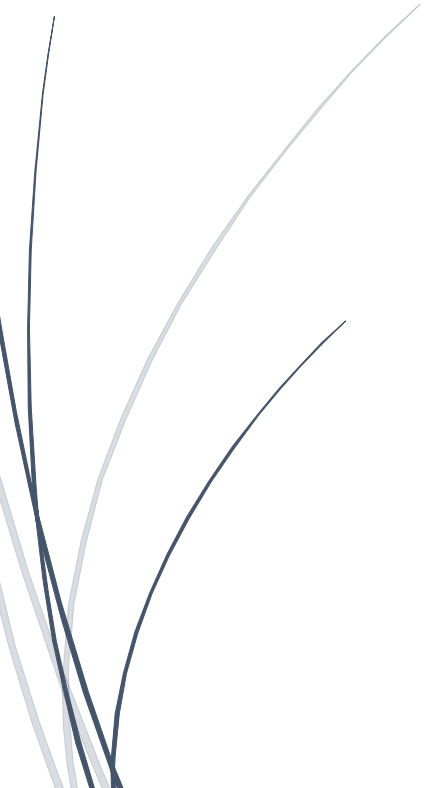
Ελληνικό Στατιστικό Ινστιτούτο
Greek Statistical Institute

2022

ΠΡΑΚΤΙΚΑ
34ου Πανελληνίου Συνεδρίου Στατιστικής

ΣΤΑΤΙΣΤΙΚΗ ΚΑΙ ΠΕΙΡΑΜΑΤΙΚΕΣ ΕΠΙΣΤΗΜΕΣ



Ελληνικό Στατιστικό Ινστιτούτο
Greek Statistical Institute

ΠΡΑΚΤΙΚΑ

34ου Πανελληνίου Συνεδρίου Στατιστικής

Αθήνα, 19-22 Μαΐου 2022

ΣΤΑΤΙΣΤΙΚΗ
ΚΑΙ ΠΕΙΡΑΜΑΤΙΚΕΣ ΕΠΙΣΤΗΜΕΣ

Οργάνωση

Ελληνικό Στατιστικό Ινστιτούτο

Γεωπονικό Πανεπιστήμιο Αθηνών | Ινστιτούτο Σχεδιασμού και Ανάλυσης Πειραμάτων

ISSN 1792-2461

Περιεχόμενα – Contents

Πρόλογος

Το *Πανελλήνιο Συνέδριο Στατιστικής* αποτελεί την κύρια Επιστημονική Εκδήλωση στους κλάδους της Στατιστικής και των Πιθανοτήτων, στην Ελλάδα. Από το 1985 το Συνέδριο συνδιοργανώνεται σε ετήσια βάση από το Ελληνικό Στατιστικό Ινστιτούτο, το οποίο είναι μέλος της Ομοσπονδίας Ευρωπαϊκών Εθνικών Στατιστικών Εταιρειών (FENStats), και ένα Ελληνικό Ακαδημαϊκό Ίδρυμα.

Η ιστορία του Συνεδρίου είναι μεγάλη και σπουδαία γιατί καταγράφει την ιστορία της επιστήμης της Στατιστικής στην Ελλάδα, αναδεικνύοντας τις νέες τάσεις των τελευταίων δεκαετιών στη θεωρία, αλλά και εφαρμοσμένα προβλήματα που εμφανίζονται συνεχώς σε άλλες επιστήμες και αναζητούν λύσεις στις Πιθανότητες και τη Στατιστική. Το Συνέδριο είναι επίσης σημαντικό γιατί δίνει βήμα, αφενός σε καταξιωμένους ερευνητές και ερευνήτριες, και ακαδημαϊκούς δασκάλους, που μοιράζονται την πολύτιμη γνώση και την εμπειρία τους και, αφετέρου σε νέους και νέες που ξεκινούν την επιστημονική καριέρα τους με τις πρώτες ανακοινώσεις τους στο Συνέδριο. Η επιτυχία του Συνεδρίου όλα αυτά τα χρόνια είναι συνδυασμός της υψηλής ποιότητας της έρευνας, της ποικιλίας των επιστημονικών θεμάτων που παρουσιάζονται αλλά και του θετικού κλίματος που έχει καλλιεργηθεί ανάμεσα στους συμμετέχοντες.

Το 34ο Συνέδριο διοργανώθηκε από κοινού από το *Ελληνικό Στατιστικό Ινστιτούτο* και το *Ινστιτούτο Σχεδιασμού και Ανάλυσης Πειραμάτων* του *Γεωπονικού Πανεπιστημίου Αθηνών*, στην πόλη της Αθήνας, από 19 έως 22 Μαΐου, 2022. Το κυρίαρχο θέμα του 34ου Συνεδρίου ήταν «*Στατιστική και Πειραματικές Επιστήμες*». Η εν λόγω εκδήλωση έσμιξε Στατιστικούς, Μαθηματικούς, Αναλυτές Δεδομένων, Πληροφορικούς, Οικονομέτρες, καθώς και επιστήμονες από το ευρύτερο πεδίο των πειραματικών Γεωπονικών Επιστημών, όπως η Βιοτεχνολογία, η Επιστήμη Τροφίμων, κ.α., με σκοπό την ανταλλαγή ιδεών και τη συζήτηση των πρόσφατων εξελίξεων στην ευρύτερη περιοχή της Στατιστικής Επιστήμης και των εφαρμογών της στις Πειραματικές Επιστήμες. Η εναρκτήρια τελετή περιλάμβανε τις κεντρικές ομιλίες : *του Καθηγητή Δελλαπόρτα Π.,* με τίτλο «*Οι επαναστάσεις της a posteriori συμπερασματολογίας», του Καθηγητή Ταραντίλη ΠΑ,* με τίτλο «*Η συνεισφορά της στατιστικής ως «εργαλείο» στην επίλυση προβλημάτων αυθεντικότητας τροφίμων, του Καθηγητή Νυχά ΙΓ,* με τίτλο *«Η επιστήμη των δεδομένων στην υπηρεσία της ασφάλειας και ποιότητα των τροφίμων»,* και του *Καθηγητή Balakrishnan N,* με τίτλο «*Accelerated life testing of one-shot devices: Data collection and analysis».* Κεντρική ομιλία έδωσε και ο *Καθηγητής Mauromoustakos A,* με τίτλο «*Good and bad practices with examples from designs and analysis of experiments».* Στο Συνέδριο, που είχε και διεθνή συμμετοχή με 18 Έλληνες και ξένους προσκεκλημένους ομιλητές, συμμετείχαν συνολικά 159 σύνεδροι, και παρουσιάστηκαν 110 εργασίες και αναρτημένες ανακοινώσεις. Παράλληλα διοργανώθηκε **Εκπαιδευτικό Επιμορφωτικό Σεμινάριο** στο λογισμικό eStat, το οποίο συντονίστηκε από τον Καθηγητή Καραγρηγορίου Α., και παρουσιάστηκε από τον Κουκούμη Χ.

