνει την τιμή 0,5 καταγράφονται αξιοσημείωτα ποσοστά επιλογής και της SAR(1) διαδικασίας. Κατά συνέπεια, για να εντοπίζεται με επιτυχία η SARMA(1,1) διαδικασία θα πρέπει ο συντελεστής λ να έχει μεγάλη τιμή και οι τιμές του συντελεστή ρ να είναι και αυτές σε απόλυτες τιμές υψηλές. Ο μεγάλος αριθμός παρατηρήσεων φαίνεται και εδώ ότι διευκολύνει στον εντοπισμό της σωστής διαδικασίας. Επιπρόσθετα, πρέπει να σχολιαστεί η μικρή πτώση στα ποσοστά σωστής επιλογής που εμφανίζεται στον Πίνακα 8 όταν ο συντελεστής ρ ισούται με

Πίνακας 9. *Ποσοστά επιλογής των τριών υποδειγμάτων για όταν ισχύει η μικτή χωρική διαδικασία SARMA(1,1) με λ =0,9 και διάφορες τιμές του ρ*

N	*ρ*	**-0,9**	**-0,8**	**-0,6**	**-0,4**	**-0,2**	**0,2**	**0,4**	**0,6**	**0,8**	**0,9**
100	AIC-SAR(1)	0	0	0	0	0	0	0	0,2	3	9,7
	AIC-SMA(1)	0	0	0,2	10,8	56,3	56,1	20,5	6	1,1	0,1
	AIC-SARMA(1,1)	**100**	**100**	**99,8**	**89,2**	**43,7**	**43,9**	**79,5**	**93,8**	**95,9**	**90,2**
	BIC-SAR(1)	0	0	0	0	0	0	0	0,3	7,9	22,6
	BIC-SMA(1)	0	0	1,4	28,3	80,6	81	45,5	18,2	4,4	0,4
	BIC-SARMA(1,1)	**100**	**100**	**98,6**	**71,7**	**19,4**	**19**	**54,5**	**81,5**	**87,7**	**77**
	HQC-SAR(1)	0	0	0	0	0	0	0	0,2	4,4	14,3
	HQC-SMA(1)	0	0	0,9	16,9	68,7	69	30,6	10	1,9	0,3
	HQC-SARMA(1,1)	**100**	**100**	**99,1**	**83,1**	**31,3**	**31**	**69,4**	**89,8**	**93,7**	**85,4**
400	AIC-SAR(1)	0	0	0	0	0	0	0	0	0	0,1
	AIC-SMA(1)	0	0	0	0	15,9	19,8	0	0	0	0
	AIC-SARMA(1,1)	**100**	**100**	**100**	**100**	**84,1**	**80,2**	**100**	**100**	**100**	**99,9**
	BIC-SAR(1)	0	0	0	0	0	0	0	0	0,1	0,6
	BIC-SMA(1)	0	0	0	0,2	48,9	59,1	4	0	0	0
	BIC-SARMA(1,1)	**100**	**100**	**100**	**99,8**	**51,1**	**40,9**	**96**	**100**	**99,9**	**99,4**
	HQC-SAR(1)	0	0	0	0	0	0	0	0	0	0,3
	HQC-SMA(1)	0	0	0	0,1	28,2	36,6	0,8	0	0	0
	HQC-SARMA(1,1)	**100**	**100**	**100**	**99,9**	**71,8**	**63,4**	**99,2**	**100**	**100**	**99,7**
325	AIC-SAR(1)	0	0	0	0	0	0	0	0	0	0
	AIC-SMA(1)	0	0	0	0	2,6	6,1	0	0	0	0
	AIC-SARMA(1,1)	**100**	**100**	**100**	**100**	**97,4**	**93,9**	**100**	**100**	**100**	**100**
	BIC-SAR(1)	0	0	0	0	0	0	0	0	0	0,1
	BIC-SMA(1)	0	0	0	0,1	17,8	27,7	0,6	0	0	0
	BIC-SARMA(1,1)	**100**	**100**	**100**	**99,9**	**82,2**	**72,3**	**99,4**	**100**	**100**	**99,9**
	HQC-SAR(1)	0	0	0	0	0	0	0	0	0	0
	HQC-SMA(1)	0	0	0	0	8	14,1	0,3	0	0	0
	HQC-SARMA(1,1)	**100**	**100**	**100**	**100**	**92**	**85,9**	**99,7**	**100**	**100**	**100**

τις τιμές 0,6, 0,8 και 0,9 και η αντίστοιχη αύξηση στα ποσοστά επιλογής της SAR(1) διαδικασίας. Η συμπεριφορά αυτή που ανάλογη δεν έχει παρατηρηθεί στον Πίνακα 5 όπου ο συντελεστής *ρ* έχει τη σταθερή τιμή 0,5 θα μπορούσε να αποδοθεί στο γεγονός ότι η SAR(1) διαδικασία εκφράζει πιο ισχυρή χωρική αυτοσυσχέτιση ενώ η SMA(1) διαδικασία πιο ασθενή. Δηλαδή, κατά κάποιον τρόπο η SAR(1) διαδικασία κυριαρχεί στη SMA(1) διαδικασία.

4. ΣΥΜΠΕΡΑΣΜΑΤΑ

Στην εργασία αυτή πραγματοποιήθηκαν μια σειρά από προσομοιώσεις για να μελετηθεί η ικανότητα των τριών πιο διαδεδομένων πληροφοριακών κριτηρίων, δηλαδή των AIC, BIC και HQC να εντοπίζουν και να επιλέγουν χωρικές διαδικασίες. Από την ανάλυση των αποτελεσμάτων είναι φανερό ότι τα πληροφοριακά κριτήρια αποτελούν εργαλεία που μπορούν να βοηθήσουν τον ερευνητή σε μια χωρική οικονομετρική ανάλυση να επιλέξει χωρικές διαδικασίες. Ωστόσο, υπάρχουν περιπτώσεις που αποτυγχάνουν και που οδηγούν σε λανθασμένα συμπεράσματα. Ειδικότερα, διαπιστώνεται ότι σε μεγάλα δείγματα η συμπεριφορά τους είναι πιο

ικανοποιητική άλλα αρκετές φορές η επιλογή της διαδικασίας επηρεάζεται από το μέγεθος των τιμών των χωρικών συντελεστών. Μεγάλες τιμές των χωρικών συντελεστών οδηγούν πιο εύκολα στην επιλογή της σωστής διαδικασίας. Στη μικτή SARMA(1,1) διαδικασία και οι δύο συντελεστές θα πρέπει να έχουν υψηλές τιμές ώστε να είναι ξεκάθαροι και ο αυτοπαλίνδρομος όρος και ο όρος κινητού μέσου. Διαφορετικά, μπορεί να κυριαρχήσει στην επιλογή αυτός με τη μεγαλύτερη τιμή και να επιλεχθεί απλή διαδικασία. Όταν οι πραγματικές διαδικασίες είναι η SAR(1) ή η SMA(1) το κριτήριο BIC έχει τις καλύτερες επιδόσεις ενώ όταν ισχύει η μικτή SARMA(1,1) διαδικασία το κριτήριο AIC δίνει καλύτερα αποτελέσματα. Φυσικά αυτό θα μπορούσε να αποδοθεί σε κάποιο βαθμό στη γνωστή τάση του BIC να επιλέγει υποδείγματα με μικρό αριθμό παραμέτρων και του AIC να επιλέγει υποδείγματα πολλών παραμέτρων.

Είναι αυτονόητο ότι όπως σε όλες τις μελέτες με προσομοίωση τα συμπεράσματα ισχύουν αποκλειστικά με βάση τις συγκεκριμένες υποθέσεις και μήτρες χωρικών σταθμίσεων. Αυτό τονίζεται διότι σε αντίθεση με τις χρονοσειρές, η χωρική εξάρτηση, που ορίζεται στα υποδείγματα από τον ερευνητή εξωγενώς, μπορεί να καθοριστεί με ποικίλους τρόπους γειτονίας. Εντούτοις, η εφαρμογή της ίδιας μεθόδου γειτονίας σε διαφορετικές χωρικές δομές μπορεί να οδηγήσει σε τελείως ανόμοιες μήτρες χωρικών σταθμίσεων. Στην εργασία επιλέχθηκε να χρησιμοποιηθεί και η χωρική δομή της Ελλάδας που παρουσιάζει αρκετές γεωγραφικές ιδιομορφίες με αποτέλεσμα να προκύψουν ιδιαίτερα ασύμμετρες μήτρες χωρικών σταθμίσεων. Το γενικό συμπέρασμα είναι ότι τα αποτελέσματα φαίνεται ότι ισχύουν όχι μόνο για τη θεωρητική γεωγραφική τοπολογία των ομαλών πλεγμάτων αλλά και για την πραγματική γεωγραφική δομή όπως εκφράζεται μέσω ενός χάρτη.

Είναι άγνωστο εάν οι εκτιμήσεις των συντελεστών των διαδικασιών επηρεάζονται από το χαρακτηριστικό των πολύ χαμηλών κάτω ορίων στα εύρη ορισμού τους που προκύπτουν με τα κριτήρια γειτονίας του πύργου και των κοντινότερων γειτόνων αλλά και από τον αλγόριθμο μεγιστοποίησης μέσω της συνάρτησης nlminb στην R. Η συνέπεια των αποτελεσμάτων με διαφορετικές μήτρες χωρικών σταθμίσεων αλλά και διαφορετικές διαδικασίες εκτίμησης θα αποτελέσουν αντικείμενα μελλοντικής εργασίας. Επιπρόσθετα, αντικείμενο μελλοντικής έρευνας θα αποτελέσει και η συμπεριφορά των πληροφοριακών κριτηρίων για περιπτώσεις μη Γκαουσιανών χωρικών διαδικασιών καθώς και για χωρικές διαδικασίες που περιλαμβάνουν χωρικές υστερήσεις ανωτέρων της πρώτης τάξεων.

ABSTRACT

In spatial data analysis the sample is considered to be a realization from a spatial process. The most common spatial processes are the first order spatial autoregressive process SAR(1), the first order spatial moving average process SMA(1) and the mixed spatial autoregressive moving average process SARMA(1,1) which are extensions of the respective time series processes to the spatial context allowing for multidirectional dependence. In this study, using simulation techniques, we examine the ability of the three most common applied in practice information criteria which are Akaike's Information Criterion (AIC), Bayesian Information Criterion (BIC) and Hannan and Quinn's Criterion (HQC) to select the best fitted model. The results indicate

that the behavior of the three criteria depends on coefficients' values and sample size and in some cases they lead to misleading selections.

ΑΝΑΦΟΡΕΣ

Akaike H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle, in B. N Petrov and Csaki F.(eds), *Proceedings of the Second International Symposium of Information Theory*, Academia Kiado, Budapest, 267-281.

Anselin, L. (1988). *Spatial Econometrics*: *Methods and Models*, Kluwer Academic Publishers, Dordrecht.

Besag J. (1974). Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society, Series B*, **36**, 192-236.

Bivand R. (2015). Spdep: Spatial Dependence: Weighting Schemes. Statistics and Models. R package version 0.5-82. http://CRAN.R-project.org/package=spdep

Bivand R., Hauke J. and Kossowski T. (2013). Computing the Jacobian in Gaussian Spatial Autoregressive Models: An Illustrated Comparison of Available Methods. *Geographical Analysis*, **45**, 150-179.

Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control*, Holden Day, San Francisco.

Cliff, A. D. and Ord, J. K. (1981). *Spatial Processes: Models and Applications*, Pion Ltd, London.

Haining R. (1978). The Moving Average Process for Spatial Interaction. *Transactions of the Institute of British Geographers*, **3**, 202-225.

Hannan E. J. and Quinn B. G. (1979). The Determination of the Order of an Autoregression. *Journal of the Royal Statistical Society, Series B*, **41**, 190-195.

Huang J. S. (1984). The Autoregressive Moving Average Model for Spatial Analysis. *Australian Journal of Statistics*, **26**, 169-178.

LeSage, J. and Pace, R. (2009). *Introduction to Spatial Econometrics*, Chapman & Hall/CRC, Boca Raton.

Mur J. and Angulo A. M. (2007). Clues for Discriminating Between Moving Average and Autoregressive Models in Spatial Processes. *Spanish Economic Review*, **9**, 273-298.

Ord J. K (1975). Estimation Methods for Models of Spatial Interaction. *Journal of the American Statistical Association*, **70**, 120-126.

Schwarz G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, **6**, 461-464.

Whittle P. (1954). On Stationary Processes in the Plane. *Biometrika*, **41**, 434-449.

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 31ου Πανελληνίου Συνεδρίου Στατιστικής (2018), σελ. 251−258

ΓΕΝΕΣΗ ΘΕΩΡΙΑΣ ΠΙΘΑΝΟΤΗΤΩΝ: CARDANO ENANTI PASCAL ΚΑΙ FERMAT

Χαράλαμπος Α. Χαραλαμπίδης
Τμήμα Μαθηματικών, Πανεπιστήμιο Αθηνών
ccharal@math.uoa.gr

ΠΕΡΙΛΗΨΗ

Η διατύπωση πιθανοθεωρητικών λύσεων στα προβλήματα της διανομής μεριδίων και του συμφερότερου στοιχήματος, από τους Pascal και Fermat, θεωρείται ευρύτατα ως η αρχή (γένεση) της Θεωρίας Πιθανοτήτων. Η εργασία του Cardano στις Πιθανότητες, η οποία προηγήθηκε εκείνης των Pascal και Fermat κατά ένα περίπου αιώνα, δημοσιεύθηκε μετά τον θάνατό του στο περίφημο δεκαπεντασέλιδο βιβλίο του με τίτλο *Liber de Ludo Aleae (The Book on Games of Chance)*. Προβλήματα πιθανοτήτων εξέτασε και στο βιβλίο του *Practica Arithmetice et Mensurandi Singularis*. Η συμβολή του Cardano στη γένεση της Θεωρίας Πιθανοτήτων θεωρείται από αρκετούς ότι δεν έχει τύχει της αναγνώρισης που πιθανόν αξίζει. Στο άρθρο αυτό παρουσιάζονται συνοπτικά οι πιθανοθεωρητικοί νόμοι και τα προβλήματα παιχνιδιών τύχης που απασχόλησαν τον Cardano. Τα στοιχεία συνηγορούν υπέρ της ευρύτατα αποδεκτής άποψης για τη γένεση της Θεωρίας Πιθανοτήτων.

Λέξεις Κλειδιά: Διανομή μεριδίων, Κλασική πιθανότητα, Νόμος μεγάλων αριθμών, Παράδοξο Αγίας Πετρούπολης, Πολλαπλασιαστικός τύπος πιθανοτήτων, Πρόβλημα Γαλιλαίου, Τύπος του Cardano.

1. ΕΙΣΑΓΩΓΙΚΑ

Η Θεωρία Πιθανοτήτων οφείλει τη γένεσή της στα παιγνίδια τύχης. Η διατύπωση μεθόδων υπολογισμού πιθανοθεωρητικών λύσεων στα προβλήματα της διανομής μεριδίων και του συμφερότερου στοιχήματος από τους Blaise Pascal (1623-1662) και Pierre de Fermat (1608-1665), θεωρείται ευρύτατα ως η αρχή (γένεση) της Θεωρίας Πιθανοτήτων. Η εργασία του Gerolamo Cardano (1501-1575) στις Πιθανότητες, η οποία προηγήθηκε εκείνης των Pascal και Fermat κατά ένα περίπου αιώνα, δημοσιεύθηκε μετά τον θάνατό του, το 1663, στο δεκαπεντασέλιδο βιβλίο του με τίτλο *Liber de Ludo Aleae*, το οποίο αποτελείται από 32 σύντομα κεφάλαια. Το βιβλίο αυτό μεταφράσθηκε στα Αγγλικά από τον S. H. Gould και με τίτλο *The Book on Games of Chance* περιλαμβάνεται στη βιογραφία του Cardano που συνέγραψε ο Ore (1953) με τίτλο *Cardano: The Gamling Scholar*. Προβλήματα πιθανοτήτων εξέτασε

και στο βιβλίο του *Practica Arithmetice et Mensurandi Singularis*.

Η εργασία του Cardano στις Πιθανότητες, αξιολογείται από αρκετούς ότι σηματοδοτεί την αρχή της ανάπτυξης της Θεωρίας Πιθανοτήτων. Έτσι, στην βιογραφία του Cardano, ο Ore σημειώνει τα ακόλουθα:

Έχω πεισθεί ότι η πρωτοποριακή αυτή εργασία στις πιθανότητες είναι τόσο ευρεία και σε ορισμένα ερωτήματα τόσο επιτυχής ώστε η διατριβή του Cardano θα μπορούσε να θεωρηθεί ότι σηματοδοτεί την αρχή της Θεωρίας Πιθανοτήτων, πολύ περισσότερο από τη συνήθη θεώρηση των συζητήσεων του Pascal με τον φίλο του παίκτη τυχερών παγνιδιών de Mere και την ακολουθήσασα αλληλογραφία με τον Fermat, που όλα αυτά συνέβησαν τουλάχιστο ένα αιώνα αφότου ο Cardano άρχισε να συνθέτει το δικό του Liber de Ludo Aleae.

Ο ιστορικός των Μαθηματικών Burton (2003) φαίνεται να μοιράζεται την ίδια γνώμη:

Για πρώτη φορά, βρίσκουμε την μετάβαση από την εμπειρική στη θεωρητική έννοια του δίκαιου κύβου. Καταφέρνοντας αυτό, ο Cardan (Cardano) πιθανώς κατέστη ο πραγματικός πατέρας της μοντέρνας θεωρίας πιθανοτήτων.

Όμως, παρά την αξιόλογη συμβολή του Cardano στο πεδίο αυτό, σε κανένα από τα προβλήματα δεν έφτασε στο επίπεδο της μαθηματικής λεπτότητας και ωριμότητας που επιδείχθηκε από τους διαδόχους του. Πολλές από τις ανακαλύψεις του ήταν είτε στοιχειώδεις είτε λανθασμένες. Από την άλλη, η εργασία των Pascal και Fermat υπήρξε αυστηρή και έδωσε την πρώτη ώθηση συστηματικής μελέτης της Θεωρίας Πιθανοτήτων. Σχετικά, ο A. W. F. Edwards (1987), σε αντίθεση με τους Ore και Burton, εκφράζει την άποψη:

Παρά την αυξηθείσα ενημέρωσή μας για την προηγηθείσα εργασία του Cardano και του Galileo είναι φανερό ότι πριν από τους Pascal και Fermat τίποτα περισσότερο από τον υπολογισμό βασικών πιθανοτήτων σε διάφορα παγνίδια με κύβους και χαρτιά δεν είχε επιτευχθεί.

Στο άρθρο αυτό παρουσιάζονται συνοπτικά οι πιθανοθεωρητικοί νόμοι και τα προβλήματα παιχνιδιών τύχης που απασχόλησαν τον Cardano και τα οποία αργότερα ερευνήθηκαν πλήρως και με σχετική μαθηματική αυστηρότητα από τους διαδόχους του. Η αξιολόγηση των στοιχείων αυτών συνηγορεί υπέρ της ευρύτατα αποδεκτής άποψης που αποδίδει στους Pascal και Fermat τη γένεση της Θεωρίας Πιθανοτήτων.

2. ΠΙΘΑΝΟΘΕΩΡΗΤΙΚΟΙ ΝΟΜΟΙ

Η θεωρούμενη σήμερα ως κλασική πιθανότητα ορίσθηκε, για πρώτη ίσως φορά, από τον Cardano στο Κεφάλαιο 14 ως εξής:

Θεωρούμε ολόκληρη την περιοχή (circuit) και τον αριθμό εκείνων των ρίψεων που παριστάνουν κατά πόσους τρόπους το ευνοϊκό αποτελέσμα μπορεί να εμφανισθεί και συγκρίνουμε το αριθμό αυτό με την υπόλοιπη περιοχή, και σύμφωνα με την πρόταση θα πρέπει το αμοιβαίο στοίχημα να τεθεί έτσι ώστε ο οποιοσδήποτε να ανταγωνίζεται με ίσους όρους.

Σημειώνουμε ότι ο Cardano καλεί *περιοχή (circuit)* τον γνωστό σήμερα ως δειγ-

ματικό χώρο. Έτσι, αν ο δειγματικός χώρος περιλαμβάνει κ ευνοϊκά αποτελέσματα και r μη ευνοϊκά αποτελέσματα για ένα ενδεχόμενο και όλα τα αποτελέματα είναι εξίσου δυνατά (ισοπίθανα), τότε σύμφωνα με τον ορισμό του Cardano ο λόγος των συμπληρωματικών πιθανοτήτων (odds) του ενδεχομένου είναι κ/r και δίνει τη σωστή πιθανότητα $\kappa/(\kappa + r)$. Ας συγκρίνουμε τον ορισμό της κλασικής πιθανότητας του Cardano με τους αντίστοιχους ορισμούς που διατυπώθηκαν στη συνέχεια.

Ο ορισμός της πιθανότητας που δόθηκε από τον Gottfried Wilheim Leibniz (1646-1716) το 1710 είναι ο ακόλουθος:

Αν μια κατάσταση μπορεί να οδηγήσει σε διαφορετικά πλεονεκτικά αμοιβαίως αποκλειόμενα αποτελέσματα, η εκτίμηση της αναμενόμενης τιμής θα είναι το άθροισμα των δυνατών πλεονεκτημάτων από το σύνολο όλων εκείνων των αποτελεσμάτων, διαιρεμένων από τον αριθμό των συνολικών αποτελεσμάτων.

Ο Jacob Bernoulli (1654-1705) διατύπωσε σχετικά το 1713 στο βιβλίο του *Ars Conjectandi* τα ακόλουθα:

Αν το ολοκλήρωμα και η απόλυτη βεβαιότητα, τα οποία συμβολίζουμε με το γράμμα α ή με την μονάδα 1*, θεωρηθούν ότι αποτελούνται, για παράδειγμα, από πέντε πιθανότητες, όπως ακριβώς πέντε τμήματα, τρία από τα οποία ευνοούν την ύπαρξη ή την πραγματοποίηση κάποιων ενδεχομένων, με τα υπόλοιπα όμως εναντίον αυτής, θα λέμε ότι το ενδεχόμενο αυτό έχει βεβαιότητα* $3a/5$*, ή* $3/5$*.*

Ο ορισμός που δόθηκε από τον Abraham de Moivre (1646-1716) το 1711 στο βιβλίο του *De Mensura Sortis* έχει ως ακολούθως:

Αν κ *είναι ο αριθμός των περιπτώσεων στις οποίες ορισμένο ενδεχόμενο μπορεί να πραγματοποιηθεί, και* r *είναι ο αριθμός των περιπτώσεων στις οποίες (το ενδεχόμενο) αποτυγχάνει (να πραγματοποιηθεί), οι πραγματοποιήσεις όπως και οι αποτυχίες έχουν τον βαθμό πιθανότητας τους. Αλλά, αν όλες οι περιπτώσεις στις οποίες ορισμένο ενδεχόμενο μπορεί το ίδιο εύκολα (equally easy) να πραγματοποιηθεί ή να αποτύχει, η πιθανότητα πραγματοποίησης ως προς την πιθανότητα αποτυχίας θα είναι όπως το* κ *ως προς το* r*.*

Σύμφωνα με τον ορισμό του de Moivre ο λόγος των συμπληρωματικών πιθανοτήτων (odds) ενδεχομένου δίνεται σωστά ως κ/r.

Ο ορισμός της κλασικής πιθανότητας διατυπώθηκε από τον Pierre-Simon Laplace (1749-1827) το 1774 στην πρώτη του πιθανοθεωρητική εργασία και είναι ο εξής:

Η πιθανότητα ενός ενδεχομένου είναι το πηλίκο του αριθμού των ευνοϊκων γι΄ αυτό περιπτώσεων προς τον αριθμό των δυνατών περιπτώσεων, όταν τίποτα δεν μας κάνει να πιστεύουμε ότι κάποια από τις περιπτώσεις μπορεί να εμφανισθεί συχνότερα από οποιαδήποτε άλλη, έτσι ώστε οι περιπτώσεις αυτές να θεωρούνται εξίσου δυνατές (equally possible).

Επισημαίνουμε την προσπαθεια ορισμού της έννοιας των εξίσου δυνατών (ισοπίθανων) περιπτώσεων ανεξάρτητα από την έννοια της πιθανότητας, γιατί διαφορετικά ο κλασικός ορισμός της πιθανότητας θα οδηγούσε σε φαύλο κύκλο. Τούτο επιτυγχάνεται στον ορισμό του Laplace με την επίκληση της *αρχής της έλλειψης επαρκούς λόγου.*

Ο *πολλαπλασιαστικός τύπος* για την εύρεση του λόγου των πιθανοτήτων υπέρ και κατά της εμφάνισης ενός ενδεχομένου σε κάθε εκτέλεση μιας σειράς εκτελέσεων ενός παιχνιδιού τύχης, αποτελεί μια από τις πιο σημαντικές συμβολές του Cardano στην ανάπτυξη της Θεωρίας Πιθανοτήτων. Συγκεκριμμένα, ας θεωρήσουμε ένα παιγνίδι τύχης με r ισοπίθανα δυνατά αποτελέσματα και έστω ότι κ από αυτά είναι ευνοϊκά για την πραγματοποίηση συγκεκριμμένου ενδεχομένου. Τότε $\kappa/(r-\kappa)$ είναι ο λόγος των πιθανοτήτων υπέρ και κατά της εμφάνισης του. Αναλύοντας τα παιγνίδια με κύβους τριών, τεσσάρων, πέντε και έξι ισοπίθανων εδρών, συμπέρανε ότι σε ν εκτελέσεις ενός παιχνιδιού τύχης, ο λόγος των πιθανοτήτων υπέρ και κατά της εμφάνισης ενός ενδεχομένου σε κάθε εκτέλεση του παιγνιδιού είναι $\kappa^\nu/(r^\nu-\kappa^\nu)$. Ο τύπος αυτός αναφέρεται και ως *τύπος του Cardano*. Θέτοντας $p=\kappa/r$, η έκφραση αυτή μετασχηματίζεται στην $p^\nu/(1-p^\nu)$, η οποία συνεπάγεται ότι η πιθανότητα εμφάνισης ενός ενδεχομένου σε κάθε εκτέλεση μιας σειράς ν ισοπίθανων δοκιμών (εκτελέσεων) ενός παιχνιδιού τύχης είναι ίση με p^ν, το γινόμενο των επιμέρους πιθανοτήτων. Ο τύπος αυτός αποτελεί μερική περίπτωση του πολλαπλασιαστικού τύπου πιθανοτήτων για ανεξάρτητα ενδεχόμενα.

Ο Cardano χρησιμοποίησε τον *Νόμο των Μεγάλων Αριθμών του Bernoulli*, αν και ποτέ δεν τον διατύπωσε εκπεφρασμένα, 150 χρόνια περίπου πριν από τον J. Bernoulli στο οποίον αποδίδεται ο νόμος αυτός. Στο βιβλίο *Cardano: The Gamling Scholar*, ο Ore (1953) γράφει σχετικά:

Είναι προφανές ότι είναι γνώστης του Νόμου των Μεγάλων Αριθμών στην πιο στοιχειώδη μορφή του. Τα μαθηματικά του Cardano ανήκουν στην περίοδο πριν από τη χρήση τύπων στις εκφράσεις, έτσι ώστε δεν μπορούσε να διατυπώσει τον νόμο εκπεφρασμένα με τύπο, αλλά τον χρησιμοποίησε ως εξής: όταν η πιθανότητα ενός ενδεχομένου είναι p, τότε σε μεγάλο αριθμό ν επαναλήψεων, ο αριθμός των φορών εμφάνισης του δεν μπορεί να κείται μακριά από την τιμή $\mu = \nu p$.

3. ΠΡΟΒΛΗΜΑΤΑ ΠΑΙΧΝΙΔΙΩΝ ΤΥΧΗΣ

Το *κλασικό πρόβλημα της διανομής μεριδίων* στη γενική περίπτωση έχει ως εξής: Έστω ότι δύο παίκτες α και β αγωνίζονται σε μια σειρά παιγνιδιών και νικητής αναδεικνύεται εκείνος που κερδίζει πρώτος ν παιγνίδια. Ας υποθέσουμε ότι για κάποιο λόγο διακόπτεται η σειρά των παιγνιδιών όταν ο α χρειάζεται κ ακόμα νίκες για να αναδειχθεί νικητής της σειράς, ενώ ο β χρειάζεται r νίκες. Στην περίπτωση αυτή πώς πρέπει να μοιρασθεί δίκαια το στοίχημα;

Ο Pacioli το 1494 στο βιβλίο του *Summa de Arithmetica, Geometria, Proportioni et Proportionalità* θεώρησε την μερική περίπτωση $\nu=6$, $\kappa=1$ και $r=4$ και πρότεινε το στοίχημα να μοιρασθεί ανάλογα με τα παιγνίδια $\nu-\kappa=5$ και $\nu-r=2$ που κάθε παίκτης κέρδισε μέχρι την διακοπή.

Ο Cardano το 1539 στο βιβλίο του *Practica Arithmetice et Mensurandi Singularis* απέρριψε την λύση του Pacioli ως περιλαμβάνουσα λάθος, που ακόμα και ένα παιδί αναγνωρίζει, και πρότεινε το στοίχημα να μοιρασθεί στους παίκτες α και β σε μερίδια ανάλογα με $r(r+1)/2$ και $\kappa(\kappa+1)/2$. Η αιτιολόγηση της λανθασμένης αυτής

λύσης είναι μάλλον δυσνόητη. Συγκεκριμμένα, εισήγαγε μια νέα σειρά παιγνιδιών στην οποία ο παίκτης α, αρχίζοντας από το μηδέν, αναδεικνύεται νικητής αν κερδίσει κ παιγνίδια πριν ο παίκτης β κερδίσει r παιγνίδια. Θεωρώντας την μερική περίπτωση $\kappa = 1$ και $r = 3$, χρησιμοποίησε επαγωγική επιχειρηματολογία και πρότεινε το στοίχημα να μοιρασθεί στους παίκτες α και β σε μερίδια ανάλογα με 6 και 1 και για την επίσης περίπτωση $\kappa = 1$ και r συμπέρανε ότι πρέπει το στοίχημα να μοιρασθεί ανάλογα με $r(r+1)/2$ και 1. Όμως δεν δικαιολόγησε την μετάβαση από την τελευταία μερική περίπτωση στη γενική περίπτωση κ και r. Ίσως να θεώρησε ότι αυτό προκύπτει από το συμμετρικό του προβλήματος ως προς τους δύο παίκτες.

Η σωστή λύση του προβλήματος δόθηκε από τους Pascal και Fermat, η οποία βασίζεται είτε στο διωνυμικό, είτε στο αρνητικό διωνυμικό στοχαστικό πρότυπο: Το στοίχημα πρέπει να μοιρασθεί στους παίκτες α και β ανάλογα με τις πιθανότητες $P(A)$ και $1 - P(A)$, όπου

$$P(A) = \frac{1}{2^{\kappa+r-1}} \sum_{i=\kappa}^{\kappa+r-1} \binom{\kappa+r-1}{i} = \sum_{j=\kappa}^{\kappa+r-1} \binom{j-1}{\kappa-1} \frac{1}{2^j}$$

είναι η πιθανότητα να κερδίσει τη σειρά των παιγνιδιών ο παίκτης α, αν η σειρά των παιγνιδιών συνεχιζόταν.

Το γνωστό ως *πρόβλημα του Γαλιλαίου*, το οποίο έθεσε στον διάσημο Μαθηματικό και Φυσικό Galileo Galilei (1564-1642) ο Μέγας Δούκας της Τοσκάνης, είναι το ακόλουθο: Ας θεωρήσουμε το παιχνίδι της ρίψης τριών κύβων στο οποίο ενδιαφέρει το άθροισμα των τριών αριθμών που εμφανίζονται. Σημειώνουμε ότι τα αθροίσματα εννέα, δέκα, ένδεκα και δώδεκα δύνανται να προκύψουν κατά έξι διαφορετικούς συνδυασμούς:

$$1+2+6 = 1+3+5 = 1+4+4 = 2+2+5 = 2+3+4 = 3+3+3 = 9,$$

$$1+3+6 = 1+4+5 = 2+2+6 = 2+3+5 = 2+4+4 = 3+3+4 = 10,$$

$$1+4+6 = 1+5+5 = 2+3+6 = 2+4+5 = 3+3+5 = 3+4+4 = 11,$$

$$1+5+6 = 2+4+6 = 2+5+5 = 3+3+6 = 3+4+5 = 4+4+4 = 12.$$

Τότε, γιατί τα αθροίσματα εννέα ή δώδεκα είναι λιγότερο πιθανά από τα αθροίσματα δέκα ή ένδεκα;

Σημειώνουμε ότι η ρίψη τριών κύβων αποτελούσε τμήμα ενός παιγνιδιού γνωστού ως *υπερδέκα (passadieci, passedix, passten)* στο οποίο απαιτείτο άθροισμα μεγαλύτερο του δέκα για τον νικητή. Ο Galileo το 1620 στην πιθανοθεωρητική εργασία του *Sopra le Scoperte dei Dadi* δίνει την ακόλουθη λύση: Αρχίζει θεωρώντας ότι οι δυνατές (ισοπίθανες) περιπτώσεις είναι οι $6^3 = 216$ διατάξεις των 6 εδρών ανά 3 με επανάληψη, το οποίο, εμμέσως, συνεπάγεται ότι οι $\binom{6+3-1}{3} = 56$ συνδυασμοί των 6 εδρών ανά 3 με επανάληψη δεν είναι ισοπίθανοι. Σημειώνουμε ότι στο μη ισοπίθανον των 56 συνδυασμών με επανάληψη, οφείλεται η έμμεση επισήμανση της ανισότητας

των πιθανοτήτων των ανωτέρω αθροισμάτων, αν και καθένα από αυτά περιλαμβάνει 6 τέτοιους συνδυασμούς. Περαιτέρω, σε κάθε συνδυασμό με τρία διαφορετικά στοιχεία αντιστοιχούν $3! = 6$ διατάξεις, ενώ σε κάθε συνδυασμό με δύο ίδια στοιχεία αντιστοιχούν 3 διατάξεις με επανάληψη και δε κάθε συνδυασμό με τρία ίδια στοιχεία αντιστοιχεί μόνο μια διάταξη με επανάληψη. Επομένως, σε κάθε ένα από τα αθροίσματα εννέα και δώδεκα αντιστοιχούν 25 ευνοϊκές περιπτώσεις, ενώ σε κάθε ένα από τα αθροίσματα δέκα και ένδεκα αντιστοιχούν 27 ευνοϊκές περιπτώσεις και οι αντίστοιχες πιθανότητες είναι $25/216$ και $27/216$.

Όμως, χωρίς να το γνωρίζει ο Galileo, το ίδιο πρόβλημα είχε επιτυχώς λυθεί από τον Cardano ένα περίπου αιώνα νωρίτερα. Το πρόβλημα αυτό περιλαμβάνεται στο Κεφάλαιο 13 του βιβλίου *Liber de Ludo Aleae*. Εξετάζει τις περιπτώσεις της ρίψης δύο και τριών κύβων και καταγράφει σε πίνακα τον αριθμό s των ευνοϊκών αποτελεσμάτων, για $s = 2, 3, \ldots, 12$ στην περίπτωση δύο κύβων και για $s = 3, 4, \ldots, 18$ στην περίπτωση τριών κύβων. Για την περίπτωση του προβλήματος του Γαλιλαίου δίνει τους σωστούς αριθμούς των ευνοϊκών αποτελεσμάτων 25, για $s = 9, 12$, και 27, για $s = 10, 11$.

Το *παράδοξο της Αγίας Πετρούπολης*, το οποίο είναι ένα από τα πλέον συζητημένα στην ιστορία της Θεωρίας Πιθανοτήτων, έχει ως ακολούθως: Σε κάποιο καζίνο παίζεται το παιγνίδι της επαναλαμβανόμενης ρίψης νομίσματος με τον όρο το καζίνο να πληρώνει στο παίκτη 2^{x-1} νομισματικές μονάδες αν εμφανισθεί για πρώτη φορά κεφαλή στη x-οστή ρίψη, για $x = 1, 2, \ldots$. Πόσο, πρέπει να είναι το ποσό συμμετοχής του παίκτη για να είναι δίκαιο το παιγνίδι;

Σημειώνουμε ότι ένα παιγνίδι θεωρείται δίκαιο αν το αναμενόμενο (μέσο) κέρδος τόσο του παίκτη όσο και του καζίνου είναι μηδενικό. Σχετικά, παρατηρούμε ότι το ενδεχόμενο εμφάνισης για πρώτη φορά κεφαλής στη x-οστή ρίψη είναι το ενδεχόμενο εμφάνισης γραμμάτων σε κάθε μια από τις $x - 1$ πρώτες ρίψεις και κεφαλής στη x-οστή ρίψη και έτσι η πιθανότητα πραγματοποίησής του είναι $f(x) = (1/2)^x$, για $x = 1, 2, \ldots$. Το μέσο κέρδος του παίκτη απειρίζεται,

$$\mu = \sum_{x=1}^{\infty} 2^{x-1} f(x) = \sum_{x=1}^{\infty} 2^{x-1} \left(\frac{1}{2}\right)^x = \frac{1}{2} + \frac{1}{2} + \cdots = \infty.$$

Επομένως, ανεξάρτητα από το ποσό συμμετοχής του παίκτη στο παιγνίδι, αυτός ή αυτή πάντοτε θα βγαίνει κερδισμένος και το παιγνίδι δεν μπορεί να καταστεί δίκαιο.

Το πρόβλημα της Αγίας Πετρούπολης προτάθηκε για πρώτη φορά από τον Nicolas Bernoulli (1687-1759) σε επιστολή του στον Pierre-Rémond de Montmort (1678-1719). Δημοσιεύθηκε το 1713 από τον Montmort στη δεύτερη έκδοση του βιβλίου του *Essay d' Analyse sur les Jeux de Hazard*. Στην εξέταση του προβλήματος αυτού, επανήλθε ο Daniel Bernoulli (1687-1759), ο οποίος το αντιμετώπισε μέσω της έννοιας της μέσης ωφέλειας (mean utility). Η εργασία αυτή δημοσιεύθηκε το 1738 στο περιοδικό St. Petersburg Academy Proceedings και από τότε κατέστη γνωστό ως παράδοξο της Αγίας Πετρούπολης (St. Petersburg's Paradox).

Μια ακόμα πιο παλιά εκδοχή του προβλήματος Αγίας Πετρούπολης, η οποία εξετάζεται από το Cardano στο βιβλίο του *Practica Arithmetice et Mensurandi Singularis*, είναι η ακόλουθη: Ένας πλούσιος και ένας φτωχός άνθρωπος παίζουν στοιχηματίζοντας το ίδιο ποσόν. Αν ο φτωχός άνθρωπος κερδίσει, την επόμενη μέρα το στοίχημα διπλασιάζεται και η διαδικασία συνεχίζεται. Αν ο πλούσιος άνθρωπος κερδίσει μια φορά, το παιγνίδι τερματίζεται. Ο Cardano διατύπωσε την άποψη ότι ο πλούσιος άνθρωπος βρίσκεται σε μειονεκτική θέση στο παιγνίδι αυτό. Ας εξετάσουμε κατά πόσον αυτή είναι η πραγματικότητα.

Ας υποθέσουμε ότι το παιγνίδι παίζεται με την επαλαμβανόμενη ρίψη ενός νομίσματος και ότι το αρχικό ποσό του στοιχήματος είναι μια νομισματική μονάδα. Επίσης, έστω ότι ο φτωχός έχει αρχικό κεφάλαιο μια νομισματική μονάδα και ο πλούσιος έχει αρχικό κεφάλαιο ν νομισματικές μονάδες. Τότε, ο μέγιστος αριθμός παιγνιδιών, έστω m, που μπορεί να παιχθούν προκύπτει θεωρώντας την περίπτωση που ο πλούσιος συνεχίζει να χάνει μέχρι να μη έχει τα απαραίτητα χρήματα για το στοίχημα. Τότε, το συνολικό ποσό $s \leq \nu$ που ο πλούσιος χάνει μέχρι και την m-οστή ρίψη είναι

$$s = \sum_{x=1}^{m} 2^{x-1} = \frac{2^m - 1}{2 - 1} = 2^m - 1,$$

οπότε το m είναι ο μέγιστος θετικός ακέραιος που ικανοποιεί την ανισότητα

$$2^m - 1 \leq \nu.$$

Παρατηρούμε ότι ο φτωχός μπορεί να χάσει το κεφάλαιο του (μια νομισματική μονάδα) στην x-οστή ρίψη με πιθανότητα $(1/2)^x$, για $x = 1, 2, \ldots, m$, ή να κερδίσει συνολικά $2^m - 1$ νομισματικές μονάδες στην m-οστή ρίψη με πιθανότητα $(1/2)^m$, οπότε το μέσο κέρδος του είναι

$$\mu_1 = \sum_{x=1}^{m} (-1)\left(\frac{1}{2}\right)^x + (2^m - 1)\left(\frac{1}{2}\right)^m = -\frac{\frac{1}{2} - \left(\frac{1}{2}\right)^{m+1}}{1 - \frac{1}{2}} + 1 - \left(\frac{1}{2}\right)^m$$

$$= -1 + \left(\frac{1}{2}\right)^m + 1 - \left(\frac{1}{2}\right)^m = 0.$$

Ανάλογα υπολογίζεται το μέσο κέρδος του πλούσιου και είναι $\mu_\nu = 0$, οπότε το παιγνίδι είναι δίκαιο.

Παρατηρούμε ότι οι πιθανότητες να κερδίσει τελικά το παιγνίδι ο φτωχός ή ο πλούσιος είναι $(1/2)^m$ και $1 - (1/2)^m$, αντίστοιχα. Επειδή

$$\left(\frac{1}{2}\right)^m < 1 - \left(\frac{1}{2}\right)^m, \quad \text{ή ισοδύναμα} \quad \left(\frac{1}{2}\right)^m < \frac{1}{2}, \quad m = 2, 3, \ldots,$$

αντίθετα με την άποψη του Cardano, ο φτωχός άνθρωπος βρίσκεται σε μειονεκτική θέση στο παιγνίδι αυτό. Ακριβέστερα, όσο πιο μεγάλο είναι το κεφάλαιο ν του πλουσίου, το οποίο συμπαρασύρει ανάλογα τον μέγιστο αριθμό m των παιγνιδιών, τόσο είναι λιγότερο πιθανόν ο φτωχός να καταφέρει να χρεοκοπήσει τον πλούσιο.

4. ΣΥΜΠΕΡΑΣΜΑ

Ο Cardano έθιξε πιθανοθεωρητικούς νόμους και αρκετά προβλήματα παιγνιδιών τύχης που αργότερα κατέλαβαν διακεκριμένη θέση στη σχετική βιβλιογραφία. Όμως, δεν κατάφερε να κάνει την εντυπωσιακή (αποφασιστική) ανακάλυψη. Αυτό κατά ένα ποσοστό θα μπορούσε να αποδοθεί στο γεγονός ότι μέχρι τότε δεν είχε αναπτυχθεί κατάλληλη συμβολική μαθηματική γλώσσα. Η ευρύτατη χρήση του μαθηματικού συμβολικού λογισμού και μεθόδων συνδυαστικής απαρίθμησης, των οποίων η ανάπτυξη ακολούθησε λίγο αργότερα, συνέβαλε αποφασιστικά στην ανάπτυξη της Θεωρίας Πιθανοτήτων. Οι διάδοχοι του Cardano και ιδιαίτερα οι Jacob Bernoulli (1654-1705), Pierre-Rémond de Montmort (1678-1719), Abraham de Moivre (1667-1754) και Pierre-Simon de Laplace (1749-1827), κατάφεραν να διατυπώσουν και αποδείξουν με σχετική μαθηματική ακρίβεια τους νόμους που διέπουν την κλασική πιθανότητα και να αντιμετωπίσουν με επιτυχία τα προβλήματα παιχνιδιών τύχης που εκείνος πρώτος διατύπωσε και εξέτασε. Συμπερασματικά, τα στοιχεία που παρουσιάσθηκαν στο άρθρο αυτό συνηγορούν υπέρ της ευρύτατα αποδεκτής άποψης ότι η Θεωρία Πιθανοτήτων οφείλει τη γένεσή της στους Pascal και Fermat.

ABSTRACT

The formulation of probabilistic solutions to the problems of division of the shares and advantageous stake by Pascal and Fermat, is widely considered as the origin (genesis) of Probability Theory. The work of Cardano on Probability, which preceded that of Pascal and Fermat by about one century, was published after his death in the famous fifteen-page book entitled *Liber de Ludo Aleae (the Book on Games of Chance)*. Probabilistic problems were also examined in his other book *Practica Arithmetice et Mensurandi Singularis*. Cardano's contribution to the genesis of Probability Theory is considered by several people that has not got the recognition that it perhaps deserves. In this article, the probabilistic rules and the problems of games of chance that Cardano touched on are briefly presented. The facts are in favour of the widely accepted opinion about the genesis of Probability Theory.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Burton, D. (2003) *The History of Mathematics: An Introduction*, 5th Edition, McGraw Hill, New York.

David, F. N. (1962) *Games, God, and Gambling*, Hafner Press, New York.

Edwards, A. W. F (1987) *Pascal's Arithmetical Triangle*, Griffin, London.

Hald, A. (2003) *A History of Probability and Statistics and Their Applications before 1750*, Wiley, New Jersey.

Ore, O. (1953) *Cardano, the Gambling Scholar*, Princeton University Press.

Todhunter, I. (1865) *A History of the Mathematical Theory of Probability*, MacMillan, Cambridge and London. (Reprinted Chelsea, New York, 1949).

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 31ου Πανελληνίου Συνεδρίου Στατιστικής (2018), σελ.259-264

ΒΕΛΤΙΣΤΟΙ 3^m ΠΑΡΑΓΟΝΤΙΚΟΙ ΣΧΕΔΙΑΣΜΟΙ ΕΚΤΙΜΗΤΙΚΗΣ ΤΑΞΗΣ IV ΜΕ $N = 0 \bmod 3$ ΠΑΡΑΤΗΡΗΣΕΙΣ

Β. Χασιώτης[1], Σ..Χατζόπουλος[1], Σ.. Κουνιάς[2]

[1]Τμήμα Μαθηματικών ΑΠΘ, [2]Τμήμα Μαθηματικών ΕΚΠΑ

chasiotisv@math.auth.gr, cstavros@math.auth.gr, skounias@math.uoa.gr

Περίληψη

Στην παρούσα εργασία μελετώνται οι 3^m παραγοντικοί σχεδιασμοί εκτιμητικής τάξης IV με $N \equiv 0$ mod 3 παρατηρήσεις. Οι παράμετροι που θα εκτιμηθούν είναι οι γραμμικές και τετραγωνικές αντιθέσεις των κύριων επιδράσεων και οι γραμμικές με γραμμικές αλληλεπιδράσεις δύο παραγόντων. Για N=18 ο φ-βέλτιστος σχεδιασμός είναι ο ορθογώνιος σχηματισμός OA(N,4,3,3) με ισχύ τρία, ο οποίος δεν υπάρχει και χρησιμοποιούνται οι ισορροπημένοι σχηματισμοί BA(18,4,3,2) και ο ορθογώνιος σχηματισμός OA(18,4,3,2) για την εύρεση των A, D, E βέλτιστων σχεδιασμών.

1.ΕΙΣΑΓΩΓΗ

Στους κλασματικούς 3^m πειραματικούς σχεδιασμούς, με m παράγοντες ο καθένας σε τρία επίπεδα, οι παράγοντες συμβολίζονται με $F_1,...,F_m$, τα επίπεδα κάθε παράγοντα με 0,1,2, και $l_i, u_i, q_i, \quad i = 1,...,m$ είναι η επίδραση του F_i όταν μετέχει στο πείραμα αντίστοιχα στα επίπεδα 0,1,2.

Η γραμμική αντίθεση είναι $(q_i - l_i)/\sqrt{2}$ και η τετραγωνική αντίθεση είναι $(q_1 - 2u_i + l_i)/\sqrt{6}$, w_{jk}^{pq} συμβολίζεται η αλληλεπίδραση των δύο παραγόντων F_j, F_k όταν είναι στα επίπεδα $p, q, \ p, q = 0,1,2$.

Χωρίς αλληλεπιδράσεις το μοντέλο είναι (Pericleous at al, 2017),

$$\mathrm{E}(y_s) = \mu_1 + \sum_{i=1}^{m}\left(\frac{q_i - l_i}{\sqrt{2}} g_{is} + \frac{(q_i - 2u_i + l_i}{\sqrt{6}} 3h_{is}\right), \ s = 1,...,N, \mu_1 = \sum_{i=1}^{m} u_i \qquad (1)$$

$$g_{is} = \frac{1}{\sqrt{2}}\begin{cases} -1 & if \quad d(i,s) = 0 \\ 0 & if \quad d(i,s) = 1 \\ 1 & if \quad d(i,s) = 2 \end{cases}, \qquad \mathrm{h}_{is} = \frac{1}{\sqrt{6}}\begin{cases} 1 & if \quad d(i,s) = 0 \\ 0 & if \quad d(i,s) = 1 \\ 1 & if \quad d(i,s) = 2 \end{cases}$$

d(i,s) είναι το επίπεδο του παράγοντα F_i στην s παρατήρηση, τα σφάλματα είναι ασυσχέτιστα με διασπορά σ^2. Σε κάθε παρατήρηση μετέχει κάθε παράγοντας σε ένα από τα τρία επίπεδα.

Όταν οι γραμμικές με γραμμικές αλληλεπιδράσεις δύο παραγόντων περιέχονται στο μοντέλο, τότε

$$E(y_s) = \mu_1 + \sum_{i=1}^{m}(v_i g_{is} + p_i 3h_{is}) + \sum_{j=1}^{m-1}\sum_{k>j}^{m} c_{qr} w_{jk}^{pq} \qquad (2)$$

$$c_{00} = 2(-g_{js})(-g_{ks}), c_{01} = -\sqrt{2}g_{js}, c_{02} = 2(-\sqrt{g_{js}})(\sqrt{g_{ks}}), c_{10} = (-\sqrt{2}g_{js}),$$

$$c_{11} = 1, c_{12} = (\sqrt{2}g_{ks}), c_{20} = 2(g_{js})(-g_{ks}), c_{21} = \sqrt{2}g_{js}, c_{22} = 2(g_{js})(g_{ks})$$

Το μοντέλο αυτό γράφεται,

$$E(y_s) = \mu + \sum_{i=1}^{m}(v_i g_{is} + p_i 3h_{is}) + \sum_{j=1}^{m-1} t_1^1 g_{is} + \sum_{k=2}^{m} t_k^2 g_{ks} + \sum_{j=1}^{m-1}\sum_{k>j}^{m} t_{jk}^3 g_{js} g_{ks} \qquad (3)$$

Όπου $\mu = \mu_1 + \mu_2, \mu_2 = \sum_{j=1}^{m-1}\sum_{k>j}^{m} w_{jk}^{11}$, $v_i = (q_i - l_i)/\sqrt{2}, p_i = (q_i - 2u_i + l_i)/\sqrt{6}$,

$$t_j^1 = \sqrt{2}\sum_{k>j}^{m}(w_{jk}^{21} - w_{jk}^{01}), t_k^2 = \sqrt{2}\sum_{j=1}^{k-1}(w_{jk}^{12} - w_{jk}^{10}), t_{jk}^3 = 2(w_{jk}^{00} - w_{jk}^{02} - w_{jk}^{20} + w_{jk}^{22}).$$

Ο πίνακας σχεδιασμού είναι, $X = [X_{me} \quad X_{ll}]$, όπου

$$X_{me} = (g_1, ..., g_m, h_1, ..., h_m), X_{ll} = (g_1 g_2, ...g_1 g_m, ..., g_{m-1} g_m).$$

Ορισμός 1. Αν σε ένα γραμμικό μοντέλο $\sigma^2 V$ είναι ο πίνακας διασποράς των εκτιμημένων παραμέτρων $\hat{\theta}$ που μας ενδιαφέρουν, ο πίνακας $Q = V^{-1}$ λέγεται πίνακας πληροφορίας των παραμέτρων θ.

Στο μοντέλο (3) μας ενδιαφέρουν όλες οι παράμετροι εκτός από το μέσο μ και ο πίνακας πληροφορίας είναι,

$$Q = X^T X - \frac{1}{N} X^T J_N X$$

J_N είναι $N \times N$ πίνακας με όλα τα στοιχεία ίσα με 1.

Στο μοντέλο (1) για $N \equiv 0 \bmod 9$ ο φ βέλτιστος σχεδιασμός είναι ένας ορθογώνιος σχηματισμός με ισχύ δύο και συμβολίζεται $OA(N, m, 3, 2)$, για τιμές $N = 9k + 3, N = 9k + 6$ χρησιμοποιούνται ισορροπημένοι σχηματισμοί BA(N,m,3,2) (Pericleous et al. 2017).

Περισσότερες πληροφορίες για τους ορθογώνιους σχηματισμούς περιέχονται στο βιβλίο Hedayat, Sloane & Stufken (1999).

Ορισμός 2. Σε ένα γραμμικό μοντέλο αν Ω είναι το σύνολο των πινάκων πληροφορίας, τότε ένας σχεδιασμός, με πίνακα πληροφορίας $Q \in \Omega$ και ιδιοτιμές $\lambda_1, ..., \lambda_k$, λέγεται φ-βέλτιστος αν ελαχιστοποιείται το $\sum_{i=1}^{k} g(\lambda_i)$ για όλες τις κυρτές συνεχείς συναρτήσεις g(x).

2. ΒΕΛΤΙΣΤΟΙ ΣΧΕΔΙΑΣΜΟΙ ΜΕ ΤΗ ΧΡΗΣΗ ΙΣΟΡΡΟΠΗΜΕΝΩΝ ΣΧΗΜΑΤΙΣΜΩΝ

Στο μοντέλο (1) με m παράγοντες , το πλήθος των παραμέτρων είναι 2m και ο φ-βέλτιστος σχεδιασμός είναι ένα $OA(18,m,3,2),\ \mathrm{m} \le 3$ (Pericleous et al, 2017).

Στο μοντέλο (3) ο πίνακας πληροφορίας περιέχει γινόμενα της μορφής $g_1 g_i g_j$ και ο φ-βέλτιστος σχεδιασμός είναι $OA(18,m,3,3)$ με ισχύ τρία, όμως τέτοιος σχεδιασμός δεν υπάρχει, ενώ ο αντίστοιχος σχεδιασμός $OA(18,m,3,2)$ με ισχύ 2 υπάρχει και είναι $m \le 7$, Hedayat, Sloane & Stufken (1999).

Θα χρησιμοποιήσουμε ισορροπημένους σχηματισμούς που ορίσθηκαν από τον Chakravarti (1961).

Οι ισορροπημένοι σχηματισμοί συμβολίζονται $BA(N,m,s,\mathrm{t})$, αποτελούνται από μία ή δύο ομάδες στηλών, έχουν N γραμμές m στήλες, στοιχεία $0,1,...,s-1$ και ισχύ t. Μας ενδιαφέρουν οι $BA(N,m,s,\mathrm{t}=2).$

Ορισμός 3. (ι) Ο σχηματισμός $BA(N,m=m_1+\mathrm{m}_2,s,\mathrm{t}=2),m_1 \ge 1,\mathrm{m}_2=0$ έχει μια ομάδα στηλών με N γραμμές και $m=m_1$ στήλες. Κάθε στήλη έχει n^p στοιχεία $p,\ p=0,1,...,s-1, n^0+n^1+\cdots+n^{s-1}=N$ και κάθε δύο στήλες έχουν $n^{pq}, p,q=0,...,s-1$ του ζεύγους pq, υπάρχουν s^2 ζεύγη pq. Στον Πίνακα 1 είναι $n^0=4,n^1=4,n^2=7,n^{00}=3,n^{12}=1$ κ.λπ.

Αν $m=1$ έχουμε το τετριμμένο $BA(N,m=1,s,1)$.

(ιι) Αν $m=m_1+m_2,m_1 \ge 1,m_2 \ge 1$,υπάρχουν δύο ομάδες στηλών $BA(N,m_1,s,2),BA(N,m_2,s,2)$, κάθε ομάδα είναι ισορροπημένος σχηματισμός, n_1^p συμβολίζει το πλήθος του στοιχείου p σε κάθε στήλη της πρώτης ομάδας και n_{11}^{pq} είναι το πλήθος του ζεύγους pq σε κάθε ζευγάρι στηλών της πρώτης ομάδας. Αντίστοιχα n_2^p, n_{22}^{pq} ισχύουν για τη δεύτερη ομάδα. Τέλος n_{12}^{pq} είναι το πλήθος του ζεύγους pq σε κάθε ζεύγος στηλών με μια στήλη από κάθε ομάδα .

Ο ορισμός (3) είναι γενίκευση του ορισμού Chakravarti (1961), που χρησιμοποιεί μια ομάδα στηλών και του ορισμού Pericleous et al.(2017 με $n^0=n^2=(N-n^1)/2$.

Πίνακας 1: $BA(18,4,3,2)$

0	0	0	0
0	0	0	2
0	0	2	0
0	1	1	1
0	2	0	2
0	2	2	0
0	2	2	2
1	0	1	1

1	1	1	2
1	1	2	1
1	2	0	0
2	0	0	2
2	0	2	0
2	0	2	2
2	1	0	0
2	2	0	1
2	2	1	0
2	2	2	2

$$n^0 = n^2 = 7, n^1 = 4, n^{00} = n^{02} = n^{20} = n^{22} = 3,$$

$$n^{01} = n^{10} = n^{12} = n^{21} = 1, n^{11} = 2$$

Για την κατασκευή του παραπάνω σχεδιασμού, χρησιμοποιήθηκε αλγόριθμος που δίνεται στο Παράρτημα.

3**) ΒΕΛΤΙΣΤΟΙ ΣΧΕΔΙΑΣΜΟΙ ΓΙΑ Ν=18**

Για την εύρεση του βέλτιστου σχεδιασμού χρησιμοποιούνται οι γεωμετρικοί ορθογώνιοι σχηματισμοί (see Hedayat et al., 1999, Section 5:10), όπως ορίζονται από τους Bird & Street (2018) και οι ισορροπημένοι σχεδιασμοί. Οι Evangelaras, Koukouvinos & Lappas (2007) μελέτησαν τους συνδιαστικά μη ισόμορφους ορθογώνιους σχηματισμούς με 18 παρατηρήσεις και 3-7 παράγοντες με τρεις στάθμες.

Αν δεν υπάρχει φ-βέλτιστος σχεδιασμός, τα πλέον χρησιμοποιούμενα κριτήρια είναι:

Ε-βέλτιστος: $\min \lambda_{\max}(V) \Leftrightarrow \max \lambda_{\min}(Q),$

Α-βέλτιστος: $\min(tr(V) = \min \sum_{i=1}^{m} \lambda_i(V) \Leftrightarrow \min(\sum_{i=1}^{m} 1/\lambda_i(\mathrm{Q}),$

D-βέλτιστος: $\min \det(V) \Leftrightarrow \max\det(\mathrm{Q})$,

Όπου $\sigma^2 V$ είναι ο πίνακας διασποράς και $Q = V^{-1}$ ο πίνακας πληροφορίας των παραμέτρων που μας ενδιαφέρουν και $\lambda_i(V)$ είναι η i ιδιοτιμή του V, $i = 1,...,m$.

Στον παρακάτω πίνακα 2, η πρώτη ομάδα είναι BA(18,2,3,2) και οι σχέσεις μεταξύ των στηλών δίνονται στον Πίνακα 1.

Η δεύτερη ομάδα είναι BA(18,3,3,2)και οι σχέσεις μεταξύ των στηλών είναι,

$$\mathrm{n}^{00} = 4, n^{01} = n^{02} = n^{10} = n^{12} = n^{21} = n^{22} = 2, n^{11} = n^{22} = 1$$

ΟΙ σχέσεις κάθε στήλης της πρώτης ομάδας με κάθε στήλη της δεύτερης ομάδας είναι,

$$n^{00} = n^{20} = 3, n^{01} = n^{02} = n^{10} = n^{21} = n^{22} = 2, n^{11} = n^{12} = 1.$$

Χρησιμοποιώντας 4 στήλες που είναι ο πίνακας σχεδιασμού 4 παραγόντων, εκτιμάμε 14 παραμέτρους που είναι: 4 γραμμικές, 4 τετραγωνικές αντιθέσεις καθώς και 6 γραμμικές-γραμμικές αλληλεπιδράσεις δύο παραγόντων

Πίνακας 2. BA(18,2+3,3,2)

0	0	0	1	2
0	0	1	2	0
0	0	2	0	1
0	2	0	2	1
0	2	1	0	2
0	2	2	1	0
0	1	0	0	0
2	0	0	2	1
2	0	1	0	2
2	0	2	1	0
2	2	0	1	2
2	2	1	2	0
2	2	2	0	1
2	1	0	0	0
1	0	0	0	0
1	2	0	0	0
1	1	1	1	1
1	1	2	2	2

Αυτοί λέγονται σχεδιασμοί εκτιμητικής τάξης IV διότι γίνεται εκτίμηση των κύριων επιδράσεων και μερικών αλληλεπιδράσεων δύο παραγόντων. Στο μοντέλο (2) οι αλληλεπιδράσεις τριών και περισσότερων παραγόντων θεωρούνται αμελητέες.

Στον πίνακα 3 παίρνουμε τις 2 πρώτες στήλες και όποιες δύο από τις 3 τελευταίες στήλες.

Στον πίνακα 3 δίνονται οι τιμές των κριτηρίων E, A, D για 4 παράγοντες, συγκρίνοντας τους τρεις πειραματικούς σχεδιασμούς: OA(18,4,3,2), BA_1 του πίνακα 1 και BA_2 του πίνακα 2.

Να σημειώσουμε ότι χρειάζεται να γίνει περαιτέρω έρευνα για την εύρεση και κατασκευή των ισορροπημένων σχηματισμών που οδηγούν σε βέλτιστους σχεδιασμούς.

Πίνακας 3. Τιμές των κριτηρίων E, A, D για τους τρεις σχεδιασμούς OA(18,4,3,2), BA_1, BA_2.

Κριτήρια	OA(18,4,3,2)	BA_1	BA_2
E	0.71612	1.09452*	-----
A	6.20348	4.82318	4.66749*
D	14431284	93710763	99300006*

Με το σύμβολο * συμβολίζεται ο καλύτερος σχεδιασμός. Σημειώνουμε ότι ο ισορροπημένος σχεδιασμούς BA_2 που αποτελείται από δύο ομάδες στηλών είναι A και D καλύτερος.

ABSTRACT

In this contribution are studied 3^m fractional factorial designs of resolution IV with $N \equiv 0 \bmod 3$ observations. The parameters to be estimated are the linear and quadratic contrasts and the linear by linear two factor interactions. For N=18 the φ optimal design is the orthogonal array OA(18,4,3,3) of strength three which does not exist. OA(18,4,3,2) of strength two and Balanced arrays BA(18,4,3,2) are employed to find the optimal E-A-D criteria.

,ΑΝΑΦΟΡΕΣ

Bird, E. & Street, D. J. (2018). Complete enumeration of all geometrically non-isomorphic three level orthogonal arrays in 18 runs. Australian journal of combinatorics, vol.71(3),pages 336-350.

Chakravarti, I.M. (1961). On some methods of construction of partially balanced arrays. *The Annals of Mathematical Statistics, 32, 1181-1185.*

Evangelaras, H., Koukouvinos, C., & Lappas, E. (2007). 18-run nonisomorphic three level orthogonal arrays. *Metrika, 66, 31-37.*

Hedayat, A. S., Sloane, N. J. A., & Stufken, J. (1999). *Orthogonal arrays: theory and Applications.* Springer Series in Statistics, New York.

Pericleous, K., Chatzopoulos, St. A., Kolyva-Machera, F., & Kounias, S. (2017). Optimal designs for estimating linear and quadratic contrasts with three level factors, the case $N \equiv 0 \bmod 3$.*Statistics, 51, 1061-1081.*

ΠΑΡΑΡΤΗΜΑ

Αλγόριθμος για την κατασκευή του BA(18,4,3,2) του πίνακα 1.

Εδώ για οικονομία χώρου αναγράφονται οι στήλες ως γραμμές.

Οι δύο πρώτες στήλες (γραμμές) είναι,

0 0 0 0 0 0 0 1 1 1 1 2 2 2 2 2 2 2

0 0 0 1 2 2 2 0 1 1 2 0 0 0 1 2 2 2

Βήμα 1:Η τρίτη γραμμή έχει στις 7 πρώτες θέσεις έναν από τις 140 μεταθέσεις των 0 0 0 1 2 2 2, στις 4 θέσεις 8,9,10,11 έχει μία από τις 12 μεταθέσεις των 0 1 1 2, στις τελευταίες 7 θέσεις έχει μία από τις 140 μεταθέσεις των 0 0 0 1 2 2 2. Συνολικά υπάρχουν $140 \times 12 \times 140 = 235200$ διανύσματα, αυτά αποτελούν το σύνολο S1.

Βήμα 2: Βρίσκουμε όλα τα διανύσματα από το σύνολο S1, που η δεύτερη γραμμή με αυτά έχει,

$$n^{00} = 3, n^{01} = 1, n^{02} = 3, n^{10} = 1, n^{11} = 2, n^{12} = 1, n^{20} = 3, n^{21} = 1, n^{22} = 3. \qquad (\Pi 1)$$

Αυτό το σύνολο συμβολίζεται με S2 και $S2 \subset S1$.

Βήμα 3: Παίρνουμε κάθε ένα διάνυσμα από το σύνολο S2 και το τοποθετούμε ως τρίτη γραμμή μετά τις δύο πρώτες γραμμές. Εξετάζουμε πόσα από τα διανύσματα του S2 ικανοποιούν οι γραμμές 2 και 3 με κάθε ένα από αυτά τα διανύσματα τις σχέσεις που δίνονται στο (Π1). Όσα ικανοποιούν αυτές τις σχέσεις αποτελούν το σύνολο S3, $S3 \subset S2 \subset S1$

Έτσι δημιουργούμε 4 γραμμές που ανά δύο ικανοποιούν τις σχέσεις που δίνονται στο (Π1).

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 31ου Πανελληνίου Συνεδρίου Στατιστικής (2018), σελ.265-276

ΠΕΙΡΑΜΑ ΣΥΜΜΕΤΟΧΙΚΗΣ ΔΙΑΒΟΥΛΕΥΤΙΚΗΣ ΔΙΑΧΕΙΡΙΣΗΣ ΘΕΜΑΤΩΝ: ΔΗΜΟΣ ΘΕΡΜΗΣ

Θ. Χατζηπαντελής, Ε. Μαρκάκη, Μ.Σωτήρογλου
Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης, Τμήμα Πολιτικών Επιστημών
chadji@polsci.auth.gr, markakie@polsci.auth.gr, misotiro@polsci.auth.gr

ΠΕΡΙΛΗΨΗ

Η διαμόρφωση ισχυρών συμμετοχικών θεσμών είναι απαραίτητη προϋπόθεση προκειμένου να ενισχυθεί η τοπική δημοκρατία, να αναπτυχθούν αποτελεσματικοί, αξιόπιστοι και διαφανείς θεσμοί και επιπροσθέτως να αναπτυχθεί μια συμμετοχική και αντιπροσωπευτική Κοινωνία των Πολιτών χωρίς αποκλεισμούς στη λήψη αποφάσεων. Η ενεργοποίηση σε τοπικό επίπεδο προκαλεί, παράλληλα, με την ροή από «πάνω» προς τα κάτω για την ανάδειξη τοπικών θεμάτων και προτάσεων πολιτικής, την ροή από «κάτω» προς τα «πάνω», δηλαδή από το κύτταρο (τοπικό διαμέρισμα ή/και γειτονιά) προς το αποφασιστικό όργανο (δημοτικό συμβούλιο). Πρόκειται για την ανάπτυξη ενός δικτύου ροής πολιτικών αποφάσεων και ένα πείραμα συμμετοχικής διαβουλευτικής διαχείρισης θεμάτων.

Το πείραμα επικεντρώνεται σε θέματα προστασίας και αναβάθμισης του φυσικού περιβάλλοντος της περιοχής του Δήμου Θέρμης, όπως θέματα ανακύκλωσης των οικιακών απορριμμάτων και διασφάλισης της ποιότητας ζωής.

Συγκεκριμένα, με τη χρήση ερωτηματολογίου σεναρίων και με τη μέθοδο της Conjoint Analysis, εξετάζονται οι συνδυασμοί παραγόντων που περιγράφουν τα προφίλ των πολιτών, καθώς και συγκρίνονται οι παράμετροι σχετικά με τη στάση των πολιτών σε θέματα φυσικού περιβάλλοντος στο Δήμο Θέρμης. Ενώ παράλληλα πραγματοποιήθηκε δειγματοληπτική έρευνα με «face-to-face» συνεντεύξεις για την αποτύπωση των στάσεων και αντιλήψεων των πολιτών σε τοπικά θέματα.

Λέξεις Κλειδιά: διαβούλευση, Δημοκρατία, συμμετοχικοί θεσμοί, conjoint analysis

1. ΕΙΣΑΓΩΓΗ

Το Τμήμα Πολιτικών Επιστημών του Αριστοτελείου Πανεπιστημίου Θεσσαλονίκης σε συνεργασία με το Δήμο Θέρμης πραγματοποίησε έρευνα κοινής γνώμης σχετικά με ζητήματα καθημερινότητας, συμμετοχής των πολιτών, προοπτικών ανάπτυξης της περιοχής καθώς και στάσεων/απόψεων των δημοτών για τους τοπικούς πολιτικούς θεσμούς.

Συγκεντρώθηκαν δεδομένα σχετικά την ικανοποίηση των δημοτών από τις παρεχόμενες υπηρεσίες καθώς και την εκτίμηση της λειτουργίας των αιρετών και των υπαλλήλων του Δήμου. Επίσης, καταγράφηκαν προβλήματα που απασχολούν τους δημότες, εκτίμηση για τα προβλήματα κα τις δράσεις κοινωνικής πολιτικής και λήψης πρωτοβουλιών καθώς επίσης η γνώμη των δημοτών σχετικά με τη Δημοτική Επιτροπή Διαβούλευσης (Floridia 2007) και τα τοπικά δημοψηφίσματα για την επίλυση κοινωνικών και άλλων θεμάτων. Παράλληλα, πραγματοποιήθηκε μια μελέτη των συμμετοχικών θεσμών με αντικείμενο την καταγραφή της σύνθεσής τους, τις θεματικές ενότητες της ατζέντας των θεσμών αυτών, τους τρόπους λήψης αποφάσεων και τη δυνατότητα επίλυσης προβλημάτων/θεμάτων ανά δημοτική ενότητα.

Η διαμόρφωση ισχυρών συμμετοχικών θεσμών είναι απαραίτητη προϋπόθεση προκειμένου να ενισχυθεί η τοπική δημοκρατία, να αναπτυχθούν αποτελεσματικοί, αξιόπιστοι και διαφανείς θεσμοί και επιπροσθέτως να αναπτυχθεί μια υπεύθυνη, συμμετοχική, αντιπροσωπευτική Κοινωνία των Πολιτών χωρίς αποκλεισμούς στη λήψη αποφάσεων σε όλα τα επίπεδα (Elster 1998, Manin 2005, Bobbio 2007). Πρόκειται για την ανάπτυξη ενός δικτύου ροής πολιτικών αποφάσεων. Αυτό αποτελεί γενικό στόχο της παρέμβασής μας με την έννοια διαμόρφωσης προτάσεων για την ανάπτυξη και λειτουργία

του. Στο πλαίσιο αυτό πραγματοποιήθηκε πείραμα συμμετοχικής διαβουλευτικής διαχείρισης θεμάτων (Elster 1998) στο Δήμο με αντικείμενο την διαχείριση των οικιακών απορριμμάτων. Το αντικείμενο μελέτης είναι το θέμα της διαχείρισης των οικιακών απορριμμάτων με βάση το Νόμο 4496/2017 που προβλέπει την εμπλοκή των Δήμων στη διαχείριση των οικιακών αποβλήτων. Επιδιώκεται, γενικότερα, η ενεργοποίηση των πολιτών σε τοπικό επίπεδο για την ανάδειξη θεμάτων και αποφάσεων γύρω από κίνητρα (π.χ. περιορισμό δημοτικών τελών) και απαιτούμενα μέτρα για την επιτυχία του εγχειρήματος. Σημειώνουμε ότι έχουν ήδη πραγματοποιηθεί διαβουλεύσεις σε τοπικά διαμερίσματα και προχωρά η υλοποίηση για το σύνολο των δημοτικών ενοτήτων.

Στα επόμενα κεφάλαια παρουσιάζονται τα αποτελέσματα της δειγματοληπτικής έρευνας σχετικά με το αίσθημα εγγύτητας/ απόστασης από τους θεσμούς και η στάση των πολιτών όσον αφορά δημοκρατικούς θεσμούς και διαδικασίες όπως η Δημοτική Επιτροπή Διαβούλευσης (ΔΕΔ) καθώς και τα αποτελέσματα της Conjoint Analysis.

2. ΔΕΙΓΜΑΤΟΛΗΠΤΙΚΗ ΕΡΕΥΝΑ

2.1 Περιγραφή δείγματος

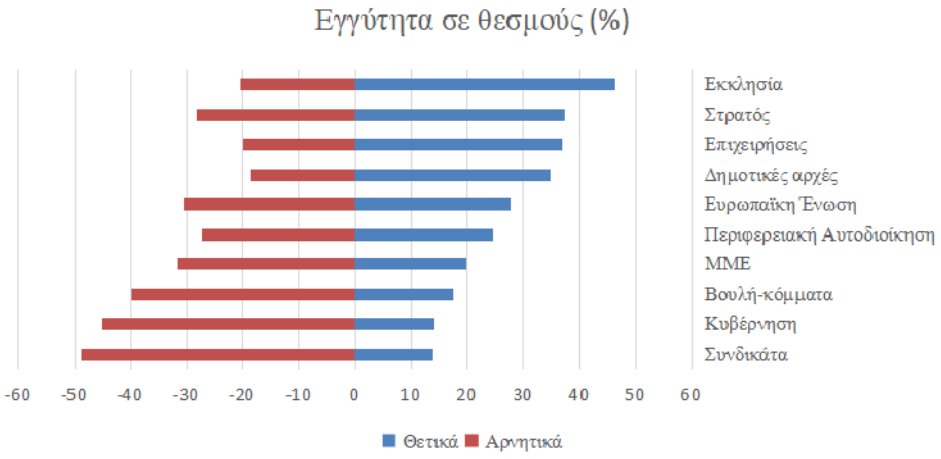
Η έρευνα πραγματοποιήθηκε στα 14 τοπικά διαμερίσματα του Δήμου Θέρμης το χρονικό διάστημα 04/12/2017 έως 12/02/2018. Από το συνολικό αριθμό ερωτηματολογίων 1.590 που συγκεντρώθηκαν αναλύονται μετά τον έλεγχο 1.500 συμπληρωμένα ερωτηματολόγια.

Το δειγματοληπτικό πλαίσιο σχεδιάστηκε σύμφωνα με τα στοιχεία της γενικής απογραφής πληθυσμού (2011). Στην έρευνα συμμετείχαν δημότες (73,5%) και μόνιμοι κάτοικοι (14,3%) που είναι εγγεγραμμένοι σε άλλο δήμο. Ποσοστό 12,2% δεν προσδιορίζει σε ποιο δήμο ανήκει. Ποσοστό 12% του δείγματος ανήκει στην ηλικιακή ομάδα «18-25 ετών», 13% «26-35 ετών», 35% «36-50 ετών», 16% «51-65 ετών» και 14% «άνω των 66 ετών».

2.2 Ανάλυση αποτελεσμάτων

Μελετήθηκε η απόσταση/εγγύτητα από τους «θεσμούς» του σημερινού πολιτικού συστήματος. Οι απαντήσεις ανά θεσμό παρουσιάζουν ιδιαίτερο ενδιαφέρον καθώς στο σύνολο του δείγματος καταγράφεται ισχυρό αίσθημα απόστασης από τους θεσμούς (βλ. γράφημα 1).

Γράφημα 1. *Απόσταση (-) /Εγγύτητα (+) από τους Θεσμούς του πολιτικού συστήματος*



Αναλυτικότερα, οι πολίτες δηλώνουν ότι νιώθουν εγγύτερα (Πάρα πολύ/Πολύ/Αρκετά) στην Εκκλησία με ποσοστό 46.3, στο Στρατό (37,4%), στις Επιχειρήσεις (36,9%) και στις Δημοτικές Αρχές (34,8%). Επίσης, παρατηρείται ότι αισθάνονται αποκομμένοι (Καθόλου)

από τα Συνδικάτα (48.7%), την Κυβέρνηση (45,1%) και τα Κόμματα-Βουλή (40,0%). Ακολουθούν: ΜΜΕ (31.6%), η Ε.Ε. (30.5%) και η Περιφερειακή Αυτοδιοίκηση (27.3%).

Στη συνέχεια μελετήθηκε αναλυτικά το αίσθημα εγγύτητας/ απόστασης από τις Δημοτικές Αρχές. Παρατηρείται ότι 34,8% αισθάνεται εγγύτερα (πάρα πολύ/πολύ/αρκετά) στις Δημοτικές αρχές, 39% αισθάνεται αποκομμένο (λίγο/καθόλου) και 26,2% αισθάνεται σχετικά εγγύτερα στις Δημοτικές αρχές. Επίσης, μελετήθηκε το αίσθημα απόστασης / εγγύτητας από τους «θεσμούς» ανά δημοτική ενότητα προκειμένου να διαπιστωθεί η στάση των πολιτών προς τις Δημοτικές Αρχές (βλ. πίνακα 1).

Συγκεκριμένα, οι πολίτες από τη δημοτική ενότητα Θέρμης αισθάνονται (πάρα πολύ/πολύ/αρκετά) εγγύτερα στις Δημοτικές αρχές σε ποσοστό 40,9%. Οι πολίτες από τη δημοτική ενότητα Μίκρας αισθάνονται (πάρα πολύ/πολύ/αρκετά) εγγύτερα στις Δημοτικές αρχές σε ποσοστό 31,8% ενώ πολίτες από τη δημοτική ενότητα Βασιλικών αισθάνονται (πάρα πολύ/πολύ/αρκετά) εγγύτερα στις Δημοτικές αρχές σε ποσοστό 21%.

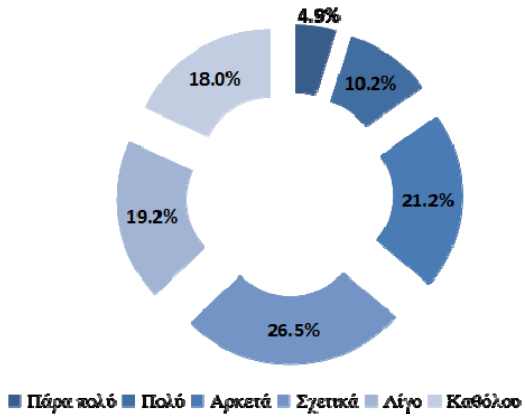
Πίνακας 1. *Δημοτική Ενότητα/ Απόσταση (-) /Εγγύτητα (+) από τις Δημοτικές Αρχές*

	Δημοτικές αρχές					
	Πάρα πολύ	**Πολύ**	**Αρκετά**	**Σχετικά**	**Λίγο**	**Καθόλου**
Δημοτική Ενότητα Θέρμης	7,7%	11,6%	21,6%	24,8%	19,4%	14,9%
Δημοτική Ενότητα Βασιλικών	1,8%	6,5%	12,7%	24,0%	25,8%	29,1%
Δημοτική Ενότητα Μίκρας	1,6%	6,4%	23,8%	30,7%	19,3%	18,2%

Ο πληθυσμός είναι μοιρασμένος όσον αφορά το αίσθημα εγγύτητας/ απόστασης από την Περιφερειακή Αυτοδιοίκηση. Οι πολίτες από τη δημοτική ενότητα Θέρμης αισθάνονται (πάρα πολύ/πολύ/αρκετά) εγγύτερα στην Περιφερειακή Αυτοδιοίκηση σε ποσοστό 26%. Οι πολίτες από τη δημοτική ενότητα Μίκρας αισθάνονται (πάρα πολύ/πολύ/αρκετά) εγγύτερα στην Περιφερειακή Αυτοδιοίκηση σε ποσοστό 25% ενώ πολίτες από τη δημοτική ενότητα Βασιλικών αισθάνονται (πάρα πολύ/πολύ/αρκετά) εγγύτερα σε ποσοστό 20,3%.

Ειδικά για όσους δηλώνουν δημότες του Δήμου (βλ. γράφημα 2) παρατηρείται ότι ποσοστό 4,9% νιώθει πάρα πολύ εγγύτερα, ποσοστό 10,2% πολύ, 21,2% αρκετά κοντά στις Δημοτικές αρχές ενώ σχετικά κοντά ποσοστό 26,5%. Το υπόλοιπο 37,2% δηλώνει ότι αισθάνεται μεγαλύτερη απόσταση από τις Δημοτικές αρχές.

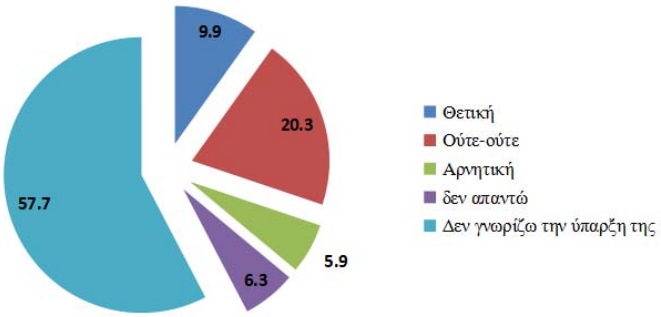
Γράφημα 2. *Απόσταση (-) /Εγγύτητα (+) από τις Δημοτικές αρχές για δημότες*



2.3 Δημοτική Επιτροπή Διαβούλευσης

Οι πολίτες (βλ. γράφημα 3) δεν εκφράζουν άποψη όσον αφορά τη ΔΕΔ. Έξι στους δέκα δεν γνωρίζουν την ύπαρξη της ΔΕΔ, ενώ δύο στους δέκα έχουν ουδέτερη θέση απέναντι στη ΔΕΔ.

Γράφημα 3. *Δημοτική Επιτροπή Διαβούλευσης*



Σχετικά με το θεσμό των τοπικών δημοψηφισμάτων παρατηρείται ότι 64,5% του δείγματος τον αξιολογεί θετικά και μόλις 8% αρνητικά, ενώ μόλις ένας στους είκοσι δηλώνει ότι δεν τον γνωρίζει.

Η στάση των πολιτών ανά Δημοτική ενότητα παρουσιάζει αξιόλογες διακυμάνσεις (βλ. πίνακα 2). Συγκεκριμένα, παρατηρείται ότι οι πολίτες της Δημοτικής ενότητας Βασιλικών είναι περισσότερο θετικοί προς το θεσμό της ΔΕΔ σε ποσοστό 20,4% ενώ ακολουθεί η Δημοτική ενότητα Θέρμης (8,9%) και Μίκρας (4,2%).

Πίνακας 2. *ΔΕΔ/ Δημοτική Ενότητα*

	Δημοτική ενότητα/ ΔΕΔ			
	Δεν γνωρίζω την ΔΕΔ	**Θετική**	**Ούτε-ούτε**	**Αρνητική**
Δημοτική Ενότητα Θέρμης	62,9%	8,9%	24,3%	3,9%
Δημοτική Ενότητα Βασιλικών	47,1%	20,4%	23,9%	8,6%
Δημοτική Ενότητα Μίκρας	54,2%	4,2%	33,3%	8,2%

Ανάλογη είναι η στάση των πολιτών ανά Δημοτική ενότητα όσον αφορά τη χρήση των τοπικών δημοψηφισμάτων για την επίλυση κοινωνικών θεμάτων. Συγκεκριμένα, παρατηρείται ότι οι πολίτες της Δημοτικής ενότητας Βασιλικών είναι περισσότερο θετικοί

προς τη χρήση τοπικών δημοψηφισμάτων σε ποσοστό 72,8% ενώ ακολουθεί η Δημοτική ενότητα Μίκρας (62,7%) και Θέρμης (62,5%).

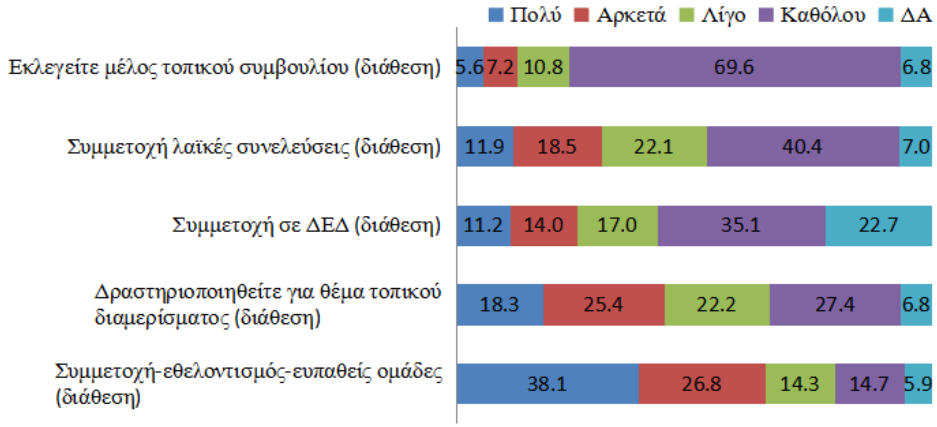
2.4 Συμμετοχή σε πρωτοβουλίες και δραστηριότητες

Στη συνέχεια εξετάστηκε η διάθεση και πρόθεση των πολιτών να συμμετάσχουν σε πρωτοβουλίες σχετικά με τοπικά θέματα/προβλήματα.

Οι πολίτες προτίθενται να συμμετάσχουν (πολύ/αρκετά) σε πρωτοβουλίες που αφορούν την ανακύκλωση (74,3%), την βοήθεια σε άτομα με ειδικές ανάγκες (72%) και την βοήθεια σε ηλικιωμένους (66%). Το ποσοστό των πολιτών που δηλώνει ότι δεν ενδιαφέρεται (καθόλου) να συμμετάσχει σε πρωτοβουλίες σχετικά με τις Δημοτικές εκλογές είναι 48%, σχετικά με τα προγράμματα άθλησης είναι 23,3%, την συγκοινωνία και την σχολική στέγη 17,9% αντίστοιχα.

Η διάθεση συμμετοχής σε δράσεις σε τοπικό επίπεδο έχει διάφορες διακυμάνσεις. Για τη συμμετοχή τους σε τοπικό επίπεδο σε δράσεις και θεσμικά όργανα, (βλ. γράφημα 4) ποσοστό 64,9% δηλώνει διατεθειμένοι «πολύ/αρκετά» να συμμετάσχει σε εθελοντικές δράσεις για την υποστήριξη των φτωχότερων κοινωνικών ομάδων, 43,7% να δραστηριοποιηθεί για θέματα του τοπικού διαμερίσματος, 30,4% να συμμετάσχει σε λαϊκές συνελεύσεις, 25,4% να συμμετάσχει σε συνεδριάσεις της ΔΕΔ και 12,8% να εκλεγεί ως μέλος του τοπικού συμβουλίου.

Γράφημα 4. *Συμμετοχή σε εθελοντικές δράσεις*



Όσον αφορά τους δείκτες πολιτικής κινητοποίησης παρατηρούμε ότι η πλειονότητα των πολιτών δεν είναι διατεθειμένοι να κινητοποιηθούν σε τοπικό επίπεδο. Ποσοστό 64% των πολιτών δεν είναι διατεθειμένο να δραστηριοποιηθεί υπέρ κάποιου συνδυασμού, 62,7% δεν είναι διατεθειμένο να ασκήσει επιρροή σε ψηφοφόρους και 51,4% επίσης δεν είναι διατεθειμένο να βοηθήσει κάποιον υποψήφιο σύμβουλο. Παρατηρείται ότι για όσα χαρακτηριστικά πολιτικής κινητοποίησης απαιτείται μεγαλύτερη «δέσμευση» και ενεργητική συμμετοχή από την πλευρά των πολιτών, μειώνεται η διάθεση συμμετοχής (βλ. δραστηριοποίηση υπέρ κάποιου συνδυασμού, άσκηση επιρροής σε ψηφοφόρους και βοήθεια υπέρ κάποιου υποψήφιου σύμβουλου) ενώ όσο περιορίζεται ο χαρακτήρας της ενεργής συμμετοχής, αυξάνεται η διάθεση συμμετοχής (βλ. παρακολούθηση τηλεόρασης για δημοτικά θέματα και παρακολούθηση συγκεντρώσεων υποψηφίων).

2.5 Συμπεράσματα

Οι πολίτες δηλώνουν ότι νιώθουν εγγύτερα στην εκκλησία, στο στρατό, στις επιχειρήσεις (οικονομικούς και οικουμενικούς θεσμούς) και στις Δημοτικές Αρχές. Επίσης, παρατηρείται ότι αισθάνονται «μακριά» από τα Συνδικάτα, την Κυβέρνηση και τα Κόμματα-Βουλή

(πολιτικούς θεσμούς). Ειδικά για τους δημότες, όσον αφορά το αίσθημα εγγύτητας/απόστασης από τις Δημοτικές αρχές, παρατηρείται ότι ποσοστό 36,3% νιώθει εγγύτερα, ενώ σχετικά κοντά ποσοστό 26,5%. Αντίθετα ποσοστό 37,2% δηλώνει ότι αισθάνεται μεγαλύτερη απόσταση από τις δημοτικές αρχές. Επίσης, μελετήθηκε το αίσθημα εγγύτητας/απόστασης από την Περιφερειακή Αυτοδιοίκηση για τους δημότες, όπου ένας στους δύο δημότες δηλώνει ότι αισθάνεται σε απόσταση από την περιφερειακή αυτοδιοίκηση (λίγο/καθόλου 49%).

Οι δείκτες πολιτικής κινητοποίησης προϋποθέτουν μεγαλύτερη «δέσμευση» και ενεργητική συμμετοχή από την πλευρά των πολιτών ωστόσο παρατηρήθηκε ότι σε αυτή την περίπτωση, μειώνεται η διάθεση συμμετοχής από τους πολίτες. Ενώ όσο περιορίζεται ο χαρακτήρας της απαιτούμενης ενεργής συμμετοχής, αυξάνεται και η διάθεση συμμετοχής των πολιτών. Γενικότερα, καταφεύγουν σε δράσεις βάσει ατομοκεντρικών προτύπων και όχι συλλογικών.

Σχετικά με το θεσμό της Δημοτικής Επιτροπής Διαβούλευσης, οι πολίτες δεν εκφράζουν άποψη όσον αφορά τη ΔΕΔ. Έξι στους δέκα δεν γνωρίζουν την ύπαρξη της ΔΕΔ, ενώ δύο στους δέκα έχουν ουδέτερη θέση απέναντι της. Σχετικά με το θεσμό των τοπικών δημοψηφισμάτων παρατηρείται ότι η πλειοψηφία την αξιολογεί θετικά ενώ μόλις ένας στους είκοσι δηλώνει ότι δεν τον γνωρίζει.

3. CONJOINT ANALYSIS

3.1 **Εισαγωγή**

Η χρήση της Conjoint Analysis (Orme 2010) επιτρέπει στους ερευνητές να διερευνούν και να αποτυπώνουν με ρεαλισμό τις επιλογές των ατόμων ως αποτέλεσμα αποφάσεων μεταξύ πολλαπλών και διαφορετικών χαρακτηριστικών ή παραγόντων (Hair, Anderson, Tatham, & Grablowsky 1998). Αποτιμάται η «σχετική βαρύτητα» κάθε χαρακτηριστικού, παράγοντα, έννοιας, κατάστασης ή αντικειμένου που διερευνάται. Κάθε φορά που διερευνούμε μία έννοια, ένα αντικείμενο ή μία κατάσταση καθορίζουμε τα χαρακτηριστικά ή τους παράγοντες που υποθέτουμε ότι συνθέτουν το ερευνούμενο μέγεθος. Στη συνέχεια με τη χρήση της Conjoint Analysis προσδιορίζουμε τη «σχετική βαρύτητα», δηλαδή τη σημασία του κάθε χαρακτηριστικού ή παράγοντα ξεχωριστά.

Τη σημασία αυτών των χαρακτηριστικών, τα οποία παρουσιάζονται ως εναλλακτικές επιλογές, τις οποίες αξιολογεί ο ερωτώμενος και τα οποία παρουσιάζονται ως παραλλασσόμενες δέσμες χαρακτηριστικών, μπορούμε να προσδιορίσουμε με τη συγκεκριμένη μέθοδο. Βασικό στοιχείο για τη χρησιμοποίηση των παραγόντων είναι η λειτουργικότητά τους. Οι παράγοντες οι οποίοι χρησιμοποιούνται στην Conjoint Analysis πρέπει να έχουν αλληλεπίδραση και σχέση μεταξύ τους, καθώς επίσης και συγκρισιμότητα.

3.2 Επιλογή Παραγόντων. Το παρόν πείραμα

Θέλουμε να μελετήσουμε ποιος συνδυασμός επιλογών (Halme and Kallio, 2011)–θα αποφέρει το καλύτερο δυνατό αποτέλεσμα απόδοσης στο Δήμο Θέρμης σχετικά με τη διαχείριση απορριμμάτων. Για να δημιουργήσουμε τον συνδυασμό επιλογών τις οποίες θα αξιολογήσουν κατά τη διάρκεια της έρευνας οι ερωτώμενοι επιλέγουμε τους παράγοντες και τα επίπεδα τους (Green and Srinivasan, 1990).

Για την παρούσα έρευνα επιλέγουμε αρχικά τον παράγοντα «συλλογικές οικονομικές επιπτώσεις». Οι τιμές της συγκεκριμένης μεταβλητής είναι τρεις: Μείωση τελών σε όλους το ίδιο (1), Μείωση ανάλογα με την ποσότητα απορριμμάτων (2), Έργα για την περιοχή (3). Ο δεύτερος παράγοντας που συμμετέχει στην ανάλυση της παρούσας μελέτης είναι το «Ατομικές Οικονομικές Επιπτώσεις». Τα επίπεδα του συγκεκριμένου παράγοντα είναι δύο: Επιβολή Προστίμου (1) , Μη επιβολή Προστίμου (2)

Ο τρίτος παράγοντας είναι η «πρόσβαση σε σημεία συγκέντρωσης». Ο παράγοντας αυτός έχει δύο επίπεδα: Μικρή απόσταση (1), Μεγάλη απόσταση (2). Ο τέταρτος παράγοντας είναι η «Σωτηρία του Πλανήτη». Ο παράγοντας αυτός διακρίνεται σε δύο επίπεδα: Αποτελεί

πρώτη προτεραιότητα (1), Δεν αποτελεί πρώτη προτεραιότητα (2). Ο πέμπτος και τελευταίος παράγοντας είναι ο «διαχωρισμός στο σπίτι». Ο παράγοντας αυτός διακρίνεται σε δύο επίπεδα: Είμαι διατεθειμένος να κάνω (1), Δεν είμαι διατεθειμένος να κάνω (2).

Με βάση τα παραπάνω προκύπτει ένα σχέδιο από τον ορθογώνιο σχεδιασμό (βλ. πίνακα 3) με οχτώ συνδυασμούς (προφίλ επιλογών) με βάση τα παραπάνω επίπεδα (στάθμες) των παραγόντων όπου αναφέρονται στον παρακάτω πίνακα.

Πίνακας 3. *Ορθογώνιος σχεδιασμός σε επίπεδα*

Παράγοντας 1	**Παράγοντας 2**	**Παράγοντας 3**	**Παράγοντας 4**	**Παράγοντας 5**
1	1	2	2	1
1	2	2	2	2
3	1	1	2	2
1	1	1	1	2
1	2	1	1	2
2	2	1	2	1
3	2	2	1	1
2	1	2	1	2

3.3 Συλλογή Στοιχείων

Η έρευνα πραγματοποιήθηκε στο Δήμο Θέρμης με τη συμμετοχή 105 πολιτών την περίοδο από Δεκέμβριο 2017 έως και Απρίλιο 2018. Οι συμμετέχοντες ρωτήθηκαν για θέματα φυσικού περιβάλλοντος και θέματα που αφορούν το Δήμο. Η διαδικασία ολοκληρώθηκε ως εξής: κλήθηκαν οι πολίτες να ενημερωθούν και να τοποθετηθούν σε θέματα διαχείρισης θεμάτων του Δήμου σε ανοιχτές λαϊκές συνελεύσεις. Με την ολοκλήρωση της διαδικασίας οι πολίτες απάντησαν σε ένα σύντομο ερωτηματολόγιο, η συμπλήρωση του οποίου δεν ξεπερνούσε τα 10λεπτά.

3.4 Ερωτηματολόγιο σεναρίων

Ζητάμε από τους μετέχοντες στην έρευνα να ταξινομήσουν τα προφίλ των επιλογών με βάση το σενάριο που παρουσιάζουμε παρακάτω. Η ιεράρχηση γίνεται από το 1 μέχρι και το 8. Στην περίπτωση μας, για το ερευνητικό ερώτημα που τίθεται ο αριθμός «1» αντιπροσωπεύει αυτό τον συνδυασμό επιλογών που θα φέρουν το καλύτερο αποτέλεσμα στο δήμο και σταδιακά φτάνουμε στο «8» που αντιπροσωπεύει το προφίλ επιλογών που κατά τη γνώμη του ερωτώμενου θα είναι λιγότερο αποτελεσματικό.

Ο Δήμος (Χ) καλείται να εφαρμόσει, το νέο θεσμικό πλαίσιο για τη διαχείριση απορριμμάτων. Καλείται να αποφασίσει για την αξιοποίηση του οικονομικού οφέλους, τη θέση των σημείων συλλογής απορριμμάτων καθώς και την επιβολή ή όχι προστίμου σε όσους δεν συμμετέχουν στη διαχείριση απορριμμάτων.
Ο πολίτης (Υ) κρίνει τη σημασία του θέματος καθώς και εκφράζει την θέση του σχετικά με τη διαλογή στο σπίτι.

«Βάλτε στη σειρά από το 1 έως το 8 τους παρακάτω συνδυασμούς στάσης του πολίτη και πολιτικής του δήμου που θα φέρει το καλύτερο αποτέλεσμα στο δήμο. Στη θέση 1 θα είναι ο συνδυασμός με το καλύτερο αποτέλεσμα ενώ στη θέση 8 ο συνδυασμός που θα αποφέρει το λιγότερο καλό αποτέλεσμα στο δήμο».

Πίνακας 4: *Ερωτηματολόγιο σεναρίων*

	Ο **Δήμος Χ** έχει αποφασίσει να μειώσει τα δημοτικά τέλη, **σε όλους το ίδιο** και **να επιβάλλει** πρόστιμο σε όσους δεν συμμετέχουν. Η απόσταση από τα σημεία συλλογής που έχει ορίσει είναι **σχετικά μεγάλη**. Ο **Υ** θεωρεί ότι το θέμα **δεν είναι πρώτης προτεραιότητας** αλλά **είναι διατεθειμένος** να κάνει διαλογή στο σπίτι ανά είδος απορριμμάτων.
	Ο **Δήμος Χ** έχει αποφασίσει να μειώσει τα δημοτικά τέλη, **σε όλους το ίδιο**, **χωρίς να επιβάλλει** πρόστιμο σε όσους δεν

Παράγοντες	Επίπεδα Παραγόντων	(Χρησιμότητα) Utility	Std. Error
Συλλογικές Οικονομικές Επιπτώσεις	Σε όλους το ίδιο	-0,31	0,252

	συμμετέχουν. Η απόσταση από τα σημεία συλλογής που έχει ορίσει είναι **σχετικά μεγάλη**. Ο **Υ** θεωρεί ότι το θέμα **δεν είναι πρώτης προτεραιότητας** και **δεν είναι διατεθειμένος** να κάνει διαλογή στο σπίτι ανά είδος απορριμμάτων.
	Ο **Δήμος Χ** έχει αποφασίσει **να αξιοποιήσει το οικονομικό όφελος για έργα στο δήμο** και **να επιβάλλει πρόστιμο** σε όσους δεν συμμετέχουν. Η απόσταση από τα σημεία συλλογής που έχει ορίσει είναι **μικρή**. Ο **Υ** θεωρεί ότι το θέμα **δεν είναι πρώτης προτεραιότητας** και **δεν είναι διατεθειμένος** να κάνει διαλογή στο σπίτι ανά είδος απορριμμάτων.
	Ο **Δήμος Χ** έχει αποφασίσει να μειώσει να μειώσει τα δημοτικά τέλη, **σε όλους το ίδιο** και **να επιβάλλει πρόστιμα** σε όσους δεν συμμετέχουν. Η απόσταση από τα σημεία συλλογής που έχει ορίσει είναι **μικρή**. Ο **Υ** θεωρεί ότι το θέμα **είναι πρώτης προτεραιότητας** και **είναι διατεθειμένος** να κάνει διαλογή στο σπίτι ανά είδος απορριμμάτων.
	Ο **Δήμος Χ** έχει αποφασίσει να μειώσει να μειώσει τα δημοτικά τέλη, **σε όλους το ίδιο**, **χωρίς να επιβάλλει πρόστιμα** σε όσους δεν συμμετέχουν. Η απόσταση από τα σημεία συλλογής που έχει ορίσει είναι **μικρή**. Ο **Υ** θεωρεί ότι το θέμα **είναι πρώτης προτεραιότητας** αλλά **δεν είναι διατεθειμένος** να κάνει διαλογή στο σπίτι ανά είδος απορριμμάτων.
	Ο **Δήμος Χ** έχει αποφασίσει να μειώσει τα δημοτικά τέλη **ανάλογα με την ποσότητα που θα παραδώσει το άτομο**, **χωρίς να επιβάλλει πρόστιμα** σε όσους δεν συμμετέχουν. Η απόσταση από τα σημεία συλλογής που έχει ορίσει είναι **μικρή**. Ο **Υ** θεωρεί ότι το θέμα **δεν είναι πρώτης προτεραιότητας** αλλά **είναι διατεθειμένος** να κάνει διαλογή στο σπίτι ανά είδος απορριμμάτων.
	Ο **Δήμος Χ** έχει αποφασίσει **να αξιοποιήσει το οικονομικό όφελος για έργα στο δήμο**, **χωρίς να επιβάλλει πρόστιμα** σε όσους δεν συμμετέχουν. Η απόσταση από τα σημεία συλλογής που έχει ορίσει είναι σχετικά **μεγάλη**. Ο **Υ** θεωρεί ότι το θέμα **είναι πρώτης προτεραιότητας** και **είναι διατεθειμένος** να κάνει διαλογή στο σπίτι ανά είδος απορριμμάτων.
	Ο **Δήμος Χ** έχει αποφασίσει να μειώσει τα δημοτικά τέλη **ανάλογα με την ποσότητα που θα παραδώσει το άτομο** και **να επιβάλλει πρόστιμα** σε όσους δεν συμμετέχουν. Η απόσταση από τα σημεία συλλογής που έχει ορίσει είναι **σχετικά μεγάλη**. Ο **Υ** θεωρεί ότι το θέμα **είναι πρώτης προτεραιότητας** αλλά **δεν είναι διατεθειμένος** να κάνει διαλογή στο σπίτι ανά είδος απορριμμάτων.

3.5. Αποτελέσματα Conjoint Analysis

Ο πίνακας 5, παρουσιάζει έναν σταθερό όρο (constant), συντελεστής ο οποίος αντιπροσωπεύει τη μέση ανταπόκριση στα διαφορετικά προφίλ ατόμων και ένα σύνολο χρησιμότητες (utilities) καθεμία από τις οποίες αντιστοιχεί στη συνεισφορά κάθε επιπέδου κάθε παράγοντα (ανεξάρτητη μεταβλητή) στη συνολική χρησιμότητα (total utility).