Το Βραβείο **Καλύτερης Εργασίας Νέου Έλληνα Στατιστικού,** το οποίο έχει δωροθετήσει ο Καθηγητής Balakrishnan (McMaster University, Canada), απονεμήθηκε, μετά από απόφαση

της Επιτροπής Βραβείου, στον κ. **Αναστάσιο Αψεμίδη** για την εργασία «*Stochastic Epidemic Modelling of COVID-19»,* ενώ εύφημος μνεία δόθηκε στην κα. **Παναγιώτα Τσαμτσακίρη** για την εργασία «*A new Conditional Auto-regressive Range model»*. Η Επιτροπή αποτελείτο από τους Καθηγητές Μπατσίδη Α., Μπουρνέτα Α., και Ψαρράκο Γ.

Στα χέρια σας κρατάτε τα Πρακτικά του Συνεδρίου στα οποία υποβληθήκαν 16 εργασίες, και έγιναν δεκτές 14 εργασίες.

Η Επιτροπή Έκδοσης Πρακτικών του ΕΣΙ εκφράζει τις ευχαριστίες της προς τους κριτές, κ.κ., Βόντα Ι., Ευαγγελάρα Χ., Καζάνα Θ., Καρλής Δ., Λύτρα Θ., Μέρκατα Χ, Μηλιένο Φ., Μηλιώνης Α., Μωυσιάδη Π, Ντζούφρα Ι, Οικονόμου Π., Παπασταμούλης Π., Πολίτης Κ., Σύψα Β, Τζαβελλά Γ., Τσακλίδη Γ., Φουσκάκη Δ., Χαλικιάς Μ.,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.

Οι Συντονιστές των Πρακτικών

Ελευθέριος Αγγελής
Μαλβίνα Βαμβακάρη
Αλέξανδρος Καραγρηγορίου
Σωτηρία Μαλεφάκη
Σωτήριος Μπερσίμης
Δημοσθένης Παναγιωτάκος
Γεώργιος Ψαρράκος

Τοπική Οργανωτική Επιτροπή

Βαμβακάρη Μ., Καθηγήτρια, *Χαροκόπειο Πανεπιστήμιο*.
Δαφνής Σ., Ακαδημαϊκός Υπότροφος, *Γεωπονικό Πανεπιστήμιο Αθηνών*.
Ευαγγελάρας Χ., Αναπληρωτής Καθηγητής, *Πανεπιστήμιο Πειραιώς*.
Ζώτος Χ., Υποψήφιος Διδάκτωρ, *Γεωπονικό Πανεπιστήμιο Αθηνών*.
Καλλιγέρης Ε.-Ν., Υποψήφιος Διδάκτωρ, *Πανεπιστήμιο Αιγαίου*.
Καλύβας Δ., Καθηγητής, *Γεωπονικό Πανεπιστήμιο Αθηνών*.
Καραγρηγορίου Α., Καθηγητής, *Πανεπιστήμιο Αιγαίου.*
Κατσιλέρος Α., Ε.ΔΙ.Π., *Γεωπονικό Πανεπιστήμιο Αθηνών*.
Κωστοπούλου Κ., Καθηγήτρια, *Γεωπονικό Πανεπιστήμιο Αθηνών*.
Μαγριπλή Ε., Επίκουρος Καθηγήτρια, *Γεωπονικό Πανεπιστήμιο Αθηνών*.
Μαλέσιος Χ., Επίκουρος Καθηγητής, *Γεωπονικό Πανεπιστήμιο Αθηνών*.
Μαλιάππης Μ., Ε.Δ.Ι.Π, *Γεωπονικό Πανεπιστήμιο Αθηνών*.
Μηλιένος Φ., Επίκουρος Καθηγητής, *Πάντειο Πανεπιστήμιο Κοινωνικών και Πολιτικών Επιστημών*.
Μπερσίμης Σ., Αναπληρωτής Καθηγητής, *Πανεπιστήμιο Πειραιώς*.
Μωυσιάδης Χ., Ομότιμος Καθηγητής, *Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης*.
Ντότσης Κ., Υποψήφιος Διδάκτωρ, *Πανεπιστήμιο Αιγαίου*.
Παναγιωτάκος Δ., Καθηγητής, *Χαροκόπειο Πανεπιστήμιο*.
Παπαδόπουλος Γ., Αναπληρωτής Καθηγητής, *Γεωπονικό Πανεπιστήμιο Αθηνών*.
Σίμου Ε., ΕΤΕΠ, *Γεωπονικό Πανεπιστήμιο Αθηνών*.
Σωτηράκογλου Κ., Καθηγήτρια, *Γεωπονικό Πανεπιστήμιο Αθηνών*.
Ψαρράκος Γ., Αναπληρωτής Καθηγητής, *Πανεπιστήμιο Πειραιώς*.