Από τα αποτελέσματα του παρακάνω πίνακα (πίνακας 5), παρατηρούμε την τιμή της χρησιμότητας (utility) σε κάθε επίπεδο παράγοντα. Η επιλογή εκείνου του επιπέδου από κάθε παράγοντα με τη μεγαλύτερη τιμή ορίζει ένα προφίλ επιλογών του οποίου οι χρησιμότητες (utilities) υπολογίζονται σε μία ενιαία αριθμητική κλίμακα ώστε αφενός η συνολική χρησιμότητα (total utility) να προκύπτει από το άθροισμα των αντίστοιχων χρησιμοτήτων (utilities) των επιπέδων και αφετέρου μεγαλύτερη αριθμητικά χρησιμότητα (utility) να σημαίνει μεγαλύτερη επίδραση του επιπέδου του αντίστοιχου παράγοντα.

Πίνακας 5. *Αποτελέσματα*

	Ανάλογα με την ποσότητα	0,126	0,296
	Για έργα στο δήμο	0,184	0,296
Πρόσβαση σε σημεία συγκέντρωσης	Μικρή	-0,302	0,189
	Μεγάλη	0,302	0,189
Ατομικές Οικονομικές Επιπτώσεις	Επιβολή Προστίμου	-0,113	0,189
	Μη επιβολή Προστίμου	0,113	0,189
Σωτηρία του Πλανήτη	Είναι προτεραιότητα	0,267	0,189
	Δεν είναι προτεραιότητα	-0,267	0,189
Διαλογή στο σπίτι	Είναι διατεθειμένος	0,044	0,189
	Δεν είναι διατεθειμένος	-0,044	0,189
Σταθερά (Constant)		**4,578**	**0,199**
	Μη επιβολή Προστίμου	0,113	0,189
Σωτηρία του Πλανήτη	Είναι προτεραιότητα	0,267	0,189
	Δεν είναι προτεραιότητα	-0,267	0,189
Διαλογή στο σπίτι	Είναι διατεθειμένος	0,044	0,189
	Δεν είναι διατεθειμένος	-0,044	0,189
Σταθερά (Constant)		**4,578**	**0,199**

Παρατηρούμε επίσης, ότι τα επίπεδα παραγόντων που συγκεντρώνουν τις υψηλότερες τιμές είναι το «για έργα στο δήμο» (0.184) ως προς τον παράγοντα «οικονομικές επιπτώσεις», η «μεγάλη απόσταση» (0.302) ως προς την «απόσταση», η «μη επιβολή προστίμου» (0.113) ως προς τις «ατομικές οικονομικές επιπτώσεις», η «προτεραιότητα» (0.267) ως προς τον παράγοντα «σωτηρία του πλανήτη» και ο «είναι διατεθειμένος» (0.444) ως προς τον παράγοντα «διαλογή στο σπίτι». Αυτό σημαίνει ότι η αποδοτικότερη ασκούμενη πολιτική για το Δήμο είναι να διαθέσει τα χρήματα για έργα, η απόσταση πρόσβασης από τα σημεία συγκέντρωσης να είναι σχετικά μεγάλη, να μην επιβάλλει ατομικά πρόστιμα, να θέτει ως προτεραιότητα την σωτηρία του πλανήτη και να καλείται ο πολίτης να κάνει διαλογή στο σπίτι.

Ο παρακάτω πίνακας (πίνακας 6) αποτυπώνει τη σημασία κάθε παράγοντα στο σύνολο της επίδρασης με τη μορφή της σχετικής βαρύτητας.

Πίνακας 6: *Σημασία παραγόντων*

ΠΑΡΑΓΟΝΤΑΣ	ΣΧΕΤΙΚΗ ΒΑΡΥΤΗΤΑ	ΚΑΤΆΤΑΞΗ
Πρόσβαση	31,045	1ος σε σημασία
Σωτηρία του Πλανήτη	27,463	2ος σε σημασία
Συλλογικές Οικονομικές Επιπτώσεις	25,373	3ος σε σημασία
Ατομικές Οικονομικές Επιπτώσεις	11,642	4ος σε σημασία
Διαλογή στο σπίτι	4,478	5ος σε σημασία

Η «πρόσβαση στα σημεία συγκέντρωσης» δηλαδή, το πόσο μικρή ή σχετικά μεγάλη θεωρεί την απόσταση ο πολίτης, έχει σημασία σε ποσοστό 31.045%. Η «σωτηρία του πλανήτη» δηλαδή, το αν κάποιος θεωρεί ή δεν θεωρεί προτεραιότητα τη διάσωση του πλανήτη επιδρά σε ποσοστό 27.463%, οι «συλλογικές οικονομικές επιπτώσεις» δηλαδή, αν τα χρήματα μοιράζονται ισομερώς στους πολίτες, αν μοιράζονται αναλογικά ή αν

χρησιμοποιούνται για έργα στο δήμο επιδρά σε ποσοστό 25.373% στη συνολική επίδραση. Στη συνέχεια, ο παράγοντας «ατομικές οικονομικές επιπτώσεις» επιδρά με ποσοστό 11.642% και τέλος, η «διαλογή στο σπίτι» δηλαδή, το αν κάποιος πολίτης είναι διατεθειμένος ή όχι να κάνει διαλογή στο σπίτι επηρεάζει το 4.478% της συνολικής αποτελεσματικότητας.

Μέσω της ανάλυσης που πραγματοποιήσαμε, μπορούν με βάση τον πίνακα των χρησιμοτήτων να αξιολογηθούν όλοι οι πιθανοί συνδυασμοί στα προφίλ επιλογών που μπορούν να προκύψουν παίρνοντας ένα επίπεδο από κάθε παράγοντα. Η μεγαλύτερη χρησιμότητα προκύπτει για το συνδυασμό: «για έργα στο δήμο» (0.184), «μεγάλη απόσταση» (0.302), «μη επιβολή προστίμου» (0.113), «προτεραιότητα» (0.267), και «είναι διατεθειμένος» (0.444). Η συνολική χρησιμότητα υπολογίζεται: 0.184+0.302+0.113+0.267+ 0.444= 1.31

Με τον ίδιο τρόπο μπορεί να υπολογιστεί η χρησιμότητα κάθε πεντάδας τιμών. Κάθε πεντάδα αντιστοιχεί σε ένα προφίλ επιλογών. Στον πίνακα 7 οι υψηλές τιμές των συντελεστών Pearson's R και Kendall's tau δείχνουν ότι οι μετρήσεις για το προφίλ του ερευνούμενου εξομοιώνονται με το αθροιστικό μοντέλο της ανάλυσης.

***Πίνακας 7**. Συντελεστές*

Correlations [a]		
	Value	Sig.
Pearson's R	**0,931**	**0**
Kendall's tau	**0,618**	**0,017**
Correlations between observed and estimated preferences		

4. ΑΛΛΑ ΑΠΟΤΕΛΕΣΜΑΤΑ ΑΠΟ ΤΗ ΔΕΙΓΜΑΤΟΛΗΠΤΙΚΗ ΕΡΕΥΝΑ

Στην ερώτηση σχετικά με το ποιος νομίζουν οι πολίτες ότι ευθύνεται για την κατάσταση με τα σκουπίδια, το 68,2% των ερωτηθέντων θεωρούν υπεύθυνο το δήμο, το 11,8% την κυβέρνηση, το 2,1% των ερωτηθέντων θεωρούν υπεύθυνο κάποιον άλλο και το 6,8% απαντά ότι δεν ξέρει.

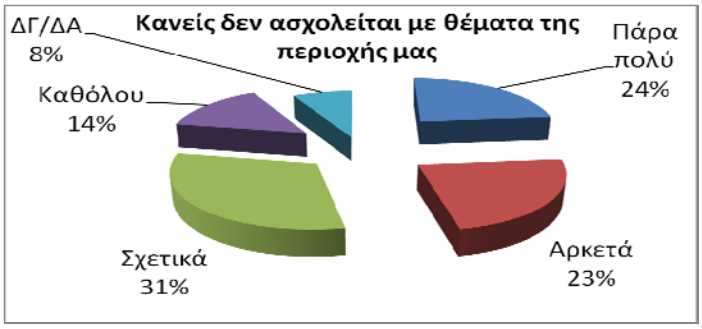
Αναφορικά με την ερώτηση σχετικά με το κατά πόσο είναι διατεθειμένος να συμμετέχει σε πρωτοβουλίες σχετικά με την ανακύκλωση και μείωση των σκουπιδιών, το 46% των ερωτηθέντων απαντάει «πολύ», το 25% «σχετικά», το 15% «λίγο», το 10% «καθόλου» και το 4% απαντά «δεν ξέρω».

Στην ερώτηση εάν οι δημοτικές αρχές ενδιαφέρονται για έργα «βιτρίνας» χωρίς να ασχολούνται με τα πραγματικά προβλήματα (βλ. γράφημα 5), το 54% των ερωτηθέντων απαντά «πάρα πολύ», το 24% «αρκετά», το 14% «σχετικά», 3% «καθόλου» και το 5% απαντά «δεν ξέρω».

***Γράφημα 5**. Γνώμη για έργα «βιτρίνας»*

Στην ερώτηση εάν στις δημοτικές αρχές ασχολείται κανείς με τα θέματα της περιοχής (βλ. γράφημα 6), το 24% των ερωτηθέντων απαντά «πάρα πολύ», το 23% «αρκετά», το 31% «σχετικά», 14% «καθόλου» και το 8% απαντά «δεν ξέρω».

***Γράφημα 6**. Γνώμη για ενασχόληση με τοπικά θέματα*

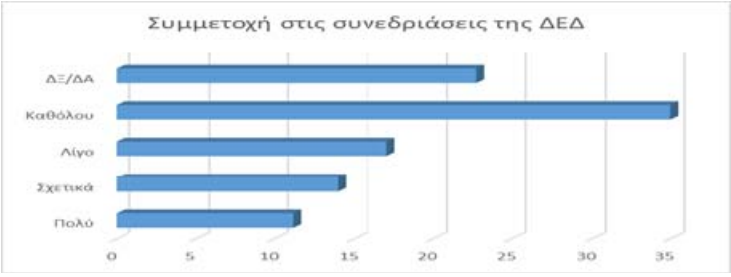


Αναφορικά με την ερώτηση εάν συμφωνούν ότι στις δημοτικές αρχές πρέπει να ξεκινήσουμε από τα μικρά καθημερινά θέματα, που είναι σημαντικά στο Δήμο μας, το 54% των ερωτηθέντων απαντά «πάρα πολύ», το 29% «αρκετά», το 11% «σχετικά», 1% «καθόλου» και το 5% απαντά «δεν ξέρω».

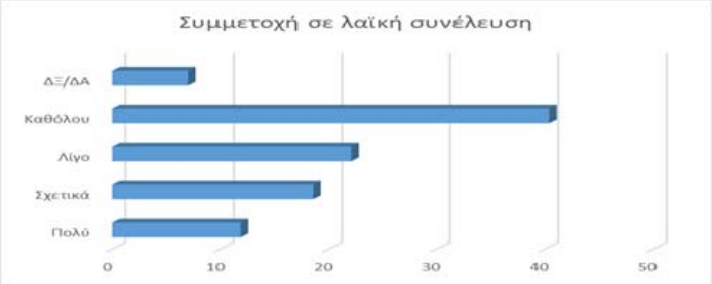
Στην ερώτηση κατά πόσο είναι διατεθειμένοι να συμμετέχουν σε συνεδριάσεις της ΔΕΔ (βλ. γράφημα 7), ένα σημαντικό ποσοστό κοντά στο 35% των ερωτηθέντων δεν είναι καθόλου διατεθειμένοι να συμμετέχουν σε συνεδριάσεις της ΔΕΔ.

Στην ερώτηση κατά πόσο είναι διατεθειμένοι να συμμετέχουν σε λαϊκές συνελεύσεις (βλ. γράφημα 8), ένα σημαντικό ποσοστό κοντά στο 40% των ερωτηθέντων δεν είναι καθόλου διατεθειμένοι να συμμετέχουν σε λαϊκές συνελεύσεις.

***Γράφημα 7**. Συμμετοχή σε συνεδριάσεις*



Γράφημα 8 *. Συμμετοχή σε λαϊκές συνελεύσεις*



Τέλος, στην ερώτηση για το αν τα τοπικά δημοψηφίσματα είναι χρήσιμα στην επίλυση κοινωνικών θεμάτων το 42,1% των ερωτηθέντων τα αξιολογεί θετικά, το 22,4% πολύ θετικά, το 19% δεν γνωρίζει, το 6,5% το αξιολογεί αρνητικά.

ABSTRACT

Strong participatory institutions are essential requirement in order to enhance local democracy, to develop reconstructive, reliable and transparent institutions. In addition, they contribute to develop a participatory and representative Civic Society without exclusions in decision-making process. Active citizens in local level promote regional issues from "bottom" (neighbourhood) to "top" (municipal board). This experiment focuses on the protection and improvement of natural environment in Municipality of Thermi. The issues referring to recycling domestic waste and improving the quality of life in the region. Deliberation offers the opportunity to use efficiently governance structures to benefit the citizens and improve municipality's services. We apply Conjoint Analysis and we examine the factors that describe citizen's profile and compare the citizens' attitude towards environmental issues in Thermi. In addition, "face-to-face" interviews has been conducted to reflect citizens' attitude towards local issues.

ΑΝΑΦΟΡΕΣ

Bobbio, L. (2007). Tipi di deliberazione. *Rivista Italiana di Scienza Politica*, **37 (3)**, 359-384.

Elster J. (1998). *Deliberative Democracy*, Cambridge: Cambridge University Press.

Floridia A. (2007). *From participation to deliberation, A critical Genealogy of Deliberative Democracy*, ECPR Press.

Green, P. E. and Srinivasan V. (1990). Conjoint Analysis in Marketing: New Developments with Implications for Research and Practice, *Journal of Marketing*, **54**, 3-19.

Hair, J.F. Jr., Anderson, R.E., Tatham, R.L. and Grablowsky, B.J. (1979). Multivariate data analysis, Tulsa, OK: Petroleum Publishing.

Halme, M. and Kallio, M. (2011). Estimation methods for choice-based conjoint analysis of consumer preferences. *European Journal of Operational Research,* **214(1)**, 160-167.

Manin, B. (2005). *Democratic Deliberation: why we should focus on debate rather than discussion*, Program in Ethics and Public Affairs Seminar. Princeton University.

Orme, B. (2010). Getting started with conjoint analysis. 2nd ed. Madison, WI: Research Publishers, LLC.

εργασίες στα αγγλικά

Greek Statistical Institute
Proceedings of the 31st Panhellenic Statistics Conference (2018), pp. 278−291

Markovian Arrival Process modeling for the detection of seismicity rate changes for the strong earthquakes in Greece

P. Bountzis[1], E. Papadimitriou[1], G. Tsaklidis[2]
[1]Aristotle University of Thessaloniki, Geophysics Department, 54124, Thessaloniki, Greece
pmpountzp, ritsa@geo.auth.gr
[2]Aristotle University of Thessaloniki, Mathematics Department, 54124, Thessaloniki, Greece
tsaklidi@math.auth.gr

ABSTRACT

The Markovian Arrival Process (MAP) model is applied to a complete earthquake catalog with magnitude $M \geq 6.5$, in the broader area of Greece, for investigating the temporal variations of seismicity. MAP is a stochastic process, that counts the regional earthquake occurrences and allows for transitions between the hidden states of a Markov Process, such that each state corresponds to a different level of seismicity rate. The inference regarding the choice of the optimal number of hidden states is made through Akaike and Bayesian information criteria, and the estimation of the model parameters is accomplished via the EM algorithm. A local decoding procedure is implemented to reveal the times of a change in the occurrence rate of earthquakes. The evaluated conditional intensity function demonstrates an alternation between higher and lower seismic activity levels, which is associated in some cases with major earthquake occurrences.

Keywords: hidden states, EM algorithm, forward-backward equations, seismicity rates.

1. INTRODUCTION

One of the most accepted feature of large earthquakes generation is the assumption of time independence (Michael, 2011; Daub et al., 2012; Parsons and Geist, 2012). Strong earthquakes occur according to a stationary Poisson process, meaning that they occur (arrive) according to a random process in time with a stable occurrence rate, μ. However, brevity of the earthquake catalogs, as well the limited global number of large earthquake occurrences reduce the robust validation of the statistical tests (Dimer de Oliveira, 2012; Lombardi and Marzocchi, 2007).

A possible non-stationarity of the time series of strong earthquakes (Bufe and Perkins, 2005; McCaffrey, 2008) is an interesting assumption. Transient perturbation caused by poroelastic relaxation probably increased the background rate during the 1997-1998 Umbria-Marche, Central Italy (Lombardi et al. 2010). Even though global interactions between large earthquakes can not be assumed, there is a significant increase of seismicity rate after major earthquakes in regional scale far beyond the rupture zone (Papazachos et al., 2000; Parsons and Velasco, 2011; Stein and Toda, 2013).

In this work, we want to analyze the temporal behavior of strong earthquakes in the broader area of Greece. MAP model is applied in order to examine whether there are long-term fluctuations in the seismicity rate. It was introduced by Neuts (1979) and a more explicit definition was given later by Lucantoni et al., (1990). It concerns a two dimensional Markov process, $\{(N_t, J_t), t \geq 0\}$, with the first marginal process being a counting one, N_t, in which the associated arrivals correspond to earthquake occurrence times, and the second one, J_t, being a Markov process with unobservable states that determine the arrival rate. MAPs provide a generalization of the Poisson process and renewal models and the variable arrival rate which is assumed, reveals the non-stationarity of the counting process. The interarrival times follow a phase type (PH) distribution and are conditionally but not completely independent (He, 2014). Hence, correlations among interevent times can be captured. In addition any stochastic point process is the weak limit of a sequence of MAPs (Asmussen and Koole, 1993). A comprehensive review on MAPs and its applications, along with some important special cases are given in Artalejo et al., (2010) and Cordeiro and Kharoufeh, (2011).

Section 2 describes the dataset used in the current study and preliminaries of MAP theory, along with the testing methodology. Section 3 presents the results of the applied method. Finally, in Section 4 the concluding remarks are drawn and some issues and recommendations are given.

2. DATA AND METHODOLOGY

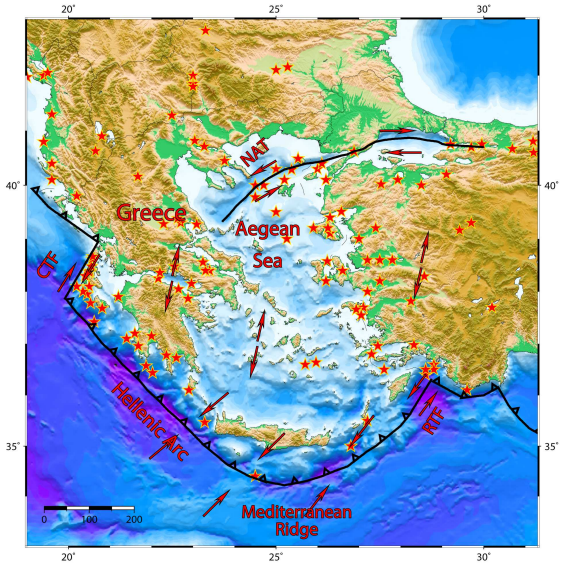
2.1 Data

Our aim is to reveal the temporal behavior of strong earthquake occurrences with magnitude $M \geq 6.5$ in the broader area of Greece, and to examine whether there are seismicity rate changes related to major events. The focal mechanisms of large events have been thoroughly investigated since 1950 (Taymaz et al., 1991a,b and Baker et al., 1997 among others), and a brief bibliography regarding events with $M \geq 6.5$, is given in Papazachos and Papazachou (2003) and the references therein.

In our study a seismicity catalog of shallow events with depth $D \leq 45$ km and with magnitude $M \geq 6.5$ is compiled, according to the most recent publishments. The catalog is assumed to be complete and homogeneous (Papazachos

and Papazachou, 2003) and includes 118 events, since 1845, within the boundaries $\phi = 33^o - 43^o$ and $\lambda = 19^o - 30^o$ (Figure 1). Paradisopoulou et al., (2010) provided evidence of interactions between strong earthquakes in Western Turkey and North Anatolian Fault (NAF), due to stress changes in the area. Since our goal is to model temporal patterns of possible earthquake interactions, four events outside the boundaries of our study area are included in the catalog. These are events like the Düzce ($M_w = 7.1$) on 12 November, 1999 which occurred almost three months after the earthquake of Izmit ($M_w = 7.6$) on 17 August, 1999.

Figure 1: *Seismotectonic properties of Greece and its surrounding area along with the epicentral distribution of earthquakes with $M \geq 6.5$ that occurred in the study area since 1845.*



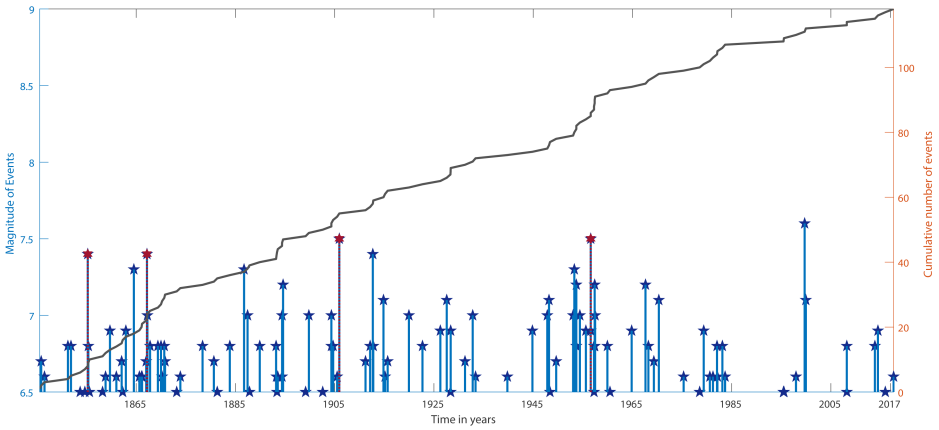
The temporal distribution in conjunction with the cumulative number of events (Figure 2), illustrates the existence of active periods like in 1980-1983 which alternate with quiescent ones (1983-1995). Moreover, in some cases these changes could be associated with the occurrence of large earthquakes, like the Amorgos $M_w = 7.5$ (Papadimitriou et al., 2005) event on 9 July, 1956 (red pentagon).

2.2 Methodolody

2.2.1 Markovian arrival process

As a first step, the definition of the Markovian arrival process (He, 2010; 2014) is given. A counting (observed) process, $\{N_t\}_{t\geq 0}$, determines the number of arrivals (earthquake occurrences) in a time interval $[0, t)$, driven by an irreducible

Figure 2: *Temporal distribution of events (left axis) and cumulative number of events (right axis) with $M \geq 6.5$ in Greece and its surrounding area, since 1845. Red pentagons are shown characteristic earthquakes of the catalog.*



finite Markov process, $\{J_t\}_{t\geq 0}$, where $E = \{1, \ldots, m\}, m \in \mathbb{N}^+$ stands for the state space. The evolution of the underlying (unobservable states) process J_t defines the seismic activity at each time $t \geq 0$; that is, its coordinates indicate the rate of the earthquake occurrences. The bivariate stochastic process $(\mathbf{N}, \mathbf{J}) := (N_t, J_t)_{t\in\mathbb{R}^+}$ is called Markovian Arrival Process (MAP), and it is a Markov process with an infinitesimal generator of the structure,

$$\mathbf{Q} = \begin{bmatrix} \mathbf{D}_0 & \mathbf{D}_1 & & \\ & \mathbf{D}_0 & \mathbf{D}_1 & \\ & & \ddots & \\ & & & \ddots \end{bmatrix}.$$

The elements of the $m \times m$ matrices $\mathbf{D}_0 = \{q_{ij}(0)\}_{i,j\in E}$ and $\mathbf{D}_1 = \{q_{ij}(1)\}_{i,j\in E}$ compile the parameter space of the model, namely the transition rates from state i to j, $i, j \in E$, without or with an arrival, respectively; their elements are non-negative, where

$$q_{ij}(k) = \lim_{h\to 0} \frac{P(N_{t+h}=k, J_{t+h}=j / J_t=i)}{h}$$

for all $i, j \in E$ and $k = 0, 1$, except the diagonal elements of $\mathbf{D}_0$, which are defined as

$$q_{ii}(0) = -\sum_{j\neq i} q_{ij}(0) - \sum_{j=1}^{m} q_{ij}(1).$$

The latter is assumed to be non-singular and the transition times are finite with probability equal to one. We assume a stationary version of the MAP, where the initial state vector after an arbitrary arrival $\boldsymbol{\pi}_{arr}$ is computed by the relation,

$$\boldsymbol{\pi}_{arr}^T = -\boldsymbol{\pi}^T \mathbf{D}_0 (\boldsymbol{\pi}^T \mathbf{D}_1 \mathbf{1}_m)^{-1},$$

where superscript ^{T} stands for the transpose of a vector or matrix, $\mathbf{1}_m$ stands for the unit column vector of dimension m and $\boldsymbol{\pi}$ is the stationary probability vector of the Markov process J_t.

The model is efficient in capturing changes of the temporal behavior of the earthquake occurrence times, namely revealing the variations of the seismicity rate. In more detail, the MAP allows for transitions between the states, where to each one of them corresponds an occurrence rate, q_i. The rate is the $i-th$ element of $\mathbf{D}_1\mathbf{1}_m$. The tractable structure of the process enables the physical interpretation of the hidden states of the Markov process J_t, assuming a correspondence between occurrence rates and seismicity periods.

For a brief description of the process let us assume that the underlying process is in state $i \in E$, whereas $J_t = i$ earthquakes occurrence follows a Poisson process with arrival rate $q_{ii}(1), i = 1, ..., m$. At the end of the sojourn time in state i, which is assumed to follow an exponential distribution, either a transition to another state occurs without designating an arrival, with probability $q_{ij}(0)/(-q_{ii}(0))$, or a transition to another state designating an arrival, with probability $q_{ij}(1)/(-q_{ii}(0))$, for $i \neq j$ and $j \in E$.

In the estimation procedure the EM algorithm is used to obtain the maximum likelihood estimates of the model parameters. It is an iterative algorithm which in each step computes the loglikelihood function with the help of the forward-backward equations (McDonald and Zucchini, 1997) and reach a local maximum after a predefined number of iterations or a converging criterion. The initial estimates of the parameters were chosen to match the mean value of the trace (interevent times). A special structure of the rate matrices is followed (Horváth and Okamura, 2013), where transitions among the hidden states are allowed only when arrivals are taking place, implying that $q_{ij}(0) = 0$, for $i \neq j$.

2.2.2 Model selection

For the determination of the best number of hidden states visited in the time interval under study, we will apply Akaike Information Criterion (AIC) (Akaike, 1974)

$$AIC = -2p + 2logL, \tag{1}$$

and the Bayesian Information Criterion (BIC) (Schwarz, 1978)

$$BIC = -p \cdot logn + 2logL, \tag{2}$$

where p is the number of parameters, $logL$ is the loglikelihood function and n the number of observations.

2.2.3 Evaluation measures

First, the goodness-of-fit (gof) to assess the adjustment of the observed interevent time series to the proposed models is provided, through the Kolmogorov-Smirnov (K-S) (Chakravarty et al., 1967) and Anderson-Darling (A-D) (Stephens, 1974) tests. We denote by τ_n the interarrival time between the $(n-1)th$ and the nth earthquake. We recall that each interarrival time, τ_n, is assumed to follow a PH distribution with representation $(\boldsymbol{\pi}_{arr}, \mathbf{D}_0)$. Hence, the hitting time distribution for an earthquake occurrence assuming that at time $t = 0$ an event occurred, is expressed by

$$F(x) = 1 - \boldsymbol{\pi}_{arr}^T e^{\mathbf{D}_0 x} \mathbf{1}_m, \ x \geq 0. \tag{3}$$

In the sequel, the distance between the empirical distribution and the fitted ones is evaluated, via the l_2 distance and the Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951). If we consider two continuous probability distributions P, Q with the same support S, then the two distance measures can be defined as

$$d_{l_2}(P,Q) = \|(P-Q)\|_2 = \sqrt{\int_S (P-Q)^2 d\tau} \tag{4}$$

$$d_{KL}(P,Q) = \int_S P \cdot log(\frac{P}{Q}) d\tau \tag{5}$$

Another useful way to assess the goodness-of-fit of the MAP is through residual analysis (Ogata, 1988). According to this technique, if we consider the transformation $\tau = \Lambda(s) = \int_0^s \lambda(t)dt$, where $\lambda(t)$ is the intensity function of a point process, then the rescaled sample τ_i is a Poisson process with unit rate. Hence, when the fitted model achieves a good approximation of the data, (only) a small declination from the bisector can be allowed.

2.2.4 Local decoding of the hidden states

The most likely path of the underlying process J_t that have generated the sequence of observations is estimated by the forward-backward equations, used in the implementation of EM algorithm. According to McDonald and Zucchini (1997) these probabilities can be evaluated by

$$P(J_{t_k} = i/H_T) = \frac{P(T_1 = \tau_1, ..., T_n = \tau_n, J_{t_k} = i)}{P(T_1 = \tau_1, ..., T_n = \tau_n)} = \frac{\boldsymbol{a}[k]_i \cdot \boldsymbol{b}[k+1]_i}{\sum_{j=1}^m \boldsymbol{a}[k]_j \cdot \boldsymbol{b}[k+1]_j} \tag{6}$$

where $H_T = \{\tau_1, ..., \tau_n\}$ is the complete history of the process, n is the number of observations and the forward-backward vectors can be expressed by

$$\boldsymbol{a}[0] = \boldsymbol{\pi}_{arr}^T, \quad \boldsymbol{a}[k] = \boldsymbol{a}[k-1]e^{\mathbf{D}_0\tau_k}\mathbf{D}_1$$
$$\boldsymbol{b}[n] = \mathbf{1}_m, \quad \boldsymbol{b}[k] = e^{\mathbf{D}_0\tau_k}\mathbf{D}_1\boldsymbol{b}[k-1].$$

They describe the evolution of the process from the first event up to the occurrence time t_k and from the latter to the last event, respectively. This procedure is often called local decoding, and the estimated sequence of the hidden states is the local decoding path.

The estimated intensity function can be easily obtained by

$$\lambda_{t_k} = \sum_{i=1}^{m} q_i \cdot \frac{\boldsymbol{a}[k]_i \cdot \boldsymbol{b}[k+1]_i}{\sum_{j=1}^{m} \boldsymbol{a}[k]_j \cdot \boldsymbol{b}[k+1]_j}, \tag{7}$$

which denotes the expected occurrence rate at time t_k, conditional on the complete history of the observations, H_T.

3. MAIN RESULTS

3.1 Optimal model

We fitted models from two to five states, since we believe that there is no need for more states in order to capture the temporal fluctuations of the earthquake occurrences with $M \geq 6.5$. For the implementation of the EM algorithm, BuTools program packages (Bodrog et al., 2014) were used in the MATLAB environment. We computed the AIC and BIC values through equations (1) and (2) and we chose the models with the lowest values (Table 1.).

Table 1. The log likelihood, *AIC* and *BIC* values of the fitted models, where the number of states varies from 2 to 5.

# of S	*LogL*	AIC	BIC
2	-156	325	341
3	-155	334	367
4	-137	315	370
5	-133	326	409

AIC suggests the four state model (AIC=315), however BIC indicates that two states (BIC=341) are sufficient for describing the temporal behavior of the data. We can see that BIC values are highly affected by the data shortage, since it takes into account the number of observations. In the sequel, we evaluate the performance of both models, in order to be more accurate about the choice of the optimal one.

3.2 Evaluation of the model

The goodness-of-fit of the proposed models to the empirical data is computed, and a comparison among the most popular theoretical distributions (11 in total: Exponential, Extreme value, Gamma, Generalized Pareto, Inverse Gaussian,

Logistic, Lognormal, Nakagami (Nakagami, 1960), Normal, Rayleigh, Weibull) is derived. For all the tests a significance level $a = 0.05$ is assumed. As a first step, in Table 2. we present the best four approximated distributions to the data (Exponential, Gamma, Weibull and Nakagami) according to the AIC, BIC and the A-D p-values.

Table 2. A-D test statistics with the corresponding p-values, and the AIC and BIC values for the best four distributions.

Distribution	$A-D$		AIC	BIC
	$cv = 2.492$	$p\ value$		
Exponential	4.04	0.008	326	329
Gamma	1.04	0.336	298	304
Weibull	1.12	0.299	303	309
Nakagami	1.38	0.207	301	307

Both Weibull and Gamma distributions are well adapted, however the values of AIC $(303, 298)$ and BIC $(309, 304)$, indicate a better approximation to data for Gamma compared to Weibull. In addition, high $p-$values ($p_{gam} = 0.336$) of the former are observed, strengthening the previous argument.

In the sequel, the hitting time distribution of the MAP with two and four states is evaluated according to equation (3), and the K-S gof test is applied. We further computed the l_2 and KL measures through equations (4) and (5) and the results are presented in Table 3. In addition, Gamma distribution was also evaluated and compared with the proposed models.

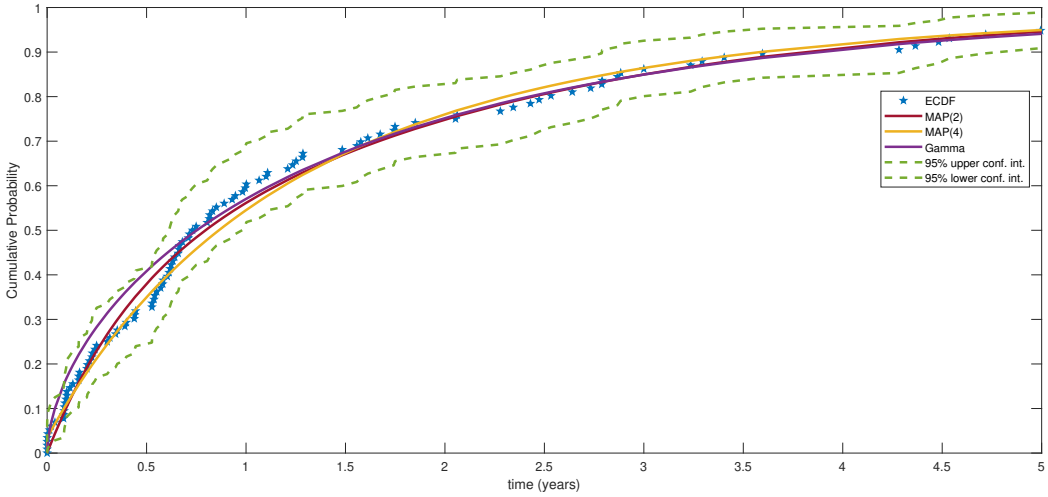
Table 3. K-S test statistics with the corresponding p-values, and the l_2 and KL values for the MAP with two and four states, as well for the Gamma distribution.

Model	$K-S$		l_2	KL
	$cv = 0.1135$	$p\ value$		
$MAP(2)$	0.0680	0.62	0.0006	0.1175
$MAP(4)$	0.0061	0.75	0.0007	0.1271
$Gamma$	0.0950	0.22	0.0959	0.1524

The gof tests for both models do not allow us to reject the null hypothesis ($ksstat_{map(2)} = 0.0668$ and $ksstat_{map(4)} = 1.0996$). Furthermore, we assume that both MAPs outperform Gamma distribution in terms of the K-S p-values, and according to the l_2, KL distance measures.

Figure 3 shows the hitting time distribution for earthquake occurrences, under the condition that an earthquake occurred at $t = 0$, for both the empirical data and the proposed models. The cumulative distribution function (cdf) of Gamma distribution was also cited, in order to illustrate the outperformance of the models. It can be seen that for almost a time period of one year, MAP(2) and MAP(4) adapt better to data compared with Gamma distribution, especially for

Figure 3: *Cumulative distributions of the data, fitted MAP(2), MAP(4) and Gamma distribution along with the 95% confidence zone.*



$t < 0.7$ years. The difference between the cdfs of the two models is not significant, however, for the rest of the parer MAP(2) is considered as the optimal model. The study period is quite short compared to the duration of the physical process of earthquake generation, hence this makes us reluctant over the acceptance of a model with many states.

For MAP(2) the transformed times τ_i are calculated, and the cumulative number of the residuals versus the transformed times is plotted (Figure 4). We conclude that the declination from the bisector $y = x$ is acceptable.

3.3 Intensity function and sequence of hidden states

The decoded path of the underlying process J_t and the estimated intensity function are evaluated by equations (6) and (7), respectively and the results are shown in Figure 5. We observe that state 2 ($q_2 = 2.6170$) (Figure 5, top plot) corresponds mainly to large shocks and in some cases to their immediate aftershocks or preshocks, since a declustering procedure has not been applied. This is the case of the saros-marmara earthquake ($M_w = 7.4$) that occurred on 9 August, 1912 (Abraseys and Finkel, 1987) and the $M_w = 6.8$ aftershock that followed just one month later. Another characteristic example is the Chirpan and Plovdiv 14 and 18 April, 1928 events with magnitudes $M_w = 6.5$ and $M_w = 6.9$ (Karakostas et al., 2006), respectively. It is important to notice, that the model recognizes this patterns and successfully makes a transition to the state of high occurrence rate. However, there are also some longer periods of relative higher seismicity rate like the 09/07/1956-26/05/1957, which include the Amorgos $M_w = 7.5$ earth-

Figure 4: *The residual process for the fitted MAP(2) model.*

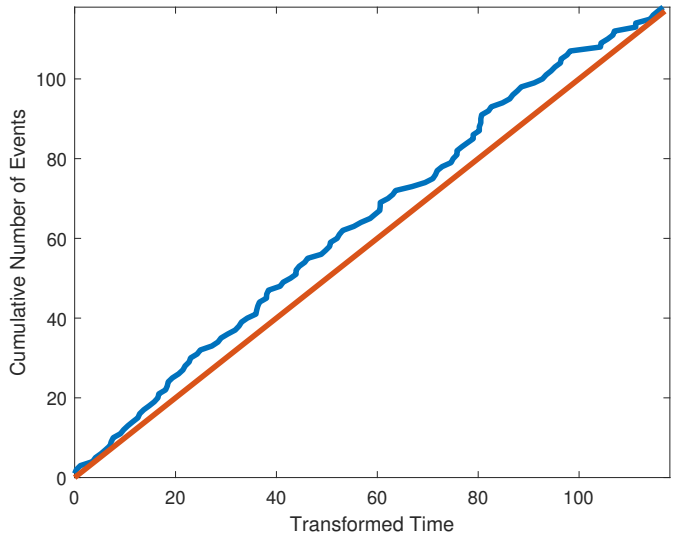
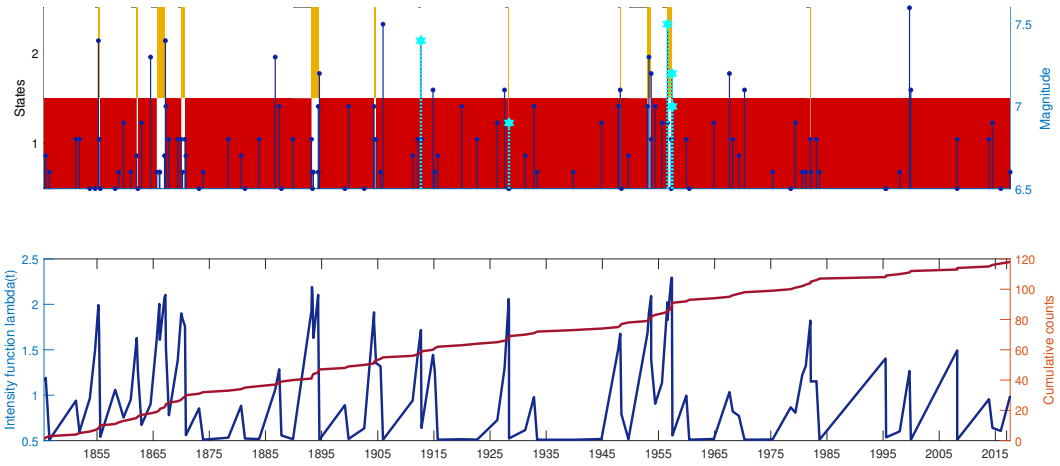


Figure 5: *Top plot: The optimal path of the hidden states of the MAP(2) given the observed sequence together with the magnitude time plot (with cyan color are shown events referenced in the text). Bottom plot: The estimated intensity function of the MAP(2) and the cumulative number of earthquakes.*



quake on 9 July, 1956, the Fethiye $M_w = 7.2$ event on 25 April, 1957 and the Abant $M_w = 7.0$ event on 26 May, 1957 (Abraseys, 1970). The model captures

this intense period with great success, and then makes a transition to state 1 ($q_1 = 0.5095$), which corresponds to a relative quiescent period with 2 earthquake occurrences in more than 7 years.

4. CONCLUSION

A Markovian arrival model was applied for investigating the temporal behavior of strong earthquakes, $M \geq 6.5$, in the broader area of Greece. The underlying Markov process J_t captures the seismicity evolution extremely well, and reveals seismicity rate changes associated in many cases with the occurrence of large events.

AIC and BIC demonstrated that two and four-state MAPs best capture the seismic activity of earthquakes with $M \geq 6.5$ in the study area, however the former was chosen, since we want to avoid an overparameterization problem. The goodness-of-fit tests verified the significance of the proposed model, which was further compared to broadly accepted theoretical distributions. The evaluated measures indicated the outperformance of the model.

The evaluation of the conditional intensity function and the optimal sequence of hidden states, showed that temporal variations characterize the frequency of strong earthquakes, $M \geq 6.5$, in the wider area of Greece. MAP with two states seems adequate in capturing patterns of the real catalog, i.e., periods of seismic quiescence alternate with seismic active ones, hence fitted models with more states do not provide additional information.

Further investigation for the existence of spatial associations between the events should be fulfilled. It is also important to apply a MAP on the declustered earthquake catalog, in order to examine the spatio-temporal behavior of the mainshocks, in the broader area of Greece. Currently, the temporal variations of large earthquakes in a global scale, is under investigation.

ΠΕΡΙΛΗΨΗ

Το μοντέλο Μαρκοβιανών Διαδικασιών Αφίξεων (ΜΔΑ) εφαρμόσθηκε σε πλήρη σεισμικό κατάλογο, με σεισμούς μεγέθους $M \geq 6.5$, στον ευρύτερο ελληνικό χώρο, με σκοπό τη μελέτη των χρονικών μεταβολών της σεισμικότητας και την εύρεση σημείων αλλαγής του ρυθμού σεισμογένεσης. Η ΜΔΑ είναι μία στοχαστική διαδικασία, που απαριθμεί το πλήθος σεισμών και μεταβαίνει μεταξύ των κρυφών καταστάσεων μίας Μαρκοβιανής Διαδικασίας, έτσι ώστε κάθε κατάσταση να αντιστοιχεί σε ένα διαφορετικό επίπεδο ρυθμού σεισμικότητας. Για τον καθορισμό του βέλτιστου αριθμού κρυφών καταστάσεων χρησιμοποιήθηκαν τα κριτήρια πληροφορίας Akaike και Bayes, ενώ η εκτίμηση των παραμέτρων του μοντέλου πραγματοποιήθηκε μέσω του αλγορίθμου ΕΜ. Η εκτίμηση των χρονικών στιγμών μετάβασης των καταστάσεων της κρυφής Μαρκοβιανής Διαδικασίας έγινε με τη μπρος-πίσω διαδικασία. Η εκτιμώμενη

υπό συνθήκη συνάρτηση έντασης αναδεικνύει εναλλαγές μεταξύ υψηλών και χαμηλών επιπέδων σεισμικής δραστηριότητας, σε μερικές περιπτώσεις σε συνδυασμό με τη γένεση κύριων σεισμών.

Acknowledgements: The financial support by the European Union and Greece (Partnership Agreement for the Development Framework 2014-2020) under the Regional Operational Programme Ionian Islands 2014-2020, for the project " Telemachus - Innovative Operational Seismic Risk Management System in the Region of Ionian Islands " is gratefully acknowledged. The software Generic Mapping Tools was used to plot the map of the study area (Wessel and Smith, 1998). Geophysics Department Contribution 914.

REFERENCES

Ambraseys, N. N. (1970). Some characteristic features of the Anatolian fault zone. *Tectonophysics* **9**, 143-165.

Ambraseys, N. N., & Finkel, C. F. (1987). The Saros-Marmara earthquake of 9 August 1912. *Earthquake engineering & structural dynamics* **15**, 189-211.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control* **19**, 716-723.

Artalejo, J. R., & Gómez-Corral, A. (2010). Markovian arrivals in stochastic modelling: a survey and some new results (invited article with discussion: Rafael Pérez-Ocón, Miklos Telek and Yiqiang Q. Zhao). *SORT-Statistics and Operations Research Transactions* **34**, 101-156.

Asmussen, S., & Koole, G. (1993). Marked point processes as limits of Markovian arrival streams. *Journal of Applied Probability* **30**, 365-372.

Baker, C., Hatzfeld, D. L., Lyon-Caen, H., Papadimitriou, E., & Rigo, A. (1997). Earthquake mechanisms of the Adriatic Sea and Western Greece: implications for the oceanic subduction-continental collision transition. *Geophysical Journal International* **131**, 559-594.

Bodrog, L., Buchholz, P., Heindl, A., Horváth, A., Horváth, G., Kolossváry, I., Mészáros A., Németh Z., Papp J., Reinecke P., Telek, M. & Vécsei M. (2014). *program packages for computations with PH, ME distributions and MAP, RAP processes,* User Manual; 2014.

Bufe, C. G., & Perkins, D. M. (2005). Evidence for a global seismic-moment release sequence. *Bulletin of the Seismological Society of America* **95**, 833-843.

Chakravarty, I. M., Roy, J. D., & Laha, R. G. (1967). *Handbook of methods of applied statistics,* New York: McGraw Hill

Cordeiro, J. D., & Kharoufeh, J. P. (2011). Batch markovian arrival processes (BMAP). *Wiley Encyclopedia of Operations Research and Management Science.* Hoboken, N.J.: John Wiley & Sons.

Daub, E. G., Ben Naim, E., Guyer, R. A., & Johnson, P. A. (2012). Are megaquakes clustered?. *Geophysical Research Letters* **39**, L06308.

Dimer De Oliveira, F. (2012). Can we trust earthquake cluster detection tests?. *Geophysical Research Letters* **39**, L17305.

He, Q. M. (2010). Construction of continuous time markovian arrival processes. *Journal of Systems Science and Systems Engineering* **19**, 351-366.

He, Q. M. (2014). *Fundamentals of matrix-analytic methods,* New York: Springer.

Horváth, G., & Okamura, H. (2013). A fast EM algorithm for fitting marked Markovian arrival processes with a new special structure. *In European Workshop on Performance Engineering,* Berlin, Heidelberg: Springer.

Karakostas, V., Papadimitriou, E., Gospodinov, D., & Ranguelov, B. (2006). Slip distribution of the 1928 Chirpan and Plovdiv main shocks and earthquake triggering. *In VIth International Conference of SGEM*, 119-127.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics* **22**, 79-86.

Lombardi, A. M., & Marzocchi, W. (2007). Evidence of clustering and nonstationarity in the time distribution of large worldwide earthquakes. *Journal of Geophysical Research: Solid Earth* **112**, B02303.

Lombardi, A. M., Cocco, M., & Marzocchi, W. (2010). On the increase of background seismicity rate during the 1997-1998 Umbria-Marche, Central Italy, sequence: apparent variation or fluid-driven triggering?. *Bulletin of the Seismological Society of America* **100**, 1138-1152.

Lucantoni, D. M., Meier-Hellstern, K. S., & Neuts, M. F. (1990). A single-server queue with server vacations and a class of non-renewal arrival processes. *Advances in Applied Probability* **22**, 676-705.

McCaffrey, R. (2008). Global frequency of magnitude 9 earthquakes. *Geology* **36**, 263-266.

MacDonald, I. L., & Zucchini, W. (1997). *Hidden Markov and other models for discrete-valued time series,* London, U.K.: Chapman & Hall.

Michael, A. J. (2011). Random variability explains apparent global clustering of large earthquakes. *Geophysical Research Letters* **38**, L21301.

Nakagami, M. (1960). The m-Distribution, a general formula of intensity of rapid fading. *Statistical Methods in Radio Wave Propagation* **47**, 3-36.

Neuts, M. F. (1979). A versatile Markovian point process. *Journal of Applied Probability* **16**, 764-779.

Ogata, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical association* **83**, 9-27.

Papadimitriou, E., Sourlas, G., & Karakostas, V. (2005). Seismicity variations in the Southern Aegean, Greece, before and after the large (M7. 7) 1956 Amorgos earthquake due to evolving stress. *Pure and applied geophysics* **162**, 783-804.

Papazachos, B. C., Karakaisis, G. F., Papazachos, C. B., & Scordilis, E. M. (2000). Earthquake triggering in the north and east Aegean plate boundaries due to the Anatolia westward motion. *Geophysical research letters* **27**, 3957-3960.

Paradisopoulou, P. M., Papadimitriou, E. E., Karakostas, V. G., Taymaz, T., Kilias, A., & Yolsal, S. (2010). Seismic hazard evaluation in western Turkey as revealed by stress transfer and time-dependent probability calculations. *Pure and applied geophysics* **167**, 1013-1048.

Papazachos, B.C. and Papazachou C.B. (2003). *The earthquakes of Greece,* Thessaloniki: Ziti Publications.

Parsons, T., & Geist, E. L. (2012). Were global $M \geq 8.3$ earthquake time intervals random between 1900 and 2011?. *Bulletin of the Seismological Society of America* **102**, 1583-1592.

Parsons, T., & Velasco, A. A. (2011). Absence of remotely triggered large earthquakes beyond the mainshock region. *Nature Geoscience* **4**, 312-316.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics* **6**, 461-464.

Stephens, M. A. (1974). EDF statistics for goodness of fit and some comparisons. *Journal of the American statistical Association* **69**, 730-737.

Stein, R. S., & Toda, S. (2013). Megacity megaquakes?Two near misses. *Science* **341**, 850-852.

Taymaz, T., Eyidoğan, H., & Jackson, J. (1991a). Source parameters of large earthquakes in the East Anatolian Fault Zone (Turkey). *Geophysical Journal International* **106**, 537-550.

Taymaz, T., Jackson, J., & McKenzie, D. (1991b). Active tectonics of the north and central Aegean Sea. *Geophysical Journal International* **106**, 433-490.

Greek Statistical Institute
Proceedings of the 31st Panhellenic Statistics Conference (2018), pp. 292−301

A Comparative Study for the Use of Periodic Regression Models for Detection of Influenza Outbreaks

E.-N. Kalligeris[1], A. Karagrigoriou[1], A. Lambrou [2], C. Parpoula[1]
[1]Lab of Statistics and Data Analysis, Department of Statistics and Actuarial–Financial Mathematics, University of the Aegean, Karlovasi, 83200 Samos, Greece
{ekalligeris, alex.karagrigoriou, parpoula.ch}@aegean.gr
[2]Hellenic Center for Disease Control and Prevention, 15123 Athens, Greece
lambrou@keelpno.gr

ABSTRACT

A rich array of statistical methodologies is available for the early detection of epidemic activity in biosurveillance systems. This alone raises the research question "Which is the most appropriate methodology to use?". It is not feasible to make detailed recommendations as to which method is "best", because this depends critically on the specific details of the application and implementation, as well as on its purpose and context. In this work, baseline periodic regression models are fitted for establishing "normal behavior", and then based on an "anomaly detection scheme", epidemic alert thresholds are produced which can be used to perform real time surveillance. Real time surveillance though, is often difficult to be fully achieved because of the seasonality involved in the time series. Indeed, whenever the correlation structure of a series depends on the season, the time series involved fails to reach stationarity with all the associated modeling consequences. In such situations, a useful class of models is that of periodic auto-regressive moving average (PARMA) models allowing parameters that depend on season. Thus, in this work, we also fit a general form as well as special cases of PARMA models, and via model selection identification and likelihood-based techniques, the optimal model is selected. In addition, several random meteorological factors associated with influenza-like illness syndrome are also incorporated into the model structure. The derived results are satisfactory in both cases examined (i.e. periodic regression and PARMA models succeeded in detecting the beginning and end of the epidemic waves). Following a model selection pathway, the periodic regression model chosen as the optimal one identified the pattern that best matches the recent influenza activity for Greece for the period $2014 - 2016$ (based on 105 weekly data for influenza-like illness) and outperformed typical forecasting models. Moreover, improved baseline modeling via PARMA models and subsequent reduction of residual errors, led to improved outbreak detection.