Επιστημονική Επιτροπή

Αγγελής Ε., Καθηγητής, *Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης*.
Βαμβακάρη Μ., Καθηγήτρια, *Χαροκόπειο Πανεπιστήμιο*.
Βασδέκης Β., Καθηγητής, *Οικονομικό Πανεπιστήμιο Αθηνών.*
Δαμιανού Χ., Αναπληρωτής Καθηγητής, *Πανεπιστήμιο Αθηνών*.
Δαφνής Σ., Ακαδημαϊκός Υπότροφος, *Γεωπονικό Πανεπιστήμιο Αθηνών*.
Ευαγγελάρας Χ., Αναπληρωτής Καθηγητής, *Πανεπιστήμιο Πειραιώς*.
Ζωγράφος Κ., Καθηγητής, *Πανεπιστήμιο Ιωαννίνων*.
Ηλιόπουλος Γ., Καθηγητής, *Πανεπιστήμιο Πειραιώς*.
Καζάνα Α., Καθηγητής, *Πανεπιστήμιο ΑΠΘ*
Καλαματιανού Α., Ομότιμη Καθηγήτρια, *Πάντειο Πανεπιστήμιο Κοινωνικών και Πολιτικών Επιστημών*.
Καλύβας Δ., Καθηγητής, *Γεωπονικό Πανεπιστήμιο Αθηνών*.
Καραγρηγορίου Α., Καθηγητής, *Πανεπιστήμιο Αιγαίου*.
Καρλής Δ., Καθηγητής, *Οικονομικό Πανεπιστήμιο Αθηνών*.
Κατέρη Μ., Καθηγήτρια, *RWTH Aachen University*.
Κατσιλέρος Α., Ε.ΔΙ.Π., *Γεωπονικό Πανεπιστήμιο Αθηνών*.
Κολυβά-Μαχαίρα Φ., Αναπληρώτρια Καθηγήτρια, *Πανεπιστήμιο Θεσσαλονίκης*.
Κουνιάς Σ., Ομότιμος Καθηγητής, *Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών* και Επίτιμος Πρόεδρος του *ΕΣΙ*.
Κούτρας Μ., Καθηγητής, *Πανεπιστήμιο Πειραιώς*.

Κωστοπούλου Κ., Καθηγήτρια, *Γεωπονικό Πανεπιστήμιο Αθηνών*.
Μαγριπλή Ε., Επίκουρος Καθηγήτρια, *Γεωπονικό Πανεπιστήμιο Αθηνών*.
Μακρή Φ., Καθηγήτρια, *Πανεπιστήμιο Πατρών*.
Μαλέσιος Χ., Επίκουρος Καθηγητής, *Γεωπονικό Πανεπιστήμιο Αθηνών*.
Μαλεφάκη Σ., Επίκουρη Καθηγήτρια, *Πανεπιστήμιο Πατρών*.
Μηλιένος Φ., Επίκουρος Καθηγητής, *Πάντειο Πανεπιστήμιο Κοινωνικών και Πολιτικών Επιστημών.*
Μηλιώνης Α., Καθηγητής, *Πανεπιστήμιο Αιγαίου*
Μπασιάκος Ι., Καθηγητής, *Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών*.
Μπατσίδης Α., Αναπληρωτής Καθηγητής, *Πανεπιστήμιο Ιωαννίνων*.
Μπερσίμης Σ., Αναπληρωτής Καθηγητής, *Πανεπιστήμιο Πειραιώς*.
Μπουρνέτας Α., Καθηγητής, Πανεπιστήμιο Αθηνών.
Μπραγουδάκης Ζ., Ερευνητής Β, *Τράπεζα της Ελλάδος* και Επισκέπτης Καθηγητής, *Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών*.
Μωυσιάδης Χ., Ομότιμος Καθηγητής, *Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης*.
Ντζούφρας Ι., Καθηγητής, *Οικονομικό Πανεπιστήμιο Αθηνών*.
Οικονόμου Π., Αναπληρωτής Καθηγητής, *Πανεπιστήμιο Πατρών*.
Παναγιωτάκος Δ., Καθηγητής, *Χαροκόπειο Πανεπιστήμιο*.
Παπαδόπουλος Γ., Αναπληρωτής Καθηγητής, *Γεωπονικό Πανεπιστήμιο Αθηνών*.
Παπαϊωάννου Τ., Ομότιμος Καθηγητής, *Πανεπιστήμιο Πειραιώς και Πανεπιστήμιο Ιωαννίνων* και Επίτιμος Πρόεδρος του *ΕΣΙ*.
Παπασταμούλης Π., Καθηγητής, *Οικονομικό Πανεπιστήμιο Αθηνών*
Πετρόπουλος Κ., Αναπληρωτής Καθηγητής, *Πανεπιστήμιο Πατρών*.
Ρακιτζής Α., Επίκουρος Καθηγητής, *Πανεπιστήμιο Πειραιώς*.
Συρακούλης Κ., Αναπληρωτής Καθηγητής, *Πανεπιστήμιο Θεσσαλίας*.
Σωτηράκογλου Κ., Καθηγήτρια, *Γεωπονικό Πανεπιστήμιο Αθηνών*.
Τριανταφύλλου Ι., Επίκουρος Καθηγητής, *Πανεπιστήμιο Πειραιώς*.
Τσακλίδης Γ., Καθηγητής, *Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης*.
Φουσκάκης Δ., Καθηγητής, *Εθνικό Μετσόβιο Πολυτεχνείο*.
Χαραλαμπίδης Χ.Α., Ομότιμος Καθηγητής, *Πανεπιστήμιο Αθηνών*.
Χατζηπαντελής Θ., Καθηγητής, *Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης*.
Χριστοφίδης Τ., Καθηγητής, *Πανεπιστήμιο Κύπρου*.
Ψαρράκος Γ., Αναπληρωτής Καθηγητής, *Πανεπιστήμιο Πειραιώς*.

Κεντρικοί Ομιλητές

- **Dellaportas** P., Professor, *Athens University of Economics and Business* and *University College London*.
- **Mauromoustakos** A., Professor, *University of Arkansas*.
- **Nychas** G.I., Professor, *Agricultural University of Athens*.
- **Tarantilis** P., Professor, *Agricultural University of Athens*.

2022 Εναρκτήρια Διάλεξη στη μνήμη του Θεόφιλου Κάκουλλου

- **Balakrishnan** N., Distinguished Professor, *McMaster University*.

Προσκεκλημένοι Ομιλητές

- **Alzahrani** A., PhD Candidate, *Cardiff University*.
- **Barbu** V.S., Associate Professor, *University of Rouen - Normandy*.
- **Consonni** G., Professor, *Universita Cattolica del Sacro Cuore*.
- **Egidi** L., Assistant Professor, *Univesity of Trieste*.
- **Eryilmaz** S., Professor, *Atilim University*.
- **Gillard** J., Professor, *Cardiff University*.
- **Piepho** H.P., Professor, *University of Hohenheim*.
- **Heuchenne** C., Professor, *University of Liege*.
- **Jones** O., Professor, *Cardiff University*.
- **Marsman** M., Assistant Professor, *University of Amsterdam*.
- **Mukherjee** A., Professor, *XLRI - Xavier School of Management, XLRI Jamshedpur*.
- **Noonan** J., Postdoc Researcher, *Cardiff University*.
- **Ombao** H., Professor, *CEMSE King Abdullah University of Science and Technology*.
- **Robert** C.P., Professor, *University of Paris Dauphine* and *University of Warwick*.
- **Rossell** D., Visiting Professor, *Universitat Pompeu Fabra* & Director of *GSE Barcelona Data Science MSC and Data Science Center*.
- **Tsui** K.L., Professor, *Virginia Polytechnic and State University*.
- **Trevezas** S., Assistant Professor, *University of Athens*.
- **Tudor** C., Professor, *University of Lille I & University of Paris I*.
- **Wagenmakers** E.-J., Professor, *University of Amsterdam*.

ΠΡΟΓΡΑΜΜΑ ΣΥΝΕΔΡΙΟΥ/CONFERENCE PROGRAMME*

Πέμπτη 19 Μαΐου 2022/Thursday May 19, 2022	
13:00 – 14:00	Εγγραφή Συνέδρων/Conference Registration Διανομή Συνεδριακού Υλικού/Distribution of Conference Material
Αίθουσα Α/Room A 14:00 – 15:00	**Έναρξη του Συνεδρίου – Χαιρετισμοί/Opening Ceremony** **Προεδρείο/Chairs: Α. Καραγρηγορίου, Γ. Παπαδόπουλος/ A. Karagrigoriou, G. Papadopoulos**
Αίθουσα Α/Room A 15:00 – 17:40	**Κεντρικές Ομιλίες/Plenary Talks** **Προεδρείο/Chairs: Ε. Αγγελής, Γ. Παπαδόπουλος/E. Aggelis, G. Papadopoulos**
15:00 – 15:40	**Π. Δελλαπόρτας** *Οι επαναστάσεις της a posteriori συμπερασματολογίας*
15:40 – 16:20	**Π.Α. Ταραντίλης** *Η συνεισφορά της στατιστικής ως «εργαλείο» στην επίλυση προβλημάτων αυθεντικότητας τροφίμων*
16:20 – 17:00	**Γ.-Ι. Νυχάς** *Η επιστήμη των δεδομένων στην υπηρεσία της ασφάλειας και ποιότητα των τροφίμων*
17:00 – 17:40	**N. Balakrishnan, 2022 Inaugural Theo Cacoullos Memorial Lecture** *Accelerated life testing of one-shot devices: Data collection and analysis*
17:40 – 18:15	***ΔΙΑΛΕΙΜΜΑ – ΚΑΦΕΣ/COFFEE BREAK***
Αίθουσα Α/Room A 18:15 – 19:45	**Προσκεκλημένη Ειδική Συνεδρία Ι/Special Invited Session I** Η συμβολή της Στατιστικής στην αντιμετώπιση της πανδημίας **Προεδρεύων/Chair: N. Δεμίρης/N. Demiris**
18:15 – 18:45	**Β. Σύψα** *Assessing the impact of vaccination prioritisation strategies and of non-pharmaceutical interventions on COVID-19 pandemic using mathematical models*
18:45 – 19:15	**Γ. Τουλούμη** *Clinical Trials during COVID-19 pandemic: Challenges and lessons*
19:15 – 19:45	**Α. Χατζηλένα** *On the statistical analysis of freely available COVID-19 data*
20:00	***WELCOME COCKTAIL*** *Κήπος Γεωργικού Μουσείου/Garden of the Agricultural Museum*

Παρασκευή 20 Μαΐου 2022/Friday May 20, 20222	
Αίθουσα Α/Room A **09:00 – 11:15**	**Προσκεκλημένη Ειδική Συνεδρία ΙΙ/Special Invited Session II** *Wet lab vs dry lab scientists: αντίπαλοι ή συμπαίκτες? Βιο-μαθηματικά και Βιο-στατιστική από το Χωράφι στο Στομάχι* **Προεδρεύων/Chair: Π. Σκανδάμης/P. Skandamis**
09:00 – 09:20	**Π.-Κ. Ρεβέλου** *Βιοπληροφορική προσέγγιση στη μεταβολομική για αυθεντικότητα/βοτανική ταυτοποίηση του ελαιόλαδου: γραμμικές και μη γραμμικές μέθοδοι επί υπερφασματικών δεδομένων και δεδομένων GC-MS*
09:20 – 09:40	**Ε.Ζ. Πανάγου** *Βιοπληροφορική προσέγγιση στη μεταβολομική για νοθεία και μικροβιολογική ποιότητα των τροφίμων: εφαρμογή γραμμικών και μη γραμμικών μεθόδων με χρήση φασματοσκοπικών δεδομένων*
09:40 – 10:00	**Π.Ν. Σκανδάμης** *Integrating omics in quantitative microbial risk assessment I: Οπλοστάσιο μαθηματικών και στατιστικής στην υπηρεσία της ανάλυσης επικινδυνότητας*
10:00 – 10:20	**Κ. Παπαδημητρίου** *Integrating omics in quantitative microbial risk assessment II: Βιοπληροφορική σε δεδομένα που προέρχονται από ομικές αναλύσεις αλληλουχιών βιομακρομορίων (genomics – metagenomics – transcriptomics)*
10:20 – 10:40	**Α. Βολουδάκης** *Δεδομένα μεγάλης κλίμακας στην επισιτιστική ασφάλεια τροφίμων – Η περίπτωση της πρωτογενούς φυτικής παραγωγής*
10:40 – 11:15	**Στρογγυλή Τράπεζα/Round Table** *Statistical power - it is necessary, but is it always feasible or what? Let's find a compromise* **Συντονιστής/Coordinator:** Π. Σκανδάμης/P. Skandamis **Συμμετέχοντες/Participants:** Α. Καραγρηγορίου, Μ.Β. Κούτρας, Φ. Μηλιένος & Ομιλητές της Συνεδρίας/A.Karagrigoriou, M. V. Koutras, F. Milienos & Session Participants
11:15 – 11:45	***ΔΙΑΛΕΙΜΜΑ – ΚΑΦΕΣ/COFFEE BREAK***
Αίθουσα Α/Room A **11:45 – 14:15**	**Προσκεκλημένη Συνεδρία Ι/Invited Session I** **Προεδρεύων/Chair: Δ. Καρλής/D. Karlis**
11:45 – 12:15	**S. Trevezas** *Modeling plant growth with the Greenlab model of deterministic organogenesis – Estimation strategies*
12:15 – 12:45	**E.-J. Wagenmakers** *Approximateobjectivebayesfactorfromp-valuesandsamplesize:The3p√nrule*
12:45 – 13:15	**H. Ombao** *Modeling spectral dependence and non-linear Granger causality*
13:15 – 13:45	**S. Eryilmaz** *Statistical aspects of wind and hybrid power systems*
13:45 – 14:15	**Y. Devrim, S. Eryilmaz** *An overview and new challenges on reliability assessment of a fuel cell stack system*
14:30 – 17:30	***ΠΕΡΙΗΓΗΣΕΙΣ – ΞΕΝΑΓΗΣΕΙΣ/GUIDED TOURS***
20:30	***ΕΠΙΣΗΜΟ ΔΕΙΠΝΟ/CONFERENCE DINNER***