Keywords: Statistical Modeling, Model Selection, Periodic Regression Analysis, Mixed Periodic Autoregressive Moving Average Models, Time Series, Outbreaks.

1. INTRODUCTION

The health of the population is a valuable commodity in the center of interest of society and health services (13). Epidemiological surveillance is a dynamic scientific activity which continuously progresses and constitutes a specific sector of biosurveillance related solely to the human population. The modern definition of epidemiological surveillance includes aspects such as controlling the validity of data, analyzing data via advanced statistical methods as well as extracting safe conclusions with scientific and methodological adequacy (4). The need for timely and accurate prediction of an epidemic wave has led to the implementation of statistical routine methods to detect outbreaks in epidemiological surveillance systems in several European countries (5) and Centers for Disease Control and Prevention (CDC).

Various statistical methodologies are available for the early detection of epidemic activity in epidemiological surveillance systems. However, detailed recommendations as to which method is the "best" is not possible, since this depends critically on the specific details of the application and implementation, as well as its purpose and context. The two main goals of this paper are firstly to develop novel regression approaches (within this framework, Serfling-type periodic regression models and mixed periodic auto-regressive moving average regression (PARMA) models that work best for monitoring processes will be developed), and secondly to compare and evaluate the developed methodologies via a retrospective analysis of time series of epidemiological data. Note that for computational purposes the automated Pelat et al. software (`https://periodic.sentiweb.fr/`) has been adopted and properly adjusted. The desired results based on 105 weekly data for influenza-like illness (ILI) for Greece for the period $2014 - 2016$, are satisfactory in both cases examined (i.e. periodic regression and mixed PARMA models succeeded in detecting the beginning and end of the epidemic waves). Following a model selection pathway, the periodic regression model chosen as the optimal one (in both cases examined) identified the pattern that best matches the recent influenza activity and outperformed typical forecasting models. Moreover, improved baseline modeling via mixed PARMA models and subsequent reduction of residual errors, led to improved outbreak detection.

The rest of this paper is organized as follows. In Section 2., we discuss the motivation of this paper, present the research methodology and describe the stages and concepts employed in this work. In Section 3., we conduct an extended empirical study including model identification and evaluation. Finally, in Section 4., the obtained results are discussed and some concluding remarks are made.

2. METHODS

2.1 Motivation

During the past decade, a widespread implementation and use of several statistical modeling techniques for epidemiological surveillance purposes is observed. In the relevant areas of biostatistics and biosurveillance, we find several detailed reviews of these statistical methods in the recent literature, such as those of Sonesson and Bock in (12), Farrington and Andrews in (3), Buckeridge et al. in (1), Shmueli and Burkom in (11), Unkel et al. in (14), and Ozonoff et al. in (7).

A commonly used fully parametric regression model for outbreak detection is based on that of Serfling in (10). Serfling made use of a trigonometric function with linear trend, assuming Gaussian white noise errors, in order to model historical baselines. Pelat et al. in (9) developed an automated version of the Serfling's model with cubic trend and three trigonometric terms for prospective and retrospective surveillance.

Parpoula et al. in (8), developed Serfling-type periodic regressions models and compared their performance to typical forecasting models, concluding that a periodic regression model with quadratic trend, annual, semi-annual and quarterly periodicity, as well as a moving average model of three terms had almost similar performance (with the former one to be slightly superior). Motivated by the aforementioned results, Kalligeris et al. in (6) incorporated in one model structure trend and seasonality components of a periodic regression model as well as moving average and autoregressive terms. Additionally, they incorporated random meteorological covariates into the aforementioned model structure, and this resulted in a novel class of models called as mixed PARMA models. In this paper, we conduct a comparative study in order to examine how these periodic regression modeling techniques perform for modeling effectively influenza morbidity and evaluate them in terms of prediction performance.

2.2 Research Methodology

There are two types of analysis for surveillance times series:

a. Prospective assuming that only observations made prior to the time of the detection are available to the detection algorithm. It is useful for monitoring incoming public health data streams in an online fashion and

b. Retrospective for which the entire set of the observations is available to the detection algorithms. It is useful primarily for offline analysis of historical data.

This study focuses on epidemic detection and quantification from time series data. Four steps are necessary to be followed in such a case:

1. Determination of the training period: following Pelat et al. in (9), we conducted a retrospective analysis, making use of all available two year historical weekly influenza-like illness (ILI) data, in order to estimate the baseline level.

2. Purge of the training period: we selected to exclude the 15% highest observations from the training period (that is the default value selected by Pelat et al. in (9)) so that the baseline level is estimated from truly non epidemic data.

3. Estimation of the regression equation: a periodic regression model is fitted to the training data. Selection of the best fitting model is made possible by an exhaustive search and selection process which is relied on Analysis Of Variance (ANOVA) comparison (significance level is chosen to be 0.05) to select between nested models, and on Akaike's Criterion (AIC), or on Schwarz's Bayesian Information Criterion (BIC), to select between non-nested models.

4. Epidemic alert notification: epidemic thresholds are obtained by taking the upper 95th percentile for the prediction distribution (2), and an epidemic is declared when 2 weekly successive observations are above the estimated threshold.

For a more detailed explanation of the above 4–step procedure, the interested reader may refer to (6) and (9).

3. MODEL IDENTIFICATION AND RESULTS

In Greece, since 1999, a system of epidemiological surveillance (sentinel surveillance system) is in operation. Through the sentinel system, the evolution of the frequency of certain diseases is recorded, based on clinical diagnoses. This paper deals with one of the main syndromes under surveillance namely, influenza-like illness (ILI). In this direction, 105 weekly ILI rate data (September 29, 2014 to October 2, 2016), were analyzed in order to determine the signaled start and end weeks for the past two seasonal influenza outbreaks. In addition, optimal empirical epidemic thresholds are being established.

In the regression equation, the trend was modeled considering a polynomial of 3rd degree, and the seasonality was modeled using *sine* and *cosine* terms with period one year, six months and three months. The following equation represents the full form of the model:

$$Y(t) = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \alpha_3 t^3 + \gamma_1 \cos(\frac{2\pi t}{n}) + \delta_1 \sin(\frac{2\pi t}{n})$$

$$+\gamma_2 \cos(\frac{4\pi t}{n}) + \delta_2 \sin(\frac{4\pi t}{n}) + \gamma_3 \cos(\frac{8\pi t}{n}) + \delta_3 \sin(\frac{8\pi t}{n}) + \varepsilon(t), \quad (1)$$

where $Y(t)$ are the observed time series values (weekly ILI rate), $\varepsilon(t)$ are centered zero-mean random variables with variance σ^2, n denotes the number of observations within on year ($n = 52.179 = (4 \text{ x } 365+1)/(4 \text{ x } 7)$) based on (9)), and model coefficients are estimated by least squares regression. It is worth to be noted here, that we also examined all models which are nested within the one presented in Eq.1.

Through an exhaustive process, using ANOVA comparisons and AIC, BIC information criteria (see (8)), the best fitting model was selected to be **M23**, with quadratic trend, and annual, semi-annual and quarterly periodic terms, described as follows:

$$Y(t) = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \gamma_1 \cos(\frac{2\pi t}{n}) + \delta_1 \sin(\frac{2\pi t}{n})$$

$$+\gamma_2 \cos(\frac{4\pi t}{n}) + \delta_2 \sin(\frac{4\pi t}{n}) + \gamma_3 \cos(\frac{8\pi t}{n}) + \delta_3 \sin(\frac{8\pi t}{n}) + \varepsilon(t). \quad (2)$$

Additionally, using evaluation criteria such as the Root Mean Squared Error (RMSE), AIC, and BIC, the forecasting capability of the selected model (**M23**) was examined in contrast to typical trend or/and seasonality detection models, i.e., a linear trend (LT), a moving average of 3 terms (MA3), an exponential smoothing (SES), a Holt's model, a Brown's model, a Winter's model, and the standard CDC algorithm for flu detection (Serfling's model).

According to this comparative study, **M23** model was found to be more accurate (in terms of RMSE values) than the rest of the models under comparison. Also, based on AIC and BIC criteria, **M23** has the smaller and second smaller value, respectively and thus, it suits best to the dataset, with the MA3 being a strong competitor (smaller BIC value, and satisfactory AIC and RMSE values). The interested reader may refer to (8) for more details about the parameters used for each forecasting model considered, and all derived performance criteria values.

The above results, motivated Kalligeris et al. in (6) to incorporate into the full form of the model described in Eq. 1, Auto-regressive Moving Average (ARMA) terms. Random meteorological covariates (12 covariates related to wind speed, direction and temperature), were also included in the model structure, thus the resulting model is a mixed PARMA(2,2), described as follows:

$$y_t = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \alpha_3 t^3 + \alpha_4 t^4 + \gamma_1 cos\left(\frac{2\pi t}{n}\right) + \delta_1 sin\left(\frac{2\pi t}{n}\right)$$

$$+\gamma_2 cos\left(\frac{4\pi t}{n}\right) + \delta_2 sin\left(\frac{4\pi t}{n}\right) + \gamma_3 cos\left(\frac{8\pi t}{n}\right) + \delta_3 sin\left(\frac{8\pi t}{n}\right)$$

$$+\phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t + \lambda_1 \epsilon_{t-1} + \lambda_2 \epsilon_{t-2}$$
$$+\zeta_1 \text{minwd} + \zeta_2 \text{maxwd} + \zeta_3 \text{medianwd} + \zeta_4 \text{meanwd}$$
$$+\theta_1 \text{minwf} + \theta_2 \text{maxwf} + \theta_3 \text{medianwf} + \theta_4 \text{meanwf}$$

$$+\omega_1 \text{mintemp} + \omega_2 \text{maxtemp} + \omega_3 \text{mediantemp} + \omega_4 \text{meantemp}. \quad (3)$$

Following the same exhaustive search process (as previously described), the model chosen as optimal was the simplest one, that is a mixed model with a linear trend, 12-month seasonal periodicity, ARMA(2,1) terms and the minimum temperature as the only significant random meteorological covariate (see also Table 1), labeled as **MXM11** and defined as follows:

$$y_t = \alpha_0 + \alpha_1 t + \gamma_1 cos\left(\frac{2\pi t}{n}\right) + \delta_1 sin\left(\frac{2\pi t}{n}\right) + \phi_1 y_{t-1}$$

$$+\phi_2 y_{t-2} + \epsilon_t + \lambda_1 \epsilon_{t-1} + \omega_1 \text{mintemp}. \quad (4)$$

Table 1: *Models Selected Through the Algorithm Pathway*

	T[a]				P[b]			ARMA			LV[c]	IC[d]	
MXM[e]	t	t^2	t^3	t^4	1 y[f]	6 m[g]	3 m	AR(1)	AR(2)	MA(1)	MT[h]	AIC	BIC
MXM11	*				*			*	*	*	*	409.31	429.29
MXM12	*				*	*		*	*	*	*	406.96	431.37
MXM13	*				*	*	*	*	*	*	*	406.56	435.42
MXM21	*	*			*			*	*	*	*	410.72	432.92
MXM22	*	*			*	*		*	*	*	*	408.49	435.11
MXM23	*	*			*	*	*	*	*	*	*	408.17	439.24
MXM31	*	*	*		*			*	*	*	*	403.82	428.23
MXM32	*	*	*		*	*		*	*	*	*	402.83	431.67
MXM33	*	*	*		*	*	*	*	*	*	*	401.04	434.33
MXM41	*	*	*	*	*			*	*	*	*	405.56	432.19
MXM42	*	*	*	*	*	*		*	*	*	*	403.90	434.97
MXM43	*	*	*	*	*	*	*	*	*	*	*	402.48	437.98

[a] "T" denotes *Trend*;
[b] "P" denotes *Periodicity*;
[c] "LV" denotes *Latent Variable*;
[d] "IC" denotes *Information Criterion*;
[e] "MXM" denotes *Mixed Model*;
[f] "y" denotes *Year*;
[g] "m" denotes *Months*;
[h] "MT" denotes *Minimum Temperature*.

Fig. 1 represents the behaviour of minimum temperature, which was calculated via averaging the minimum daily measurements of each week for a total of 105 weeks for $2014-2016$ in Greece (daily data provided by the Hellenic National Meteorological Service).

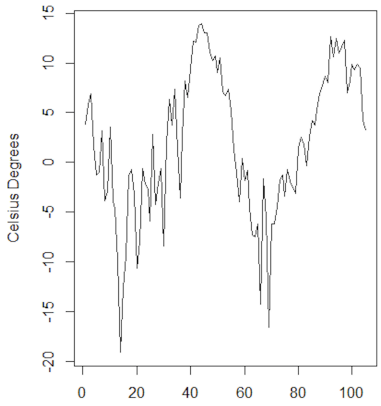


Figure 1: *Weekly minimum temperature measurements for Greece, week40/2014-week39/2016.*

We then conducted a further comparative study in order to evaluate the three models identified to perform well for modeling influenza morbidity, i.e., MXM11, M23 and MA3. Table 2 shows that MXM11 clearly outperforms M23 and MA3, since it succeeds a significant improvement of RMSE, AIC and BIC values.

Table 2: *Comparative Performance of Forecasting Models*

Model	RMSE	AIC	BIC
MXM11	4.30	409.31	429.29
M23	12.02	547.89	572.77
MA3	13.51	554.72	565.33

Table 3: *Retrospective Evaluation of the Excess Influenza Morbidity, Greece 2014-16*

SW [a]	EW [a]	Cases	Expected cases	Excess cases	Excess percentage
201501	201512	1151	891	260	29%
201605	201608	316	225	91	40%

[a] SW and EW denote the signaled start and end weeks for epidemics, respectively.

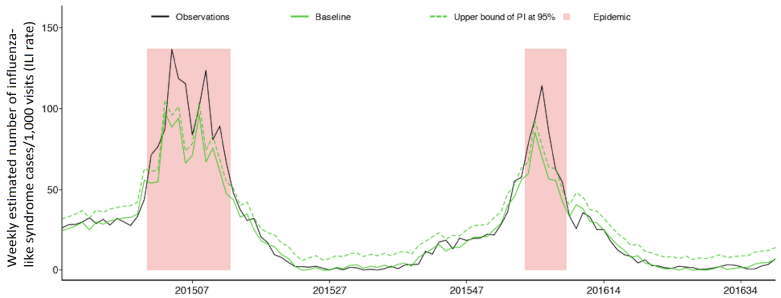


Figure 2: *Detected epidemics for influenza-like syndrome cases per 1,000 visits (ILI rate) in Greece* $2014 - 2016$.

Fig. 2 illustrates the plots of the observed ILI rate, the predicted baseline level and the estimated epidemic threshold for each time period under study. The epidemics detected by the selected model (***MXM*11**) appear in light red polygons. Table 3, presents the dates and the results of the retrospective evaluation of the excess influenza morbidity in Greece for $2014-2016$ along with excess percentages, using the $MXM11$ periodic regression mixed model. The excess morbidity is defined as the cumulative difference between observations and baseline over the entire epidemic period. Excess percentages were calculated as the observed size divided by the sum of expected values throughout each epidemic.

4. CONCLUSION

Conclusively, in this study, we conducted a retrospective analysis of epidemiological time series data (week40/2014 to week39/2016) for Greece. We developed an alternative approach in order to model seasonality of influenza morbidity, based on a periodic regression model which incorporates autoregressive and moving average components into Serfling's model, including also random meteorological covariates (known to be related to influenza activity) with the ultimate aim of the early and accurate outbreak detection. The model finally selected (via an exhaustive search process) as the optimal one succeeded in estimating accurately the influenza-like syndrome morbidity burden in Greece for the period $2014-2016$ as well as the duration of the epidemic waves. Improved baseline modeling via mixed PARMA models' approach, led to subsequent reduction of residual errors and improved outbreak detection. Within this framework, the present work provided general recommendations to serve critical needs of Public Health for the very early and accurate detection of epidemic activity.

ΠΕΡΙΛΗΨΗ

Μια πλούσια ποικιλία στατιστικών μεθοδολογιών υφίσταται για την έγκαιρη ανίχνευση της επιδημικής δραστηριότητας στα συστήματα βιοεπιτήρησης. Έτσι τίθεται το ερευνητικό ερώτημα 'Ποια είναι η πλέον ενδεδειγμένη μεθοδολογία;'. Δεν είναι εφικτό να κάνουμε λεπτομερείς συστάσεις ως προς την 'καλύτερη' μέθοδο καθώς αυτό εξαρτάται κυρίως από τις λεπτομέρειες της εκάστοτε εφαρμογής, καθώς και από το σκοπό και το πλαίσιό της. Σε αυτή τη μελέτη, βασικά μοντέλα περιοδικής παλινδρόμησης προσαρμόζονται για την καθιέρωση της κανονικής συμπεριφοράς και στη συνέχεια βασιζόμενοι σε ένα 'σχήμα ανίχνευσης ανωμαλιών', παράγονται όρια επιδημικής προειδοποίησης τα οποία μπορούν να χρησιμοποιηθούν για την επιτήρηση σε πραγματικό χρόνο. Ωστόσο, η παρακολούθηση σε πραγματικό χρόνο είναι συχνά δύσκολο να επιτευχθεί πλήρως λόγω της εποχικότητας που εμπλέκεται στις χρονολογικές σειρές. Πράγματι, κάθε φορά που η δομή συσχέτισης μιας σειράς εξαρτάται από την εποχή, οι χρονολογικές σειρές αποτυγχάνουν να φτάσουν σε στάσιμη κατάσταση με ο,τι αυτό συνεπάγεται ως προς τη μοντελοποίησή τους. Σε τέτοιες καταστάσεις, μια χρήσιμη κλάση μοντέλων είναι αυτή των περιοδικών μοντέλων αυτοπαλίνδρομου κινητού μέσου (PARMA) που επιτρέπουν παραμέτρους που εξαρτώνται από την εποχή. Έτσι, σε αυτή την εργασία, προσαρμόζουμε μια γενική μορφή καθώς και ειδικές περιπτώσεις των PARMA

μοντέλων, και μέσω της μεθόδου επιλογής μοντέλων και τεχνικών βασιζόμενων στην πιθανοφάνεια, επιλέγεται το βέλτιστο μοντέλο. Επιπλέον, τυχαίοι μετεωρολογικοί παράγοντες που σχετίζονται με το σύνδρομο της γριπώδους συνδρομής ενσωματώνονται επίσης στη δομή του μοντέλου. Τα προκύπτοντα αποτελέσματα είναι ικανοποιητικά και στις δύο περιπτώσεις που εξετάστηκαν (δηλαδή περιοδική παλινδρόμηση και τα μοντέλα PARMA κατάφεραν να ανιχνεύσουν την αρχή και το τέλος των επιδημικών κυμάτων). Ακολουθώντας ένα 'μονοπάτι' επιλογής μοντέλου, το μοντέλο περιοδικής παλινδρόμησης που επιλέχθηκε ως το βέλτιστο, αναγνώρισε με βάση 105 εβδομαδιαία δεδομένα γρίπης για την περίοδο 2014-2016 το πρότυπο που ταιριάζει καλύτερα στην πρόσφατη δραστηριότητα της γρίπης για την Ελλάδα και υπερέχει τυπικών προβλεπτικών μεθόδων. Επιπλέον, η βελτιωμένη μοντελοποίηση της γραμμής βάσης (baseline) μέσω των μοντέλων PARMA και η επακόλουθη μείωση των καταλοίπων οδήγησαν στη βελτίωση της ανίχνευσης των επιδημικών εξάρσεων.

Acknowledgments: The authors would like to thank the Hellenic National Meteorological Service (HNMS) for providing the real meteorological data as well as the Department of Epidemiological Surveillance and Intervention of the Hellenic Center for Disease Control and Prevention (HCDCP) for providing the influenza-like illness (ILI) rate data, collected weekly through the sentinel surveillance system. Finally, the authors wish to express their appreciation to an anonymous referee whose comments and suggestions greatly improved the quality and presentation of the manuscript.

REFERENCES

[**1**] Buckeridge, D.L., Burkom, H.S., Campell, M., Hogan, W.R. and Moore, A. (2005). Algorithms for rapid outbreak detection: a research synthesis. *Journal of Biomedical Informatics* **38**, 99–113.

[**2**] Costagliola, D., Flahault, A., Galinec, D., Garnerin, P., Menares, J. and Valleron, A.-J. (1991). A routine tool for detection and assessment of epidemics of influenza-like syndromes in France. *American Journal of Public Health* **81**, 97–99.

[**3**] Farrington, C.P. and Andrews, N. (2004). Outbreak detection: Application to infectious disease surveillance. In Monitoring the Health of Populations: Statistical Principles & Methods for Public Health Surveillance, eds. Brookmeyer, R., and Stroup, D.F., *Oxford University Press: Oxford*, 203–231.

[**4**] Fleming, D.M. and Rotar-Pavlic, D. (2002). Information from primary care: its importance and value. A comparison of information from Slovenia and England and Wales, viewed from the "Health 21" perspective. *European Journal of Public Health* **12**, 249–253.

[**5**] Hulth, A., Andrews, N., Ethelberg, S., Dreesman, J., Faensen, D., van Pelt, W. and Schnitzler, J. (2010). Practical usage of computer-supported outbreak detection in five European countries. *Eurosuveillance* **15**(36), 1–6.

[**6**] Kalligeris, E.-N., Parpoula, C. and Karagrigoriou, A. (2018). On Mixed PARMA Modeling of Epidemiological Time Series Data. *Communications in Statistics-Case Studies, Data Analysis and Applications* (submitted).

[**7**] Ozonoff, A., Sukpraprut, S. and Sebastiani, P. (2006). Modeling seasonality of influenza with Hidden Markov Models. Proceedings of the American Statistical Association, Section on Statistics in Defense and National Security.

[**8**] Parpoula, C., Karagrigoriou, A. and Lambrou, A. (2017). Epidemic intelligence statistical modeling for biosurveillance. *J. Blömer et al. (Eds.): MACIS 2017*, **LNCS 10693**, 1–15.

[**9**] Pelat, C., Boëlle, P.−Y., Cowling, B.J., Carrat, F., Flahault, A., Ansart, S. and Valleron, A.−J. (2007). Online detection and quantification of epidemics. *BMC Medical Informatics and Decision Making* **5**, 29.

[**10**] Serfling, R. (1963). Methods for current statistical analysis of excess pneumonia-influenza deaths. *Public Health Reports* **78**, 494–506.

[**11**] Shmueli, G. and Burkom, H. (2010). Statistical challenges facing early outbreak detection in biosurveillance. *Technometrics* **52**, 39–51.

[**12**] Sonesson, C. and Bock, D. (2003). A review and discussion of prospective statistical surveillance in public health. *Journal of the Royal Statistical Society Series A* **166**, 5–21.

[**13**] Starfield, B., Shi, L. and Mackinko, J. (2005). Contribution of primary care to health systems and health. *Milbank Quarterly* **83**, 457–502.

[**14**] Unkel, S., Farrington, C.P., Garthwaite, P.H., Robertson, C. and Andrews, N. (2012). Statistical methods for the prospective detection of infectious disease outbreaks: a review. *Journal of the Royal Statistical Society: Series A* **175**, 49–82.

Greek Statistical Institute
Proceedings of the 31st Panhellenic Statistics Conference (2018), pp.302-313

TIME SERIES WITH INTERDEPENDENT LEVEL AND SECOND MOMENT: TESTING; CONSEQUENCES FOR MODELLING; APPLICATIONS (part II)

A.E.Milioniss[1], N.G.Galanopoulos[2]
[1]Bank of Greece and Department of Mathematics, University of the Aegean
amilionis@bankofgreece.gr
[2]Department of Statistics and Actuarial – Financial Mathematics, University of the Aegean
sasd18001@aegean.gr

ABSTRACT

In Part I of this work, as presented at the 30th Panhellenic Statistical Conference, a new procedure regarding the statistical testing for the existence and the identification of the character of time-varying second moment in its dependence on a non-constant mean level in time series was introduced. It was argued that the new approach has some considerable advantages over the existing ones and this argument was also backed by the empirical analysis of time series on Greek external trade. In the course of that analysis it was also observed that the TSW programme, a software specialized on time series analysis, occasionally appeared to be biased, favouring the log transformation of the data. TSW stands for TRAMO-SEATS for Windows, a Windows version of the DOS programmes TRAMO and SEATS (TRAMO is the acronym for: Time Series Regression with ARIMA Noise, Missing Observations and Outliers; SEATS is the acronym for Signal Extraction in ARIMA Time Series). Extending the empirical analysis performed in Part I, in Part II, simulated data, as well as the time series of Consumer Price Index (CPI) and Harmonized Index of Consumer Price (HICP) for Greece were used aiming at: a) providing further evidence on the TSW bias; b) focusing on the consequences of falsely transforming an already variance stationary time series. The results using the simulated data show that there is a bias in TSW, which depends on initial conditions. Furthermore, it is established that the consequences of falsely transforming a time series, which is originally variance stationary, do exist, but are less severe than the consequences of falsely not transforming an originally variance non-stationary series. This empirical evidence is supported by theoretical arguments.

Key Words: applied time series analysis, seasonal adjustment, variance non-stationary time series, detection of outliers.

1. INTRODUCTION

In Part I of this work (see Milionis and Galanopoulos, 2017) a formal statistical approach was introduced which allows the detection of non-stationary variance in time series and is suggestive of the transformation necessary to correct for it. The proposed statistical procedure is robust to the particular partition of a time series (a step necessary for the aforementioned procedure, as explained in section 3) and the possible existence of outliers.

Using time series of monthly external trade statistics (Total Imports of Goods and Total Exports of Goods excluding fuels and ships) from the Balance of Payments for Greece, the possible advantages of the application of this methodology on some crucial elements of time series modelling such as forecasting, seasonal adjustment and outlier detection were documented.

The effect of the application of this methodology on the aforementioned elements of time series modelling is examined through a comparative analysis, using both the new proposed methodology and TSW programme.

More specifically, it was shown that: (a) the widely used specialized time series software TSW appeared to be biased towards the log transformation of the original data; (b) the consequences of not properly transforming a non variance stationary time series were very serious particularly with respect to univariate modelling and outlier detection.

The aim of this work is to extend the empirical analysis performed in Part I, using simulated data as well as the time series of Consumer Price Index (CPI) and Harmonized Index of Consumer Price (HICP) for Greece aiming at: (1) providing further evidence on the TSW bias; (2) focusing on the consequences of log transforming an already variance stationary time series.

The paper is structured as follows: Details about the data set are described in section 2; in section 3 a brief reproduction of the statistical testing procedure developed in Part I is presented; in section 4 the results are presented and commented upon; section 5 concludes the paper.

2. DATA

The data set comprises of the time series of Greek CPI and HICP covering the time period from January 2003 to June 2015 and consist of one hundred and fifty (150) monthly observations (source: Hellenic Statistical Authority). The particular data were selected in conjunction with the data of Part I due to their importance in the economy of the country at that period. In addition, simulated time series data of the same length will be constructed and used.

3. STATISTICAL TESTING APPROACH

3.1 Description

In this section the statistical testing approach is briefly reviewed. For further details see Part I of this work (Milionis and Galanopoulos, 2017). At first, time series are partitioned into segments (subsamples) of equal length and for each subsample the (local) mean (LM) and the (local) standard deviation (LSD) are calculated. Local standard deviation is assumed to be functionally dependent on local mean in a non-linear fashion as follows:

$$LSD = \alpha LM^{\beta} e^{u}, \quad (1)$$

where α, β are model parameters, e is the base of natural logarithms and u the stochastic disturbance. Model parameters α, β are estimated via Ordinary Least Squares (henceforth OLS) using the corresponding log-log model. The estimated value of β ($\hat{\beta}$) provides the necessary information for the existence (or non-existence) and the type of data transformation needed to ensure variance stationarity (e.g. for the most popular transformations, namely the log-transformation and the square root one, correspond to β =1, and β =0.5, respectively).

3.2 Notation

Index (k) indicates the ascending number of a subsample in a partition;
index (j) indicates the ascending number of the particular partition, $j=1,2,\ldots,j_{max}$;
index i_j represents the maximum value of k (number of subsamples) in partition j;
N is the total length (size) of the initial time series;
n_{ij} represents the size of subsamples in partition j;

$\hat{\beta}_j$ is the estimate of the exponent β using subsamples derived from partition with ascending number j;

$\hat{u}_{jk}, \hat{\varepsilon}_j, \hat{u}^*_{jk}$ are independent of each other regression residuals;

An asterisk (*) over a symbol denotes the corresponding transformed data, or the corresponding estimate derived from the transformed data.

3.3 Equations

$i_j = (N/n_{ij})$, if (N/n_{ij}) is an integer; $n_{ij}>5$;

$i_j = \text{int}(N/n_{ij}) + 1$, if (N/n_{ij}) is not an integer, $n_{ij}>5$ and the residual of the division is ≥ 5;

$i_j = \text{int}(N/n_{ij})$, if (N/n_{ij}) is not an integer, $n_{ij}\geq 5$ and the residual of the division is <5;

$\hat{\beta}_j$ is estimated for each partition j, j=1,2,…, j_{max} via OLS from the model (First stage regression):

$$\ln(LSD_{jk}) = \ln(\alpha_j) + \hat{\beta}_j \ln(LM_{jk}) + \hat{u}_{jk}, \quad (2)$$

$\hat{\beta}$ is estimated via OLS as the constant term of the model (Second Stage regression):

$$\hat{\beta}_j = \hat{\beta} + \hat{d}j + \hat{\varepsilon}_j, \quad (3)$$

Model using the transformed data (Third stage regression):

$$\ln(LSD^*_{jk}) = \ln(\alpha^*_j) + \hat{\beta}^*_j \ln(LM^*_{jk}) + \hat{u}^*_{jk}. \quad (4)$$

3.4 Statistical Hypotheses and comments

Throughout the research three different statistical hypotheses are tested which are the following:

1) **H_a:** β_j**=0** $\forall$ **j** (or at least the majority of β_j s,).

   This hypothesis can be tested from the first stage regression (Equation 2) and is utilized to ensure that there exists a dependence of local standard deviation on local mean. Failure to reject **H_a** means that there is no such dependence, hence, no variance instability exists, and therefore, the algorithm stops.

2) **Robustness test. H_b:** d **=0**

   The dependent variable in Eq. 3 (second stage regression) is the estimate of β derived from the partition of ascending number j ($\hat{\beta}_j$), while the independent variable is the ascending number of the partition itself (j). Therefore, $\hat{d}$ is the estimate of the slope of the regression. This hypothesis states that the slope $\hat{d}$ should not be statistically significant, and non-rejection of it, means that $\hat{\beta}$ is robust to any particular partition of the series, or outliers. Additionally, non-rejection of H_b also ensures a better estimate of β by making more efficient use of information available in all partitions.

3) **Under-transformation test. H_c:** β^*_j**=0** $\forall$ **j,** where β^*_j are the corresponding β_j for the transformed data estimated from the third stage regression (Equation 4). The third hypothesis is a kind of an "under-transformation" test. If $\hat{\beta}^*_j$ are statistically significant it means that there will still be a dependence between local mean and local standard deviation in the already transformed data, consequently the chosen transformation did not succeed in removing variance instability. Therefore, H_c states

that there is no remaining dependence of local mean on local standard deviation in the transformed data. Non-rejection of this hypothesis ensures that the chosen transformation is sufficient and has adequately rendered an unconditionally stable variance.
The addition of robustness and under-transformation tests to the above methodology offers advantages over existing methods, as the latter are devoid of these features.

4. RESULTS-DISCUSSION

The possible need for a data transformation of the original time series data will be examined using the new method, as well as the TSW routine. Furthermore, each decision derived from the new methodology and the corresponding one derived from the TSW routine will be compared. Once a decision about the proper data transformation is made by the two methods, TSW will be used for both cases for further analysis on statistical forecasting, outlier detection and seasonal adjustment.
Specifically, for outlier detection the Chen and Liu (1993) approach is used. This approach distinguishes three types of outliers depending on their effect in a times series: Additive Outliers (AO); Transitory Changes (TC); Level Shifts (LS). For seasonal adjustment the Burman-Wilson algorithm (Burman, 1980) is used to decompose a time series in unobserved components [for further details see Maravall (1995), Gómez and Maravall (1996)].

4.1 Comparative analysis of the time series of "CPI" and "HICP"

4.1.1 Analysis using of the new statistical testing approach

The results for the first stage regressions for the time series "CPI" are presented in Table 1. It can be seen that in almost all subsample pairs the $\hat{\beta}_j$ estimates are statistically insignificant at 5% significance level (exceptions exist only for the partitions with subsample size 5 and 8). Hence, $\mathbf{H_a}$ is not rejected and therefore no transformation of the original data is suggested.

Table 2 presents the results of the first stage regressions for the series of HICP. From these results it is evident that $\hat{\beta}_j$ is not statistically significant in the 5% significance level. Thus, according to the new testing approach, in both cases no transformation of the original data is suggested as the series variance is (unconditionally) stationary.

4.1.2 Analysis using exclusively the TSW testing approach

The same series were reanalyzed following the standard TSW procedure. The way TSW tests, whether or not, the data need to be transformed in order to stabilize the variance is based on a variant of the so-called range-mean regression (see Gomez and Maravall, 1996). It is noted that the only alternatives with TSW is either the log-transformation, or no transformation.

Using the TSW procedure for both series, TSW suggested the logarithmic transformation of the original data for CPI as well as HICP as the output from TSW that refers to this test stated that: "*LOG-LEVEL PRETEST: SSlevels/(SSlog*Gmean(levels)^2)=1.1253168* ***LOGS ARE SELECTED***" and "*LOG-LEVEL PRETEST: SSlevels/(SSlog*Gmean(levels)^2)=1.1096478* ***LOGS ARE SELECTED***", respectively.

Table 1. *Estimates of $\hat{\beta}_j$ for the various partitions for the first stage regression (CPI)*

Subsample size (n_{ij})	5	6	8	10	12	14	16	18	20

Number of subsample pairs (i_j)	30	25	19	15	13	11	10	9	8
$\hat{\beta}_j$	1.604	0.681	1.148	0.933	0.898	0.565	0.377	-0.629	0.011
t-statistic	2.112	1.390	2.166	1.315	1.656	0.891	0.325	-0.667	0.010
p-value	0.047	0.178	0.048	0.211	0.126	0.396	0.753	0.526	0.992

Table 2. *Estimates of $\hat{\beta}_j$ for the various partitions for the first stage regression (HICP)*

Subsample size (n_{ij})	5	6	8	10	12	14	16	18	20
Number of subsample pairs (i_j)	30	25	19	15	13	11	10	9	8
$\hat{\beta}_j$	1.003	0.076	0.639	0.363	0.399	0.039	0.018	-0.957	-0.503
t-statistic	1.383	0.146	1.266	0.529	0.689	0.057	0.016	-1.010	-0.460
p-value	0.178	0.885	0.223	0.606	0.505	0.956	0.988	0.346	0.662

Therefore, TSW falsely suggests the logarithmic transformation of the data in both cases and so there is a bias in favour of the logarithmic transformation, whereas no transformation of the original data needs to be performed as shown using the new method.

4.1.3. The effect of data transformation on univariate modelling and outlier detection

The possible effect of log transforming an already variance stationary time series on the specification of the univariate ARIMA model and the detection of outliers is examined through Tables 3-6.

Table 3. *Univariate ARIMA modelling (CPI)*

ARIMA model with linearized original data (new method) [Model (1)]
ARIMA$(0,1,1)$ $(1,1,0)_{12}$ MODEL SPECIFICATION
$(1-0.221B^{12})(1-B)(1-B^{12})Y_t = (1-0.208B)\varepsilon_t$
ARIMA model with linearized transformed data (TSW) [Model (2)]
ARIMA$(0,1,1)$ $(0,1,1)_{12}$ MODEL SPECIFICATION
$(1-B)(1-B^{12})\log Y_t = (1-0.206B)(1+0.223B^{12})\varepsilon_t$

Table 4. *Outlier Detection (CPI)*

Outliers with original data (New Method)	Outliers with log-transformed data (TSW)
105 LS (9 2011), 118 TC (10 2012), 131 A0 (11 2013)	105 LS (9 2011), 118 TC (10 2012), 131 A0 (11 2013)

Table 5. *Univariate ARIMA modelling (HICP)*

ARIMA model with linearized original data (New Method) [Model (3)]
ARIMA(0,1,1) $(0,1,1)_{12}$ MODEL SPECIFICATION
$(1-B)(1-B^{12})Y_t = (1-0.188B)(1+0.203B^{12})\varepsilon_t$
ARIMA model with linearized transformed data (TSW) [Model (4)]
ARIMA(0,1,1) $(0,1,1)_{12}$ MODEL SPECIFICATION
$(1-B)(1-B^{12})\log Y_t = (1-0.213B)(1+0.175B^{12})\varepsilon_t$

Table 6. *Outlier Detection (HICP*)

Outliers with original data (new method)	Outliers with log-transformed (TSW)
105 LS (9 2011), 131 AO (11 2013)	25 AO (1 2005), 105 LS (9 2011), 131 AO (11 2013)

For the time series of CPI, Table 3 presents the results on univariate ARIMA modelling with and without the log-transformation, while the estimation details are quoted in the Appendix 1. From the results of Table 3 it is obvious that different ARIMA models are proposed for the linearized transformed data and linearized original data (see Part I for the details about "linearization" of a time series). The results on the detection of outliers are presented in Table 4. As is evident both approaches identify the same outliers in terms of both time and type.

For the series of HICP, Table 5 presents the results on univariate ARIMA modelling with and without the log-transformation (for estimation details see Appendix 1). From the results of Table 5 it is apparent that the differences in the two univariate models are of minor character, as in both cases the univariate model is the so-called "airline model" of Box and Jenkins (1976). The differences, weak as they are, are confined only to the estimated values of the parameters of the two models. The results on the detection of outliers with both the original and the log-transformed data are quoted in Table 6. As is evident, TSW falsely identifies one more outlier (25 AO) due to the scale "squizing" caused by the logarithmic transformation. These results should be contrasted to those of Part I in which it was documented that the consequences of not transforming a variance non-stationary time series were indeed severe, at least as far as univariate time series modelling is concerned (see Milionis and Galanopoulos, 2017). The above conclusions, regarding the severity of the consequences of a wrong decision about the transformation of the original values of a time series, may be backed by theoretical argumentation. Indeed, in the case of unnecessarily (over) transforming an already variance stationary series, the original and the over-transformed series are both variance stationary. In contrast, when the original series is variance non-stationary and is not transformed (as it should) it is evident that the usual univariate analysis, which legitimately can be applied strictly to variance stationary series only, is falsely applied to a variance non stationary series.

Hence, it is natural for sharp differences with the analysis of the properly transformed series to appear. Obviously, this conclusion is of much practical importance.

4.2.2 Further Analysis

The examination of the forecasting performance of the univariate models accordingly with the data transformation is assessed using the Mean Absolute Percentage Error (MAPE) statistic. Table 7 presents the results and as is evident twelve-step-ahead forecasts with no transformation (as suggested by the new method) are superior in terms of the MAPE value, as compared to the corresponding forecasts with the data log-transformed, as suggested by TSW. As regards the one-step-ahead forecasts, the forecasting performance of CPI is better with the new method in terms of the MAPE value, while the MAPE value is exactly the same for HICP.

Table 7. *Values of Mean absolute percentage error of forecasts*

		MAPE (%)	
	Forecasts	Original Data (New Method)	Log-transformed data (TSW)
CPI	Twelve step ahead	1.294	1.457
HICP	Twelve step ahead	0.848	0.865
CPI	One step ahead	0.395	0.467
HICP	One step ahead	0.452	0.452

The Mean Absolute Percentage Difference (MAPD) statistic was employed to evaluate the differences in the seasonally adjusted series produced from original data versus transformed data. The results are presented in Table 8. It is remarked that, as expected according to the previous conclusion argumentation, the values of MAPD themselves are small in the statistical sense, although with possibly higher economic importance as they refer to consumer price indices. MAPD is calculated as:

$$MAPD = \frac{1}{J}\sum_{j=1}^{N}\left(\frac{\left|\tilde{X}_j^{\ln} - \tilde{X}_j^{lev}\right|}{\tilde{X}_j^{\ln}}\cdot 100\right) \qquad (5)$$

where $\tilde{X}_j^{\ln}$ is the seasonally adjusted value using log-transformed data and $\tilde{X}_j^{lev}$ the corresponding seasonally adjusted data using the original data themselves.

4.3 Further analysis using simulated data

It should be recalled that in the normal course of an analysis of a time series, the test for the possible need to transform the original data in order to stabilize the variance precedes all other actions such as the creation of the univariate ARIMA model, seasonal adjustment etc. Therefore, it is obvious that the outcome of those actions is affected by the decision on data transformation. That lends even more importance to that decision and thus further evidence was sought to back our previous finding, i.e. that TSW suggests the log-transformation far too often due to bias. To this end, use was made of simulated data. After some experimentation, it was observed that with series following ARIMA(p,1,q) processes, the initial conditions influence (although they should not) the decision of TSW regarding the log-transformation.

In sequence, 40 time series were artificially created. The first 20 were built simulating a simple Gaussian random walk model:

$$X_t = X_{t-1} + e_t \text{ with } e_t \approx N(0,1) \text{ and initial condition } X_o = 0 . \qquad (6)$$

Table 8. *Differences in Seasonally Adjusted Series produced from original data versus transformed data*

	MAPD (%)	Minimum Percentage Error (%)	Maximum Percentage Error (%)
CPI	0.018	-0.057	0.037
HICP	0.017	-0.316	0.035

The other 20 were created by the same manner, except that X_o was set equal to 1000. All simulated data are available from the authors on request. All these time series are by construction homogeneously non-stationary so no data transformation is needed.

Applying the new testing procedure, the outcome was indeed no transformation in all 40 cases (in the simulated series with initial condition the presence of negative values was observed in every single series. Therefore, before log-transforming the series the usual practice of "shifting" their values upwards by adding a constant so that the minimum value of each series becomes slightly positive was adopted). Applying TSW, the outcome was no transformation only for the cases with $X_o = 0$. In all the cases with $X_o = 1000$ TSW, falsely, suggested the log transformation.

The results of the testing procedure for each case using both methods are presented in Appendix 2. Indeed the results of TSW are, once again, biased in favour of data transformation where no transformation is needed and in the particular case at hand this bias is clearly related to the initial condition of the simulated series. Moreover, even after the log-transformation, when the 20 log-transformed series with X_0 = 1000 were tested again by TSW, the result was biased again (suggestion to log-transform the already log-transformed series)**.** Those results are presented in Appendix 3.

5. SUMMARY AND CONCLUSIONS

In this work a new statistical testing procedure for variance non-stationary time series introduced by Milionis and Galanopoulos (2017) was further tested at first using real data. The empirical analysis was expanded using simulated data as well. The analysis of CPI-HICP data provided evidence for the TSW bias towards transforming the original data. Further, the results using simulated data were conductive to that finding. Moreover, with simulated homogeneously non-stationary processes, it was possible to identify that the bias of TSW depends on the initial conditions. It was also established that the consequences of falsely transforming a time series, which is originally variance stationary, do exist, but are less severe than the consequences of falsely not transforming an originally variance non-stationary series, a conclusion of much practical importance.

APPENDIX 1

Parameter estimation of Univariate ARIMA models

Table A.1. *Parameter estimation of model (1)*

Parameter	Value	t-statistic
θ_1	0.208	2.49
Φ_1	0.221	2.65

Table A.2. *Parameter estimation of model (2)*

Parameter	Value	t-statistic
θ_1	0.206	2.46
Θ_1	-0.223	-2.68

Table A.3. *Parameter estimation of model (3)*

Parameter	Value	t-statistic
θ_1	0.188	2.23
Θ_1	-0.203	-2.42

Table A.4. *Parameter estimation of model (4)*

Parameter	Value	t-statistic
θ_1	0.213	2.55
Θ_1	-0.175	-2.08

Appendix 2

Test results for data transformation using TSW

LOG-LEVEL PRETEST :	
1st simulated series with initial value 1000	SSlevels/(SSlog*Gmean(levels)^2)= 1.0004275 LOGS ARE SELECTED
2nd simulated series with initial value 1000	SSlevels/(SSlog*Gmean(levels)^2)= 1.0004200 LOGS ARE SELECTED
3rd simulated series with initial value 1000	SSlevels/(SSlog*Gmean(levels)^2)= 0.99937392 LOGS ARE SELECTED
4th simulated series with initial value 1000	SSlevels/(SSlog*Gmean(levels)^2)= 0.99821483 LOGS ARE SELECTED
5th simulated series with initial value 1000	SSlevels/(SSlog*Gmean(levels)^2)= 1.0029646 LOGS ARE SELECTED
6th simulated series with initial value 1000	SSlevels/(SSlog*Gmean(levels)^2)= 0.99932366 LOGS ARE SELECTED
7th simulated series with initial value 1000	SSlevels/(SSlog*Gmean(levels)^2)= 0.99998323 LOGS ARE SELECTED
8th simulated series with initial value 1000	SSlevels/(SSlog*Gmean(levels)^2)= 0.99976469 LOGS ARE SELECTED
9th simulated series with initial value 1000	SSlevels/(SSlog*Gmean(levels)^2)= 0.99899563 LOGS ARE SELECTED
10th simulated series with initial value 1000	SSlevels/(SSlog*Gmean(levels)^2)= 1.0009120 LOGS ARE SELECTED
11th simulated series with initial value 1000	SSlevels/(SSlog*Gmean(levels)^2)= 0.99940611 LOGS ARE SELECTED

12th simulated series with initial value 1000	SSlevels/(SSlog*Gmean(levels)^2)= 0.99783903 LOGS ARE SELECTED
13th simulated series with initial value 1000	SSlevels/(SSlog*Gmean(levels)^2)= 0.99956026 LOGS ARE SELECTED
14th simulated series with initial value 1000	SSlevels/(SSlog*Gmean(levels)^2)= 0.99996242 LOGS ARE SELECTED
15th simulated series with initial value 1000	SSlevels/(SSlog*Gmean(levels)^2)= 1.0009673 LOGS ARE SELECTED
16th simulated series with initial value 1000	SSlevels/(SSlog*Gmean(levels)^2)= 0.99844882 LOGS ARE SELECTED
17th simulated series with initial value 1000	SSlevels/(SSlog*Gmean(levels)^2)= 0.99780523 LOGS ARE SELECTED
18th simulated series with initial value 1000	SSlevels/(SSlog*Gmean(levels)^2)= 1.0011250 LOGS ARE SELECTED
19th simulated series with initial value 1000	SSlevels/(SSlog*Gmean(levels)^2)= 1.0010450 LOGS ARE SELECTED
20th simulated series with initial value 1000	SSlevels/(SSlog*Gmean(levels)^2)= 1.0001285 LOGS ARE SELECTED

LOG-LEVEL PRETEST :	
21st simulated series with initial value 0	LEVELS are Selected
22nd simulated series with initial value 0	LEVELS are Selected
23rd simulated series with initial value 0	LEVELS are Selected
24th simulated series with initial value 0	LEVELS are Selected
25th simulated series with initial value 0	LEVELS are Selected
26th simulated series with initial value 0	LEVELS are Selected
27th simulated series with initial value 0	LEVELS are Selected
28th simulated series with initial value 0	LEVELS are Selected
29th simulated series with initial value 0	LEVELS are Selected
30th simulated series with initial value 0	LEVELS are Selected
31st simulated series with initial value 0	LEVELS are Selected
32nd simulated series with initial value 0	LEVELS are Selected
33rd simulated series with initial value 0	LEVELS are Selected
34th simulated series with initial value 0	LEVELS are Selected
35th simulated series with initial value 0	LEVELS are Selected
36th simulated series with initial value 0	LEVELS are Selected

37th simulated series with initial value 0	LEVELS are Selected
38th simulated series with initial value 0	LEVELS are Selected
39th simulated series with initial value 0	LEVELS are Selected
40th simulated series with initial value 0	LEVELS are Selected

Appendix 3

Test results for data transformation using TSW and log-transformed series

LOG-LEVEL PRETEST :	
1st simulated series in log-transformed data with initial value 1000	SSlevels/(SSlog*Gmean(levels)^2)= 1.0000550 LOGS ARE SELECTED
2nd simulated series in log-transformed data with initial value 1000	SSlevels/(SSlog*Gmean(levels)^2)= 1.0000529 LOGS ARE SELECTED
3rd simulated series in log-transformed data with initial value 1000	SSlevels/(SSlog*Gmean(levels)^2)= 0.99989907 LOGS ARE SELECTED
4th simulated series in log-transformed data with initial value 1000	SSlevels/(SSlog*Gmean(levels)^2)= 0.99973569 LOGS ARE SELECTED
5th simulated series in log-transformed data with initial value 1000	SSlevels/(SSlog*Gmean(levels)^2)= 1.0003951 LOGS ARE SELECTED
6th simulated series in log-transformed data with initial value 1000	SSlevels/(SSlog*Gmean(levels)^2)= 0.99989544 LOGS ARE SELECTED
7th simulated series in log-transformed data with initial value 1000	SSlevels/(SSlog*Gmean(levels)^2)= 0.99999495 LOGS ARE SELECTED
8th simulated series in log-transformed data with initial value 1000	SSlevels/(SSlog*Gmean(levels)^2)= 0.99996343 LOGS ARE SELECTED
9th simulated series in log-transformed data with initial value 1000	SSlevels/(SSlog*Gmean(levels)^2)= 0.99984226 LOGS ARE SELECTED
10th simulated series in log-transformed data with initial value 1000	SSlevels/(SSlog*Gmean(levels)^2)= 1.0001277 LOGS ARE SELECTED
11th simulated series in log-transformed data with initial value 1000	SSlevels/(SSlog*Gmean(levels)^2)= 0.99990892 LOGS ARE SELECTED
12th simulated series in log-transformed data with initial value 1000	SSlevels/(SSlog*Gmean(levels)^2)= 0.99967945 LOGS ARE SELECTED
13th simulated series in log-transformed data with initial value 1000	SSlevels/(SSlog*Gmean(levels)^2)= 0.99993525 LOGS ARE SELECTED
14th simulated series in log-transformed data with initial value 1000	SSlevels/(SSlog*Gmean(levels)^2)= 0.99999001 LOGS ARE SELECTED
15th simulated series in log-transformed data with initial value 1000	SSlevels/(SSlog*Gmean(levels)^2)= 1.0001311 LOGS ARE SELECTED
16th simulated series in log-transformed data with initial value 1000	SSlevels/(SSlog*Gmean(levels)^2)= 0.99976445 LOGS ARE SELECTED
17th simulated series in log-transformed data with initial value 1000	SSlevels/(SSlog*Gmean(levels)^2)= 0.99966162 LOGS ARE SELECTED
18th simulated series in log-transformed data with initial value 1000	SSlevels/(SSlog*Gmean(levels)^2)= 1.0001710 LOGS ARE SELECTED
19th simulated series in log-transformed data with initial value 1000	SSlevels/(SSlog*Gmean(levels)^2)= 1.0001444 LOGS ARE SELECTED
20th simulated series in log-transformed data with initial value 1000	SSlevels/(SSlog*Gmean(levels)^2)= 1.0000163 LOGS ARE SELECTED

ΠΕΡΙΛΗΨΗ

Στο πρώτο μέρος της εργασίας αυτής, που παρουσιάστηκε στο 30ο Πανελλήνιο Στατιστικό Συνέδριο, προτάθηκε μια νέα διαδικασία που αφορά τον στατιστικό έλεγχο για την ύπαρξη και τον προσδιορισμό του χαρακτήρα της χρονικά μεταβαλλόμενης δεύτερης ροπής όταν αυτή εξαρτάται από ένα μη σταθερό μέσο επίπεδο. Αποδείχθηκε ότι η νέα διαδικασία έχει σημαντικά πλεονεκτήματα έναντι των υφιστάμενων και αυτό το επιχείρημα υποστηρίχθηκε και από την εμπειρική ανάλυση των χρονοσειρών που αφορούν το Ελληνικό εξωτερικό εμπόριο. Στα πλαίσια της ως άνω έρευνας, παρατήθηκε ότι το πρόγραμμα TSW, το οποίο είναι ένα λογισμικό εξειδικευμένο στην ανάλυση χρονοσειρών, είναι μεροληπτικό υπό την έννοια της υπερβολικής απόρριψης της μηδενικής υπόθεσης ότι μία χρονοσειρά δεν εμφανίζει μη στασιμότητα στη δεύτερη ροπή. Επεκτείνοντας την εμπειρική ανάλυση που εφαρμόστηκε στο πρώτο μέρος της εργασίας, στο παρόν δεύτερο μέρος, χρησιμοποιήθηκαν προσομοιωμένα δεδομένα καθώς και οι χρονοσειρές του Δείκτη Τιμών Καταναλωτή (ΔΤΚ) και του Εναρμονισμένου Δείκτη Τιμών Καταναλωτή (ΕνΔΤΚ) για την Ελλάδα με σκοπό: α) την περαιτέρω επιβεβαίωση της συγκεκριμένης μεροληψίας του TSW, β) την εστίαση στις συνέπειες ενός λανθασμένου μετασχηματισμού μίας ήδη στάσιμης ως προς την διακύμανση χρονοσειράς. Τα αποτελέσματα χρησιμοποιώντας τα προσομοιωμένα δεδομένα αποδεικνύουν ότι πράγματι υπάρχει μεροληψία στο TSW, η οποία εξαρτάται από τις αρχικές συνθήκες. Επιπλέον, από την εμπειρική ανάλυση προκύπτει ότι οι συνέπειες ενός λανθασμένου μετασχηματισμού μιας χρονοσειράς που είναι εξαρχής στάσιμη ως προς την διακύμανση όντως υπάρχουν, αλλά είναι λιγότερο σοβαρές από τις συνέπειες της λανθασμένης μη μετατροπής μιας χρονοσειράς που είναι εξαρχής μη στάσιμη ως προς την διακύμανση. Τα εμπειρικά αυτά ευρήματα υποστηρίζονται και από θεωρητικά επιχειρήματα.

REFERENCES

Box G.E.P. and Jenkins, G.M. (1976). *Time Series Analysis: Forecasting and Control*. Revised Edition, San Francisco: Holden-Day.

Burman, J. P. (1980). 'Seasonal Adjustment by Signal Extraction', *Journal of the Royal Statistical Society*, **143**, 321-337.

Chen, C. and Liu, L.-M. (1993). Forecasting Time Series With Outliers, *Journal of Forecasting*, **12**, 13-35.

Gomez, V. and Maravall, A. (1996). Programmes SEATS and TRAMO: instructions for the User. *Working Paper No 9628, Bank of Spain*.

Maravall, A. (1995). *Unobserved components in economic time series,* In Handbook of Applied Econometrics (Eds) M. H. Pesaran and M. Wickens, Blackwell, Oxford.

Milionis E.A. and Galanopoulos G.N. (2017). 'Time series with interdependent level and second moment: testing, consequences for modelling; preliminary results', *Proceedings of the 30th Panhellenic Statistics conference,* 109-123.

Greek Statistical Institute
Proceedings of the 31st Panhellenic Statistics Conference (2018), pp.314-328

SYSTEMATIC RISK AND MARKET MICROSTRUCTURE IN A MARKET UNDER CONDITIONS OF ECONOMIC CRISIS. THE CASE OF THE ATHENS STOCK EXCHANGE.

Milionis E. Alexandros[1], Karagiota V. Elisavet[2]
1 Bank of Greece and Department of Mathematics, University of the Aegean
amilionis@bankofgreece.gr
2 Department of Mathematics, University of the Aegean
sasm15004@sas.aegean.gr

ABSTRACT

The purpose of this work is to examine the estimation of systematic risk (beta coefficients) in its relation to some market microstructure phenomena related to friction in the trading process, in a capital market under conditions of economic crisis. The data are daily closing prices of stocks traded at the Athens Stock Exchange (ASE) covering the period 25/09/2012 – 22/09/2017. That was indeed a turbulent period for ASE owing to the acute economic crisis over Greece for all that period. It is therefore of much interest, amongst others, to examine, whether or not, some stylized facts relating to the effect of friction in the trading process on the estimation of betas may be reconfirmed. More specifically, friction in the trading process causes beta estimates not to be invariant but to change systematically with respect to the differencing interval over which asset returns are calculated ("intervalling effect"). It is also known that a different starting observation (here day) may affect beta estimates for the same differencing interval ("starting observation effect"), but the importance of the phenomenon yet has not been sufficiently assessed quantitatively. In relation to the above, this work provides evidence for the following: (a) The starting observation effect is intense and fortified further as the differencing interval is lengthened; (b) The intervalling effect is extremely weak, if any, although it was found to be strong for the same market at previous time periods; (c) low cap stocks expose clearly stronger starting observation effect as compared to large cap stocks; (d) The starting observation effect overwhelmingly dominates over the intervalling effect and it is the determining factor of changes in betas for the same stock; (e) The importance of the starting observation effect has not been properly assessed in the published research thus far, as it is either ignored completely, or smoothed away by taking average betas for the same differencing interval but different starting observation.

Key Words: systematic risk, beta coefficient, market microstructure, intervalling effect, different starting observation effect, friction in the trading process.

1. INTRODUCTION

One of the most important issues in finance is the relation between risk and return. The most celebrated way to express the relation between them is through the so-called Capital Asset Pricing Model (CAPM), according to which (Sharpe, 1964; Lintner, 1965)

$$E(R_j) = R_F + \beta_j[E(R_m) - R_F] \quad (1)$$

where, E represents the mathematical expectation operator, R_j is the return on the security j, R_F is the risk-free rate of return, R_m is the return on the market index and β_j is the beta coefficient for security j.

Beta is a measure of the risk arising from exposure to general market movements, the so called systematic risk. Large changes to macroeconomic variables such as interest rates, inflation, foreign exchange, wars, natural disasters, broad changes to government policies and general events that cannot be planned for, or avoided, are changes that impact the broader market and cannot be avoided or reduced through diversification. The beta coefficient of an asset j is formally defined as:

$$\beta_j = \frac{Cov(R_j, R_m)}{Var(R_m)}$$

Betas may be estimated from the CAPM model in Equation (1). The most direct way though of predicting the value of betas is estimates using historical data. These estimates result with the Ordinary Least Square (OLS) method using the so-called Market Model (MM). The Market Model (MM) or Single Index Model (SIM) is a simple asset pricing model. Mathematically the SIM is expressed as (Elton et. Al, 2007):

$$R_j - R_F = a_j + \beta_j(R_m - R_F) + u_j$$

where $u_j$represents the stochastic disturbance.

For realistic changes in the value of the risk free rate (R_F), there is very little difference in the estimations of beta, typically in the third decimal place. So, the MM is expressed as follows:

$$R_j = a_j + \beta_j R_m + u_j \quad (2)$$

Therefore, practically betas are obtained by regressing R_j on R_m.

The basic difference between CAPM model and MM model is that the first is formulated in terms of expectations, which are not observable, while the second one is based on observations. Assuming that expected returns are on average equal to the observed ones, observed returns may be used for convenience, as in Equation (2), for estimating betas.

It must be stressed at this point that there is a voluminous literature on the ability of CAPM to explain the observed returns and there are serious specification issues about the explanatory power of the single – factor model to explain expected returns. Therefore multifactor, and/or conditional versions of the CAPM have been proposed by several researchers (see amongst others Fama and French, 1992, 1993, 1996, Milionis 2011). However, for several reasons (Jaganathan and Wang , 1996; Elton et al., 2007; Milionis, 2010; Milionis and Patsouri, 2011) the beta coefficient (the historical measure of the systematic risk which is used in the CAPM) is still the most widely used measure of systematic risk by analysts and financial managers.

In a frictionless market and under the usual MM assumptions, beta estimates should be invariant with the length of the time (differencing) interval over which market and security returns are calculated. However, a plethora of empirical evidence (Corhay,1991; Cohen et al.,1983a; Scholes and Williams,1977; Dimson,1979) have established that MM beta estimates vary systematically with the length of the differencing interval. This phenomenon is known as the "intervalling effect".

Another issue related to the differencing interval is the "starting observation effect". As such is meant the fact that to a given security and a certain differencing interval L correspond L different estimates of beta, as the calculation of returns may start using any of the L observations included in a differencing interval of length L. Thus, to calculate returns the selection of observations for a given differencing interval L is a repeating procedure performed L times. Each time the starting observation is shifted by one (observation). Therefore, for a single differencing interval of size L, L estimates of a beta may be obtained (for further details see Corhay, 1992).

The purpose of this work is to examine the measurement of systematic risk (beta coefficients) in relation to some market microstructure phenomena, such as the ones described above in conjunction with securities' capitalization, in a market under conditions of economic crisis.

The structure of the paper is as follows: section 2 provides a skeleton review of the basic theory and describes the methodological steps to be followed; section 3 describes the market and the data; section 4 presents the results and commending; section 5 concludes the paper.

2. BRIEF REVIEW OF THEORY AND METHODOLOGY

The characteristic of market microstructure responsible for the effect of the differencing interval on beta estimation is friction in the trading process and the entailed price adjustment delays (Cohen et al., 1983b). Due to friction in the trading process, true returns, which would be observed in a frictionless market, are not observable. Changes in true returns are faster than in observed returns, so in each time

period only a part of the true returns is incorporated in the observed ones. The abovementioned can be expressed quantitatively by the following model which connects the observed returns (R_j^o) with the true returns (R_j) and was proposed by Cohen et al. (1983b):

$$R_{jt}^o = \sum_{n=0}^{N} \gamma_{j,t-n,n} R_{j,t-n}$$

with $$E(\sum_{n=0}^{\infty} \gamma_{j,t,n}) = 1 \quad \forall j, t$$

where, $\gamma_{j,t-n,n}$ are random variables representing the portion of true returns of security j at time $t-n$ which is reflected in observed returns at time t.

Therefore, the MM for true and observed returns respectively is written as:

$$R_{jt} = a_j + \beta_j R_{mt} + u_{jt}$$

$$R_{jt}^o = a_j^o + \beta_j^o R_{mt}^o + u_{jt}^o$$

where β_j and β_j^o express the systematic risk for true and observed returns respectively. β_j and β_j^o are related through the following equation (Cohen et al., 1983b):

$$\beta_j^o = \beta_j(1 + 2\sum_{n=1}^{N} \beta_{m+n}^o) - \sum_{n=1}^{N}\left(\beta_{j+n}^o + \beta_{j-n}^o\right) \qquad (3)$$

where,

$$\beta_{m+n}^o = \frac{Cov(R_{m,t+n}^o, R_{m,t}^o)}{Var(R_{m,t}^o)}, \quad \beta_{j\pm n}^o = \frac{Cov(R_{j,t\pm n}^o, R_{m,t}^o)}{Var(R_{m,t}^o)}$$

From equation (3) it is obvious that betas for the observed returns depend on betas for the true returns, on the autocorrelations of stocks' and index's observed returns up to order N and on the cross-correlations among the observed returns of the security j and the market index m up to order N. Consequently, the OLS estimator of beta for observed returns is a biased estimator of the systematic risk. From equation (3) it is evident that the bias of the beta estimator will tend to decrease as the differencing interval is lengthened.

In this work at first the importance of the intervalling effect in a market under conditions of crisis will be assessed. The importance of the starting observation effect in its own right, as well as in comparison to the intervalling effect, will also be assessed. Furthermore, the potential effect of the capitalization value on the above will be evaluated. Finally, other features of beta estimates related to market microstructure and the possible influence of the extreme conditions during the sampling period will be examined.

3. DATA AND THE MARKET

3.1 Data description

Two time scales important for the empirical determination of beta coefficients are: (a) the total time interval covered by the data, which will be called the "sampling time" (T); (b) the differencing interval which is used for the calculation of returns (L), which will be called "smoothing time". From a purely statistical point of view in order to obtain accurate measurements we need a sampling time as long as possible and a smoothing time as short as possible. From the financial point of view though, the opposite is required. Indeed, taking a large T increases the probability during this time period for events such as acquisitions, mergers and other that affect the riskiness of a security, hence, the corresponding beta coefficient, to occur. Furthermore, a small smoothing time increases the bias of the estimations due to increased microstructure friction. The most common approach of most financial institutions is to use data for a five year period and monthly returns in order to estimate beta coefficients.

The data to be used in this work are daily closing prices of all stocks traded in the Main Market of the Athens Stock Exchange (ASE). The sampling time T is a five year period from 25/09/2012 to 22/09/2017. Returns are expressed as the logarithmic difference of price relatives, using the unit differencing interval. The smoothing time is varied from one to thirty days. The Athens General Index (GEN) is used as the market index. The data were collected from the website www.naftemporiki.gr. As the frequency of observations is days-1, the starting observation effect will be called "starting day" effect hereafter.

3.2 The market

ASE started trading in 1876. From May 31 2001, the ASE joined the Mature Financial Markets according to the classification of Morgan Stanley Capital International (MSCI). Until that date ASE belonged to the European Emerging Markets. In November 2013 the same institution downgraded ASE in the Emerging Markets status. ASE was also downgraded by FTSE on March 21 2016, in the category of Advanced Developing Markets.

It is remarkable that until the late 1980s ASE was a small market with low capitalization value and a thin volume of trade, while participants were almost exclusively local investors. However, the changes in the legislative and regulatory framework, during the 1990s, towards harmonization with the standards of the more developed mature financial markets have brought significant changes for ASE (for details, see for instance Milionis et al., 1998). As a consequence of these reforms the total capitalization of ASE increased from 13.6 million USD by the end 1993 to 145 million USD by the end 2005 (source: World Federation of Exchanges). Furthermore,

foreign investors owned less than 10% of the total capitalization in 1993 (approximate indirect assessment, as no exact official figure exists,) but more that 37% of the total capitalization at end of April 2005, while their share in the average daily volume of transactions at April 2005 exceed 54% (source: Central Securities Depository of ASE, Monthly Statistical Bulletin, April 2005).

The current capitalization of ASE, as compared to that of 2005, has been significantly decreased as a consequence of the economic crisis to approximately 60 billion USD, while the majority of share capital is now owned by foreign investors. Further details about the ASE and its institutional and operational evolution are given in many papers (see for instance Milionis and Papanagiotou, 2009, 2011).

The period under study is a very condensed one in terms of economic – political events. Some of the most important ones are the National Parliamentary elections, the failed attempts to elect a President of the Republic, and the closure of the banks and of the ASE in July 2015. The extreme volatility of the General Index is evidenced in Figure 1 which shows the General Index in comparison to three other indices, namely the FTSE100 of the London Stock Exchange, the DAX of the Frankfurt Stock Exchange and the DOW JONES of the New York Stock Exchange, all expressed in standardized values. Supplementary, many crucial events over the sampling time are presented in Table 1.

Figure 1. The variation of the Athens Stock General Index in comparison to FTSE100, DAX and DJIA, expressed in standardized values.

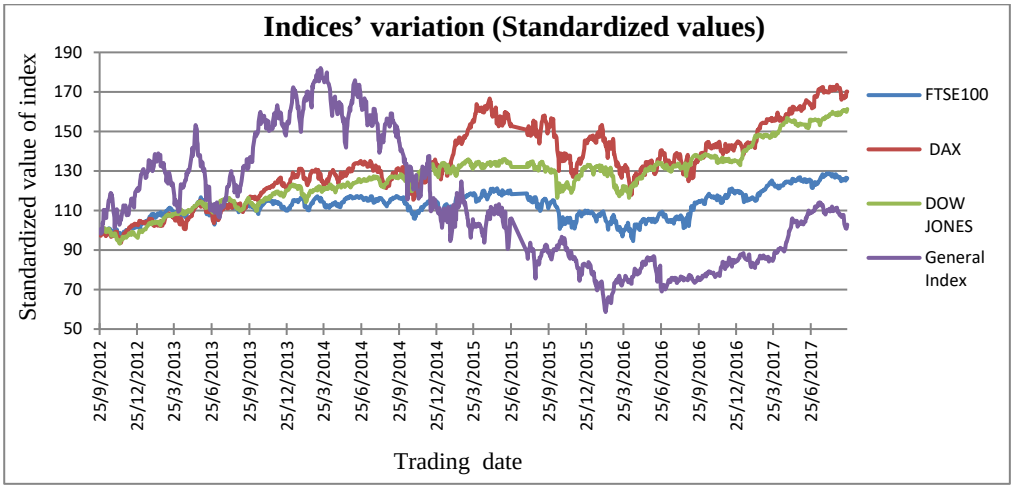


Table 1. *The most important economic – political events during the sampling time.*

Events during the sampling time *	
Date	Event
7/11/2012	Approval of the medium term rescue program 2013-2016
11/2013	Downgrade of the Athens Stock Exchange by MSCI (Emerging market)
25/5/2014	European elections (win of SYRIZA)
17/12/2014	Parliament failed to elect President of the Republic (First time)
23/12/2014	Parliament failed to elect President of the Republic (Second time)
29/12/2014	Parliament failed to elect President of the Republic (Third and last time)
25/1/2015	National Parliamentary elections
20/2/2015	First agreement between the new (SYRIZA-ANEL) government and the lenders
29/6/2015	Imposition of bank holiday and restrictions on capital movements
5/7/2015	Referendum, Closure of banks and imposition of a take-over cash limit. Closure of the Athens Stock Exchange
14/8/2015	Resolution of the third memorandum
20/11/2015	Recapitalization of the banks
21/3/2016	Downgrade of the Athens Stock Exchange by FTSE (Advanced Developing market)

*The list is indicative but not exhaustive

4. METHODOLOGICAL APPROACH, RESULTS AND DISCUSSION

4.1 MM model

The empirical results of previous studies have shown that the effect of the differencing interval used to calculate the securities' and market's returns on the estimation of systematic risk for ASE is quite intense (e.g. Milionis 2010, 2011).

In the present study for the given sampling time the total number of stocks is 122 stocks traded continuously in the Main Market of ASE. As discussed previously, the estimates of beta coefficients resulting for the same differencing interval (smoothing time) L, but for a different starting day of returns calculation, are different. This will be called "starting day" or "Corhay effect" henceforth. Therefore, to obtain all estimates of betas of each stock it is necessary to estimate L regressions for every differencing interval L, with L varying from one to thirty days. For the same L the data for each regression will start each time one day later. Subsequently, the average value of beta estimates for every L will be calculated. To this end a procedure in R programming environment was developed to perform this tedious task in an automatic manner. For each regression the values of the t-statistic and the corresponding p-value were saved in order to subsequently test the following hypothesis:

$$H_0: Beta = 0 \;\; H_1: Beta \neq 0$$

The results show that from the whole sample of 122 stocks 93 of them have statistically significant beta values for returns calculated using L=1.

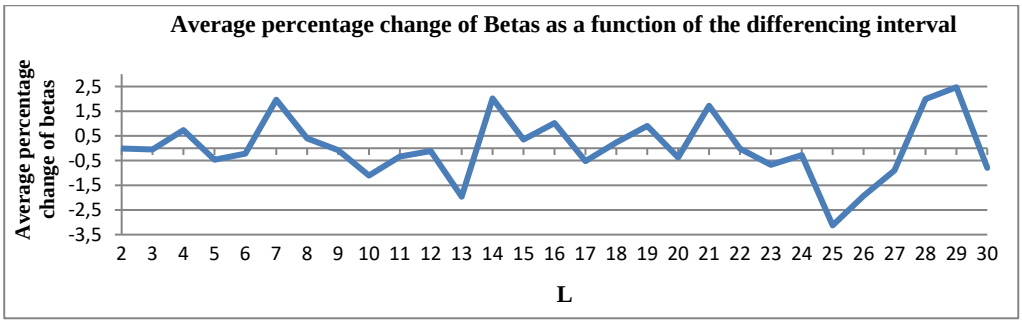
4.2 Intervalling effect

In order to evaluate the existence and strength of the intervalling effect on beta estimations the percentage change of beta is calculated as:

$$\frac{avBeta(L) - Beta(1)}{Beta(1)} * 100 \qquad (4)$$

where, $Beta(1)$ is the beta estimate for L=1 and $avBeta(L)$ is the average beta estimate for every L (L 1 to 30). The results are presented in Figure 2.

Figure 2. The average percentage change of betas as a function of the differencing interval.



As is evident from Figure 2, not only the intervalling effect during the sampling period is very weak, if any at all, but also there exist negative values of average percentage changes, contrary to the thus far empirical evidence from past published work (e.g. Cohen et al., 1983; Milionis, 2011).

4.3 Starting day (Corhay) effect

Another issue to be examined regarding beta estimates is the starting day effect. In view of the fact that the empirical work on that issue in the published literature is very limited, an assessment of the character and intensity of this effect, especially in conjunction with the intervalling effect, is of much interest.

An ad-hoc measure for the intensity of the Corhay effect (ICE(L)) is introduced as follows:

$$\text{ICE(L)} = \frac{Stdev(Beta(L))}{Median\,(Beta(L))} * 100 \qquad (5)$$

where *Beta(L)* represents each of the L beta estimates corresponding to a difference interval L, and $Stdev(Beta(L))$ and $Median\,(Beta(L))$ are the corresponding standard deviation and median.

Figure 3 shows the dependence of ICE(L) on L and (indirectly) on capitalization value. The latter is introduced by using the Large Cap and Mid. Cap Stock Indices of ASE to differentiate securities according to capitalization. Securities not belonging to these indices are those of smaller capitalization value and were grouped together as "rest of stocks".

As is evident from Figure 3 there is a very pronounced starting day effect. It is also clear that the effect is more pronounced for smaller cap stocks, which are included in the "rest of stocks" category.

To examine further the dependence of the staring day effect on capitalization, a log-log model between ICE(L) and capitalization was considered. The results of the regression of log(ICE(L)) against log(capitalization) together with a scatterplot are presented in Figure 4. The Coefficient of Determination of the regression model exceeds 60% indicating a strong dependence of the starting day effect on the capitalization value.

4.4 Intervalling versus starting day effect

There is little doubt that the importance of the intervalling effect relative to the starting day effect is of interest. To this end the ratio of $\frac{avBeta(L)-Beta(1)}{Beta(1)} * 100$ over ICF(L) against L is presented in Figure 5. As is evident from Figure 5, the starting day effect overwhelmingly dominates over the intervalling effect at all differencing intervals.

Figure 3. *The intensity of starting day (Corhay) effect as a function of the differencing interval for the various stock categories.*

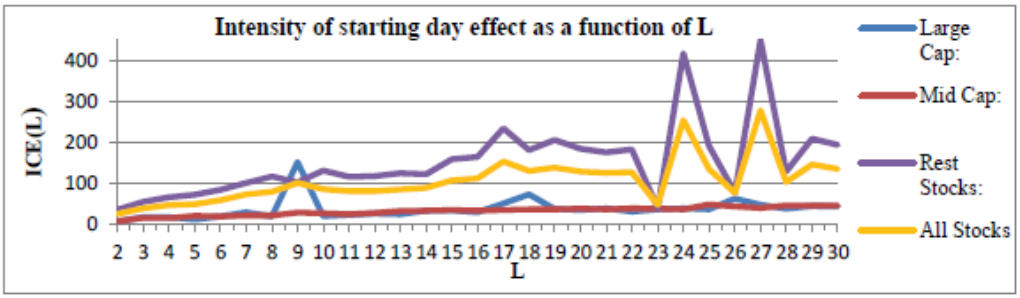


Figure 4*. Strength of starting day (Corhay) effect in relation to capitalization.*

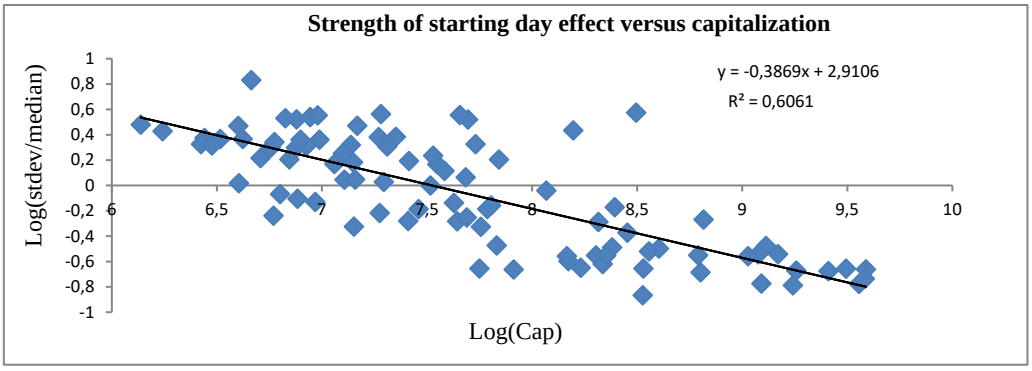
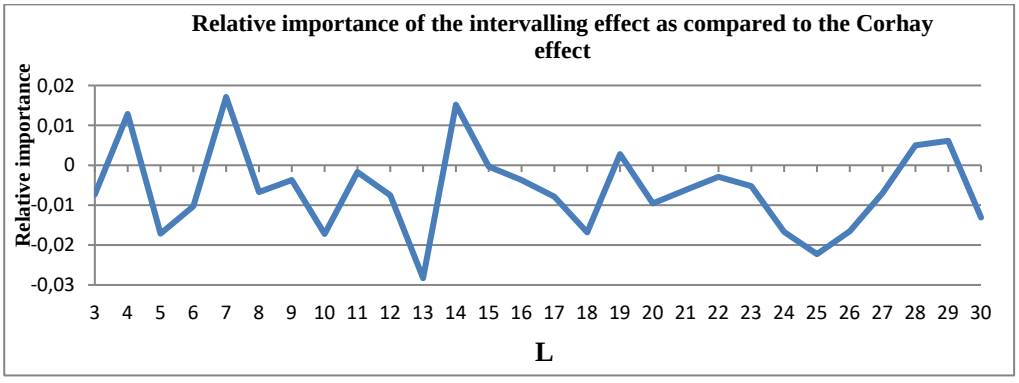


Figure 5. *The relative importance of the intervalling effect as compared to the starting day (Corhay) effect.*



The above empirical finding is also non-standard and deserves further analysis. To this end at first the maximum beta estimate for each security and each L was selected. Then the average of these maximum betas for each L was calculated and regressed against L. The regression model together with the corresponding graph is shown in Figure 6. As is evident, if the maxima of beta estimates were selected, an extremely strong intervalling effect would appear. In contrast, by doing the same thing, but selecting the minima instead of the maxima of the beta estimates for each security and each L, the situation is completely reversed, indicating the opposite of what would normally be expected, i.e. a consistent increase of average beta with L (see Figure 7). The overall disappearance of the intervalling effect is a reflection of the acute variation of beta estimates of a security for a particular differencing interval, i.e. of a very intense starting day effect.

Figure 6*. The average percentage change of betas due to the intervalling effect using the maximum of beta estimates for every security and every L.*

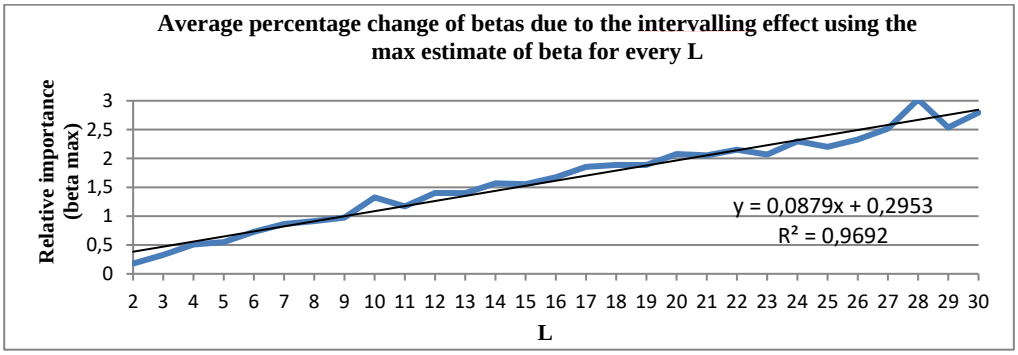
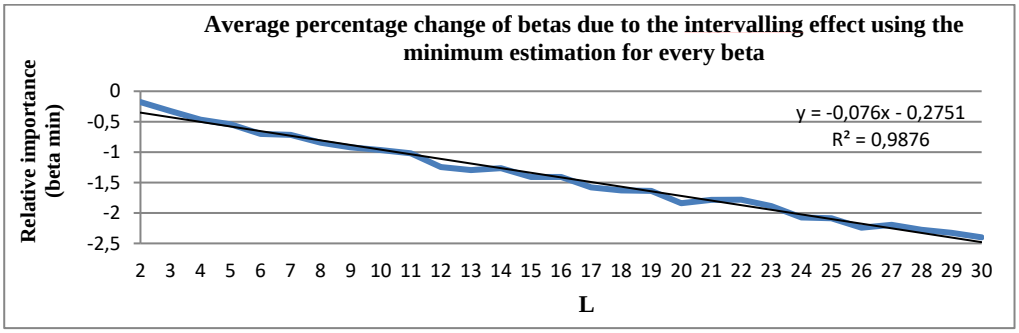


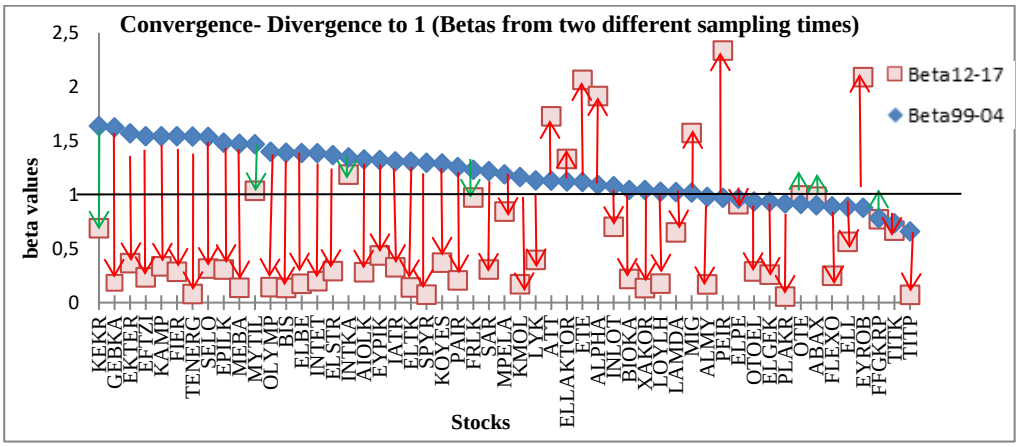
Figure 7*. The average percentage change of Betas due to the intervalling effect using the minimum of beta estimates for every L.*



4.5 Convergence – Divergence of beta estimates

Yet another aspect to be examined is the temporal variation of beta estimates for two different sampling periods over the same market with reference to unity. As found by Blume (1971, 1975) and Vasicek (1973), who estimated betas over two adjacent sampling periods, beta estimates far from unity in one sampling period tend towards unity in the next sampling period. For this task, beta estimates for the sampling time considered in this work, as well for the period 1999 – 2004, taken from the published work of Milionis (2010), were considered. The results are presented in Figure 8.

Figure 8. *Convergence – Divergence of Beta estimation from unity for two sampling periods.*



It is apparent that for betas of the second period there is mostly divergence from 1. Also, the values of beta estimates for the second period for the majority of securities are smaller than their values of the first sampling period. There are some prominent exceptions however, and such betas correspond mainly to securities of the banking sector (ALPHA BANK, ETE BANK, EUROBANK, PEIRAIOS BANK, ATTIKA BANK). This reflects the fact that the monetary-financial institutions were those affected the most by the turbulence caused by the economic crisis. It should be stressed, however, that the sampling times used in this comparison are not adjacent time periods, as were the periods used in the studies of Blume and Vasicek.

5. SUMMARY AND CONCLUSIONS

The main purpose of this empirical study was to examine systematic risk (beta) estimation in relation to some phenomena caused by market microstructure in a market under unusual circumstances. The results provide counterevidence for many stylized facts about beta estimation. More specifically, it was found that the starting

day (Corhay) effect is intense and strengthened further as the differencing interval is lengthened, while the intervalling effect is extremely weak. In fact, Corhay effect overwhelmingly dominates over the intervalling effect and it is actually the determining factor of changes in betas for the same stock, caused by different smoothing time and starting observation. Furthermore, low cap stocks expose clearly stronger Corhay effect as compared to large cap stocks. Generally, beta values, as compared to those of the past, are quite low for most stocks except banks. It is also evident that the importance of the Corhay effect has been undervalued in the published research thus far, as it is either ignored completely, or smoothed away by taking average betas for the same differencing interval but different starting observation. It is speculated that over the particular sampling period a single factor model, as MM, may not sufficiently capture the structure of the security returns. This is in line with other studies in which pricing models with additional to beta risk premia have been suggested. For ASE in particular, such a risk premium has been derived from a GARCH-in-mean-model (Milionis 2010; Milionis 2011).

ΠΕΡΙΛΗΨΗ

Σκοπός αυτής της μελέτης είναι η εξέταση της εκτίμησης του συστηματικού κινδύνου (συχνά παρατίθεται ως συντελεστής βήτα) σε σχέση με μερικά από τα φαινόμενα που σχετίζονται με την τριβή στις συναλλαγές, όπως η μικρή εμπορευσιμότητα των μετοχών, το επονομαζόμενο «intervalling effect» και το φαινόμενο που προκαλείται από την αλλαγή της παρατήρησης (εν προκειμένω ημέρας) εκκίνησης στον υπολογισμό των αποδόσεων κρατώντας σταθερό το μέγεθος αυτού του διαστήματος, σε μία αγορά υπό συνθήκες κρίσης. Τα δεδομένα που χρησιμοποιήθηκαν είναι ημερήσιες τιμές κλεισίματος μετοχών που διαπραγματεύονταν συνεχώς στην Κύρια Αγορά του Χρηματιστηρίου Αξιών Αθηνών (ΧΑΑ) κατά την περίοδο 25/09/2012 – 22/09/2017. Αυτή η περίοδος υπήρξε ταραχώδης για το Χρηματιστήριο Αθηνών, αντανακλώντας την έντονη οικονομική κρίση στην οποία βρισκόταν η χώρα κατά την ως άνω περίοδο. Αποτελεί αντικείμενο ιδιαίτερου ενδιαφέροντος, μεταξύ άλλων, η εξέταση φαινομένων που έχουν παρατηρηθεί σε προηγούμενες μελέτες και σχετίζονται με το πως επιδρά η τριβή στις συναλλαγές στις εκτιμήσεις των συντελεστών βήτα, ως προς την ισχύ τους και για το χρονικό διάστημα που διεξάγεται η συγκεκριμένη μελέτη. Πιο συγκεκριμένα, οι καθυστερήσεις στην αναπροσαρμογή των τιμών που προκύπτουν από την τριβή στις συναλλαγές έχουν ως επακόλουθο τη συστηματική μεταβολή των τιμών των συντελεστών βήτα με το διάστημα υπολογισμού των αποδόσεων («intervalling effect»). Επιπλέον, η μεροληψία αυτή στις εκτιμήσεις είναι μεγαλύτερη για μετοχές με μικρότερη εμπορευσιμότητα. Ένα άλλο πρόβλημα που σχετίζεται με την εκτίμηση του συστηματικού κινδύνου είναι αυτό που προκύπτει επειδή για το ίδιο διάστημα υπολογισμού των αποδόσεων αλλά για διαφορετική ημέρα εκκίνησης (Corhay effect) προκύπτουν διαφορετικές εκτιμήσεις βήτα, η σημασία του οποίου όμως δεν έχει μέχρι στιγμής εκτιμηθεί επαρκώς. Η ερευνητική συμβολή της συγκεκριμένης μελέτης

συνοψίζεται στα ακόλουθα ευρήματα-συμπεράσματα: (α) Το Corhay effect είναι ιδιαίτερα έντονο και ενισχύεται περαιτέρω καθώς μεγαλώνει το διάστημα υπολογισμού των αποδόσεων. (β) Το intervalling effect είναι ιδιαίτερα αδύναμο, πρακτικά δεν υπάρχει, παρ' όλο που ήταν έντονο σε άλλες μελέτες για την ίδια αγορά σε προηγούμενες χρονικές περιόδους. (γ) Οι μικρότερης κεφαλαιοποίησης μετοχές παρουσιάζουν εντονότερο Corhay effect σε σχέση με τις μετοχές μεγαλύτερης κεφαλαιοποίησης. (δ) Το Corhay effect κυριαρχεί έναντι του intervalling effect και αποτελεί τον καθοριστικό παράγοντα για τις αλλαγές στο συντελεστή κινδύνου για την ίδια μετοχή. (ε) Τα παραπάνω ευρήματα οδηγούν στο συμπέρασμα ότι η σημασία του Corhay effect δεν έχει εκτιμηθεί κατάλληλα μέχρι στιγμής στις ήδη δημοσιευμένες μελέτες (κατ' ακρίβεια η σημασία του έχει υποτιμηθεί), καθώς είτε αγνοείται εντελώς, είτε εξομαλύνεται λαμβάνοντας το μέσο όρο των εκτιμήσεων για το ίδιο διάστημα υπολογισμού των αποδόσεων αλλά για διαφορετική παρατήρηση εκκίνησης.

REFERENCES

Blume M., (1971). On the assessment of risk. *Journal of Finance,* **VI(**1), 1-10.

Blume M., (1975). Betas and their regression tendencies. *Journal of Finance,* **X**(3), 785-795.

Cohen C., Hawawini G., Maier S., Schwartz R. and Whitcomb D., (1983a). Estimating and adjusting for the intervalling effect bias in beta. *Management Science* vol. **29**, 1.

Cohen C., Hawawini G., Maier F., Schwartz R. and. Whitcomb D. (1983b). Friction in the trading process and the estimation of systematic risk. *Journal of Financial Economics,* **12**, pp. 263-278.

Corhay A., (1992). The intervalling effect bias in beta: A note. *Journal of Banking and Finance* **16**, 61-73.

Dimson E., (1979). Risk measurement when shares are subject to infrequent trading. *Journal of Financial Economics,* **7**, 197-226.

Elton E., Gruber M. and Brown S., Goetzmann W., (2007). Modern portfolio theory and investment analysis. 7^{th} edition, Wiley, NY.

Fama E. and French K., (1992). The cross section of expected stock returns. *Journal of Finance,* **47***(*1), 427-466.

Fama E. and French K., (1993). Common risk factors in the returns on bonds and stocks. *Journal of Finance,* **33**(1), 3-56.

Fama E. and French K., (1996). Multifactor explanations of asset pricing anomalies. *Journal of Finance,* Vol. **51**(1), 55-84.

Jaganathan R. and Wang Z., (1996). The conditional CAPM and the cross section of expected returns. *Journal of Finance,* **51**(1), 3-53.

Lintner J., (1965). The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets, *The Review of Economics and Statistics,* **47**(1), 13-37.

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,* 14(1-4), 140-156.

Milionis A. E. and Papanagiotou E. (2009). A study of the predictive performance of the moving average trading rule as applied to NYSE, the Athens Stock Exchange and the Vienna Stock Exchange: sensitivity analysis and implications for weak-form market efficiency testing. *Applied Financial Economics,* **19**(14), 1171-1186.

Μηλιώνης Α., (2010), in Greek. Αμερόληπτες Εκτιμήσεις Κινδύνου στο Χρηματιστήριο Αξιών Αθηνών. *ΟΠΑ, Μελέτες για το ελληνικό χρηματοπιστωτικό σύστημα,* 295- 308.

Milionis A. E. and Papanagiotou E. (2011). A test of significance of the predictive power of the moving average trading rule of technical analysis based on sensitivity analysis: application to the NYSE, the Athens Stock Exchange and the Vienna Stock Exchange. Implications for weak-form market efficiency testing, *Applied Financial Economics* **21**(6), 421-436.

Milionis A. E. and Patsouri D., (2011). A conditional CAPM; implications for the estimation of systematic risk. *Working Paper*, 131, *Bank of Creece.*

Milionis A. E. (2011). A conditional CAPM; implications for systematic risk estimation, *The Journal of Risk Finance*, 12(4), 306-314.

Scholes M. and Williams J., (1977). Estimating beta from non-synchronous data. *Journal of Financial Economics,* **5**, 309-327.

Sharpe W., (1964). Capital Asset Prices: A theory of market equilibrium under conditions of risk. *Journal of Finance,* **XIX 3** 425-442.

Vasicek O., (1973). A note on using cross sectional information in Bayesian estimation of security betas. *Journal of Finance,* VIII(5), 1233-1239.

Greek Statistical Institute
Proceedings of the 31st Panhellenic Statistics Conference (2018), pp.329-336

REPRODUCIBILITY CRISIS IN SCIENCE: A STATISTICIAN'S PERSPECTIVE

T. Papaioannou
Universities of Piraeus *and Ioannina, Greece*
takpap@unipi.gr

ABSTRACT

Reproducibility crisis in science is a frequent topic in scientific journals in the last ten years. It refers to the fact that several or many usually published studies fail to replicate, i.e., when repeated again by the same or independent researcher or researchers usually under the same or similar conditions, lead to different conclusions than the original study. The conclusions drawn in all these studies are heavily or exclusively statistical. This of course tarnishes the image of statistics, theory and practice. I this paper we review the problem, provide a historical account, present its various approaches and points of view as well as our position on the problem. Again, proper and correct use of statistical methodology emerges as part of the problem.

1.Introduction

In the sixties/seventies while working on my PhD, I was introduced to the *scientific method* by my advisor Oscar Kemthorne at Iowa State. Kempthorne at that time, besides his interests in Statistical Genetics and Design and Analysis of Experiments, was interested in the foundations of probability and inference and other philosophical issues. He was a fan and devout supporter of Fisher. It was the time of the big conflict between frequentists and Bayesians.

Through Kempthorne and reading various books (Fisher, Lehmann, Scheffe, Cramer, Fitz) I learned that statistics and in particular tests of significance, tests of hypothesis, randomized or not, and confidence intervals provide the means to empirically verify or not scientific hypotheses, to make decisions etc. It was the essence of the scientific method, the main method to "prove" empirically hypotheses or theories in science. Science not like in Physics where theories must be experimentally verified, had used and was heavily using the scientific method to empirically investigate, prove or disprove hypotheses etc.

Soon after came the explosion of Biostatistics where numerous medical investigations started to take place with use and misuse of statistics. The misuse of statistics was common and articles like 'what doctors know about statistics' started to appear in the literature. Numerous papers where the materials and methods and analysis sections contained tools ranging from simple descriptive statistics to survival and multivariate analyses were being published. A few, very few Bayesian papers started apperaing in the litterature.

Statisticians were concerned about the misuse of statistics and voices were heard asking papers containing statistical analyses to be refereed by statistical experts as well. Editors of some scientific journals added statisticians to their editorial boards. The statistical quality of published papers started to show signs of improvement.

2. 2005

In 2005 while I was a visiting professor at the University of Cyprus I noticed the article "Why most published research is false" by John P. A. Ioannidis published in *PloS Medicine*, Aug 2005, Vol 2, #8, e124, 696-701. Ioannidis was making some provocative statements, including the title, as for example:

- "Published research findings are sometimes refuted by subsequent evidence, with ensuing confusion and disappointment".

-"False findings may be the majority or even the vast majority of published research claims".
-"It can be proven that most claimed research findings are false".
-"There is a high rate of non replication (lack of confirmation) of research discoveries".
-"The acceptance (believability) of statistically significant results can be undermined by bias, small effect sizes, and small sample sizes".

Here bias is used in a peculiar way meaning the probability that a study, for which the null hypothesis H_0 is true, is declared as statistically significant after the experiment.

Ioannidis argument is based on positive predictive values (PPV's) widely used in diagnostic tests. They are the proportion of positive results, that are true positive results, i.e. the proportion of true positive results among the positive results :

$$P(\text{ Ha is true} \mid \text{test is significant}),$$

where Ha is the alternative hypothesis. PPV's are widely used in Biostatistics (diagnostic tests) and have the following expression:

$$PPV = \frac{sensitivity \times prevalence}{sensitivity \times prevalence + (1 - specificity) \times (1 - prevalence)}.$$

In terms of tests of hypotheses this translates into

$$\begin{aligned} PPV &= P(\text{ Ha is true} \mid \text{test is significant}) \\ &= \frac{(1-\beta) \times prior}{(1-\beta) \times prior + a \times (1 - prior)} \\ &= \frac{power \times prior}{power \times prior + a \times (1 - prior)} \end{aligned} \qquad (2.1)$$

Thus, PPV depends on α, the prior for Ho and the power. Studies should have high PPVs. Note, however, that with ignorance (non-informative prior) and power 80% (the usual standard in sample size calculations)

$$PPV = 0.9411 \text{ when } \alpha = 5\% \quad \text{and}$$

$$PPV = 0.9876 \text{ when } \alpha = 1\%,$$

i.e., we achieve high PPVs.

Ioannidis gave the following expression for PPV

$$PPV = \frac{(1-\beta)R}{R - \beta R + a} \qquad (2.2)$$

where R is the ratio of the number of "true relationships" to "no relationships" among those tested in the field. R is the odds of the prior for Ha and R / (R + 1) is the prior.
Ioannidis gave no mathematical proof of his major thesis "most claimed research findings are false". He arrived at expression (2.2) not by Bayes formula, but through a 2x2 table with rows : research findings (yes, no) and columns : true relationships (yes, no).
From (2.2) comparing PPV with 1 – PPV, he arrived at the statement

"a research finding is more likely to be true than false if $(1 - \beta) R > \alpha$ "

Also from (2.2) PPV is an increasing function of the power $1 - \beta$, and thus Ioannidis concluded

"the smaller the studies the less likely the research findings to be true".

Ioannidis gave some additional corollaries:

1. "The smaller the effect sizes in a scientific field, the less likely the research findings to be true" (power is related to the effect size)

2. "The greater the number and the lesser the selection of tested relationships in a scientific field, the less likely the research findings are to be true."

3. "The greater the flexibility in designs, definitions, outcomes, and analytical modes in a scientific field, the less likely the research findings are to be true"

4. "The greater the financial and other interests and prejudices in a scientific field, the less likely the research findings are to be true."

5. "The hotter a scientific field (with more scientific teams involved), the less likely the research findings are to true."

These corollaries have stirred many discussions and comments and lack mathematical or statistical verification. The title of Ioannidis's paper was provocative, but on the other hand the title was one of the reason for the essay to be noticed and read. Apparently, the statement that most (the majority) of published research findings are false. Apparently, this covers all scientific fields where empirical research is conducted (the essay does not restrict itself to the medical sciences). Also the arguments, the justification of "why most published research findings are false", are not convincing and contain misunderstandings and misinterpretations.
The essay addresses itself to published research findings and presents several lines of reasoning as to why most of them are false. Naturally the same lines of reasoning apply to most unpublished research findings and therefore most research findings are false.
C.R. Rao, a famous and well renowned statistician, in a seminar in Athens, Greece said that one of the greatest contributions of Statistics to humanity was the establishment that smoking is hazardous to people´s health or, in other words, smoking "causes" cancer. This finding was established with published research, among others, using standard, common statistical practice.
Initially one wonders what it means "a published research finding is false", besides the apparent meaning of the words. Soon he discovers that what it actually means is "a (published) research finding **given** the results of a statistical analysis (usually statistical tests or inference)". There are many research findings which do not fall in this category (see, for example the field of Physics. The former consideration of a finding given a statistical analysis immediately leads to the Bayesian paradigm or approach in statistics and to positive predictive values (PPV) widely used in medicine (diagnostic and screening tests, epidemiology, etc).
Let H_0 be a widely known null statistical hypothesis, e.g. there is no relationship or association between a prognostic factor and a disease and H_a the alternative statistical hypothesis, also widely known, that there is a relationship between the factor and the disease. Usually in most research efforts we are interested to see if empirical data, well designed and planned and obtained data, support or not H_0. This is done by tests of significance and p-values or confidence intervals or by tests of hypothesis of Neyman - Pearson type (where the concept of power comes in) or by Bayesian methodology which requires the specification of

The seed of doubt and of questioning the integrity and validity of statistical methodology had been thrown.

3. Impact and Reproducibility

Ioannidis's article had a tremendous impact. Many comments, reactions, letters, rebuttals and even papers have been written analysing and criticizing the paper. Here is a sample:
Goodman and Greenland (April 2007) leading Bayesian methodologists from the John Hopkins Bloomberg School of Medicine wrote to PLoS Medicine: Unfortunately, while we agree that there are more false claims that many would suspect – based both on poor study design, misinterpretation of p-values, and perhaps analytic manipulation – the mathematical argument in the PLoS Medicine article underlying the "proof" of the title's claim has a degree of *circularity* (indirectness, roundaboutness)
Three months later Ioannidis (2007) responded: If a researcher doubts or is skeptical about the validity of his research findings or the findings of other researchers, it is natural or may be tempted to think 'let's repeat the experiment, the clinical trial, etc to see if we get the same or similar results'.
This of course is against the principle of the *scientific method* or the foundations of statistical inference. The scientific method has characterized natural sciences since the 17th century, consisting in systematic observation, measurement, and experiment, and the formulation, testing, and modification of hypotheses. It has been extremely successful in bringing the world out of medieval times, especially once it was combined with industrial processes. However, when the scientific method employs statistics as part of its arsenal, there are a number of both mathematical and practical issues that can have a deleterious effect on the reliability of the output of the scientific methods.
Skepticism about the research findings may be due poor experimentation, poor or incorrect application of statistical methods, biases, and many other reasons. In other words, poor data analysis.
Nevertheless, in practice, in science, (some) experiments or clinical trials have been repeated and in some cases the results were surprising. They led to what we now call the reproducibility crisis in science.
In 2015 a two-day workshop on Reproducibility was held in Washington D.C. organized by the National Academies of Science, Engineering and Medicine. Its theme was to discuss and examine the statistical challenges in assessing and fostering the reproducibility of scientific results. Here is how they defined reproducibility .
Reproducibility: It is the ability of a researcher to duplicate the results of a prior study using the same materials as were used by the original investigator. A second researcher might use the **same raw data** to build the same analysis files and implement the same statistical analysis in an attempt to yield the same results. If the same results are not obtained, the discrepancy could be due to differences in processing of the data, differences in the application of statistical tools, differences in the operations performed by the statistical tools, accidental errors by an investigator, and other factors…. Reproducibility is a minimum necessary condition for a finding to be believable and informative (cf. S. Goodman, NSF, 2015 and National Academies of Sciences, Engineering, and Medicine, 2016, p.4)
Besides reproducibility we also have replicability which is the ability of a researcher to duplicate the results of a prior study with the same procedures but with *new* **data**. A failure to replicate a scientific finding is commonly thought to occur when one study documents statistically significant relations between two or more variables and a subsequent attempt to implement the same operations fails to yield the same statistically significant relations (cf. S.Goodman, NSF, 2015 and National Academies of Sciences, Engineering, and Medicine, 2016, p.4)
So,reproducibility has several forms: use the same raw data, use new data ,use same methods or similar methods , use same or different sample size etc.We list some factors affecting reproducibility.