Σάββατο 21 Μαΐου 2022/Saturday May 21, 2022	
Αίθουσα Α/Room A **09:00 – 09:40**	**Κεντρική Ομιλία/Plenary Talk** **Προεδρεύων/Chair: Δ. Παναγιωτάκος/D. Panagiotakos**
09:00 – 09:40	**A. Mauromoustakos** *Good and bad practices with examples from designs and analysis of experiments*
09:40 – 10:00	***ΣΥΝΤΟΜΟ ΔΙΑΛΛΕΙΜΑ/SHORT BREAK***
Αίθουσα Α/Room A **10:00 – 11:20**	**Ειδική Συνεδρία Ι/Special Session I** Πιθανότητες και Αναλογισμός/Probability and Actuarial **Προεδρεύων/Chair: Γ. Ψαρράκος/G. Psarrakos**
10:00 – 10:20	**Λ. Κανελλόπουλος** *Στοχαστικές διατάξεις σύνθετων γεωμετρικών κατανομών*
10:20 – 10:40	**Π. Βλιώρα, Γ. Ψαρράκος** *Μείξεις και ελαστικότητα σταθμισμένων ασφαλίστρων*
10:40 – 11:00	**Κ. Πολίτης** *Μελέτη της ευαισθησίας για την πιθανότητα χρεοκοπίας στο κλασικό πρότυπο της θεωρίας κινδύνων*
11:00 – 11:20	**Γ. Ψαρράκος, A. Toomaj, Π. Βλιώρα** *Μια οικογένεια μέτρων μεταβλητότητας που βασίζονται στην αθροιστική υπολειπόμενη εντροπία και στις στρεβλές συναρτήσεις*
Αίθουσα Β/Room B **10:00 – 11:20**	**Εφαρμοσμένη Στατιστική Ι/Applied Statistics I** **Προεδρεύων/Chair: Ε. Αγγελής/E. Aggelis**
10:00 – 10:20	**Χ. Ευαγγελάρας** *Σχεδιασμοί για πειράματα Order-of-Addition*
10:20 – 10:40	**V. Chasiotis, D. Karlis** *Subdata selection for big data regression: An improved approach*
10:40 – 11:00	**Κ. Γεωργίου, Κ. Χαρμάνας, Ν. Μήττας, Ε. Αγγελής** *Ανάλυση πατεντών μηχανικής μάθησης: Αξιολόγηση τάσεων και ευρημάτων*
11:00 – 11:20	**Γ. Παπαγεωργίου, Σ. Μπερσίμης, Π. Οικονόμου** *Ταξινόμηση μηνυμάτων ηλεκτρονικού ταχυδρομείου χρησιμοποιώντας μηχανική μάθηση*
Αίθουσα Γ/Room C **10:00 – 11:20**	**Στοχαστικές Διαδικασίες και Εφαρμογές/Stochastic Processes and Applications** **Προεδρεύουσα/Chair: Α. Καλαματιανού/A. Kalamatianou**
10:00 – 10:20	**Σ.Μ. Τζανίνης** *Μία τεχνική αλλαγής μέτρου για σύνθετες ανανεωτικές διαδικασίες και εφαρμογές της στην θεωρία χρεοκοπίας*
10:20 – 10:40	**Δ. Λυμπερόπουλος, Ν. Μαχαιράς** *Μία γενίκευση της χρηματοοικονομικής αποτίμησης της ασφάλισης*
10:40 – 11:00	**Ρ. Λύκου, Γ. Τσακλίδης** *Κρυφά ανοιχτά ομογενή Μαρκοβιανά συστήματα*