(i) Insufficient training in experimental design and generally in organizing and conducting experiments, surveys, collecting data etc.
(ii) Poor data management and analysis.
(iii) Inappropriate understanding and interpretation of statistical concepts
(iv) Biases

4. SOME WEEL-KNOWN EXAMPLES OF REPRODUCIBILITY FAILURES

1) Potti and his co-workers at Duke University, using genomic microarray data published a paper in *Nature Medicine* (2006) which give an algorithm that predicted which cancer patients would respond to chemotherapy. Two statisticians at MD Anderson Cancer Center , using the publicly available data, attempted to reproduce the Duke University analysis and results and found many errors not only in the handling the data but also that the analysis was wrong. The original study was finally retracted from *Nature Medicine* in 2011.
2) Reinhart and Rogoff published a paper in the *American Economic Review*, 2010, claiming that countries with high debt to GOP ratio , like Greece, suffer from *low* growth. They even suggested that with a 90% or above ratio a country would have a *drop* in economic growth. This had a major impact in economic policy as it favours economic austerity. Herndon (a graduate student) and Pollin obtained the data from the authors and reproduced their analysis (Herndon et al, 2013). They found many data editing errors and that an unusual form of data weighting led to the 90% figure. Had they used a more standard weighting form, they would have obtained a smoother relationship between the ratio and growth.
3) A recent study, 2011, found that 65% of medical studies were "inconsistent' when tested, and only 6% were completely reproducible (cf. Prinz et al, 2011)
4) In 2012 a survey published in *Nature* found that 47 out of 53 medical research papers on the subject of cancer were irreproducible (Begley and Ellis, 2012)
5) Gelman's wrong sign in one of the variables: Gelman and Cai (2008) published a paper in the *Annals of Applied Statistics* showing that Democratic candidates in the United States could get more votes by moving slightly to the right on economic policy. An error in the data analysis was discovered in 2013. In trying to replicate the work, an undergraduate student had discovered that Gelman had used the wrong sign in one of the variables. Gelman published right away a three-sentence statement declaring that everything in the paper's crucial section should be considered wrong until proven otherwise.
6) Ioannidis and his team in 2009 managed to fully replicate only 2 data out of 18 microarray-based-gene expression studies. Here the replication failure was more due to unavailability of data and incomplete data annotation rather to statistical reasons.
7) In pharmaceutical research there exist several examples of non- reproducibility. For example, the therapeutic power of highly successful second-generation antipsychotic drugs, such as Zyprexa, Seroquel and Abilify, developed in the nineties, fifteen years later was shown to be steadily waning having an effect less than half of the original (Lehrer, 2010)).Also, Bayer Healthcare reported in 2011 that its scientists could not reproduce about 75% of published findings in cardiovascular disease, cancer and women's health.

There is a long history of 'statistical doubts' in Psychology. By doubts we mean conclusions, usually cause/effect relations, based on statistical investigations which are not accepted by many psychologists.

The Open Science Collaboration project (cf. Nosek, 2015) with 350 researchers operating collectively, replicated 100 experiments, reported in papers published in 2008 in three high-ranking psychology journals.The aim was to compare replication results of these 100 experiments with the original findings according to several criteria as for example p-values, effect sizes, etc. Ninety seven percent of the original studies had a p-value less than 5%. But only 36% of the replicated did so. Also, the effect sizes were found to be smaller than in the original papers. In general, the results in more than 50% of the papers were not replicated. They also reported some factors causing low reproducibility rate such as designs with low

statistical power, publication bias etc. In addition, reproducibility was negatively correlated with the original p-values and effect sizes .

5. OTHER POINTS AND ISSUES

The high profile journals *Science* and *Nature* are calling on statisticians to help with the reproducibility problem. Articles with emphatic titles have appeared in the literature: *Odds Are, It's Wrong* (Siegfried, 2010) which claims that science has failed to face the shortcomings of statistics, with feedback comments and response by the President of the American Statistical Association (ASA), the President of the International Statistical Institute and the Editor of the Journal of ASA (JASA) Pantula, Teugels and Stefanski (2010) and others; *Love affair with statistics gives science a significant problem* (Siegfried, 2011); *Publish and be Wrong* (Economist 2008); *Can you repeat that?* (Significance News, 2014); *I don't believe that, do it again* (Significance, 2012); *Fooling Ourselves* Nuzzo(2015) which claims that humans are good at self-deception, but growing concern about reproducibility is driving many researchers to seek ways to fight their own instincts; *Do clinical trials work?* (Leaf 2013), etc.
The common denominator in all these criticisms is the failure to see that the causes of non reproducibility are poor experimental and observational study design, poor statistical analysis, biases and misunderstanding of statistical concepts.
A statistical counterattack has been recently published by (Peng, 2015) where it is clearly stated that a comparative lack of analytical skills has resulted in scientific findings that are neither replicable nor reproducible. It is time to invest in statistics education he says.

An issue related with reproducibility is the use of p-values. Lately strong criticism on the use of p-values in scientific research has appeared in the literature. P-values are widely used, misused and misinterpreted. Two highlights are the ban on the use of p-values by a psychology journal (*Journal of Basic Applied Social Psychology*) and the statement of the American Statistical Association on the use of p-values (Wasserstein and Lazar, 2016).
When a study is replicated one should not expect p-values to remain constant. As a matter of fact, p-values are variable and their standard error should be evaluated. This would give a feeling of the chances of reproducibility. Boos and Stefanski (2011) state that for published studies with p-values in the range of 0.005 to 0.05 the probability of non replication is roughly 33%. The variability of p-values could account for an important fraction, but not all, of the observed lack of replication in studies reported in the literature, they say. Of course, p-value should not be the only measure of evidence against a null hypothesis.
Lowering the threshold of significance to 0.005 from 0.05 has been proposed by a large coalition of 72 methodologists as a way to cope with the problem of reproducibility and/or the use of p-values (Benjamin et al, 2018). Some additional suggestions are: Use other metrics besides p-values since they are the wrong metric to use. Use effect sizes as well. Calculate a 95% prediction interval for the replication effect size given the original effect size

6. CONCLUSION AND MY OWN POSITION

The concept of statistical significance still remains unclear to the layman. Some believe that a significant result is one in which data points were produced by chance less than 5% of the time (Lehrer, 2010). Others consider the p-value as the probability of the null hypothesis being true.

Data nowadays are collected at an amazing rate. In the sixties we conducted surveys by mailing letters. Nowadays our studies can be based on billions of instant messages collected over a short time. The cost of collecting data has dropped dramatically. There is a delude of data. One result of this is an epidemic of poor data analysis. Our conclusion is summarized with the following points.

1. The principles of statistics and the statistical methodologies are crystal clear. The problem arises when one applies and interpret them in science.

2. Regarding tests of significance or tests of hypothesis or p-values, statistical theory does give a rule how to choose the p-value. It does not indicate that the p value should be 5% or 1% or some other value. The researcher should pre select significance (α) and power (1-β) to make a decision or to declare a research finding. Several other rules should apply such as appropriateness and assumptions of the model, goodness of fit, sample selection etc. p-values arose from the application of statistical theory and the proliferation of statistical programs

3. All these examples are troubling and cast doubt on the importance and value of statistical methodology. Many of them are due to errors and poor analysis

4. Unconscious biases can explain part of reproducibility problems.

5. It appears that there is a problem. The problem lies with science not with statistics. It is the duty of Statistics and Statisticians to collaborate with science to resolve the issue

Many researchers ignore the effect of multiple testing which raises the size of Type I error and reduces the confidence level.

Some scientists advise to make research reproducible. Some people claim the reproducibility is the foundation of modern research. However, there is no clear way on how to achieve reproducibility

Two years have passed since the ASA statement on p-values and nothing has changed, business is as usual. We would only say that there is increased awareness of the problem.

ΠΕΡΙΛΗΨΗ

Η κρίση αναπαραγωγησιμότητας στην επιστήμη είναι ένα θέμα που συχνά εμφανίζεται στα επιστημονικά περιοδικά τα τελευταία δέκα-δεκαπέντε χρόνια. Αναφέρεται στο γεγονός ότι αρκετές ή πολλές συνήθως δημοσιευμένες μελέτες αποτυγχάνουν να αναπαραχθούν, δηλαδή εάν οι μελέτες επαναληφθούν από τον ίδιο ή άλλους ανεξάρτητους ερευνητές, συνήθως κάτω από τις ίδιες ή παρόμοιες συνθήκες, καταλήγουν σε διαφορετικά συμπεράσματα από οτι η αρχική μελέτη. Τα συμπεράσματα που εξάγονται από τις μελέτες αυτές κατά κύριο λόγο ή σχεδόν αποκλειστικά είναι στατιστικά. Αυτό αμαυρώνει την εικόνα της Στατιστικής, θεωρία και εφαρμογές.

Στο παρόν άρθρο κάνουμε μια επισκόπηση του προβλήματος, παρέχουμε μια ιστορική αναδρομή, παρουσιάζουμε τις διάφορες προσεγγίσεις και απόψεις καθώς επίσης και τη δική μας θέση. Και πάλι η κατάλληλη και ορθή χρήση της στατιστικής μεθοδολογίας αναδεικνύεται ως μέρος του προβλήματος.

REFERENCES

Begley, C.G. and Ellis, L.M. (2012) Drug development: Raise standards for preclinical cancer research. *Nature*, 483,531-533.

Benjamin, D.J., Berger, J.O., Johnson, V.E. et al (2018). Redefine statistical significance, *Nat. Hum. Behav.,* 2, 6-10.

Boos, D.D. and Stefanski, L.A. (2011). P-value, Precision and Reproducibility, *The American Statistician*, 85:4, 813-821.

Economist (2008) Publish and be wrong, Oct. 9, 2008.
Gelman, A. and Cai, C.J. (2008). *Ann. Appl. Stat.*2, 536-549.
Goodman, S. and Greenland, S. (2007), Why most research findings are false: Problems in the analysis*, PLoS Medicine,* Vol 4, Isue 4, 773.
Herndon, T., Ash, M. and Pollin, R. (2013). Does high-profile data consistently stifle economic growth? A critique of Reinhart and Rogoff. *Political Economy Research Working Paper Series* no. 322.PERI, University of Massachusetts, Amherst.
Ioannidis, J.P.A. (2005). Why most published research findings are false, *PLoS Medicine*, Vol 2. Issue 8, 696 – 701.
Ioannidis, J.P.A. (2007). Why most research findings are false: Author's reply to Goodman and Greenland. Plos Medicine, vol 4, issue 6, 1132-1133.
Ioannidis, J.P.A. et al (2009). Repeatability of published microarray gene expression analyses, *Nature Genet.* 41,149-155.
Leaf, Clifton (2013), Do Clinical trials work? *The New York Times*, Jul. 13, 2013.
Lehrer,J. (2010). The truth wears off – Is there something wrong with the scientific method? *The New Yorker*, Dec. 13,2010,1-10.
National Academies of Sciences, Engineering, and Medicine (2016). Statistical Challenges in Assessing and Fostering the Reproducibility of Scientific Results: Summary of a Workshop.
Washington D.C.: The National Academic Press, doi: 10.17226/21915.
Nosek, B.A. (2015). Estimating the reproducibility of psychological science, *Science*, vol. 349 issue 6251.
Nuzzo, Regina (2015). Fooling Ourselves, *Nature*, Vol 526, Oct. 8,2015 p.182-185.
Pantula, S.G., Teugels, J. and Stefanski, L. (2010), A statistical education, Letters, *Science News*, May 8, 2010, Vol. 177 #10 (p.32).
Peng, Roger (2015) The reproducibility crisis in science – A statistical counterattack, *Significance*, June 2015, 30-32.
Potti, A., Dressman, H.K., Bild, A.et al 92006). Genomic signatures to guide the use of chemotherapeutics. *Nature Medicine*, 12,1294-1300.
Prinz, F., Schlage, T. and Asadullah, K(2011), Believe it or not: how much can we rely on chemotherapeutics. *Nature Medicine*, 12,1294-1300.
Reinhart, C.M. and Rogoff,K.S. (2010). Growth in a time of debt, *American Economic Review*,100,573-578.
Siegfried,Tom(2010) Odds Are, It's Wrong, *Science News*, March 27, 2010, Vol 177,#7 (p. 26).
Siegfried,Tom(2011). Randomness. *Science News*, Nov.7, 2011,Vol 1 No 20. Significance News (2012), I don't believe that, do it again, *Significance*, Oct. 2012, p.2
Significance News (2014) Can you repeat that?, *Significance*, Oct 2014, p.4.
Wasserstein, R.L. and Lazar, N.A. (2016). The ASA's statement of p-values: content, process and purpose. *The* American *Statistician*, 70:2,129-133.

Greek Statistical Institute
Proceedings of the 31st Panhellenic Statistics Conference (2018), pp. 337−352

LOW PRICE EFFECT AND POST MODEL CORRECTION OF VALUE AT RISK

G-J. Siouris, D. Skilogianni, A. Karagrigoriou
Laboratory of Statistics and Data Analysis, Department of Statistics and Actuarial - Financial Mathematics, University of the Aegean, Samos, Greece
siouris@aegean.gr, skilogiannides@outlook.com, alex.karagrigoriou@aegean.gr

ABSTRACT

After extensive investigation on the statistical properties of financial returns, a discrete nature has surfaced when low price effect is present. In order to model the discrete nature of returns the discretization of the tail density function is applied. This is a rather logical approach as the market always operates on a specific accuracy. As a result of this discretization process, it is now possible to improve the Percentage Value at Risk (PVaR) and Expected Percentage Shortfall (EPS) estimations on which we are focusing on this work. Finally, since in risk analysis it is apparent that the returns of an asset do not all bare the same amount of information and consequently, do not all carry the same degree of importance, Adjusted Evaluation Measures are presented for foreacastig purposes.

Keywords: low price effect; PVaR; EPS; low price correction; violation ratios; normalised shortfall; adjusted evaluation measures

1. INTRODUCTION

The quantification of risk is an important issue in finance that becomes even more important during periods of financial crises. The estimation of volatility is the main financial characteristic associated with risk analysis and management since in financial modelling, it has been consolidated the view that the asset returns are preferred over prices due to their stationarity ([1], [2]). Moreover, the volatility of returns is the one that can be successfully forecasted and that is essential for risk management ([3], [4]).

The increase in complexity of both the financial system and the nature of the financial risk, results in models with limited reliability. Hence, in extreme economic events, such models become less reliable and in some instances, fully fail to measure the underlying risk. Consequently, they are producing inaccurate

risk measurements which has a great impact on many financial applications, such as asset allocation and portfolio management in general, derivatives pricing, risk management, economic capital and financial stability (based on Basel III accords).

Stock exchange markets operate with certain accuracy. In most European markets the accuracy is 0.1 cents (0.001 euros), while U.S. markets operate under the same accuracy except for securities that are priced less than USD 1.00 for which the market accuracy or minimum price variation (MPV) is USD 0.0001. Despite the readjustement for extremely low price assets, the associated fluctuation (variation) is considerable and the corresponding volatility is automatically increased. The phenomenon is magnified considerably in periods of extreme economic events (economic collapses, bankruptcies, depressions, etc.) and as a result typical market accuracy fails to handle smoothly assets of extremely low price.

Any attempt to increase the reliability of the model identification procedure for volatility estimation results unavoidably in even more complex models which still will be unable to identify and fully capture the governing set of rules of the global economy. The more complicated the model we use for volatility estimation, the larger the number of parameters that needs to be estimated. Hence, in order to have larger data sets we go further back in the past, to collect and use for the analysis, older observations, which though may be less representative of reality. Lastly, risk models utilize only the returns of an asset while ignoring the prices.

In order to answer all the above, we discuss, in this work, the concept of *low price effect (lpe)* ([5]) and recommend a *low price correction (lpc)* for improved forecasts. The low price effect is the increase in variation for stocks with low price due to the existence of a minimum possible return produced when the asset price changes by the MPV. Lpe is frequently overlooked and the main reason for that is the lack of theoretical background related to the reasons resulting in this phenomenon which surfaces primarily in periods of economic instability or extreme economic events. The pioneering of the proposed correction is that it does not require any additional parameters and takes into account the asset price. The proposed correction is associated with the rationalization of the estimated asset returns since it is rounded to the next integer multiple of the minimum possible return. Except of the proposed correction in this work we also provide a mathematical reasoning for the increase of volatility.

Inspired from the above, we came to the same conclusion as many before, that in risk analysis the returns of an asset do not all bare the same amount of information and do not all carry the same degree of significance ([6], [7], [8], [9]). In the absence of a formal mathematical notion, the term *asymmetry in the importance* decribes relatively satisfactory the above phenomenon which is also apparent in other scientific areas, related to medical, epidimiological, climatological, geophysical or meteorological phenomena. For a mathematical interpretation, one may consider a proper cost function that takes different values depending on different regions of the dataset or the entire period of observation. For instance,

in biosurveillance, the importance (and the cost) associated with an illness incidence rate is much higher in the so called epidemic periods associated with an extreme rate increase. In the case of risk analysis, risk measures like the Value at Risk (VaR) and Expected Shortfall (ES), concentrate only on the left tail of the distribution of returns with the attention being paid to the days of violation rather than to those of no violation. Consequently, failures in fitting a model on the right tail are not considered to be important. Thus, the asymmetry in the importance of information is crucial in both choosing the most appropriate model by assigning a proper weight to each region of the dataset and evaluating its forecasting ability.

In order to judge the forecasting quality of the proposed methodology applied to typical risk measurements like VaR and ES, we have appropriately adjusted a number of popular evaluation measures that take into account the *asymmetry in the importance* of information bared in the data. Since the proposed low price correction, is applied in a subset of the dataset with a specific property (in this case a very low price), it is logical for the adjusted evaluation measures to be computed on the same subset. Thus, the risk estimations can be evaluated with and without the implementation of low price correction and then, compared. The decrease in the values of the adjusted evaluation measures will show the improved forecasting quality of the proposed methodology.

In addition, for the evaluation of the proposed correction, backtesting methods such as the violation ratio (VR) for Value at Risk and normalised shortfall (NS) for Expected Shortfall can also be used (see for example, [10]). For real life examples and applications see [11], [12], [13].

2. THE LOW PRICE EFFECT AND CORRECTION

Typical market accuracy fails to handle smoothly assets of extremely low price resulting in a considerable fluctuation (variation) and increased volatility. As expected, the phenomenon is magnified considerably in periods of extreme economic events (economic collapses, bankruptcies, depressions, etc.). Such extreme events have occurred during the last decade across the globe, starting with 2007 in the US market and split over to the EU markets. Indeed, since all possible (logarithmic) returns on a specific day are integer multiples of the minimum possible return, the stock movement will fluctuate nervously so that violations will occur more frequently and forecasts will turn out to be irrational in the sense that such returns cannot be materialized. The resulting volatility increase is quite often overlooked primarily due to the fact that one takes into consideration the returns, neglecting entirely the asset prices.

In order to accommodate different accuracies, we introduce below a broad definition of the minimum possible return.

Definition 1. Let p_t be the asset value at time t and $c(p_t)$ be the minimum

price variation (market accuracy) associated with the value of the asset at time t. Then *the minimum possible return* of an asset at time t denoted by mpr_t is the logarithmic return that the asset will produce if its value changes by $c(p_t)$ and is given by:

$$mpr_t = \log\left(\frac{p_t + c(p_t)}{p_t}\right).$$

Note that mpr_t is the same for both upward and downward movements due to the symmetry of logarithmic returns. For the special case, that a market has a constant accuracy, say c, irrespectively of the stock price, Definition 1 can be simplified to $mpr_t = \log\left((p_t + c)/p_t\right)$ (see [11]).

For instance, let us assume that the value of an asset is equal to 0.19€ and the market operates under 0.001€ accuracy. Then the minimum possible return for the asset is 0.5%. Also, all possible (logarithmic) returns for the asset in this specific day are the integer multiples of the minimum possible return. This will have as a result, the stock movement to become even more nervous and models' failures to increase. Consequently, PVaR violations will occur more frequently and our model will almost always derive forecasts that are irrational since the stock cannot produce such returns.

Definition 2. *Low price effect (lpe)* is the inevitable increase of variance in stocks with low prices due to the existence of a minimum possible return.

Considering that the probability mass will be concentrated to the next integer multiple of the minimum possible return, we have that the volatility, provided that it exists, is increased to

$$\sum_{k=-\infty}^{+\infty} \sup_{R_t \in [k\cdot mpr,(k+1)\cdot mpr]} (R_t - \mu)^2 \int_{k\cdot mpr}^{(k+1)\cdot mpr} f(R_t)\, dR_t$$

where R_t is the return at time t and $f(\cdot)$, μ are the density and the mean of returns, respectively.

Although Definition 2 is of the same philosophy, it is slightly different from the original concept introduced by [5] who observed that low-priced stocks were characterized by higher returns and higher volatility. Such high volatility whenever observed, was attributed to increased returns although the exact opposite is the case, namely high volatility combined with the already low price results in higher returns. According to [14] the low price effect (or low price anomaly) is attributed to the low quality of stocks perceived by investors. Most results on lpe ([15], [16], [17], etc) deal with the US market although in the literature one can find some examples on other markets. Such markets include the Warsaw Stock Exchange, WSE ([18]), the Johannesburg Stock Exchange, JSE ([19] and [20]) and the Athens Stock Exchange, ASE [11]). A general conclusion that can be derived

from these results is that the decision on the cut-off point for a low-price share maybe a subjective one, but the researcher is free to consider and explore various cut-off points in search for an idealistic threshold, if it exists. For both practical and theoretical purposes any value can be adapted provided that above a predefined threshold, the market is assumed or believed to operate not as efficiently as it should have been. For the purposes of the current work, the value $\Theta = 0.001$ or 0.1% is arbitrarily chosen to play the role of the pre-assigned threshold or cut-off point.

Remark 1. It should be noted that in JSE the lpe was present for shares priced below 30 cents but the performance was not equally good for "super-low priced shares" (0 to 19 cents). For WSE the analysis was based on the 30% of stocks with the lowest prices.

Definition 3. Low price effect area is the range of prices for which the *mpr* is greater than a pre-specified threshold Θ.

It must be noted that the minimum possible return is a result of the combination of a low price with a minimum possible variation for the stock. As a result, a minimum possible return occurs.

Example 1. *We present below two examples for the low price effect area which is easily defined by applying Definition 1 for specific values of the accuracy and the threshold.*

(a) For a stock exchange market which operates under accuracy $c = c(p_t) = 0.001$*, and a predefined threshold* $\Theta = 0.001$*, we get* $p_t = 0.999001$*. Thus, the low price effect area is given by* $p_t \leq 0.999$ *and an appropriate technique to resolve the low price effect should be implemented.*

(b) Equivalently, for stock exchange markets like NYSE and NASDAQ, we would have for the same threshold $\Theta = 0.001$*, two different cases:*

- *for* $p_t \geq 1$ *implying that* $p_t = 9.995002$ *and*
- *for* $p_t < 1$ *implying that* $p_t = 0.09995002$.

Thus, for $1 \leq p_t \leq 10$ *and for* $p_t \leq 0.100$ *the low price effect is present and proper actions should be taken to minimize it. While in the range of 1 to 10 dollars numerous real life examples can be found, the same is not true for prices under 0.1 dollars, in which case, the stock has already been halted from the stock market, and hence, lpe makes only theoretical sense and becomes a theoretical concept.*

For the resulting low price effect area, the models considered should be appropriately adapted. This adaptation is done through the so called *low price correction* of both the estimation of percentage value at risk denoted by PVaR and the expected percentage shortfall (EPS) to be introduced respectively in the following two subsections.

2.1 Percentage Value at Risk (PVaR) and Low Price Correction

The Percentage Value at Risk which is a risk measure for estimating the possible percentage losses from trading assets, within a set time period, is defined as follows:

Definition 4 (a) Percentage Value at Risk p (PVaR(p)) is the 100pth percentile of the distribution of returns.
(b) Percentage Value at Risk p at time t (PVaR$_t(p)$) is the abovementioned risk measure at time t.

The probability of a PVaR$_t(p)$ violation, namely the p-th percentile of the distribution, is given by

$$p = \Pr[R_t \leq -PVaR_t(p)] = \int_{-\infty}^{-PVaR_t(p)} f(x)dx.$$

where R_t is the random variable of returns at time t, and $f(\cdot)$ is the probability density function of returns.

Since we defined PVaR$_t(p)$ based on the return distribution, no additional specification of the equation is needed on whether simple or logarithmic returns are available. Usually, the computation is done over standardized returns. Thus,

$$p = \Pr\left(\frac{R_t}{\sigma} \leq -\frac{PVaR_t(p)}{\sigma}\right)$$

where the distribution of standardized returns with standard deviation σ, (R_t/σ), is denoted by $F(\cdot)$. Hence, the PVaR of an asset takes the form $PVaR_t(p) = -\sigma F^{-1}(p)$ where $F^{-1}(p)$ is the $100p^{th}$ percentile of the assumed distribution.

Note that for the evaluation of PVaR$_t(p)$, one may consider any econometric model and then apply an estimation technique for the model parameters. Consider for instance, the general Asymmetric Power ARCH (APARCH) model $R_t = \sigma_t \varepsilon_t$ ([21]) with

$$\sigma_t^\delta = a_0 + \sum_{i=1}^{p} a_i \left(|\varepsilon_{t-i}| - \gamma_i \varepsilon_{t-i}\right)^\delta + \sum_{j=1}^{q} b_j \sigma_{t-j}^\delta,$$

where R_t is the logarithmic return at day t, ε_t a series of iid random variables, σ_t^2 the conditional variance of the model, $a_0 > 0$, $a_i, b_j, \delta \geq 0, i = 1, \ldots, p, j = 1, \ldots, q$ and $\gamma_i \in [-1, 1], i = 1, \ldots, p$.

Observe that for $\gamma_i = 0, \forall i$ and $\delta = 2$ the model reduces to GARCH model ([22]) which in turn, reduces further to EWMA model for $p = q = 1$ and for special values of the parameters involved. For the distribution F of the series $\{\varepsilon_t\}$ we consider in this work (see Section 3), the normal, the Student-t and the skewed Student-t distribution ([23]). Based on the available data, the estimator $\hat{\sigma}_t$ of the conditional standard deviation σ_t is obtained by numerical maximization of the log-likelihood according to the distribution chosen. If without loss of generality we

assume that the mean of the conditional distribution of the logarithmic returns is zero (otherwise the mean-corrected logarithmic returns could be used), then the estimator of PVaR at day t is given by $PVaR_t = q_p(F)\hat{\sigma}_t$ (see for instance [24]), where $q_p(F)$ is the $100p$-th percentile of the assumed conditional distribution F of ε_t.

For the distribution F usual choices are the normal, the Student-t and the skewed Student-t distribution ([23]). In case where the standardized returns are generated from a Student-t distribution with ν degrees of freedom, variance is equal to $\nu/(\nu-2)$, hence, it is never equal to one. If that sample variance is used in the calculation of PVaR, the PVaR would be overestimated. Volatility effectively shows up twice, both in $F^{-1}(p)$ and in the estimation $\hat{\sigma}$ of σ which is obtained by numerical maximization of the log-likelihood. Hence, we need to scale volatility as follows: $\sigma^2 \equiv \frac{\nu}{\nu-2}\tilde{\sigma}^2$ where $\tilde{\sigma}^2$ is the variance in excess of that implied by the standard Student-t.

We conclude this section with the definition of the low price correction for the Percentage Value at Risk. A $\text{PVaR}_t(p)$ estimate can take *any* real value not necessarily equal to integer multiples of mpr_t. Under the low price effect, stock movements become inevitably *more* nervous and *any* model used, would produce forecasts that are irrational in the sense that stocks cannot produce such returns, more often than it should. To resolve this "inconsistency" we propose the low price correction by rounding the $\text{PVaR}_t(p)$ estimate to the closest legitimate value, namely the next integer multiple of the mpr_t.

Definition 5. Let $PVaR_t(p)$ be the estimation of the PVaR on day t for a specific asset. The low price correction of the estimation denoted by $\widetilde{PVaR_t}(p)$ is given by:

$$\widetilde{PVaR_t}(p) = \begin{cases} \left(\left\lfloor \frac{PVaR_t(p)}{mpr_t} \right\rfloor + 1\right) \cdot mpr_t, & \text{if} \quad mpr_t \geq \Theta \\ PVaR_t(p), & \text{if} \quad mpr_t < \Theta \end{cases}$$

where $\lfloor w \rfloor$ is the floor function (integer part) of w.

Note that we prefer to deal with the percentage value at risk for comparative reasons between stocks of the same portfolio with different allocation. Observe that under the low price correction the market's accuracy is passed on to the evaluation of the PVaR resulting in a reasonable number of violations.

2.2 Expected Percentage Shortfall (EPS) and Low Price Correction

After we have obtained the $\text{PVaR}_t(p)$, we calculate now the conditional expectation under $\text{PVaR}_t(p)$, which is given by:

Definition 6. The expected percentage loss conditional on $\text{PVaR}_t(p)$ being violated, is defined by:

$$EPS_t = -E[R_t | R_t \leq -PVaR_t(p)].$$

Observe that the area under $f(\cdot)$ in the interval $[-\infty, -PVaR_t(p)]$ is less than one, implying that $f(\cdot)$ is not a proper density function any more. This can be resolved by defining the tail density (right-truncated) function $f_{PVaR}(\cdot)$, obtained by truncation on the right, so that the area below this density, becomes exactly equal to 1. Thus:

$$1 = \frac{1}{p}\int_{-\infty}^{-PVaR_t(p)} f(x)dx = \int_{-\infty}^{-PVaR_t(p)} f_{PVaR}(x)dx.$$

The EPS is then given by:

$$EPS_t = -\int_{-\infty}^{-PVaR_t(p)} x f_{PVaR}(x)dx = -\frac{1}{p}\int_{-\infty}^{-PVaR_t(p)} x f(x)dx.$$

For providing the discrete expression of EPS, we present below the discretization $f_{DPVaR}(\cdot)$ of $f_{PVaR}(\cdot)$. $f_{DPVaR}(x)$ is the probability function of a discrete random variable, having as values all the possible returns under $\widetilde{PVaR_t(p)}$, and probabilities given as follows:

$$f_{DPVaR}(x) = \frac{1}{p}\int_{x}^{x+mpr} f(x)\,dx = \frac{1}{p}[F(x+mpr) - F(x)],$$

$x = -\widetilde{PVaR_t(p)} - mpr, -\widetilde{PVaR_t(p)} - 2mpr, \ldots$.
Then, the definition of the discretization of EPS follows naturally:

Definition 7. Let EPS_t be the estimation of the EPS on day t for a specific asset. The discrete approximation of the estimation EPS_t, denoted by $DEPS_t$, is given by:

$$DEPS_t = -\sum_{x \le -\widetilde{PVaR_t(p)}} x f_{DPVaR}(x)$$

where $f_{DPVaR}(\cdot)$ is the discretization of $f_{PVaR}(\cdot)$.

Remark 2. Even though, on the previous definition we call $DEPS_t$ the discrete approximation of EPS_t, the truth is that the nature of f and by extension f_{PVaR}, is discrete, since there always exists a minimum possible return. We may treat returns as a continuous random variable when mpr is extremely small, but the discrete nature of returns still exists.

Definition 8. Let EPS_t be the estimation of EPS on day t for a specific asset. The low price correction of the estimation EPS_t denoted by $\widetilde{EPS}_t$, is given by:

$$\widetilde{EPS}_t = \begin{cases} DEPS_t, & \text{if} \quad mpr_t \ge \Theta \\ EPS_t, & \text{if} \quad mpr_t < \Theta \end{cases}$$

Remark 3. For historical simulation we take $DEPS_t = -\sum_{x \leq -\widetilde{PVaR_t(p)}} x f_{EPVaR}(x)$, where $f_{EPVaR}(\cdot)$ is the discrete empirical probability function of $f_{PVaR}(\cdot)$ given by

$$f_{EPVaR}(x) = \frac{\#\text{ of historical observations such that } x \leq R_i < x + mpr_t, \text{ for } i \leq t}{\#\text{ of historical observations such that } R_i < -PVaR_t(p)},$$

$x = -\widetilde{PVaR_t(p)} - mpr_t, -\widetilde{PVaR_t(p)} - 2mpr_t, \ldots, -\widetilde{PVaR_t(p)} \left(\left[\frac{\min_{i \leq t} R_i}{mpr_t} \right] + 1 \right) \cdot mpr_t$, where R_i are the realizations of returns.

2.3 Adjusted Evaluation Measures

As we explained earlier evaluation measures should take into account the "asymmetry in the importance" of information bared in the data. The aforementioned PVAR low price correction, since it is a percentile estimation process cannot be judged by the standard evaluation measures. Appropriate adjustment should be introduced in order to evaluate whether there is an improvement or not. In order to do so, adjusted evaluation measures are proposed that in contrast to the standard ones that have a constant cost for all estimations independently to the region that they belong, will adjust the cost based to their nature. Whether the costs will be defined a-priory by the researcher based on nature of the problem, or a-posteriory based on the distribution of the estimations that is left to the researcher. For modelling such data, we form the following setting. Let Ω be the available dataset, and A_i be a partition of Ω based on our criteria. Any family of non-empty sets is a partition of Ω, if and only if the following conditions hold.

1. $\Omega = \bigcup_{i \in I} A_i$ where A_i are non-empty subsets of Ω, $\forall i \in I = \{1, 2, \ldots, n\}$
2. $A_i \cap A_j = \emptyset$ for $i \neq j$.

The partition of the dataset will be done in two main ways, which we will call Type 1 and Type 2. Type 1 is a fixed partition, based on prior information for the dataset or the goal of the forecasting procedure. Type 2 is a dynamically adjusted partition of unit sets, based on their time proximity to the present. Both types, with respective examples, are presented below.

2.3.1 Type 1 Evaluation Measures

In order to measure the accuracy of the lpc procedure, one cannot rely on popular evaluation measures, such as the Mean Square Error (MSE), the Mean Absolute Error (MAE), the Mean Absolute Percent Error (MAPE), and the Heteroskedasticity Mean Square Error (HMSE), since PVaR estimations are percentile estimations. The fact that the proximity of the realised returns to the estimated percentile does not provide any information, creates the need for introducing appropriately adjusted accuracy measures. Note though that this is not the case for

backtesting procedures, since by concentrating no the underlining risk, only the proximity in days of violation provide information on the quality and accuracy of the proposed methodology. It is the decrease in their values which will show and verify the improved forecasting quality of the low price correction.

Let $R_{t=1}^{T}$ be a sample of a time series corresponding to daily logarithmic losses on a trading portfolio, T the length of the time series and PVaR_t the estimation of PVaR on day t. If on a particular day the logarithmic loss exceeds the PVaR forecast, then the PVaR limit is said to have been violated. For a given PVaR_t we define the indicator η_t as follows:

$$\eta_t = \begin{cases} 1 & \text{if} \quad R_t > PVaR_t \\ 0 & \text{if} \quad R_t \leq PVaR_t \end{cases}. \tag{1}$$

Under the above setting, it is easily seen, that the MSE of violation days, is defined as follows:

$$MSE = \frac{1}{\sum_{t \leq T} \eta_t} \sum_{t \leq T \& \eta_t = 1} (R_t - PVaR_t(p))^2. \tag{2}$$

One can easily observe that the above is a special weighted MSE expression. Indeed, notice that for $v(T) = \sum_{t \leq T} \eta_t$, the above can be written as

$$MSE = \sum_{t \leq T} \frac{\eta_t}{v(T)} \cdot (R_t - PVaR_t(p))^2.$$

All popular evaluation measures used in the literature, can be adjusted in order to fit our proposed methodology by generalizing the special weights used above. The general definition of the adjusted MSE is provided below.

Definition 9: Let y_i be a sample observation, $\hat{y}_i$ be the respective forecast and n the sample size. Then, the Adjusted MSE (AMSE) is given by

$$AMSE = \frac{1}{w} \sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2$$

where w_i is a pre-specified weight assigned to the i-th observation based on its importance, , and $w = \sum_{i=1}^{n} w_i$.

We also provide, under the general setting of Definition 9, the respective formulas for the Adjusted Mean Absolute Error (AMAE), the Adjusted Mean Absolute Percent Error (AMAPE), and the Adjusted Heteroskedasticity MSE (AHMSE).

$$AMAE = \frac{1}{w} \sum_{i=1}^{n} w_i |y_i - \hat{y}_i|, \quad AHMSE = \frac{1}{w} \sum_{i=1}^{n} w_i \left[\frac{y_i}{\hat{y}_i} - 1 \right]^2 \quad \text{and}$$

$$AMAPE = \frac{1}{w} \sum_{i=1}^{n} w_i \left| \frac{100(y_i - \hat{y}_i)}{y_i} \right|.$$

The adjusted measures presented in this work, provide a more general and flexible framework, regarding "asymmetry in importance data" since each observation y_i of a data set, is allowed to be assigned a pre-specified weight w_i directly related to its importance as it is recognized either by the investigator or by prior information. Consequently, the error associated with an estimation is proportional to the importance of the observation as it is specified by its weight.

Examples of the applicability of these measures can be found in risk analysis, although other examples can be presented in biosurveillance, among other settings. In the case of risk analysis, and specifically in the case of PVaR estimations, days in which a violation occurred, are obviously more important than the rest. The same logic holds for EPS. Also, we may want to give different weights to days with positive returns than the days with negative, since days with positive returns are not so important from the point of view of a risk analyst. In this case, the partition can be done as follows:

$$A_1 = \{i \in I | y_i \leq 0\},\ A_2 = \{i \in I | 0 < y_i \leq PVaR_i(p)\},\ A_3 = \{i \in I | y_i > PVaR_i(p)\}$$

where y_i are the daily logarithmic losses of a stock, and $PVaR_i(p)$ is the Percentage Value at Risk, as defined in [12]. Under this partition, weights can be given by $w_i = g_j, i \in A_j, j = 1, 2, 3$. As it is clear the weights are arbitrarily chosen which is the main weakness of Type 1 adjustment of evaluation measures. A possible choice for the weights is $g_1 = 0$, $g_2 = 1$ and $g_3 = 10$ under the assumption that days of PVaR violation, are 10 times more important than days of positive losses that did not violate PVaR, and days of negative losses (days of positive returns) are not taken into account. Analogously, in biosurveillance we have epidemic and non epidemic periods, and the failures of the model should not be evaluated in the same way for these two periods.

Remark 4. For equal weights in all observations, we obtain the standard evaluation measures. As a result, a standard evaluation measure constitutes a special case of the adjusted measures defined above.

2.3.2 Type 2 Evaluation Measures

In the second type of partition, weights are time dependent and exponentially decreasing as we go further back in time. In this case, models that are generally good but fail in the latest part of the time series, should be replaced by more appropriate ones. Weights can be given in the same fashion as in the EWMA model:

$$w_{t+1} = \lambda w_t = \lambda^2 w_{t-1} = \ldots = \lambda^{n+1} w_{t-n}$$

Under the assumption that the weights sup up to one, namely $\sum_{t=1}^{\infty} w_t = w_1 \sum_{t=1}^{\infty} \lambda^t = 1$ and for $|\lambda| < 1$, we have that $w_1 = 1-\lambda$. For a finite dataset, which is usually the case, $w_1 = \frac{1-\lambda}{\lambda - \lambda^{n+1}}$, where n is the sample size. In contrast to Type 1 adjustment, the weights in Type 2 case, are not arbitrarily chosen.

2.4 Backtesting and Method's Advantages

Another way to evaluate the forecast ability of the model, is backtesting. In this work, we will use the violation ratio VR to backtest PVaR estimations, and normalised shortfall NS for backtesting EPS estimations. The comparison between the observed frequency and the expected number of violations provides the primary tool for backtesting, known as the violation ratio ([25]).

Note that if the corrected version of PVaR is used then PVaR_t should be replaced by $\widetilde{PVaR_t}$ in (1). A PVaR violation is said to have occurred whenever the indicator η_t is equal to 1.

The comparison between the observed frequency and the expected number of violations provides the primary tool for backtesting, known as violation ratio ([25]) and is defined by:

Definition 10. Let $v(T)$ be the observed number of violations, p the probability of a violation and W_T the testing window. Then, the violation ratio VR is defined by

$$VR = \frac{\text{Observed number of violations}}{\text{Expected number of violations}} = \frac{v(T)}{p \times W_T}. \tag{3}$$

Intuitively, if the violation ratio is greater than one the risk is *underforecasted* while if smaller than one the risk is *overforecasted*. Acceptable values for VR according to Basel III accords lie in the region $(0.8 - 1.2)$ while values over 1.5 or below 0.5 indicate imperfect modeling (see [26]).

For each day t when PVaR is violated, The backtesting EPS is based on the normalized shortfall NS is calculated by $NS_t = R_t/EPS_t$ where EPS_t is the observed EPS on day t. From the definition of EPS, the expected return, given that PVaR is violated, is:

$$\frac{E[R_t | R_t \leq PVaR_t(p)]}{EPS_t} = 1.$$

Therefore, the average NS, denoted by $\overline{NS}$, given by:

$$\overline{NS} = \frac{1}{n}\sum_{t=1}^{n} NS_t = \frac{1}{n}\sum_{t=1}^{n} \frac{R_t}{EPS_t}$$

should be equal to one which in turn, formulates the null hypothesis: $H_0 : \overline{NS} = 1$. With EPS, we are testing whether the mean of returns on days when PVaR is violated is the same as the expected EPS on these days. As it is clear, it is much harder to create a formal test in order to ascertain, whether normalized EPS equals to one or not. Such a test would have to simultaneously test the accuracy of PVaR and the expectation beyond PVaR. This means that the reliability of any EPS backtest procedure is likely to be much lower than that of PVaR backtest procedures.

Remark 5. For a better understanding of the proposed methodology, the reader may refer to examples for the Athens Exchange (ATHEX) and the American Stock Exchange (NYSE MKT) analyzed and discussed in <https://labstadaaegean.wixsite.com/labstada-aegean/publications> ([12]; [13]) which due to space limitations cannot be provided here.

Remark 6. The importance of the proposed technique lies on the fact that under the low price correction fewer violations are expected to occur. Note that a VaR estimate can take *any* real value not necessarily equal to integer multiples of mpr since it is *not* controlled by the market's accuracy. This observation simply implies that *any* model almost always, derives forecasts that are irrational in the sense that stocks cannot produce such returns. Thus, due to this "inconsistency" between PVaR forecasts and stock movements, PVaR violations occur more frequently than they should.

Indeed, assume mpr_t is greater that the threshold Θ. Then, the PVaR estimate is plausible, as the asset's logarithmic return, as long as it coincides with an integer multiple of mpr. Otherwise, if $k \cdot mpr_t < PVaR_t < (k+1) \cdot mpr_t$ for some $k \in Z$, the PVaR estimate is inadmissible. Then, by definition, whenever $y_t = (k+1)mpr_t$, a violation takes place which though would have not been occurred if PVaR was measured according to the market's accuracy. Intuitively this "inconsistency" could be resolved if the PVaR estimate is corrected to the closest legitimate value, namely to a value that is in accordance with the market's accuracy. For this purpose we recommend the *low price (PVaR) correction* (see Definition 5) by rounding the PVaR estimation to the *next* integer multiple of the mpr (i.e. $\widetilde{PVaR_t} = (k+1)mpr_t$). In other words, the market's accuracy is enforced on the evaluation of PVaR and consequently fewer violations are expected to occur. Note that the rounding to the *previous* integer although legitimate, is not acceptable because in such a case the probability associated will be higher than p. On the other hand, the correction proposed is both legitimate and acceptable since it is associated with a probability *at most equal* to p.

3. CONCLUSION

In this work we have provided an overall presentation of our previously proposed low price correction technique for improved forecasts of risk measures like Value at Risk and Expected Shortfall. These improved forecasts are achieved via improved volatility estimations that were derived by taking into consideration the asset prices. The method turns out to be both simple and straightforward. EPS and PVaR estimations after the implementation of the low price correction, either remain unaffected or are improved.

Among the advantages of the proposed method is that it is versatile since it can be applied to both PVaR and EPS and any estimating method chosen by the investigator while in addition it does not require estimation of any extra

parameters. Furthermore, it should be pointed out that the violation ratio is always improved (to be specific decreased) in all scenarios. Although it is always applicable, as expected, it is extremely useful and efficient in extreme economic scenarios and severe stock collapses. It should also be noted that the estimates produced become admissible, in the sense that their values can be realized by the returns due to the fact that they are multiples of the minimum possible return. Consequently, unacceptable models become acceptable due to the fact that lpc yields the best results when models fail the most. In this case also, the error measurements are always decrease in days of violation.

Our methodology that pays attention not only to the asset return but also to the asset price, provides sufficient evidence that prices could contain important information which could, if taken under consideration, result in improved forecasts of risk estimation. Last but not least, we also provide the solution to the problem. The asset under investigation must be reverse split if possible in order to correct the above problems. Apart from any theoretical considerations, the practical implications of the low price correction entails are the key issues in the contributions of the present work.

We have also presented, adjusted evaluation measures, applicable to situations in which not all observations are equally important. The phenomenon is quite common in finance. The proposed adjusted evaluation measures are categorized in two general classes, according to the method of partition of the data. Type 1 is a fixed partition, based on prior information or the goal of the forecasting procedure while Type 2 is a dynamically adjusted partition of unit sets, based on the model failures, and their time proximity to the present.

For equal weights, Type 1 adjusted evaluation measures are simplified to standard measures. In other words, a standard evaluation measure constitutes a special case of the class of Type 1 adjusted measures. As shown in the application section, adjusted evaluation measures of Type 1, are quite useful offering additional information about the forecasting ability of models. On the other hand though, their main weakness is associated with the fact that weights are arbitrarily chosen. If though, our main goal is to narrow the evaluation on a subset of the available values, as it is needed for the correct evaluation of low price correction, then the choice of weights is straight forward and the results are quite clear.

Contrary to Type 1, weights of Type 2 decrease at an exponential rate as time goes back, which resembles the autocorrelation behaviour of assets returns. These time dependent weights provide a framework under which the most appropriate model in the latter part of the time series, is chosen. Note that the applicability of the proposed measures is not limited to finance but can be easily extended to other scientific areas where data do not necessarily carry the same amount of information, such as in physical (earthquakes, floods etc.) or climatological (heat or cold waves) phenomena. To sum up, the proposed low price correction technique possesses a number of advantages, namely:

1. The method is simple and straightforward.
2. It is general since it is applied *after* the forecast and it can be applied for every risk measure and estimation method.
3. It either leaves estimations intact when $mpr_t < \Theta$ or it improves the associated VR.
4. After the correction, we no longer observe irrational values.
5. It does not produce new extreme overestimations.
6. It is useful especially in extreme economic scenarios when stocks collapse.
7. The bigger the collapse, the better the improvement. Thus, when models tend to fail the most, the proposed method provides its best results.
8. For PVaR backtesting, VR is always improved while VaR volatility does not seem to be affected.

ΠΕΡΙΛΗΨΗ

Στα χρηματοοικονομικά, ένας από τους κύριους στόχους είναι η εκτίμηση της μεταβλητότητας, δεδομένης της καθοριστικής σημασίας της στην ανάλυση και διαχείριση των κινδύνων. Μετά από εκτενή μελέτη των στατιστικών ιδιοτήτων των χρηματοοικονομικών αποδόσεων, διαπιστώνεται η ύπαρξη διακριτής φύσης όταν παρατηρείται το φαινόμενο χαμηλής τιμής. Για τη μοντελοποίησή της εφαρμόζουμε την διακριτοποίηση της συνάρτησης πυκνότητας της ουράς με αποτέλεσμα να καθίσταται δυνατή η βελτίωση των εκτιμήσεων της Ποσοστιαίας Αξίας σε Κίνδυνο και της Αναμενόμενης Ποσοστιαίας Έλλειψης. Λόγω του ότι οι αποδόσεις δεν φέρουν την ίδια ποσότητα πληροφορίας και άρα ούτε την ίδια βαρύτητα, η αξιολόγηση της ποιότητας πρόβλεψης της προτεινόμενης μεθοδολογίας, βασίζεται σε κατάλληλα προσαρμοσμένα δημοφιλή μέτρα.

Acknowledgements: This work was completed as part of the research activities of the *Laboratory of Statistics and Data Analysis* of the University of the Aegean. The authors wish to express their appreciation to an anonymous reviewer for comments and suggestions that greatly improved the quality and presentation of the manuscript.

REFERENCES

[1]A Meucci. *Risk and Asset Allocation*. Springer Science & Business Media, 2009.

[2]A N Shiryaev. *Essentials of Stochastic Finance: Facts, Models, Theory*, volume 3. World Scientific, 1999.

[3]S-H Poon and C WJ Granger. Forecasting volatility in financial markets: A review. *Journal of Economic Literature*, 41(2):478–539, 2003.

[4]Torben G Andersen, Tim Bollerslev, Francis X Diebold, and Heiko Ebens. The distribution of realized stock return volatility. *Journal of Financial Economics*, 61(1):43–76, 2001.

[5]N H Fritzmeier. Relative price fluctuations of industrial stocks in different price groups. *Journal of Business*, 9(2):133–154, 1936.

[6]Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437, 2009.

[7]Mehdi Rajabi Asadabadi, Morteza Saberi, and Elizabeth Chang. Targets of unequal importance using the concept of stratification in a big data environment. *International Journal of Fuzzy Systems*, pages 1–12.

[8]Suliman Al-Hawamdeh. *Knowledge Management: Competencies and Professionalism*, volume 7. World Scientific, 2008.

[9]Cagdas Hakan Aladag, Erol Egrioglu, Suleyman Gunay, and Murat A Basaran. Improving weighted information criterion by using optimization. *Journal of computational and applied mathematics*, 233(10):2683–2687, 2010.

[10]S A Broda and M S Paolella. Expected shortfall for distributions in finance. In *Statistical tools for Fiance and Insurance*, pages 57–99. Springer, 2011.

[11]G-J Siouris and A Karagrigoriou. A low price correction for improved volatility estimation and forecasting. *Risks*, 5(3):45, 2017.

[12]G-J Siouris, D Skilogianni, and A Karagrigoriou. Adjusted evaluation measures for asymmetrically important data. *TRS04/2018, Lab of Statistics and Data Analysis, Univ. of the Aegean, https://labstadaaegean.wixsite.com/labstada-aegean/publications, Under Revision*, 2018.

[13]G-J Siouris, D Skilogianni, and A Karagrigoriou. Post model correction in value at risk and expected shortfall. *TRS03/2018, Lab of Statistics and Data Analysis, Univ. of the Aegean, https://labstadaaegean.wixsite.com/labstada-aegean/publications, Submitted*, 2018.

[14]J C Clenterin. Quality versus price as factors influencing common stock price fluctuations. *Journal of Finance*, 6(4):398–405, 1951.

[15]A A Christie. The stochastic behavior of common stock variances: Value, leverage and interest rate effects. *Journal of Financial Economics*, 10:407–432, 1982.

[16]H Desai and P C Jain. Long-run common stock returns following stock splits and reverse splits. *Journal of Business*, 70:409–433, 1997.

[17]S Hwang and C Lu. Is share price relevant? http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1341790. 2008.

[18]A Zaremba and R Żmudziński. Low price effect on the polish market. *e-Finanse: Financial Internet Quarterly*, 10(1):69–85, 2014.

[19]R A C Gilbertson, J F Affeck-Graves, and A H Money. Trading in low priced shares: an empirical investigation 1968-1979. *Investment Analysts Journal*, 11(19):21–29, 1982.

[20]K Waelkens and M Ward. The low price effect on the johannesburg stock exchange. *Investment Analysts Journal*, 26(45):35–48, 1997.

[21]Z Ding, C W J Granger, and R F Engle. A long memory property of stock market returns and a new model. *Journal of Empirical Finance*, 1:83–106, 1993.

[22]T Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327, 1986.

[23]P Lambert and S Laurent. Modelling financial time series using garch-type models with a skewed student distribution for the innovations (discussion paper no. 0125). *Institute de Statistique, UniversitG• Catholique de Louvain, Louvain-la-Neuve, Belgium*, 2001.

[24]M Braiome and N K Scholtes. Forecasting value-at-risk under different distributional assumptions. *Econometrics*, 4(1):1–27, 2016.

[25]S D Campbell. A review of backtesting and backtesting procedures. *The Journal of Risk*, 9(2):1, 2006.

[26]J Danielsson. *Financial Risk Forecasting*. Wiley, 2011.

Greek Statistical Institute
Proceedings of the 31st Panhellenic Statistics Conference (2018), pp. 353−365

ON A NEW APPROACH FOR IMPROVED MULTIVARIATE IMPUTATIONS

M. Spitieris
University of Aegean
michail.spitieris@ntnu.no

ABSTRACT

Missing data is a common problem that almost every statistician has to deal with in multivariate data sets. As a common approach one could choose either to omit the observations with missing data (complete case analysis), where we lose valuable information, or to replace them with estimated values from a statistical procedure (imputation methods). In this paper we propose a new methodology which relies on constrained clustering algorithms and improves the accuracy of existing imputation methods.

Keywords: Missing Data, Imputation Methods, Constraint Clustering, Linear Regression imputation, KNNimpute, missForest.