Σάββατο 21 Μαΐου 2022/Saturday May 21, 2022	
11:00 – 11:20	**Γ. Βασιλειάδης** *Το ομογενές Μαρκοβιανό σύστημα διακριτού χρόνου με τυχαίες χωρητικότητες. Εφαρμογή στη μελέτη ουράς με τυχαίο αριθμό εξυπηρετητών*
11:20 – 11:45	***ΔΙΑΛΕΙΜΜΑ – ΚΑΦΕΣ/COFFEE BREAK***
Αίθουσα Α/Room A 11:45 – 13:25	**Ειδική Συνεδρία ΙΙ/Special Session II** Statistical modeling for plant growth and development – Case studies **Προεδρεύων/Chair: Σ. Τρέβεζας/S. Trevezas**
11:45 – 12:05	**K. Florakis, S. Trevezas** *Prediction of water consumption of a drip irrigation system for greenhouses – A case study in Nigrita, Serres*
12:05 – 12:25	**E. Panagiotopoulos, I. Oikonomidis, S. Trevezas** *Large scale corn phenological stage prediction with random forests – The US corn belt*
12:25 – 12:45	**I. Oikonomidis, S. Trevezas** *Modeling the variability of an arabidopsis thaliana population with the Greenlab model*
12:45 – 13:05	**Ν.Α. Μαριόλης, Ε.Χ. Βλάχος, Ν.Γ. Ζανάκης** *Συγκριτική αξιολόγηση μεθόδων παραγωγικής σταθερότητας σε καλαμπόκι και βαμβάκι*
13:05 – 13:25	**Δ. Παπαδόπουλος** *Επιπτώσεις της κλιματικής αλλαγής στη γονιμότητα του εδάφους*
Αίθουσα Β/Room B 11:45 – 13:45	**Στατιστική Ι/Statistics I** **Προεδρεύουσα/Chair: Σ. Μαλεφάκη/S. Malefaki**
11:45 – 12:05	**Α. Μπατσίδης, S. Bar-Lev, J. Einbeck, X. Liu, P. Ren** *Έλεγχος καλής προσαρμογής για την οικογένεια των φυσικών εκθετικών κατανομών*
12:05 – 12:25	**Π. Οικονόμου, Α. Μπατσίδης, Γ. Τζαβελάς, Δ. Μπάγκαβος** *Στατιστικοί έλεγχοι για τη μέση τιμή και τη διασπορά χρησιμοποιώντας r-μεγέθους μεροληπτικά δείγματα*
12:25 – 12:45	**Μ. Μασούρα, Ε. Νικολοπούλου, Σ. Μαλεφάκη** *Παράγοντες που επηρεάζουν τις εκπομπές του διοξειδίου του άνθρακα και την περιβαλλοντική πολιτική στην ΕΕ*
12:45 – 13:05	**Χ.Θ. Νάκας, A.M. Franco-Pereira, M.C. Pardo, Χ.Γ. Αθανασίου** *Έλεγχοι υποθέσεων σε μη-μονότονες διατάξεις διωνυμικών ποσοστών με εφαρμογή στην ανθεκτικότητα εντόμων αποθηκών σε διαφορετικές δόσεις φωσφίνης*
13:05 – 13:25	**Κ. Σκαρλάτος, Σ. Μπερσίμης, Π. Οικονόμου** *Ανίχνευση σημείων αλλαγής και ακραίων σημείων σε πολυδιάστατες ροές δεδομένων*
13:25 – 13:45	**L. Egidi, I. Ntzoufras** *A unified Bayesian model for predicting volleyball games*

Σάββατο 21 Μαΐου 2022/Saturday May 21, 2022	
Αίθουσα Γ/Room C 11:45 – 13:45	Εφαρμοσμένη Στατιστική ΙΙ/Applied Statistics II Προεδρεύουσα/Chair: Χ. Παρπούλα/C. Parpoula
11:45 – 12:05	**Μ. Χαλικιάς** *Σχεδιασμοί επαναλαμβανόμενων μετρήσεων: Παραδείγματα σχεδιασμών για 5 χρονικές περιόδους*
12:05 – 12:25	**V. Polimenis** *The lepto-variance of stock returns*
12:25 – 12:45	**Χ. Χατζημιχαήλ, Β. Καραγιάννης** *Ανάλυση συστάδων στις ετήσιες δαπάνες του συστήματος υγείας επιλεγμένων χωρών της ευρωπαϊκής ένωσης από το 2004 ως και το 2018*
12:45 – 13:05	**Θ. Μωυσιαδής, Δ. Κοπαράνης, Κ. Λιάπης, Κ. Φωκιανός, Ι. Κοτσιανίδης** *Εξατομικευμένη πρόβλεψη στη χρόνια λεμφοκυτταρική λευχαιμία*
13:05 – 13:25	**Ι. Ανδρεάδης** *Η βιβλιοθηκη της R survey data quality*
13:25 – 13:45	**Ch. Meselidis, A. Karagrigoriou.** *The use of the modified (Φ,a) – power divergence family in statistical inference*
13:45 – 16:00	***ΜΕΣΗΜΒΡΙΝΗ ΔΙΑΚΟΠΗ/MIDDAY BREAK***
Αίθουσα Δ/Room D 15:00 – 17:40	Εκπαιδευτικό Επιμορφωτικό Σεμινάριο eStat (Part A)/Training Seminar eStat (Part A), Προεδρεύων/Chair: Α. Καραγρηγορίου/ A. Karagrigoriou
15:00 – 15:30	**Χαιρετισμοί/Welcome** • **Χ. Χρήστου,** Κοινότητα ΕΛ/ΛΑΚ Αιγαίου (fossaegean) • **Η. Βαρλάμης,** ΔΣ ΕΕΛΛΑΚ • **Ε. Αγγελής,** Πρόεδρος ΕΣΙ • **J. J. Lee,** eSTAT
15:30 – 17:40	**Χ. Κουκούμης (Εισηγητής)** eStat Μέρος Α/eStat Part A
Αίθουσα Α/Room A 16:00 – 17:40	Εφαρμοσμένη Στατιστική στις Γεωπονικές Επιστήμες/Applied Statistics in Agricultural Sciences, Προεδρεύων/Chair: Κ.Σωτηράκογλου/ K.Sotirakoglou
16:00 – 16:20	**Α. Αντωνόπουλος, Ο. Γούναρη, Π. Χαριζάνης, Κ. Καράντζαλος** *Χαρτογράφηση κάλυψης γης δασικών μελισσοκομικών φυτών από δορυφορικά πολυφασματικά δεδομένα*
16:20 – 16:40	**Σ. Γούναρη, Ν. Προύτσος, Χ.Ε. Ζώτος, Σ.Δ. Δαφνής, Γ.Κ. Παπαδόπουλος** *Η επίδραση μετεωρολογικών παραγόντων στον βιολογικό κύκλο του Marchalina Hellenica*

Σάββατο 21 Μαΐου 2022/Saturday May 21, 2022	
16:40 – 17:00	**Δ. Παπαχρήστου, Π. Κουτσούλη, Γ. Λαλιώτης, Α. Τσαπραΐλής, I. Medugorac, I. Μπιζελής** *Ανίχνευση διασταυρωμένων ατόμων σε πληθυσμούς της βραχυκερατικής φυλής βοοειδών*
17:00 – 17:20	**Α. Κουνάνη, Ν. Λαβάζος, Α. Τσιμπίρης, Δ. Βαρσάμης** *Εντοπισμός φυτικών ασθενειών με τεχνικές βαθιάς μηχανικής μάθησης*
17:20 – 17:40	**Ε. Ζωίδης, Α.Χ. Παππάς, Μ. Γκολιομύτης, Π. Σιμιτζής, Κ. Σωτηράκογλου, Σ. Ταυριζέλου, Γ. Δανέζης, Κ. Γεωργίου** *Επιδράσεις της διατροφικής προσθήκης φλαβονοειδών και στεμφυλών χυμοποιίας πορτοκαλιού (citrus sinensis) σε συγκεκριμένα στοιχεία του αυγού με χημειομετρική ανάλυση*
Αίθουσα B/Room B 16:00 – 17:40	**Στατιστική ΙΙ/Statistics II** **Προεδρεύων/Chair: Α. Μπουρνέτας/A. Burnetas**
16:00 – 16:20	**A. Apsemidis, N. Demiris** *Stochastic epidemic modelling of COVID-19*
16:20 – 16:40	**P. Tsamtsakiri, D. Karlis** *A novel conditional autoregressive range model*
16:40 – 17:00	**Χ. Χατζημιχαήλ, Μ. Μπατσιάκα, Ν. Φαρμάκης** *Σχέδια δειγματοληψίας χαμηλού προϋπολογισμού σε υποσύνολα του επιπέδου R^2*
17:00 – 17:20	**T. Tsiampalis, D.B. Panagiotakos** *The identification, impact, and management of missing values in nutritional epidemiology: Insights from the ATTICA epidemiological study*
17:20 – 17:40	**Μ. Τσιφουτίδου, Χ. Κόνσουλας** *Μελέτη θνητότητας και νοσηρότητας πληθυσμών Drosophila σε διάφορα τροφικά περιβάλλοντα*
17:40 – 18:15	***ΔΙΑΛΕΙΜΜΑ – ΚΑΦΕΣ/COFFEE BREAK***
Αίθουσα A/Room A 18:15 – 20:45	**Προσκεκλημένη Συνεδρία ΙΙ/Invited Session II** **Προεδρεύων/Chair: Α. Αρτεμίου & K. Ντότσης/A. Artemiou & K. Ntotsis**
18:15 – 18:45	**O. Jones** *Synthesising data with missing values using generative adversarial networks*
18:45 – 19:15	**A. Alzahrani, A. Artemiou** *Projection-based classification*
19:15 – 19:45	**J. Noonan** *Improving the power of a test for detecting «missing not at random"*
19:45 – 20:15	**J. Gillard** *Structured matrix completion and application to time series analysis*
20:15 – 20:45	**K. L. Tsui** *Systems health monitoring and management*

Κυριακή 22 Μαΐου 2022/Sunday May 22, 2022	
Αίθουσα Δ/Room D 08:30 – 11:15	Εκπαιδευτικό Επιμορφωτικό Σεμινάριο eStat (Part B)/Training Seminar eStat (Part B) Προεδρεύων/Chair: Α. Καραγρηγορίου/A. Karagrigoriou
08:30 – 11:15	Χ. Κουκούμης (Εισηγητής) eStat Μέρος Β/eStat (Part B)
Αίθουσα Ε/Room E 08:30 – 11:00	Προσκεκλημένη Συνεδρία ΙΙΙ/Invited Session III Προεδρεύων/Chair: Ι. Τριανταφύλλου/I. Triantafyllou
08:30 – 09:00	**A. Mukherjee** *On a more realistic way of comparing various extensions of EWMA with Original EWMA in the light of current controversies – An illustration using Lepage-type schemes*
09:00 – 09:30	**V.S. Barbu** *Testing statistical hypotheses for semi-Markov processes*
09:30 – 10:00	**H. Annoye, A. Beretta, C. Heuchenne, I.-M. Jensen** *Statistical matching using KCCA, Super-OM and Autoencoders-CCA*
10:00 – 10:30	**O. Assad, J. Gamain, C. Tudor** *Statistical inference for the stochastic wave equation*
10:30 – 11:00	**H.-P. Piepho, M. Boer, E.R. Williams** *Two-dimensional P-spline smoothing for spatial analysis of plant breeding trials*
11:15 – 11:30	***ΔΙΑΛΕΙΜΜΑ – ΚΑΦΕΣ/COFFEE BREAK***
Αίθουσα Ε/Room E 11:30 – 14:00	Προσκεκλημένη Συνεδρία IV/Invited Session IV Προεδρεύων/Chair: Ι. Ντζούφρας/I. Ntzoufras
11:30 – 12:00	**L. Egidi** *A Bayesian fairy tale: The mysteries of the mixtures, priors, likelihood and other 'multi-headed' monsters*
12:00 – 12:30	**M. Marsman, J. Haslbeck** *Bayesian analysis of a Markov random field model for ordinal variables*
12:30 – 13:00	**D. Rossell, M. Torrens, O. Papaspiliopoulos** *Confounder importance learning for treatment effect inference*
13:00 – 13:30	**F. Castelletti, G. Consonni** *Bayesian graphical modelling for heterogeneous causal effects*
13:30 – 14:00	**C. P. Robert** *Bayesian model choice in finite and infinite mixtures*
14:00 – 16:00	***ΜΕΣΗΜΒΡΙΝΗ ΔΙΑΚΟΠΗ/MIDDAY BREAK***
Αίθουσα Α/Room A 16:00 – 18:00	Ειδική Συνεδρία ΙΙΙ/Special Session III Εφαρμοσμένη Οικονομική και Υποδείγματα/Applied Economics and Models Προεδρεύων/Chair: Ζ. Μπραγουδάκης/Z. Bragoudakis