1. INTRODUCTION

Missing data is a common problem that can greatly affect the outcome of every statistical procedure. In the literature of missing data many methods have been proposed (Little & Rubin, 2002, Chapter 1.4), though their usefulness depends on the mechanisms that generate this missingness (Little & Rubin, 2002, Chapter 1.3). Imputation methods replace missing data with estimated values from a statistical procedure. The most traditional (but not always reliable) imputation methods are mean imputation and linear regression imputation. In mean imputation, the sample mean of each variable with missing values is calculated using only the recorded elements and the missing values are replaced with the corresponding mean value. The drawback of this method is that imputing values from the center of the distribution reduce the variance and consequently the covariance (Enders, 2010, Chapter 2.6), though this problem can be solved by adding a residual term randomly drawn from a normal distribution with mean 0 and variance equal to the variance of the corresponding variable. Linear Regression imputation is a method with many variations. In the simplest form, a regression model is estimated for the complete cases (observations without missing data) and the missing values are predicted by the model. The drawback of this method is that imputes values in a line, thus it creates highly correlated data (Enders, 2010, Chapter 2.7). A solution to this problem is to add a residual term randomly drawn from the resid-

uals of the estimated regression model. This method is referred to as *stochastic regression imputation*.

In recent years several reliable imputation methods have been developed. The main idea of these methods is to estimate the missing values using a non parametric regression method. This estimation can be achieved by the KNN (K Nearest Neighbors) algorithm or by a Random Forest (Breiman, 2001). The simplest performance of KNN regression is to identify the k nearest neighbors of the observation target using the Euclidean distance, where the estimation of the regression is the average of these k observations. Random Forest is a method that builts many regression trees in order to solve the problem of overfitting that a single regression tree has. The estimation of Random Forest is the average estimations of the estimation of all the regression trees. In Section 2 will describe the modifications of these two algorithms in order to estimate the missing values.

Clustering is a multivariate analysis technique which divides the dataset in k groups in such a way that observations in the same group are more similar than observations in others groups. K-means (MacQueen, 1967) is one of the most popular clustering algorithms. It is an iterative procedure that begins with k randomly chosen initial centroids and assigns each observation to the cluster that has the minimum distance from it. To calculate this distance we use the square of the Euclidean distance, thus the problem of distance minimization is transformed to a problem of variation minimization inside the clusters. It should be noted though that this algorithm can not be used in the presence of missing values. To solve this problem we often use an imputation method or we use only the variables that are fully observed (variables marginalization), thus we loose valuable information. KSC algorithm (Wagstaff, 2004) is a modified K-means which has a built in routine for dealing with missing data without using imputation or variables marginalization. The algorithm belongs to the family of the Constraint Clustering algorithms and uses the fully observed variables for Clustering and the missing variables (variables with missing values) to built pairwise constraints between observations.

The motivation of this paper is to introduce a new methodology that improves the accuracy of most of the existing imputation methods. The idea is to divide the data set in k homogeneous groups/clusters with the minimum possible variation. The data partition can be achieved by the KSC algorithm. Our recommendation relies on the fact that each group could be treated as an individual (homogeneous) data set and therefore existing imputation methods could be proved to be both reliable and accurate. The homogeneity ensures the accuracy of the proposed methodology which could be confirmed by the evaluation of any accuracy measure. In this work we apply the mean absolute error (MAE) but others work equally well. The results presented in this work are based on the simple Mean Imputation as well as the Linear Regression Imputation which are compared with two of the most prevalent methods, namely KNNimpute (Troyanskaya, 2001) and missForest

(Stekhoven, 2011). The results are based on both simulated and real data.

2. METHODS

In this section we briefly discuss imputation techniques and clustering algorithms to be combined in the next section for improved imputations. The imputation techniques to be discussed are the classical Mean and Linear Regression Imputations and the advanced KNN and Random Forest. For the clustering techniques we discuss hard- and soft-constrained clustering algorithms which allow for missing data without imputation.

Imputation Methods

K Nearest Neighbors (KNN) is a non parametric method that can be used for both classification and regression. In KNN regression the candidate observation is estimated by the mean of the k closest observations, where the closest observations can be calculated by the Euclidean distance. More specifically, KNNimpute identifies the k most "similar" completed observations, $\mathbf{x}_i^{com}$ to each missing observation and then imputes each missing value by a weighted mean of these k neighbors. To identify the neigbors the distance is calculated only for the fully observed variables, V_j^{com}. The weights are taken to be equal to the reciprocal of the distance of each missing observation. The drawback of this method is that the choice of k can greatly affect the possible outcome.

Random Forest like KNN is a non parametric method that can be used for both classification and regression. Miss Forest is a modified Random Forest algorithm which begins with a simple Mean Imputation and continues with sorting the variables in ascending order according to the number of missing values. The missing values of each missing variable V_j^{mis} are imputed by fitting a random forest. When all the variables are imputed we have the newly imputed data matrix. This procedure is repeated again until a stopping criterion is met. The stopping criterion is met when the difference between the newly imputed matrix and the previous matrix increases for the first time.

Mean imputation is the simplest imputation method, in which the sample mean of each missing variable is calculated and the missing elements are replaced by the corresponding mean value.

Linear Regression imputation is a method with many variations. In this paper we use a chained equation process (Azur, M. J., 2011). The procedure begins with a simple imputation for each missing value, for each missing variable V_j^{mis} and replaces the missing elements with a random draw from the observed values. For each replaced value we store the position. The recorded values of V_j^{mis} are regressed on the other variables in the imputed model, and the missing elements of V_j^{mis} are predicted by this model. These two steps are repeated for each missing variable. The completion of this process constitutes an iteration. The iterations

of the process which are being updated at each cycle, are usually between 5 and 10.

Constraint clustering algorithms are semi supervised learning algorithms, which they incorporate the background knowledge that we have for the data set in order to improve the accuracy of the clustering outcome. This knowledge can be incorporated as a set of pairwise constraints. Constraints are divided into two categories, hard constraints that must be satisfied by the final partition of the algorithm and soft constraints that are preferences about the final partition of the algorithm. A hard–constrained version of K–means, COP–KMEANS (Wagstaff, 2001) has been used in many real data sets, where the accuracy has reached high levels of accuracy in comparison with the simple K–means. A soft–constraints version of K–means is called SCOP–KMEANS (Wagstaff, 2002), where the constraints are created by a strength factor s which ranges from -1 to 1. Another soft– constrained version of K–means is the KSC algorithm (Wastaff, 2004). This algorithm can handle missing values without using imputation or marginalization. It divides the set of variables into V_o, the set of the fully observed variables and V_m the set of variables with missing values. In this algorithm we perceive V_m as an additional form of information that is used only for the creation of the constraints. We create a constraint $\langle d_i, d_j, s\rangle$ between the observations d_i, d_j with values v in V_m, where s is the negative Euclidean distance

$$s = -\sqrt{\sum_{v \in V_m} (d_i.v - d_j.v)^2}. \tag{1}$$

These constraints are not created for the observations that have missing values. The value of s is negative due to the fact that it shows the degree that two observations d_i, d_j should be separated. The modified objective function of KSC is

$$f := (1 - w)\frac{Var}{Var_{\max}} + w\frac{CV_d}{CV_{\max}}, \tag{2}$$

$Var = \sum_{d \in D} dist(d, \mu_i)^2$ where μ_i is the centroid of cluster i, $Var_{\max}$ is the total variance of the dataset (or the variance obtained by assigning all items in the same cluster) and it used to normalize the variance of the partition. CV is the sum of the squared strengths of violated constraints in the set of Soft Constraints (SC). $CV_{\max}$ is the sum of all squared constraints, which normalizes the quantity CV and w is a weighting factor with values ranging from 0 to 1. Following Wagstaff et al. (2004) there is no general rule for selecting the best w value, though it can be estimated by clustering with several different values for w on a small labeled subset of the data. The general idea is to choose a weight proportional to the number of variables with missing values, though if we know that one or more variables contribute more to the separation of the final partition, they should be assigned a larger weight.

In this work we incorporate the simplest imputation techniques, namely Mean Imputation and Linear Regression Imputation independently and separately for each cluster obtained via the KSC algorithm. More specifically, the proposed methodology is applied for the simple Mean Imputation in several simulated 3-dimensional data sets where the missing values have been created randomly for one variable at the 10% of the observations. Furthermore, the Linear Regression Imputation is applied on the iris data set, where the missing values have been created randomly for four different combinations in two out of four variables for 5% up to 20% proportion of missing values. These two approaches are compared with the KNNimpute method and missForest method.

3. APPROACH

Let $X_{n\times p}$ be a rectangular $n \times p$ data matrix where n denotes the number of observations and p denotes the number of variables. We assume that some elements x_{ij} of the matrix are missing and denoted by x_{ij}^{mis}, $i = 1, \ldots, n$, $j = 1, \ldots, p$, and at least one variable is complete for all elements. We denote the completely observed variable by V_j^{com} and if some elements are missing, V_j^{mis}.

Clustering algorithms try to divide the data set into k groups in such way that observations in the same group are more "similar". Our approach is to divide the data set D into k homogeneous non overlapping subsets $D_1, \ldots, D_k$, $D_i \cap D_j = \emptyset$, $\cup_{i=1}^{k} D_i = D$, and for each subset D_i, $i = 1, \ldots, k$ use an imputation method and then place the imputed values back to their initial positions in the data matrix. This partition of the data set can be achieved by the KSC algorithm (Wagstaff, 2004), which is a modified K-means that uses the missing values as an external source of information, without using imputation or marginalization. This new methodology will be demonstrated for two of the most traditional methods, Mean imputation or Regression imputation.

An issue with K-means algorithm is that the accuracy of the final partition depends on the initial choice of the centroids, that is usually random. A solution to this problem is to run K-means 20–100 times and select the partition with the smallest total within cluster variation. In KSC algorithm the objective function is modified, thus we run KSC algorithm 20–100 times and we select the partition with the smallest total within cluster objective function value.

3.1 Partition Mean Imputation

In *Partition Mean imputation* we divide the subset as described before and in D_i for each variable with missing values we calculate the average of the recorded elements and we replace the missing elements with this value. The first step to the algorithmic procedure of Partition Mean Imputation is to run KSC 20-100 times with w proportional to the number of missing variables to find the best partition. Then, each D_i is treated as an individual data set, and for $i = 1, \ldots, k$

we perform simple mean imputation. We identify the variables with missing values V_j^{mis} in each D_i, if they are any, we calculate the average of the observed values $\mu_j^{mis} = \sum_i x_{ij}^{com}/(\text{number of complete i})$ for the recorded i's and we replace the missing elements with these values. The last step is to place the imputed values back to their initial positions on the data matrix.

Algorithm 1 Partition Mean imputation

1: **Require:** $X_{n\times p}$ a data set matrix with at least one completed variable V_j^{com}
2: Run KSC-algorithm 20–100 times and identify the best partition $\{D_1, D_2, \ldots, D_k\}$
3: **for** $i \leftarrow 1$ to k **do**
4: Store the initial data matrix $X_{n\times p}$ positions for each missing element in D_i
5: Perform simple Mean Imputation in D_i
6: **end for**
7: Place the imputed data to their initial position in data matrix $X_{n\times p}$
8: **return** the imputed data matrix $X_{n\times p}^{imp}$

3.2 Partition Regression Imputation

Linear Regression imputation is one of the most popular methods with many variations. The classic form of this method is to estimate a Linear Regression model over the complete observations X^{com} and then predict the missing values X^{mis} with the estimated model. A more sophisticated version of a Linear Regression imputation is to use chained equations.

Suppose now that we have the two dimensional data set depicted in Figure 1. If we use Linear Regression imputation we see that the predictions will be far away from the real values. This new approach divides this data set into k data sets (here $k = 3$) and estimates the missing values with k different Linear Regressions. As we can see in Figure 2 we can divide the data set into 3 non overlapping data sets and estimate three different Linear Regression models. It's easy to see that this approach increases the accuracy of the Linear Regression imputation method.

This partition of the data set D into three non overlapping subsets D_1, D_2 and D_3 can be achieved by the KSC algorithm. In Partition Regression Imputation the data set D is divided into k non overlapping subsets $D_1, D_2, \ldots D_k$, each subset D_i is treated as an individual data set where the missing values are imputed by a Linear Regression model, and specifically a chained equation approach of this method. The imputed values are placed back to their initial positions in the data matrix. This algorithmic procedure is described below.

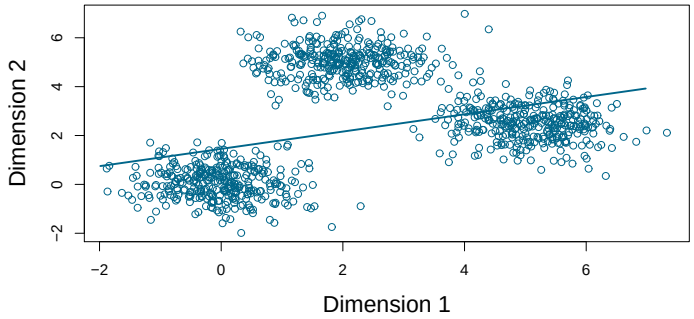


Figure 1: *Whole data set Linear Regression*

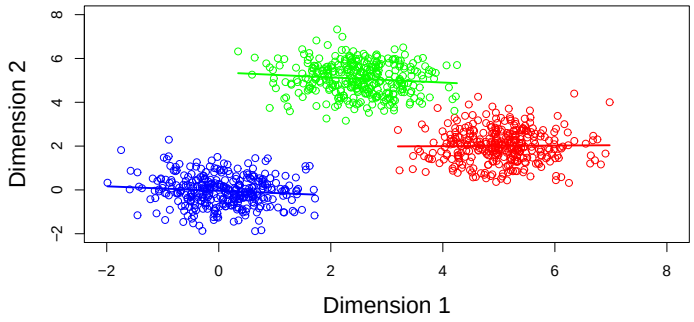


Figure 2: *Subsets Linear Regression*

Algorithm 2 Partition Regression imputation

1: **Require:** $X_{n\times p}$ a data set matrix with at least one completed variable V_j^{com}
2: Run KSC-algorithm 20–100 times and identify the best partition $\{D_1, D_2, \ldots, D_k\}$
3: **for** $i \leftarrow 1$ to k **do**
4: Store the initial data matrix $X_{n\times p}$ positions for each missing element in D_i
5: Perform chained equation Linear Regression Imputation in D_i
6: **end for**
7: Place the imputed data to their initial position in data matrix $X_{n\times p}$
8: **return** the imputed data matrix $X_{n\times p}^{imp}$

Note that when we divide the data set into k parts we estimate k different Linear Regression equations, this new approach can fix the problem of high correlated data, compared with the use of a single Linear Regression Imputation model for the whole data set. Furthermore we can add an error term as described above.

4. MAIN RESULTS

4.1 Experimental Results for Partition Mean imputation

For the experimental results we created various data sets consisted of three 3-dimensional normal distributions. We compare our method with two of the most accurate imputation methods, namely kNNimpute and missForest, for four different sample sizes. For evaluation purposes we use the mean absolute error (MAE), which is the average distance from the real values.

The data sets in Figure 3 have all 10% missing values only for one variable. In almost all cases Partition Mean Imputation is the most accurate method. Furthermore we see that as the sample size increases, the accuracy of our method in comparison with the other two methods, increases too.

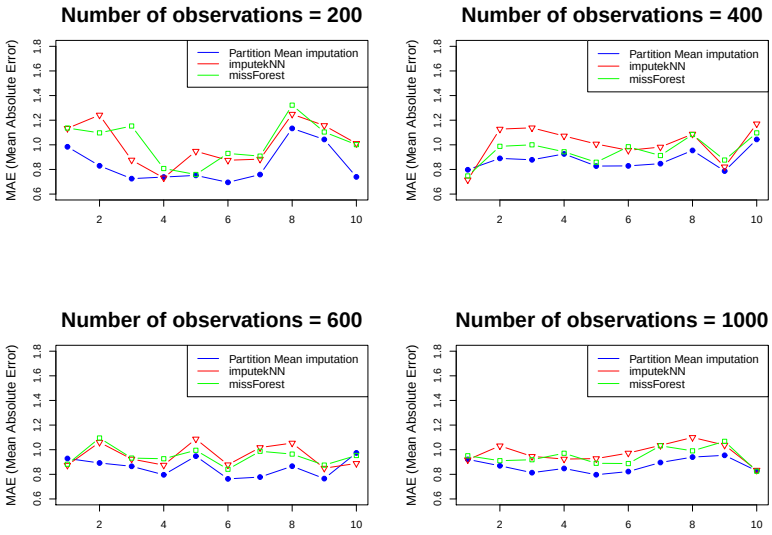


Figure 3: *Comparison of Partition Mean imputation with kNNimpute and missForest*

A common problem with clustering is that in many cases the optimal number of clusters is unknown. There are many methods for choosing the optimal k, though they are not always reliable. In order to adopt them we can use a simple imputation method to the initial data set. We found that the number of clusters is not an important issue for our method. Indeed, we selected one data set of those compared in Figure 3 and run the algorithm for various k's with the results depicted in Figure 4.

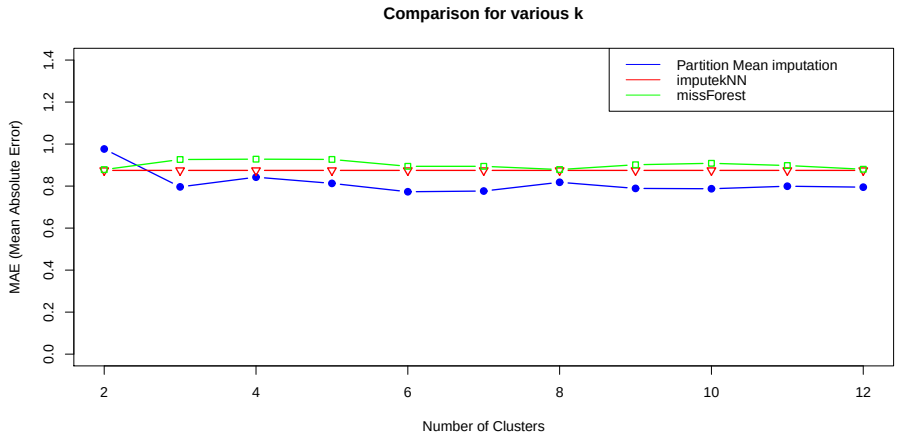


Figure 4: *Robustness to the choice of k*

Note that for all values of k that are greater than the real number of clusters (3 clusters), this method performs better, except for the value 2. Based on our experiments we concluded that all values greater than or equal to the true number of clusters are equally good choices. This is due to the fact that clustering algorithms are not used for classifying observations correctly, but instead for dividing the data set into groups with the smallest possible variation in order to achieve low variance estimations.

4.2 Experimental Results for Partition Regression Imputation

For the experimental results in this section, we use the Iris data set (consists of 150 observations and 4 variables of three distinct iris plant species). We created randomly increasing proportion of missing values for every set of two out of four variables involved. We compare Partition Regression Imputation with KNNimpute and miss Forest using as evaluation method the Mean Absolute Error (MAE). For the KSC algorithm we used a $k = 3$ and $w = 0.5$ (equal to the proportion of missing variables).

In Figure 5 we compare the Partition Regression imputation method with impute kNN and missForest using as evaluation criterion the Mean Absolute Error. In the fist plot of the Figure where the missing variables are Petal Length and Petal Width we see that for all proportions of missing values our method is the most accurate. Furthermore for the other three plots in almost all the cases the MAE of our method is smaller and consequently the accuracy is greater.

In Figures 6 and 7 we present the densities of the original (blue) and the imputed data (red) for a 10% proportion of missing values for Petal Length and Petal Width respectively.

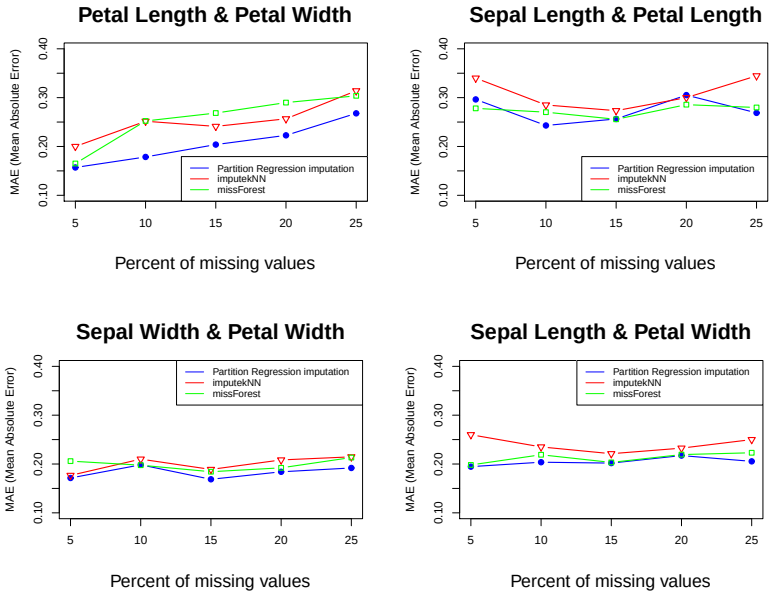


Figure 5: *Comparison of Partition Regression imputation with KNNimpute and missForest for two missing variables*

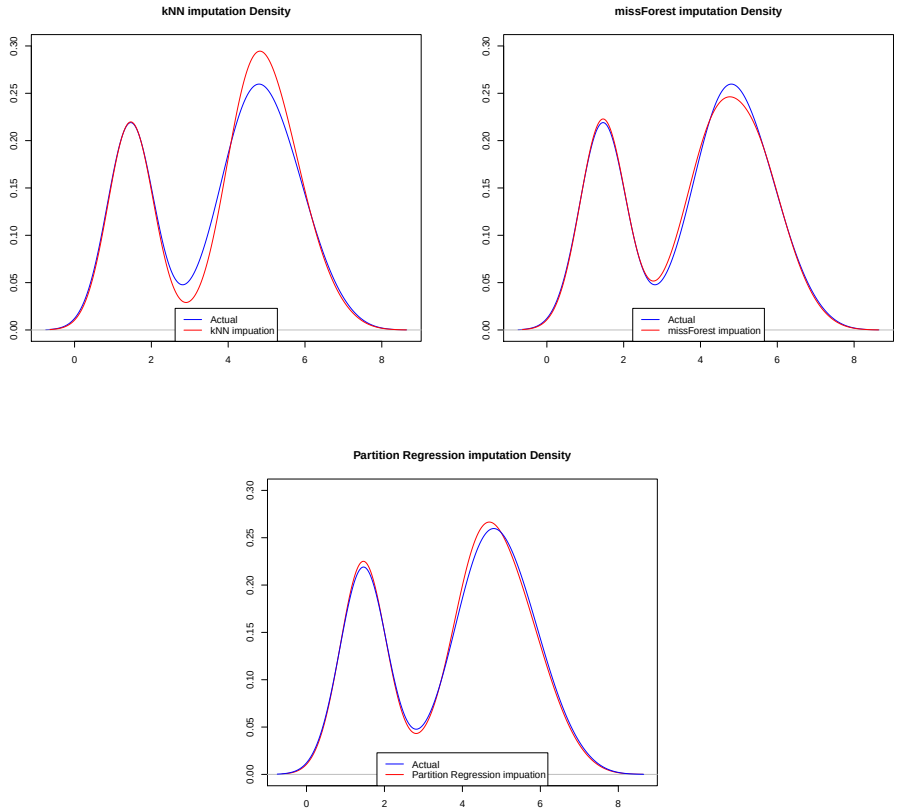


Figure 6: *Densities of the imputed data for the variable Petal Length*

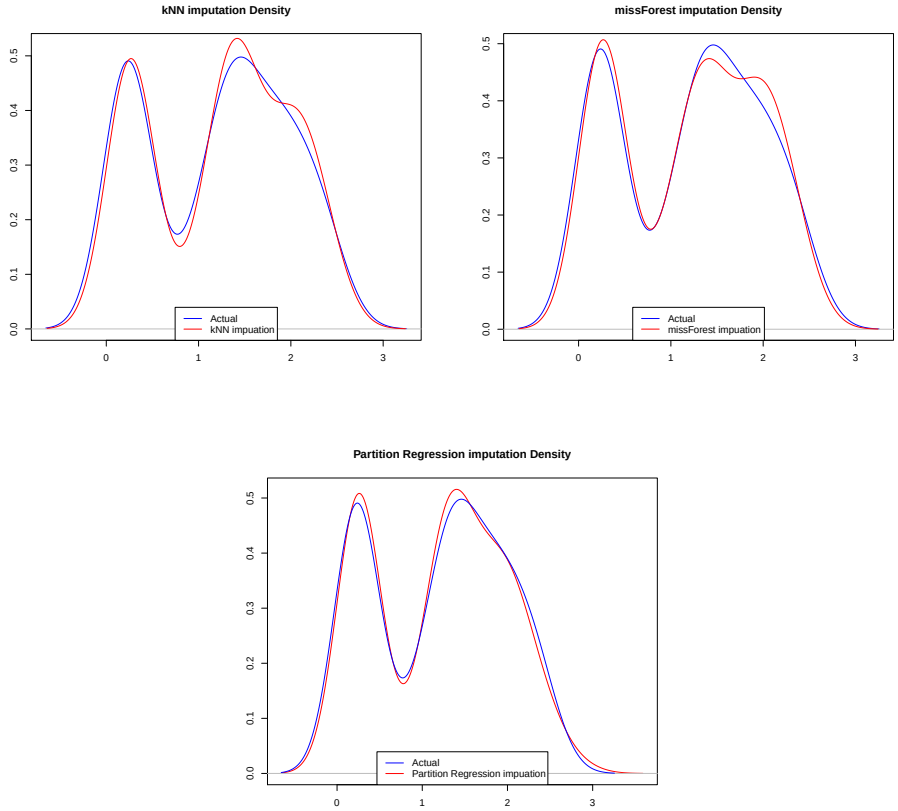


Figure 7: *Densities of the imputed data for the variable Petal Width*

5. CONCLUSION & FUTURE WORK

The methodology proposed in this work which combines a constrained clustering algorithm with imputation techniques for improved imputation is accurate enough even with the use of traditional imputation methods like mean imputation and Linear Regression imputation. Experimental results based on both simulated and real data show that both approaches outperform two of the most prevalent methods, namely KNNimpute and missForest. It should be pointed out that for our purposes associated with improved imputations, the optimal number of clusters plays a rather minor role due to the fact that the clusters are built so that the data are divided in groups with the smallest possible variation for achieving low variance estimations. Furthermore the complexity of K-means based algorithms is usually low, thus it can be used in huge data sets. In future work, we intent to explore alternative imputation methods and propose, for evaluation and comparative purposes, an accuracy measure/criterion for choosing the most accurate imputation method in each cluster separately.

ΠΕΡΙΛΗΨΗ

Τα Ελλιπή Δεδομένα αποτελούν κοινό πρόβλημα το οποίο σχεδόν κάθε ερευνητής καλείται να αντιμετωπίσει κυρίως όταν διαχειρίζεται πολυμεταβλητά δεδομένα. Οι κλασσικές τεχνικές επίλυσης του προβλήματος είτε επιλέγουν να απομακρύνουν από τα δεδομένα όλες τις παρατηρήσεις με ελλιπή δεδομένα (complete case analysis), χάνοντας με τον τρόπο αυτό πολύτιμη πληροφορία, είτε επιλέγουν να τις αντικαθιστούν με εκτιμήσεις οι οποίες προέρχονται από μία στατιστική διαδικασία (imputation methods). Στην εργασία αυτή προτείνεται μία νέα μεθοδολογία η οποία βασίζεται σε constrained clustering αλγόριθμους και βελτιώνει την ευστοχία των μεθόδων που υπάρχουν ήδη.

Acknowledgements: This work is a part of my Master Thesis that was conducted at the University of the Aegean under the supervision of Professor Alex Karagrigoriou. I would like to thank him for the suggestion of the topic; without him this Thesis would have been impossible. Working with him has been a great opportunity for expanding my knowledge and learning new things. A full copy of this Thesis is available in the library of the University of the Aegean at the following link `http://hellanicus.lib.aegean.gr/bitstream/handle/11610/18305/Master%20Thesis%20Michael%20Spitieris.pdf?sequence=1&isAllowed=y` with author name Michael Spitieris.

REFERENCES

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work?. *International Journal of Methods in Psychiatric Research,* **20(1)**, 40-49.

Breiman, L. (2001). Random forests. *Machine Learning,* **45(1)**, 5-32.

Enders, C. K. (2010). Applied missing data analysis. *Guilford Press.*

Little, R. J., & Rubin, D. B. (2002). Statistical analysis with missing data (2nd ed.). *New York: Wiley.*

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *In Proceedings of the fifth Berkeley symposium on Mathematical Statistics and Probability,* **14(1)**, 281-297.

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology,* **27(1)**, 85-96.

Stekhoven, D. J., & Bühlmann, P. (2011). MissForest–non-parametric missing value imputation for mixed-type data. *Bioinformatics,* **28(1)**, 112-118.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., ... & Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics,* **17(6)**, 520-525.

Wagstaff, K., Cardie, C., Rogers, S., & Schrödl, S. (2001). Constrained k-means clustering with background knowledge. *ICML,* **1**, 577-584.

Wagstaff, K. L.(2002). Intelligent clustering with instance-level constraints. *PhD Thesis, USA: Cornell University.*

Wagstaff, K. (2004). Clustering with missing values: No imputation required. *In Classification, Clustering, and Data Mining Applications. Springer, Berlin, Heidelberg.*, 649-658.

Wagstaff, K. L., & Laidler, V. G. (2005). Making the most of missing values: Object clustering with partial data in astronomy. *Astronomical Data Analysis Software and Systems XIV,* **347**, 172.

Greek Statistical Institute
Proceedings of the 31th Panhellenic Statistics Conference (2018), pp. 366−378

BAYESIAN ESTIMATION FOR INARCH MODELS

P.Tsamtsakiri, D.Karlis

Athens University of Economics and Business
{ptsamtsak, karlis}@aueb.gr

ABSTRACT

There is an increasing interest in models for discrete valued time series. Among them the integer autoregressive heteroskedacity (INARCH) is a model that has found several applications at the past. In the present paper we study the problem of model selection for such a model. Namely, we consider that observations conditional on the past follow a Poisson distribution where its mean depends on its past mean values and on past observations. First, parameters' estimation is done, considering a model with order (p, q). For the parameters' estimation, we use a new Bayesian approach and considering stationarity conditions as information, we define truncated normal prior distributions to satisfy them. Second, a comparison between models of different orders is done by using DIC allowing thus model selection. A real data application is given. Other related questions such as model's prediction and adjustment are discussed.

Keywords: INARCH, Poisson distribution, truncated normal, DIC, prediction, adjustment

1. INTRODUCTION

Discrete valued time series data occur in several disciplines as for example, finance, criminology, epidemiology, sports, just to name few. There is also an increasing interest for such models, see for example the recent book of Weiss (2018).

Models for univariate count time series can be split into two main categories (see, e.g. Cox (1981)). The first one is known as parameter driven models introduced by Zeger (1988) where the time autocorrelation comes from an underlying latent process for the mean of the discrete valued process. While such approaches allow to connect the models to the well known time series models for Gaussian data, the latent process is hard to be estimated and hence application of such models, despite their flexibility, is hard. The second category is the so called observation driven models where the current observations are mapped to the past observations. This allows for easier construction and estimation of the model but

due to the discreteness and the positivity of the counts special treatment is needed. In this category popular models are the Integer Autoregressive (INAR) models of Al-Zaid and Al-Osh (1987). In this class of models, the Integer Autregressive Conitional Heteroscedastic Model (INARCH) plays an important role. The model was introduced by Ferland et al. (2006) and developed further in Fokianos et al. (2009). The model has a feedback mechanism for the mean process which is related deterministically with its past values together with past observations. INARCH models and their extended versions have been proven to be very successful in modeling the volatility of count time series.

As mentioned already there are some differences between the two main families of models for discrete valued time series. Parameter driven models while produce a more flexible band of models are hard to estimate and thus observation driven models have gained more applications in practice. On the other hand observation driven models can have again a wide range of different models and properties mainly on the way the past observations influence the current ones. For a review on different approaches you can see Fokianos (2001). Fokianos et al. (2009) considered that the conditional distribution of Y_t given the past values is Poisson. Other alternative approaches where the parameter random variables $Y_1, Y_2, ..., Y_t$ are independent and the conditional distribution of Y_t is specified by a negative binomial, a generalized Poisson or a COM-Poisson distribution are proposed by Zhu (2011, 2012a, 2012b). For a zero-inflated Poisson INGARCH see Zhu (2012c). Those studies investigate the estimation of the model's parameters via classical approaches such as maximum likelihood and conditional least squares.

In this paper we apply a full Bayesian approach for estimating the Poisson INARCH model using a Markov Chain Monte Carlo approach. First, we consider the stationarity conditions. In our case, conditions of stationarity are deterministic and used for the definition of the prior distributions. Second, for a given number of nested INARCH/INGARCH models we provide Bayesian estimation via Metropolis-Hasting steps. Third, by using DIC criterion we make a choice for the best model. Implementations for the INARCH(1) and INGARCH(p, q) models where $p, q = 1, 2$, are presented.

2. The INGARCH model

Consider a time series Y_t, $t = 1, \ldots, T$, where Y_t takes values on the set of integers $\{0, 1, \ldots\}$. We consider that Y_t given the past values follows a Poisson distribution with parameter λ_t linked linearly with its past values and its past observations, namely

$$Y_t \mid \mathcal{F}_t^Y \sim Poisson(\lambda_t) \tag{1}$$

where $\mathcal{F}_t^Y$ is the σ-field generated by $\{Y_s : s \leq t\}$ and

$$\lambda_t = b_0 + \sum_{i=1}^{p} \alpha_i \lambda_{t-i} + \sum_{j=1}^{q} b_j y_{t-j}, \tag{2}$$

where b_0, b_i and α_j, $i = 1, \ldots, p$ and $j = 1, \ldots, q$ are assumed to be positive. Initial values Y_0 and λ_0 are fixed. Considering Poisson distributions, conditional mean $E[Y_t \mid \mathcal{F}_{t-1}]$ and conditional variance $Var[Y_t \mid \mathcal{F}_{t-1}]$ are equal to the parameter λ_t. The formula in equation (2) is a recurrent formula and is complicated when computational methods are used. The model is observation driven since the lambdas are not stochastic.

According to the work of Bollerslev(1986) and using the conditional expectation properties for discrete parameters

$$E[E[Y_t \mid \mathcal{F}_{t-1}]] = E[Y_t]. \tag{3}$$

We need conditions of stationarity of INGARCH models. Based on the results of Goldberg (1958) a necessary and sufficient condition for an INGARCH process to be stationary is given by

$$\sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} b_j < 1. \tag{4}$$

This equation with the combination of positivity of parameters α_i and b_j provide conditions of stationarity. This assumption is useful as a-priori information for the Bayesian estimation of the parameters when Metropolis-Hastings steps are used. Equation (2) is computationally intensive when MCMC methods are used for the parameter's estimation of low models' order because each current value of λ_t is based on its past values and observations. An alternative formula was proposed by Fokianos and Fried (2010) but is inefficient when computational methods are used. This is given by

$$\lambda_t = b_0 \frac{1 - \alpha_1^t}{1 - \alpha_1} + b_1 \sum_{i=1}^{t-1} \alpha_1^i Y_{t-(i+1)}. \tag{5}$$

3. Bayesian estimation

For the estimation of the parameters we apply a Bayesian approach. The posterior distribution can be derived as usual by

$$\pi(\theta \mid y) = \frac{f(y \mid \theta)\pi(\theta)}{f(y)} \propto f(y \mid \theta)\pi(\theta) \tag{6}$$

where denominator is the marginal likelihood and is calculated by equation

$$f(y) = \int f(y \mid \theta)\pi(\theta) d\theta \quad , \tag{7}$$

$f(y \mid \theta)$ is likelihood function and $\pi(\theta)$ is prior distribution.

The likelihood of the model was given by Fokianos et al. (2009), Fokianos and Fried (2010). It is easy to see that

$$L(\theta \mid y) = \prod_{t=1}^{T} \frac{e^{-\lambda_t} \lambda_t^{y_t}}{y_t!} \tag{8}$$

where λ_t is expressed via equation (2) or (5) and $\theta = (\theta_1, \ldots, \theta_{p+q}) = (\alpha_1, \ldots, \alpha_p, b_1, \ldots, b_q)$ is the vector of the unknown parameters. The loglikelihood function is

$$l(\theta \mid y) = \log L(\theta \mid y) = \sum_{t=1}^{T} [Y_t \log \lambda_t - \lambda_t] = \sum_{t=1}^{T} l_t(\theta). \tag{9}$$

Differentiations of this function give the score function

$$\frac{\partial l(\theta)}{\partial \theta_i} = \left(\frac{\partial \lambda_t}{\partial \theta_i} \right) \left(\frac{Y_t}{\lambda_t} - 1 \right). \tag{10}$$

In our case evaluation of $\pi(\theta \mid y)$ is analytically untractable so numerical methods are necessary for the calculation of the posterior densities. Conditions of stationarity are used for the definition of the prior distributions. So considering that $\sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} b_j < 1$ we choose as priors truncated normal distributions so as to satisfy the stationarity criteria. Namely we assume that

- $\alpha_i \sim truncatednormal(\mu_i, \sigma_i), \;\; 0 < \sum_{i=1}^{p} \alpha_i < 1$
- $b_j \sim truncatednormal(\mu_j, \sigma_j), \;\; 0 < \sum_{j=1}^{q} b_j < 1 - \sum_{i=1}^{p} \alpha_i$
- $b_0 \sim truncatednormal(\mu_0, \sigma_0), \;\; b_0 > 0$

Then the joint posterior distribution is

$$\begin{aligned} \pi(\theta \mid y) \;\; \propto \;\; & \prod_{t=1}^{T} \frac{e^{-\lambda_t} \lambda_t^{y_t}}{y_t!} \\ \times \;\; & e^{\frac{-1}{2\sigma_1^2}(\alpha_1 - \mu_1)^2} \times \cdots \times e^{\frac{-1}{2\sigma_p^2}(\alpha_p - \mu_p)^2} \\ \times \;\; & e^{\frac{-1}{2\sigma_0^2}(b_0 - \mu_0)^2} \times \cdots \times e^{\frac{-1}{2\sigma_q^2}(b_q - \mu_q)^2} \end{aligned}$$

Because joint posterior density is complex, we calculate full conditional posterior densities. For example full conditional posterior density for the parameter α_1 is

$$\pi(\alpha_1 \mid y, b_j, \alpha_i) \propto \prod_{t=1}^{T} e^{-\lambda_t} \lambda_t^{y_t} \times e^{\frac{-1}{2\sigma_1^2}(\alpha_1 - \mu_1)^2}, i = 2, \ldots, p, \;\; j = 0, 1, \ldots, q \tag{11}$$

where λ_t is expressed by (2) or (5). Those distributions have complicated forms so Metropolis-Hastings steps are used to simulate from using a proposal distribution $q(\theta_i^{can})$.

In order to implement the Mtropolis-Hastings steps, a crucial problem is the choice of proposal densities. In our work we choose as prior densities Normal distributions $N(\mu, u^2)$

1. Propose a candidate value of $\theta_i^{can} = \theta_i^{j+1}$ from a distribution with density

$$q(\theta_i^{can} \mid \theta_1^j, \theta_2^j \dots \theta_{p+q}^j) = N(\theta_i^j, u^2),$$

where u^2 is the Metropolis-Hasting step and is appropriate for the convergence of the algorithm. Considering that the interval is (0,1), in all cases $N(\theta_i^j, u^2) = N(\theta_i^j, 0.02^2 < u^2 < 0.2^2)$.

2. Next value is chosen as

$$\theta_i^{j+1} = \begin{cases} \theta_i^{can}, & p \\ \theta_i^j, & 1-p \end{cases}$$

3. Probability p is calculated as

$$p = \min\left\{1, \frac{\pi(\theta_1^{can} \mid \theta_2^j, \dots, \theta_{p+q}^j)}{\pi(\theta_1^j \mid \theta_2^j, \dots, \theta_{p+q}^j)} \frac{q(\theta_1^j \mid \theta_1^{can}, \theta_2^j, \dots, \theta_{p+q}^j)}{q(\theta_1^{can} \mid \theta_1^j, \theta_2^j, \dots, \theta_{p+q}^j)}\right\}.$$

In our case for the reason of symmetry of proposal density probability p is calculated as

$$p = min\left\{1, \frac{\pi(y \mid \theta_1^{can})\pi(\theta_1^{can})}{\pi(y \mid \theta_1)\pi(\theta_1^j)}\right\}.$$

4. Application

We illustrate the theory by presenting a real data application to estimate the parameters for each of six INGARCH models and to compare those models.

INGARCH(1,0)	$\lambda_t = b_0 + \alpha_1\lambda_{t-1}$
INGARCH(0,1)	$\lambda_t = b_0 + b_1 y_{t-1}$
INGARCH(1,1)	$\lambda_t = b_0 + \alpha_1\lambda_{t-1} + b_1 y_{t-1}$
INGARCH(1,2)	$\lambda_t = b_0 + \alpha_1\lambda_{t-1} + b_1 y_{t-1} + b_2 y_{t-2}$
INGARCH(2,1)	$\lambda_t = b_0 + \alpha_1\lambda_{t-1} + \alpha_2\lambda_{t-2} + b_1 y_{t-1}$
INGARCH(2,2)	$\lambda_t = b_0 + \alpha_1\lambda_{t-1} + \alpha_2\lambda_{t-2} + b_1 y_{t-1} + b_2 y_{t-2}$

Table 1: *All models to be used in this study*

The data consist of monthly counts of poliomyelitis cases in the United States from 1970 to 1983 (168 observations) reported by the Center for Disease Control and discussed in Zeger (1988) among others. Those data have been discussed also by Zhu (2010) for INGARCH models. So there is literature in low order time series parameter driven models. Figures 1-2 present the original series,the ACF and the PACF of the series.

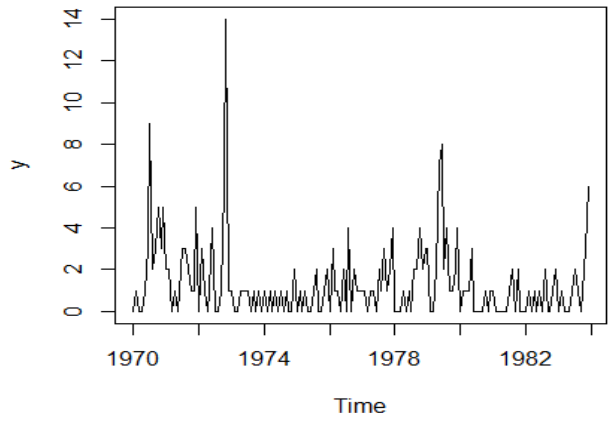


Figure 1: *Series of polio dataset*

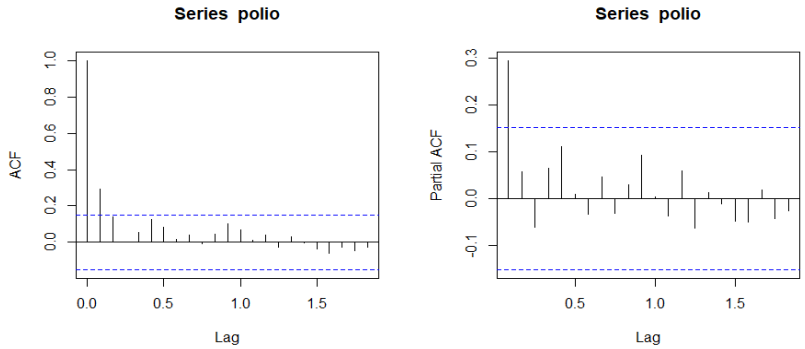


Figure 2: *ACF (left) and PACF (right) for the polio datset*

4.1 Chains' Convergence

We applied all the models from Table 1 using the MCMC described in the previous section. To save space we present convergence diagnostics and namely the ergodic mean for the parameters of the INGARCH(2,2) model in Figure 3.

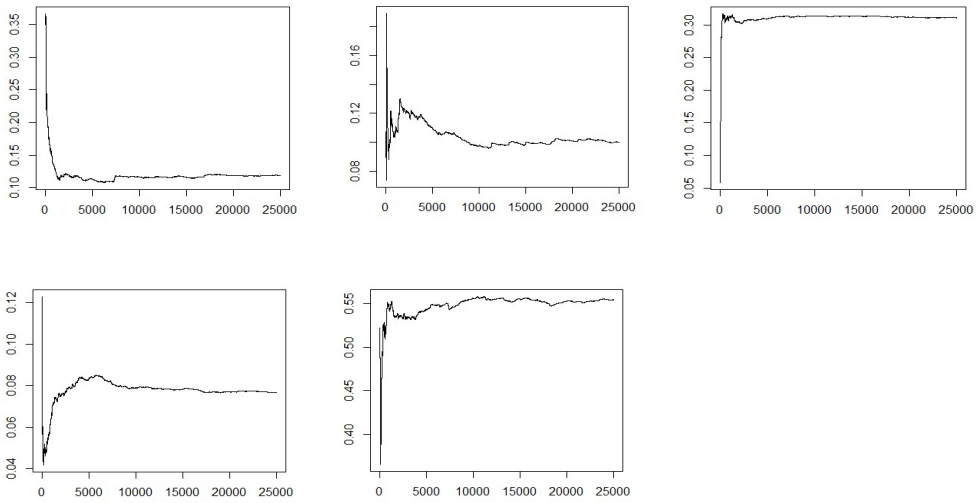


Figure 3: *Ergodic means for the parameters of the INGARCH(2,2) model*

Table 2: *Estimated posterior means, posterior medians and posterior sd for each of six INGARCH models*

Model		b_0	b_1	α_1	b_2	α_2
INGARCH(0,1)	mean	0.856	-	0.371	-	-
	median	0.853	-	0.370	-	-
	sd	0.097	-	0.065	-	-
INGARCH(1,0)	mean	0.894	0.331	-	-	-
	median	0.914	0.311	-	-	-
	sd	0.278	0.205	-	-	-
INGARCH(1,1)	mean	0.348	0.622	0.200	-	-
	median	0.347	0.623	0186	-	-
	sd	0.067	0.149	0.117	-	-
INGARCH(1,2)	mean	0.654	0.338	0.117	0.076	-
	median	0.657	0.338	0.090	0.067	-
	sd	0.131	0.068	0.099	0.052	-
INGARCH(2,1)	mean	0.546	0.325	0.192	-	0.087
	median	0.538	0.324	0.184	-	0.066
	sd	0.147	0.066	0.110	-	0.076
INGARCH(2,2)	mean	0.554	0.310	0.118	0.076	0.100
	median	0.554	0.310	0.097	0.068	0.076
	sd	0.139	0.069	0.096	0.053	0.088

The characteristics from the posterior distributions as obtained with the MCMC for all the models fitted can be seen in Table 2. The role of its parameter is somewhat different and thus no direct comparison is possible. For some of them the posterior standard deviation is large perhaps implying that the parameter can be ignored as zero is a plausible value.

4.2 Posterior densities

We depict histograms from the posterior distributions for model INGARCH(2,2) in Figure 4 to give an idea on the form of the posteriors. The dotted line is the kernel density estimate while the vertical lines are the posterior means. For parameters a the posterior is concentrated close to 0, and it has a decreasing shape. For b_2 we have much less variance than other parameters but again the posterior draws are concentrated towards 0. Parameter b_0 that relates directly to the mean and not on previous values has larger variance but a more normal shape.

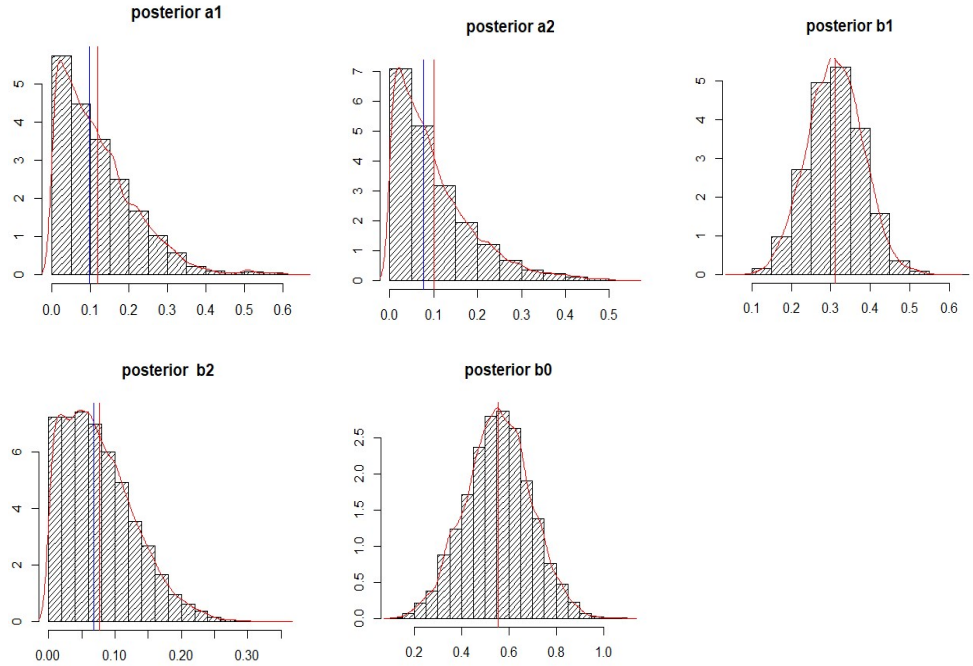


Figure 4: *Posterior densities for each parameter of INGARCH(2,2) model*

Figure 5 shows the scatter-matrix of the draws from the posterior distribution. As expected due to the stationarity conditions there is negative relationship between the parameters, some of them being close jointly to zero in many cases.

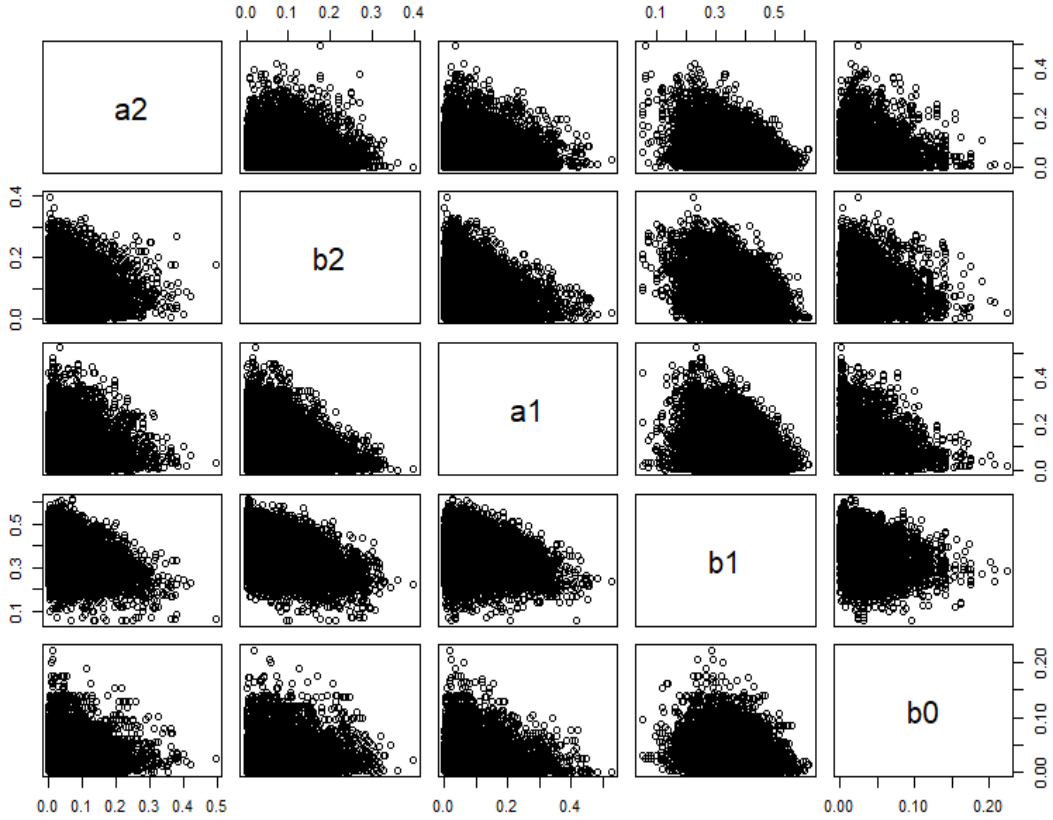


Figure 5: *Scatter-matrix based on the posterior draws for all the parameters of INARCH(2,2) model*

4.3 Model Choice

A very important aspect in time series modelling is to determine the order of the model among several candidates. This aspect has been neglected mostly for discrete valued time series. Here based on the output from several INGARCH type models we will attempt model selection based on information criteria. The idea is to identify the more suitable model among a series of candidates with different values of p and q implying different effects in the specification of the mean.

For the selection of the best model there are several criteria, perhaps the most well known being AIC, BIC, DIC criteria. In Bayesian estimation two of those are appropriate. BIC criterion is appropriate because include Bayes factor and is given by

- $BIC = k\log(T) - 2 * \log(L(y \mid \hat{\theta})) = k\log(n) - 2D(\hat{\theta})$
- $DIC = \overline{D(\theta)} + p_D = "goodness\ \ of\ \ fit" + "complexity"$

where

- k is the number of parameters
- T is the sample size
- $D(\theta) = -2\log L(y \mid \theta)$
- $p_D = E_{y|\theta}(D(\theta)) - D(E_{y|\theta}(\theta)) = \overline{(D(\theta))} + D(\overline{\theta})$

There are some differences between those criteria. BIC attempts to define the "true " model and requires specification of the number of parameters. On the other hand DIC criterion concerns predictive ability and it estimates the effective number of parameters. In our work, we are interested to make prediction as it is commonly the case for time series. Therefore we use DIC criterion for the choice of the appropriate model.

Values of DIC for all the models can be seen in Table 3. DIC selects the INGARCH(1,2) as the best model but note that INGARCH(1,1) is too close taken into account the uncertainty due to Monte Carlo. Looking back in Table 2 one can notice that the extra parameter between the two models is very small in value perhaps implying that it can be considered as 0.

Table 3: *Results based on DIC criterion, * indicates the selected model.*

Model	DIC	$\overline{D(\theta)}$	p_D
INGARCH(0,1)	283.28	279.09	1.929
INGARCH(1,0)	320.92	318.84	1.03
INGARCH(1,1)	282.11	276.81	2.65
INGARCH(1,2)	282.01*	276.73	2.63
INGARCH(2,1)	283.28	277.901	2.689
INGARCH(2.2)	283.26	278.260	2.501

4.4 Prediction

Forecasting a future (or missing) value of an observation y_f is based on the predictive distribution

$$f(y_f \mid y) = \int f(y_f \mid \theta)\pi(\theta \mid y)d\theta. \tag{12}$$

This integral is analytically untractable. Therefore computational methods could be used for this calculation.

In this work we use posterior means as estimators and we fit them at the equation of λ_t. In Figure 6 we plot the estimated means based on (2) from five candidate models. The reported values are the averages over all the draws from the posterior for each model. The differences are very small as one can see in this Figure. The means are following closely the data indicating a rather good fit for most of the model (recall that their DIC were quite close) apart from some models with few parameters that give different predictions.

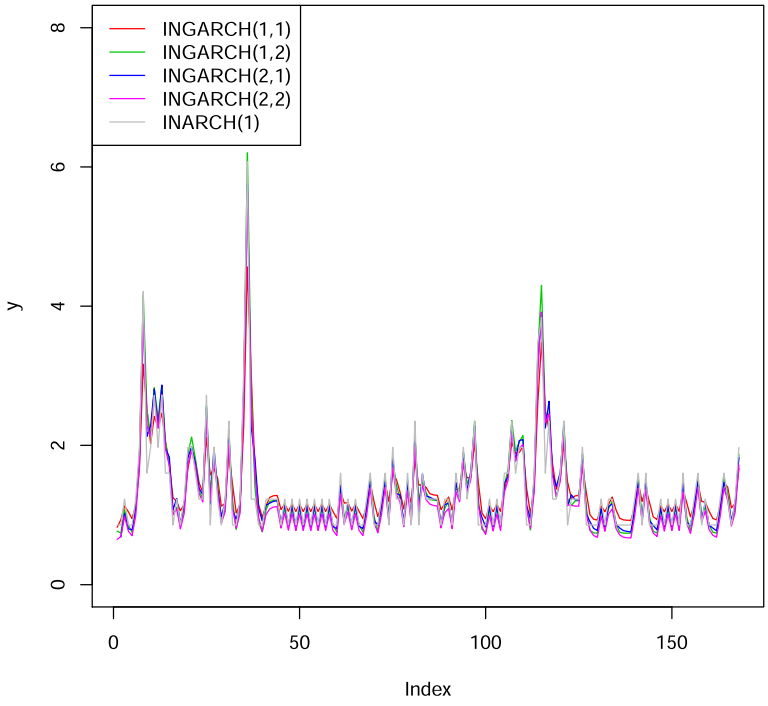


Figure 6: *Fitted mean values from five INGARCH(p,q) values. Red= INGARCH(1,1), Green= INGARCH(1,2), Blue = INGARCH(2,1), Purple=INGARCH(2,2), Grey= INARCH(1)*

5. Further Work

In this paper we presented Bayesian estimation and model selection for the family of INGARCH models using MCMC approach. The existing papers used either maximum likelihood or conditional least square approaches, so our method is novel to this extend. Bayesian approach allows to incorporate prior information, including stationarity conditions. A real data application to the well known polio

data set is given performing also model selection based on DIC.

There are quite a lot of interesting issues for further exploitation. In our future work we aim to contribute into several topics. For example, in this paper we treated only the linear model for the feedback equation. Fokianos et al. (2011) proposed a logarithmic model that removed the assumption of positive parameters and thus making the model and its estimation more realistic. The model assumes for the simple case $q = p = 1$ that

$$Y_t \mid \mathcal{F}_t^Y \sim Poisson(\lambda_t)$$

$$\nu_t = \log \lambda_t = b_0 + \alpha \nu_{t-1} + b \log(1 + Y_{t-1}) \tag{13}$$

and hence is log-linear in the sense that it models the logarithm of the mean in a suitable way. In a certain sense is more natural because of the discrete and positive nature of the count time series data. For this model, for example, we do not need to restrict the parameters a and b to be positive as in our model and this perhaps can be quite helpful to set up the MCMC approach (no need to enforce constraints in the prior). On the other hand the choice to inflate the zero values in order to be able to take the logarithm can be arbitrary.

In addition, further computational improvements are needed to speed up the proposed MCMC. Furthermore, in the present paper model selection was based on DIC and a series of fitted models. As an alternative one may implement a trans-dimensional MCMC. A trans-dimensional MCMC that could visit all candidate models and jumps between them could perform model selection in a more concrete way. Finally extension to bivariate INGARCH models like that in proposed by Liu(2016) is relatively straightforward.

6. ΠΕΡΙΛΗΨΗ

Στην παρούσα εργασία μελετάται το πρόβλημα μοντελοποίησης ενός αυτοπαλινδρομούμενου ετεροσκεδαστικού μοντέλου ($INARCH$) χρονοσειρών που μπορούν να πάρουν μόνο διακριτές τιμές. Στο συγκεκριμένο μοντέλο θεωρούμε ότι οι παρατηρήσεις ακολουθούν την κατανομή Ποισσον και ότι η μέση τιμή βασίζεται σε προηγούμενες τιμές της αλλά και σε προηγούμενες παρατηρήσεις. Αρχικά γίνεται εκτίμηση των παραμέτρων του μοντέλου, θεωρώντας ένα μοντέλο τάξεως (p, q), όπου τα p, q παίρνουν διάφορες τιμές έως 2. Για την εκτίμηση των παραμέτρων χρησιμοποιούμε την Μπευζιανή προσέγγιση και λαμβάνοντας υπόψη τις συνθήκες στασιμότητας ως πληροφορία, ορίζουμε ως $a - priori\ \ truncated\ \ normal$ κατανομές. Στην συνέχεια γίνεται σύγκριση των μοντέλων διαφορετικών τάξεων μεταξύ τους με βάση την εφαρμογή τους σε πραγματικά δεδομένα και με χρήση του κριτηρίου DIC γίνεται η επιλογή του καταλληλότερου μοντέλου για τα συγκεκριμένα δεδομένα. Θέματα σχετικά με την πρόβλεψη αλλά και την προσαρμογή του μοντέλου επίσης περιγράφονται.

Λέξεις κλειδιά: $INARCH$, κατανομή $Poisson$, $truncatednormal$, DIC, πρόβλεψη, προσαρμογή

7. References

1. Al-Osh M.A., Alzaid A.A.. (1987). First-Order Integer-Valued Autoregressive Process, Journal of Time Series Analysis, 8, 261–275.
2. Bollerslev T. (2016). Generalized autoregressive conditional heteroskedasticity, Journal of Econometrics, 31, 307–327.
3. Cox, D. R., Gudmundsson, G., Lindgren, G., Bondesson, L., Harsaae, E., Laake, P., & Lauritzen, S. L. (1981). Statistical analysis of time series: Some recent developments [with discussion and reply]. Scandinavian Journal of Statistics, 93–115
4. Doukhan P., Fokianos, K., Stove, B., Tjostheim, D. (2017) Multivariate Count Autoregression. arXiv preprint arXiv:1704.02097
5. Ferland, R., Latour, A., & Oraichi, D. (2006). Integer-valued GARCH process. Journal of Time Series Analysis, 27(6), 923-942.
6. Fokianos, K. (2012) Count time series models. Time Series Methods and Applications, In Handbook of Statistics, Volume 30, 315-347.
7. Fokianos,K., Rahbek, A. & Tjostheim, D. (2009). Poisson autoregression, Journal of the American Statistical Association, 104, 1430–1439.
8. Fokianos, K. & Tjostheim, D. (2011). Log-linear Poisson autoregression. Journal of Multivariate Analysis, 102, 563–578.
9. Goldberg S. (1958). Introduction to Difference Equations Wiley, New York.
10. Liu, H. (2012). Some models for time series of counts. PhD Thesis, Columbia University
11. McCabe, B.P.M. & MartinG.M. (2005). Bayesian predictions of low count time series,International Journal of Forecasting, 21, 315–330
12. Weiss, C. H. (2018). An Introduction to Discrete-valued Time Series. John Wiley & Sons.
13. Zeger, S. L. (1988). A regression model for time series of counts. Biometrika, 75(4), 621-629.
14. Zhu, F. (2011). A negative binomial integer-valued GARCH model, Journal of Time Series Analysis, 32, 54–67
15. Zhu, F. (2012a). Modeling time series of counts with COM-Poisson INGARCH models, Mathematical and Computer Modelling, 56, 191–203
16. Zhu, F.(2012b). Modeling overdispersed or underdispersed count data with generalized Poisson integer-valued GARCH models, Journal of Mathematical Analysis and Applications, 389, 58–71
17. Zhu, F. (2012c). Zero-inflated Poisson and negative binomial integer-valued GARCH models. Journal of Statistical Planning and Inference, 142(4), 826–839.
18. Zhu, F. (2012d). Modeling time series of counts with COM-Poisson INGARCH models. Mathematical and Computer Modelling, 56(9-10), 191–203.

Greek Statistical Institute
Proceedings of the 31st Panhellenic Statistics Conference (2018), pp. 379−392

SYMMETRICALLY ADAPTED DATA SUMMARIES

M. Viana
University of Illinois at Chicago
viana@uic.edu

ABSTRACT

The methodology for deriving the symmetrically adapted summaries of certain types of measurements is reviewed and illustrated with applications to measurements of visual fields.

Keywords: Automorphisms, projections, group algebra, permutations, invariance.

1. INTRODUCTION

Data analysis often requires some form of reduction or summary of sets of observations. For example, all forms of analyses of variance lead to a decomposition of the sum of squares $||x||^2$ into components associated with the different experimental effects and error terms. This is the result of a decomposition

$$\mathcal{V} = \mathcal{V}_1 \oplus \ldots \oplus \mathcal{V}_r$$

of the data (real vector) space $\mathcal{V}$ into a direct-sum of invariant subspaces $\mathcal{V}_\xi$, in the sense that

$$\rho(\mathcal{P}_\xi x) = \mathcal{P}_\xi(\rho x),$$

for all $x \in \mathcal{V}$, where the $\mathcal{P}_\xi$ are projection operators associated with the subspaces $\mathcal{V}_\xi$ and ρ is any one of permutation symmetries defined by experimental design and sampling considerations.

The commutativity $\rho\mathcal{P} = \mathcal{P}\rho$ of projection operators and their defining symmetries implied above, and the association of such operators to properties of experimental interest are of fundamental importance in the natural sciences e.g., Weyl (1952, p.81), Schatz and Ratner (1993, p.27), Omnès (1994, p.83), as well as in many statistical arguments, e.g., Dawid (1985), Dawid (1988). The role of group-theoretical methods in statistics and probability, as we know, has a long history dating back to contributions from Fisher (1942), James (1954), and James (1957) among others. A brief account of contributions from that initial period can be seen in Viana and Richards (2001) and also in Viana and Wynn (2009).

Decompositions of the data space into invariant subspaces can be obtained from a number of, essentially equivalent, versions of the canonical projections theorem e.g., Serre (1977, p.21) stating that for an arbitrary representation ρ of a finite group G into real or complex vector space $\mathcal{V}$ in corresponding dimension of n with irreducible characters ξ, the linear $n \times n$ operators

$$\mathcal{P}_\xi = \frac{n_\xi}{g} \sum_{\tau \in G} \xi_{\tau^{-1}} \rho_\tau, \tag{1}$$

where g is the order of G and n_ξ is the dimension of ξ, are algebraically orthogonal projections decomposing the $n \times n$ identity operator. That is, $\mathcal{P}_\xi \mathcal{P}_\eta = \mathcal{P}_\eta \mathcal{P}_\xi = 0$ for $\eta \neq \xi$, $\mathcal{P}_\xi^2 = \mathcal{P}_\xi$ and $I = \sum \mathcal{P}_\xi$, allowing for the decomposition

$$||x||^2 = x'x = x'Ix = \sum x'\mathcal{P}_\xi x$$

of the sum of squares into algebraically orthogonal components which, under certain parametric assumptions will also be statistically independent (χ^2 distributions), thus allowing for the usual F-tests for the experimental main effects and other terms in the decomposition, relative to the error components. This is the usual analysis of variance setup.

Conversely, as effectively suggested by Weyl (1952, p. 144), *Whenever you have to do with a structure-endowed entity try to determine its group of automorphisms, the group of those element-wise transformations which leave all structural relations undisturbed. You can expect to gain a deep insight into its constitution this way.*

In quantum chemistry, these projections, indexed by the irreducible representations ξ, play a central role in determining whether a chemical bonding can take place in a molecule. The bonding, points Riley et al. (2002, p. 948), is strongly dependent on whether the wavefunction of the two atoms forming a bond transforms according to the same (irreducible) representation. Typically, these reductions take place in a infinite dimension Hilbert space $\mathcal{H}$, in such a way that the invariant subspaces in the reduction of $\mathcal{H}$ define the **properties** of the quantum system. Consequently, properties are identified with the corresponding projections - one for each irreducible representation - which follow from the physical basis of the system.

When the projection I corresponds to the sure property, whereas the projection $I - \mathcal{P}$ corresponds to the negation of the property associated with $\mathcal{P}$. The properties associated with two commuting projections $\mathcal{P}$ and $\mathcal{F}$ are said to be compatible, in which case the projection $\mathcal{PF} = \mathcal{FP}$ represents the conjunction of $\mathcal{P}$ and $\mathcal{F}$, whereas the projection $\mathcal{PF} + \mathcal{P}(I - \mathcal{F}) + (I - \mathcal{P})\mathcal{F} = \mathcal{P} + \mathcal{F} - \mathcal{PF}$ is associated with the disjunction of the two properties. If, in addition, $\mathcal{PF} = 0$, the properties are compatible, mutually exclusive, and the disjunction is given by the sum $\mathcal{P} + \mathcal{F}$. See, for example, Omnès (1994) or Faris (1996) for a review of Omnès' work.

Unit norm vectors y in the Hilbert space are associated with the *states* of the system, and determine a mathematical specification of probabilities for all properties. These probabilities are obtained from the fact that associated to each set of mutually exclusive properties $\mathcal{P}_1, \ldots, \mathcal{P}_h$ whose disjunction is sure, i.e: $I = \mathcal{P}_1 + \ldots \mathcal{P}_h$, there is a decomposition

$$1 = ||y||^2 = y'y = y'\mathcal{P}_1 y + \ldots y'\mathcal{P}_h y,$$

which is interpreted as a probability distribution among the corresponding properties $\mathcal{P}_1, \ldots, \mathcal{P}_h$. Each state $y \in \mathcal{H}$ then provides a probabilistic description,

$$y \to (y'\mathcal{P}_1 y, \ldots, y'\mathcal{P}_h y) = (||\mathcal{P}_1 y||^2, \ldots, ||\mathcal{P}_h y||^2),$$

of the system, where $y'\mathcal{P}y$ is interpreted as an appropriate inner product in $\mathcal{H}$. The projections on one-dimensional subspaces are *pure states*, characteristic of the only one-dimensional representations of the symmetric groups: the symmetric (stable under all permutations) and the anti-symmetric (stable under even permutations and changing sign under odd permutations) representation. For example, $(x, y, z) \mapsto xyz$ reduces symmetrically whereas $(x, y, z) \mapsto (x - y)(x - z)(y - z)$ reduces anti-symmetrically when acted upon by all permutations of the components $\{x, y, z\}$. This simple fact has fundamental implication in the classification of elementary particles.

This group-theoretic approach, although far-reaching and appealing, is not necessary and rarely used to introduce the methods of analysis of standard experimental designs. The reason is that the underlying reductions are mostly the symmetric ones and those associated with their orthogonal complement in the representation space. The symmetries are the full set of permutations for rows, columns, and sampling units in the experiment. Formally, what we have here are full symmetric groups (S_n) acting on the components (labels) of experimental designs and on their sampling units in a way that the resulting ANOVAs decompositions are then tensor products of standard (full-symmetric) decompositions

$$I = \mathcal{A} + \mathcal{Q} \tag{2}$$

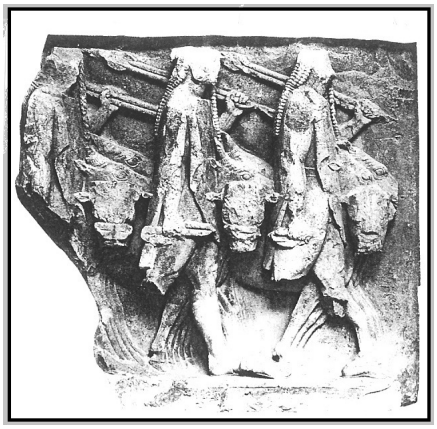
in which the $\mathcal{A}$ matrices are $m \times m$ matrices with all entries equal to $1/m$ and $\mathcal{Q} = I - \mathcal{A}$. Although (1) with $G = S_n$ and an arbitrary representation ρ would lead to a richer decomposition

$$I = \mathcal{P}_1 \oplus \mathcal{P}_2 \ldots \oplus \mathcal{P}_r,$$

it so happens that when the representation is that of S_n acting on $L = \{1, 2, \ldots n\}$, as in the sampling case, all projections $\mathcal{P}$ arising from (1) but $\mathcal{A}$ and $\mathcal{Q}$ are null. In this simplest design possible, the action of S_n on $L = \{1, 2, \ldots, n\}$ leads to

$$||x||^2 = x'\mathcal{A}x + x'\mathcal{Q}x = n(\bar{x})^2 + \sum (x_i - \bar{x})^2.$$

Figure 1: The cattle raid of the Dioskouroi. Detail from the Treasury of the Sikyonians (Delphi, c. 560 BC.



This is why, we could say, the sample mean and the sample variance are the *twin gods* ($\Delta\iota o\sigma\kappa o\upsilon\rho o\iota$) of statistics!

There is in any statistical analysis, the potential for design symmetries ($\mathcal{G}_d$), sampling symmetries ($\mathcal{G}_s$), and measurement symmetries ($\mathcal{G}$). In this brief review only the reduction of **measurement data** by symmetrically adapted methods will be emphasized, independently of sampling and design symmetry considerations, which are often derived from products of permutation groups in the full symmetric group. We only briefly address the subsequent aspects of actual inference, the elements of which should be familiar to the reader. More specifically, we shall not derive the distribution of the symmetrically adapted summaries, which will depend on the parametric assumptions imposed on the data. The summaries, however, do NOT depend on such considerations.

The focus here is in the role of certain symmetries $\mathcal{G}$ suggested by the measurements (and instrumentation) in generating a decomposition

$$\mathcal{G}_{\text{inst}} \implies I = \mathcal{P}_1 \oplus \mathcal{P}_2 \ldots \oplus \mathcal{P}_r$$

for measurements in which the salient symmetries can be identified.

In general, then, once the corresponding decompostions

$$\mathcal{G}_{\text{design}} \implies I_{\text{design}} = \mathcal{P}_1 \oplus \mathcal{P}_2 \ldots \oplus \mathcal{P}_m;$$

$$\mathcal{G}_{\text{sampling}} \implies I_{\text{sampling}} = \mathcal{A} \oplus \mathcal{Q};$$

$$\mathcal{G}_{\text{inst.}} \implies I_{\text{inst.}} = \mathcal{P}_1 \oplus \mathcal{P}_2 \ldots \oplus \mathcal{P}_r,$$

are obtained, they can be combined to decompose the overall data space identity matrix $\mathcal{I}$

$$\mathcal{I} = I_{\text{inst.}} \otimes I_{\text{design}} \otimes I_{\text{sampling}}$$

in conformity with the canonical projections theorem. The resulting product decomposition will be a more detailed one and yet retaining the canonical properties: a decomposition of the overall identity matrix into a sum of algebraically orthogonal projections. In the following sections we will emphasize the components derived from specific measurements, with the understanding that the arguments would apply to any other class of measurements.

2. MEASUREMENTS AND SYMMETRY

Measurements, such as the base quantities and dimensions used in the International System of Units, such as length, mass, time, duration, electric current, thermodynamic temperature, amount of substance, and luminosity, are always the quantification of an object or a person under certain conditions and specific instruments. A scalar (univariate) measurement can be viwed as a mapping

$$x : \ell \in L \mapsto x_\ell \in \mathbb{R}$$

in which its image x_ℓ is the observed **data** and its domain L is a set of reference **labels** (ℓ).

When statisticians underplay the role of measurements, they *face the risk of becoming like geometers who study space and never measure it*, as remarked by D. Lindley in a plenary address to the 4th Valencia International Meeting on Bayesian Statistics, Peñiscola, Spain, in 1991.

Every measurement, often through its instrumentation, is potentially subjected to two basic types of functional **actions**:

$$\Lambda \xrightarrow{\sigma} L \xrightarrow{x} \mathbb{R} \xrightarrow{s} \mathbb{R}.$$

The action $(sx)(\ell) = s(x_\ell)$ is the subject of measurement theory, e.g., Krantz et al. (2007), Suppes et al. (2007), Luce et al. (2007) whereas the action $(\sigma x)(\ell) = x_{\sigma\ell}$ is of interest to symmetry studies as in, for example Viana (2008), Viana and Lakshminarayanan (2013), Viana and Lakshminarayanan (2019) and will be reviewed here. Moreover, two basic conditions are to be distinguished: One in which the symmetries $\sigma \in G$ act on set Λ and the other, to be considered here, in which $\Lambda = L$ so that the symmetries are the elements $\sigma \in \mathrm{Aut}(L)$, that is, certain permutation groups suggested by the set L of labels. We shall also restrict our review to commutative groups and to global symmetries. Local symmetries are discussed in Viana and Lakshminarayanan (2019) and the non-commutative cases in, for example, in Viana and Lakshminarayanan (2013).

Again, we find it useful to remark that it is only when the set L has a certain *structure* recognizable or made available to the statistician that the arguments presented here can be considered. Querying the history of the underlying measurements before analyzing the ensuing data is therefore a good data analysis strategy, in particular because the set L of measurement labels is often hidden

in such history. In the following section the relevance of this set of labels will be demonstrated.

3. VISUAL FIELD MEASUREMENTS

Measurement of visual field is a common clinical measurement called perimetry in which the subjects record their own responses to certain visual stimuli. Perimetry is used to assess the integrity of the visual pathways (retina to brain) using non-invasive psychological methods of measurement. The two dimensional map of the measured visual field can then be used in localizing an eventual lack of integrity.

Automated perimetry measures the sensitivity (what you measure) of the retina at various locations (where you measure), which often form pattern of points (the labels) as in (3) yielding sensitivity measurements (in dB units) such as those in (4).

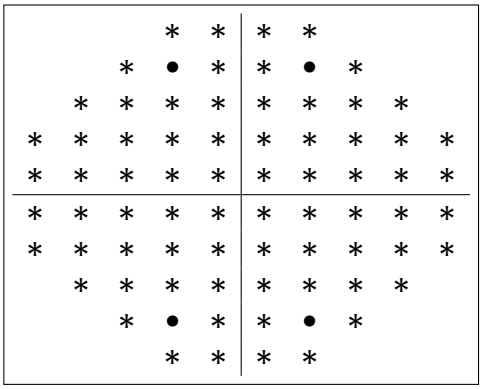
$$L = \begin{array}{|ccccccccccc|}\hline & & & * & * & * & * & & & & \\ & & * & * & * & * & * & * & & & \\ & * & * & * & * & * & * & * & * & & \\ * & * & * & * & * & * & * & * & * & * & \\ * & * & * & * & * & * & * & * & * & * & \\ * & * & * & * & * & * & * & * & * & * & \\ * & * & * & * & * & * & * & * & * & * & \\ & * & * & * & * & * & * & * & * & & \\ & & * & * & * & * & * & * & & & \\ & & & * & * & * & * & & & & \\ \hline \end{array} \tag{3}$$

$$x(L) = \begin{array}{|ccccccccccc|}\hline & & & 26 & 27 & 24 & 21 & & & & \\ & & 23 & 23 & 26 & 28 & 25 & 22 & & & \\ & 27 & 29 & 27 & 27 & 28 & 29 & 29 & 28 & & \\ 28 & 29 & 29 & 29 & 31 & 20 & 20 & 28 & 28 & 29 & \\ 26 & 26 & 29 & 30 & 33 & 34 & 28 & 27 & 28 & 27 & \\ 28 & 30 & 30 & 31 & 32 & 33 & 33 & 4 & 27 & 28 & \\ 27 & 29 & 31 & 31 & 32 & 32 & 30 & 29 & 28 & 27 & \\ & 30 & 31 & 31 & 28 & 29 & 32 & 30 & 30 & & \\ & & 28 & 32 & 29 & 28 & 29 & 28 & & & \\ & & & 26 & 27 & 28 & 29 & & & & \\ \hline \end{array} \tag{4}$$

The equivalent concept in optics is the field of view of the optical system. The result, namely visual field or field of view, has a particular geometry determined by the design of the instrument. The measurement domain therefore is obtained from a pattern of labels L such as shown by the 76 points in (3) leading to the field data $x(L)$ in (4).

The set L is amenable to a number of symmetries when embedded as a lattice of planar points. Set L is amenable (invariant or stable) to the action of a vertical (relative to the readers view of this page) line reflection (v), as well as to a horizontal (h) line reflection, and to the resulting central inversion or 180 deg rotation (r):

$L =$



The resulting automorphism group of L is then

$$\mathcal{G} = \{1, v, h, r\},$$

in which the multiplication is composition of the corresponding planar reflections and inversion. Its multiplication table is clearly

$\mathcal{G}$	1	v	h	r
1	1	v	h	r
v	v	1	r	h
h	h	r	1	v
r	r	h	v	1

, (5)

where we observe that the composition of two symmetries is also a symmetry, the composition is associative, the identity element 1 is an element in G and each element in G has an inverse element also in G. The commutative is evident.

$x(L) =$

			26	27	24	21			
		23	23	26	28	25	22		
	27	29	27	27	28	29	29	28	
28	29	29	29	31	20	20	28	28	29
26	26	29	30	33	34	28	27	28	27
28	30	30	31	32	33	33	4	27	28
27	29	31	31	32	32	30	29	28	27
	30	31	31	28	29	32	30	30	
		28	32	29	28	29	28		
			26	27	28	29			

(6)

The data $x(L)$ along with an $\mathcal{G}$-orbit $\mathcal{O}_x$ is shown in (6), where we recall that

$$\mathcal{O}_x = \{(x_{\sigma\ell})_\ell : \sigma \in \mathcal{G}\}.$$

Therefore,

$$\mathcal{O}_x = (x_1, x_v, x_r, x_h) \simeq (23, 25, 29, 32),$$

is a **set of data indexed or labeled by the symmetries in** $\mathcal{G}$, allowing us to write

$$\mathcal{O}_x = \sum x_\sigma \sigma \in \mathbb{R}\mathcal{G}$$

as a point in the group algebra $\mathbb{R}\mathcal{G}$. This structure is central in the modular approach to algebra and group theory in particular. The group algebra is a vector space (over the real field in this case) of symbolic linear combinations of scalars (the data) indexed by $\mathcal{G}$ the group elements (the symmetries) in which, along with the vector space operations, a multiplication of two elements in $\mathbb{R}\mathcal{G}$ is induced by the multiplication in $\mathcal{G}$, e.g., Alperin and Bell (1995).

By its definition it should be clear that group algebras are of natural interest to data analysis. We also remark here that the set of labels L is also amenable to additional symmetries, such as 90 deg and 270 deg rotations, as well as reflections along the 45 deg-rotated x-y axes. The resulting (dihedral) group, however, is no longer commutative, and is discussed elsewhere, e.g., Viana (2008). The group $\mathcal{G} = \{1, v, h, r\}$ is the well-known Klein small group also referred to as the dihedral group D_2. The larger set of symmetries indicated above would define the dihedral group D_4 as the group of automorphisms for the labels, and the symmetrically adapted summaries would follow accordingly.

4. SYMMETRICALLY ADAPTED SUMMARIES

The (scalar) **symmetrically adapted** summaries of data $x \in \mathbb{R}\mathcal{G}$ are certain linear combinations

$$T_\xi(x) = \sum x_\sigma \xi_\sigma$$

defined for all $\xi : \mathcal{G} \mapsto GL_1(\mathbb{R})$ satisfying $\xi_\sigma \xi_\tau = \xi_{\sigma\tau}$ for all symmetries $\sigma, \tau \in \mathcal{G}$. These are the one-dimensional representations of group $\mathcal{G}$, and hence irreducible. The summary T_ξ then **reduces as** ξ in the sense that

$$(\tau T_\xi)(x) = \sum (\tau x)_\sigma \xi_\sigma = \sum x_{\tau\sigma}\xi_\sigma = \sum x_\gamma \xi_{\tau^{-1}\gamma} = \xi_{\tau^{-1}} \sum x_\gamma \xi_\gamma = \xi_{\tau^{-1}} T_\xi(x). \quad (7)$$

That is $\tau T_\xi = \xi_{\tau^{-1}} T_\xi$, for all $\tau \in \mathcal{G}$.

5. SYMMETRICALLY ADAPTED SUMMARIES FOR THE VISUAL FIELD DATA.

The elementary theory of representations tells us that there are exactly $r = 4$ non-equivalent irreducible real representations of D_2, all of which are one-dimensional due to its commutativity. The corresponding summaries are described

in what follows: Under

$$T_1(x) = x_1 + x_v + x_h + x_r, \quad \xi_1 = (1,1,1,1),$$

we obtain

$$T_1 = \begin{bmatrix} 0 & 0 & 0 & 102 & 106 & 106 & 102 & 0 & 0 & 0 \\ 0 & 0 & 101 & \boxed{109} & 111 & 111 & \boxed{109} & 101 & 0 & 0 \\ 0 & 115 & 119 & 119 & 112 & 112 & 119 & 119 & 115 & 0 \\ 111 & 114 & 117 & 110 & 115 & 115 & 110 & 117 & 114 & 111 \\ 109 & 111 & 90 & 122 & 132 & 132 & 122 & 90 & 111 & 109 \\ 109 & 111 & 90 & 122 & 132 & 132 & 122 & 90 & 111 & 109 \\ 111 & 114 & 117 & 110 & 115 & 115 & 110 & 117 & 114 & 111 \\ 0 & 115 & 119 & 119 & 112 & 112 & 119 & 119 & 115 & 0 \\ 0 & 0 & 101 & \boxed{109} & 111 & 111 & \boxed{109} & 101 & 0 & 0 \\ 0 & 0 & 0 & 102 & 106 & 106 & 102 & 0 & 0 & 0 \end{bmatrix}$$

The second summary is

$$T_2(x) = x_1 - x_v + x_h - x_r, \quad \xi_2 = (1,-1,1,-1),$$

leading to the summary

$$T_2 = \begin{bmatrix} 0 & 0 & 0 & 2 & 2 & -2 & -2 & 0 & 0 & 0 \\ 0 & 0 & 1 & \boxed{1} & -1 & 1 & \boxed{-1} & -1 & 0 & 0 \\ 0 & -1 & 1 & -3 & -2 & 2 & 3 & -1 & 1 & 0 \\ -1 & 2 & 3 & 10 & 11 & -11 & -10 & -3 & -2 & 1 \\ -1 & 1 & 28 & 0 & -2 & 2 & 0 & -28 & -1 & 1 \\ -1 & 1 & 28 & 0 & -2 & 2 & 0 & -28 & -1 & 1 \\ -1 & 2 & 3 & 10 & 11 & -11 & -10 & -3 & -2 & 1 \\ 0 & -1 & 1 & -3 & -2 & 2 & 3 & -1 & 1 & 0 \\ 0 & 0 & 1 & \boxed{1} & -1 & 1 & \boxed{-1} & -1 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 & -2 & -2 & 0 & 0 & 0 \end{bmatrix},$$

which is horizontally symmetric and vertically and centrally anti-symmetric. This anisotropy of the D_2-symmetrically adapted summaries is such that they either remain the same or change of sign when the field is re-orientated by the symmetries of D_2. What we obtained here, therefore, is a way to statistically assess the (planar) anisotropy of the visual field (as pointed out above, the actual inferential component of the methodology is only briefly discussed here- it would follow from multivariate, joint models for the field arrays). The isotropy assessment is the consequence of the invariance property shown above in (7). The same can be observed in the remaining summaries;

$$T_3(x) = x_1 + x_v - x_h - x_r, \quad \xi_3 = (1,1,-1,-1),$$

$$T_3 = \begin{bmatrix} 0 & 0 & 0 & -8 & -4 & -4 & -8 & 0 & 0 & 0 \\ 0 & 0 & -11 & \boxed{-13} & -3 & -3 & \boxed{-13} & -11 & 0 & 0 \\ 0 & -5 & -3 & -7 & -2 & -2 & -7 & -3 & -5 & 0 \\ 3 & 0 & -3 & -12 & -13 & -13 & -12 & -3 & 0 & 3 \\ -3 & -3 & 22 & -6 & 2 & 2 & -6 & 22 & -3 & -3 \\ 3 & 3 & -22 & 6 & -2 & -2 & 6 & -22 & 3 & 3 \\ -3 & 0 & 3 & 12 & 13 & 13 & 12 & 3 & 0 & -3 \\ 0 & 5 & 3 & 7 & 2 & 2 & 7 & 3 & 5 & 0 \\ 0 & 0 & 11 & \boxed{13} & 3 & 3 & \boxed{13} & 11 & 0 & 0 \\ 0 & 0 & 0 & 8 & 4 & 4 & 8 & 0 & 0 & 0 \end{bmatrix}$$

is vertically symmetrically and horizontally and centrally anti-symmetric, whereas

$$T_4(x) = x_1 - x_v - x_h + x_r, \quad \xi_4 = (1, -1, -1, +1),$$

$$T_4 = \begin{bmatrix} 0 & 0 & 0 & 8 & 4 & -4 & -8 & 0 & 0 & 0 \\ 0 & 0 & 1 & \boxed{-5} & -3 & 3 & \boxed{5} & -1 & 0 & 0 \\ 0 & -1 & -1 & -1 & 0 & 0 & 1 & 1 & 1 & 0 \\ -1 & 0 & -1 & 8 & 11 & -11 & -8 & 1 & 0 & 1 \\ -1 & -5 & -24 & 4 & 0 & 0 & -4 & 24 & 5 & 1 \\ 1 & 5 & 24 & -4 & 0 & 0 & 4 & -24 & -5 & -1 \\ 1 & 0 & 1 & -8 & -11 & 11 & 8 & -1 & 0 & -1 \\ 0 & 1 & 1 & 1 & 0 & 0 & -1 & -1 & -1 & 0 \\ 0 & 0 & -1 & \boxed{5} & 3 & -3 & \boxed{-5} & 1 & 0 & 0 \\ 0 & 0 & 0 & -8 & -4 & 4 & 8 & 0 & 0 & 0 \end{bmatrix}$$

is centrally symmetric and vertically and horizontally anti-symmetric. The D_2-adapted summaries correspond to the direct-sum decomposition

$$\mathcal{V} = \mathbb{R}^4 = \mathcal{V}_1 \oplus \mathcal{V}_2 \oplus \ldots \oplus \mathcal{V}_4$$

of the measurement space $\mathcal{V}$ of each orbit into invariant subspaces, all of which in $\mathbb{R}$-dimension of 1, in correspondence with the associated canonical projections

$$I = \mathcal{P}_1 + \underbrace{\mathcal{P}_2 + \mathcal{P}_3 + \mathcal{P}_4}_{\text{anisotropy components}} .$$

The canonical projections for the problem follow from the regular representation of D_2, which, from (5), are

$$\rho_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \rho_v = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix},$$

$$\rho_h = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}, \quad \rho_r = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}. \tag{8}$$

From (1) and the irreducible (characters, which in this case are the) one-dimensional representations $\xi_1, \ldots, \xi_4$ introduced above, the corresponding canonical projections follow:

$$\mathcal{P}_1 = \frac{1}{4}\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}, \quad \mathcal{P}_2 = \frac{1}{4}\begin{bmatrix} 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 \end{bmatrix},$$

$$\mathcal{P}_3 = \frac{1}{4}\begin{bmatrix} 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ -1 & -1 & 1 & 1 \\ -1 & -1 & 1 & 1 \end{bmatrix}, \quad \mathcal{P}_4 = \frac{1}{4}\begin{bmatrix} 1 & -1 & -1 & 1 \\ -1 & 1 & 1 & -1 \\ -1 & 1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}.$$

Applying the decomposition of the (4×4) identity matrix canonical property to the orbit $x = (x_1, x_v, x_r, x_h) = (23, 25, 29, 32)$ highlighted in the arrays, we obtain (as also seeing in the corresponding numerical arrays) the symmetrically adapted summaries:

$$\mathcal{P}_1 x = (109, 109, 109, 109)/4, \quad \mathcal{P}_4 x = (1, -1, 1, -1)/4,$$

$$\mathcal{P}_3 x = (-13, -13, 13, 13)/4, \quad \mathcal{P}_2 x = (-5, 5, -5, 5)/4.$$

The ensuing analysis of variance is, therefore,

$$\begin{bmatrix} \xi & x'\mathcal{P}_\xi x & \text{SA-Summary} \\ \hline 1 & 2970.25 & \text{constant} \\ 2 & 0.25 & \text{1+r-v-h} \\ 3 & 42.25 & \text{1+v-r-h} \\ 4 & 6.25 & \text{1+h-v-r} \\ \hline ||x||^2 & 3019 & \end{bmatrix}.$$

The degrees of freedom are the dimension of the projection subspaces and are, as indicated above, all equal to

$$df_\xi = \text{tr } \mathcal{P}_\xi = 1.$$

Although here we do not have the components from sampling error, one might consider the inspection of one component relative to the remaining (anti-symmetric) ones. For example, since

$$13.07 = \frac{x'\mathcal{P}_3 x/1}{(x'\mathcal{P}_2 x + x'\mathcal{P}_4 x)/2} \simeq F_{1,2}$$

is somewhat high in the $F_{1,2}$ distribution, were we to assume the points in the orbit as drawings from normal independent distributions. A plausible interpretation would be that of a relatively less *vertical anisotropy* in that visual field orbit.

Note, briefly, that were we to assume that orbit x is normally distributed with a 4×4 covariance matrix Σ and vector of means μ, then the symmetrically adapted summary $\mathcal{P}x$ associated with projection $\mathcal{P}$ would be distributed as $N(\mathcal{P}\mu, \mathcal{P}\Sigma\mathcal{P}')$. Moreover, since the projections are symmetric matrices it follows that

$$\mathcal{P}\Sigma\mathcal{P}' = \mathcal{P}\Sigma\mathcal{P} = \mathcal{P}^2\Sigma = \Sigma,$$

provided that Σ and $\mathcal{P}$ commute. The algebra tells us that this is the case if and only if Σ has the symmetry of the labels, that is, if and only if Σ commutes with all the permutation

matrices ρ constructed above. As it turns out, the commuting symmetrically adapted covariance matrix should then be patterned as

$$\Sigma = \begin{bmatrix} F & B & C & D \\ B & F & D & C \\ C & D & F & B \\ D & C & B & F \end{bmatrix}.$$

More generally, given a representation ρ of G on $\mathbb{R}^n$ and Λ a real $n \times n$ matrix, then

$$\Sigma = \sum_{\tau \in G} \rho_\tau \Lambda \rho_{\tau^{-1}} \tag{9}$$

has the symmetry of (or is centralized by) ρ in the sense that $\rho_\sigma \Sigma = \Sigma \rho_\sigma$ for all $\sigma \in G$. Equivalently, Σ should be a point in the (commutator) algebra of the canonical projections, e.g., Naĭmark and Štern (1982, p. 55), Weyl (1952, p.81).

To conclude, we briefly illustrate the inclusion of an error term in the previous analysis by sampling 6 of the 19 distinct orbits (starting with x_1 in the top left quadrant) from the original visual field shown in (6), yielding

$\mathcal{O}_j$	1	v	r	h
1	23	25	29	32
2	29	29	30	31
3	28	29	27	27
4	29	28	29	31
5	27	28	29	28
6	29	27	4	30

,

so that $x = (\mathcal{O}_1, \ldots, \mathcal{O}_6)$ is the joint measurement data, in length of 24. The resulting decomposition of $||x||^2$ is shown in (10), where $\mathcal{A}$ and $\mathcal{Q}$ are the 4×4 sampling canonical projection components described in (2) and $\mathcal{P}_\xi$ are the measurement components described above. The error terms $\mathcal{Q} \otimes \mathcal{P}_\xi$ add to $\mathcal{Q} \otimes I$ with $df = 20$, whereas the main effects $\mathcal{A} \otimes \mathcal{P}_\xi$ and overall mean have $df = 1$. Were we to assume that each orbit is identically and independently normally distributed we would assess the effects from

$$f_j = \frac{x'(\mathcal{A} \otimes \mathcal{P}_j)x/1}{x'(\mathcal{Q} \otimes I)x/20} \sim F_{1,20}.$$

We observe that now, in the presence of the introduced sampling errors even the largest observed F-value (1.46087), derived from the $1+r-v-h$ effect is not sufficiently noticeable in its distribution.

$\mathcal{P}$	$x'\mathcal{P}x$	tr $\mathcal{P}$	SA-Summary
$\mathcal{A} \otimes \mathcal{P}_1$	18040.20	1	constant
$\mathcal{A} \otimes \mathcal{P}_2$	42.66	1	1+r-v-h
$\mathcal{A} \otimes \mathcal{P}_3$	0.66	1	1+v-r-h
$\mathcal{A} \otimes \mathcal{P}_4$	37.50	1	1+h-v-r
$\mathcal{Q} \otimes \mathcal{P}_1$	133.83	5	error term
$\mathcal{Q} \otimes \mathcal{P}_2$	108.33	5	error term
$\mathcal{Q} \otimes \mathcal{P}_3$	170.33	5	error term
$\mathcal{Q} \otimes \mathcal{P}_4$	162.50	5	error term
$\mathcal{Q} \otimes I$	575.00	20	error sum
$\|\|x\|\|^2$	18696.00	24	

(10)

6. CONCLUSION

In this paper we have outlined the construction of symmetrically adapted summaries of observations. We emphasized that such summaries are intrinsic to the measurements and independent of design and sampling symmetries. The construction outlined here for the commutative regular $L = G$ applies equally to any finite group and to G-sets as well. These cases, as well as corresponding applications, are discussed in the cited references.

ΠΕΡΙΛΗΨΗ

Στο άρθρο αυτό γίνεται επισκόπηση της μεθοδολογίας για την εύρεση συμμετρικά προσαρμοσμένων περιλήψεων για κάποια είδη μετρήσεων. Επιπλέον, η μεθοδολογία επεξηγείται με εφαρμογές σε μετρήσεις οπτικών πεδίων.

Acknowledgements: The author would like to express his gratitude to the members of the organizing committee for the opportunity to present parts of this work at the meeting.

REFERENCES

Alperin, J. L. and Rowen B. Bell. 1995. *Groups and representations*, Graduate Texts in Mathematics, vol. 162, Springer-Verlag, New York. MR1369573 (96m:20001)

Dawid, A. P.. 1985. *Probability, symmetry and frequency*, The British Journal for the Philosophy of Science **36**, no. 2, pp. 107–128 (English).

_____. 1988. *Symmetry models and hypotheses for structured data layouts*, Journal of the Royal Statistical Society. Series B (Methodological) **50**, no. 1, pp. 1–34 (English).

Faris, W.. 1996. *Review of Roland Omnès, the interpretation of quantum mechanics*, Notices of the American Mathematical Society **43**, no. 11, 1328–39.

Fisher, R. A.. 1942. *The theory of confounding in factorial experiments in relation to the theory of groups*, Ann. Eugenics **11**, 341–353. MR0007382 (4,127i)

James, A. T.. 1954. *Normal multivariate analysis and the orthogonal group*, Annals of Mathematical Statistics **25**, 40–75.

_____. 1957. *The relationship algebra of an experimental design*, The Annals of Mathematical Statistics **28**, no. 4, pp. 993–1002 (English).

Krantz, David H., R. Duncan Luce, Patrick Suppes, and Amos Tversky. 2007. *Foundations of measurement. Vol. I*, Dover Publications Inc., Mineola, NY. Additive and polynomial representations, Reprint of the 1971 original. MR2310074 (2007m:00007)

Luce, R. Duncan, David H. Krantz, Patrick Suppes, and Amos Tversky. 2007. *Foundations of measurement. Vol. III*, Dover Publications Inc., Mineola, NY. Representation, axiomatization, and invariance, Corrected reprint of the 1990 original. MR2348439 (2008f:00004)

Naĭmark, M. A. and A. I. Štern. 1982. *Theory of group representations*, Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. 246, Springer-Verlag, New York. Translated from the Russian by Elizabeth Hewitt, Translation edited by Edwin Hewitt. MR793377 (86k:22001)

Omnès, R.. 1994. *The interpretation of quantum mechanics*, Princeton Press, Princeton, NJ.

Riley, K. F., M. P. Hobson, and S. J. Bence. 2002. *Mathematical methods for physics and engineering*, 2nd ed., Cambridge University Press, New York, NY.

Schatz, G. C. and M. A. Ratner. 1993. *Quantum mechanics in chemistry*, Dover, NY.

Serre, Jean-Pierre. 1977. *Linear representations of finite groups*, Springer-Verlag, New York. Translated from the second French edition by Leonard L. Scott, Graduate Texts in Mathematics, Vol. 42. MR0450380 (56 #8675)

Suppes, Patrick, David H. Krantz, R. Duncan Luce, and Amos Tversky. 2007. *Foundations of measurement. Vol. II*, Dover Publications Inc., Mineola, NY. Geometrical, threshold, and probabilistic representations, Corrected reprint of the 1989 original.

Viana, Marlos. 2008. *Symmetry studies*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge. An introduction to the analysis of structured data in applications. MR2419845 (2010f:62192)

Viana, Marlos and Vasudevan Lakshminarayanan. 2019. *Symmetry in optics and vision studies*, CRC Press, Boca Raton, FL, USA. To appear.

_____. 2013. *Dihedral Fourier analysis*, Lecture Notes in Statistics, vol. 206, Springer, New York. Data-analytic aspects and applications. MR3013391

Viana, Marlos and Donald St. P. Richards (eds.) 2001. *Algebraic methods in statistics and probability*, Contemporary Mathematics, vol. 287, American Mathematical Society, Providence, RI. Papers from the AMS Special Session on Algebraic Methods in Statistics held at the University of Notre Dame, Notre Dame, IN, April 8–9, 2000. MR1873661 (2002h:60002)

Viana, Marlos and Henry Wynn (eds.) 2009. *Algebraic methods in statistics and probability II*, Contemporary Mathematics, vol. 516, American Mathematical Society, Providence, RI. Papers from the AMS Special Session on Algebraic Methods in Statistics held at the University of Illinois, Urbana, IL, March 27–29, 2009.

Weyl, H.. 1952. *Symmetry*, Princeton U. Press, Princeton, NJ.

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 31ου Πανελληνίου Συνεδρίου Στατιστικής (2018), σελ393 -395

ΕΥΡΕΤΗΡΙΟ ΣΥΓΓΡΑΦΕΩΝ

(AUTHOR INDEX)

ΑΓΙΑΚΟΓΛΟΥ Χ.	236
ΑΝΑΣΤΑΣΟΠΟΥΛΟΥ Μ.	27
ΒΕΡΟΠΟΥΛΟΥ Γ.	204
ΓΕΡΟΘΑΝΑΣΗ Κ.	105
ΓΕΩΡΓΑΚΗΣ Α.	42,120
ΖΗΜΕΡΑΣ Σ.	105
ΘΕΟΔΟΣΙΑΔΟΥ Ο.	53
ΚΑΚΟΥΛΛΟΣ Θ.	63
ΚΟΚΚΙΝΙΩΤΗ Θ.Κ.	78
ΚΟΥΝΙΑΣ ΣΤΡ.	68,259
ΚΟΥΤΡΑΣ Μ.Β.	78,90
ΚΥΡΑΝΑ Ε.	105
ΚΥΡΑΝΑ Ζ.	105
ΛΕΛΗΣ Α.	120
ΜΑΡΚΑΚΗ Ε.	265
ΜΑΥΡΟΥΤΣΟΣ Ν.	90
ΜΕΛΦΟΣ Ι.	42
ΜΕΣΣΑΡΙΤΑΚΗΣ ΣΠ.	120
ΜΠΑΝΤΟΥΝΑΣ Ι.	132
ΜΠΟΖΙΚΑΣ Α.	146
ΜΠΡΑΓΟΥΔΑΚΗΣ Ζ.	161
ΠΑΝΑΓΙΩΤΟΥ Ν.	179
ΠΑΠΑΓΕΩΡΓΙΟΥ Μ.	191
ΠΙΤΣΕΛΗΣ Γ.	132,146
ΡΑΚΙΤΖΗΣ Α.	27

ΣΕΡΑΦΕΤΙΝΙΔΟΥ Ε. 204
ΣΤΑΜΑΤΕΛΟΣ Γ. 42
ΣΤΑΜΠΟΥΡΗΣ Α. 220
ΣΩΤΗΡΟΓΛΟΥ Μ. 265
ΤΡΙΑΝΤΑΦΥΛΛΟΥ Ι. 179
ΤΣΑΚΛΙΔΗΣ Γ. 53,278
ΤΑΜΠΟΥΡΗΣ Ε. 220
ΤΑΡΑΜΠΑΝΗΣ Κ. 220
ΤΣΙΜΠΑΝΟΣ Α. 236
ΦΑΡΜΑΚΗΣ Ν. 191
ΧΑΡΑΛΑΜΠΙΔΗΣ Χ. 251
ΧΑΣΙΩΤΗΣ Β. 259
ΧΑΤΖΗΠΑΝΤΕΛΗΣ Θ. 265
ΧΑΤΖΟΠΟΥΛΟΣ Σ. 259

BOUNTZIS P. 278
GALANOPOULOS N. 302
KALLIGERIS E. 292
KARAGIOTA E. 314
KARLIS D. 366
LABROY A. 292
MILIONIS A. 302,314
PAPADIMITRIOY E. 278
PAPAIOANNOY T. 329
PARPOULA C. 292
SIOURIS G. 337
SKILOGIANNI D. 337
SPITIERIS M. 353
TSAMTSAKIRI P. 366
VIANA M. 379