Κυριακή 22 Μαΐου 2022/Sunday May 22, 2022	
16:00 – 16:20	**Z. Georganta** *Latent variable models as a combination of reflective and formative structural formations*
16:20 – 16:40	**Χ. Μαλέσιος, Α. Κανταρτζής, Π. Λεμονάκης, Γ. Αραμπατζής** *Στάσεις και απόψεις των πολιτών για τη συμβολή των μονοπατιών στην προστασία του περιβάλλοντος και την περιφερειακά ανάπτυξη*
16:40 – 17:00	**D. Sideris, G. Pavlou** *Disaggregate income and wealth effects on private consumption in Greece*
17:00 – 17:20	**Z. Bragoudakis, I. Krompas** *Greek GDP forecasting using Bayesian multivariate models*
17:20 - 17:40	**Α. Τσιμπάνος, Χ. Αγιακλόγλου** *Η συμπεριφορά των πληροφοριακών κριτηρίων στην επιλογή χωρικών οικονομετρικών υποδειγμάτων*
17:40 – 18:00	**Δ. Παπαδόπουλος** *Η σχέση της πώλησης παράνομων ουσιών με την οικονομία. Το παράδειγμα των ΗΠΑ*
Αίθουσα Β/Room B **16:00 – 18:00**	**Εφαρμοσμένες Πιθανότητες και Αναλογισμός/Applied Probability and Actuarial** **Προεδρεύουσα/Chair: Μ. Βαμβακάρη/M. Vamvakari**
16:00 – 16:20	**M. Cheng, D.G. Konstantinides, D. Wang** *Uniform asymptotic estimates in a bidimensional time-dependent risk model with proportional reinsurance and general investment returns*
16:20 – 16:40	**A. Bozikas, G. Pitselis** *A mortality forecasting method based on non-linear credibility regression*
16:40 – 17:00	**I.S. Triantafyllou, M.V. Koutras** *Compound sooner waiting time problems in multistate trials*
17:00 – 17:20	**Σ.Δ. Δαφνής, Μ.Β. Κούτρας, Φ.Σ. Μακρή** *Διωνυμική κατανομή τάξης k σε τροποποιημένη δυαδική ακολουθία*
17:20 – 17:40	**Χ. Κουτσαντώνη, Ι. Δημητρίου** *Το M/G/1 σύστημα αναμονής με διακοπές και μεταβλητό ρυθμό αφίξεων*
17:40 – 18:00	**Β.Ε. Παπαγεωργίου, Γ. Τσακλίδης** *Ένα μεικτό SEIHCRDV-UKF μοντέλο για την πρόβλεψη του COVID-19. Εφαρμογή στις ημερήσιες καταγραφές πανδημίας στη Γαλλία*
Αίθουσα Γ/Room C **16:00 – 18:00**	**Posters (Αναρτημένες Εργασίες)** Τα posters θα παραμείνουν αναρτημένα καθ' όλη τη διάρκεια του Συνεδρίου
	Μ. Γανοπούλου, Α. Μπούτσικα, Μ. Μιχαηλίδης, Ι. Μελλίδου, Α. Δαλακούρας, Χ. Μπαζάκος, Θ. Μωυσιάδης, Λ. Αγγελής, Ι. Γανόπουλος, Α. Ξανθοπούλου *Ανάπτυξη αιτιατών μοντέλων για τον καθορισμό του μοριακού προφίλ της πατάτας Νάξου*

Κυριακή 22 Μαΐου 2022/Sunday May 22, 2022	
	Π. Ε. Τζαβάρας *Έρευνα για την επίδραση των Μέσων Κοινωνικής Δικτύωσης (ΜΚΔ) στους χρήστες, σε συσχέτιση με την ηλικία, την εκπαίδευση και την απασχόλησή τους*
	E.-N. Kalligeris, A. Makrides *Statistical process control techniques for monitoring the dynamics of SARS-COV-2*
	A. Makrides *Reliability inference for actuarial-financial mathematics*
	D. E. Pavlidis, G.-J. E. Nychas *Exploring the dynamics of a 12-MOS electronic nose sensor for meat analysis purposes*
18:00 – 18:30	***ΔΙΑΛΕΙΜΜΑ – ΚΑΦΕΣ/COFFEE BREAK***
Αίθουσα Α/Room A **18:30 – 19:50**	**Ειδική Συνεδρία IV/Special Session IV** Στατιστικές και Υπολογιστικές Μέθοδοι στις Κοινωνικές Επιστήμες/ Statistical and Computational Methods in Social Sciences **Προεδρεύων/Chair: Φ.Σ. Μηλιένος/F.S. Milienos**
18:30 – 18:50	**P. Papastamoulis, I. Ntzoufras** *Post-processing MCMC outputs of Bayesian factor analytic models*
18:50 – 19:10	**S. Nikolakopoulos** *Misuse of the sign test in narrative synthesis of evidence*
19:10 – 19:30	**C. Parpoula** *Change-point analysis methods for Public Health decision-making*
19:30 – 19:50	**F.S. Milienos** *Some properties of a new class of cure models*
Αίθουσα Β/Room B **18:30 – 20:10**	**Εφαρμοσμένη Στατιστική III/Applied Statistics III** **Προεδρεύων/Chair: Σ. Μπερσίμης/S. Bersimis**
18:30 – 18:50	**K. Bourazas, F. Sobas, P. Tsiamyrtzis** *Predictive Ratio Cusum (PRC): A Bayesian approach in online change point detection of short runs*
18:50 – 19:10	**T. Tsiampalis, D.B. Panagiotakos** *A Bayesian approach to inference for defective cure rate models under the assumption of right censoring mechanism*
19:10 – 19:30	**E.-N. Kalligeris, A. Karagrigoriou, C. Parpoula** *Changepoint detection and modelling of incidence data*
19:30 – 19:50	**T. Gkelsinis, V.S. Barbu** *A new class of test statistics for Markov chains with prior information on the transitions*
19:50 – 20:10	**D. Fouskakis, I. Ntzoufras** *Power-Expected-Posterior priors as mixtures of g-priors in normal linear models*

Κυριακή 22 Μαΐου 2022/Sunday May 22, 2022	
Αίθουσα Γ/Room C 18:30 – 20:10	Εφαρμοσμένες Πιθανότητες και Στατιστική/Applied Probability and Statistics Προεδρεύων/Chair: Σ.Δ. Δαφνής/S.D. Dafnis
18:30 – 18:50	**Α. Σκαμνιά, Ε. Μπεκρή, Π. Οικονόμου** *Ανάλυση περιφερειακών μετρήσεων βροχοπτώσεων περιφερειών: Η περίπτωση της Πελοποννήσου και των νησιών του Ιονίου*
18:50 – 19:10	**T. Perdikis** *An exponentially weighted moving average control chart based on signed ranks for finite horizon processes*
19:10 – 19:30	**Χ. Ευαγγελάρας, Β. Τραπουζανλής** *Definitive screening σχεδιασμοί και προβολικές ιδιότητες*
19:30 – 19:50	**Α.Χ. Ρακιτζής, N. Kumar, S. Chakraborti, T. Singh** *Διαγράμματα ελέγχου με κανόνες ροών για χρόνους μεταξύ συμβάντων*
19:50 – 20:10	**Σ.Δ. Δαφνής, Φ.Σ. Μακρή** *Η κατανομή του συνολικού αριθμού των επιτυχιών σε ασθενείς ροές περιορισμένου μήκους*
20:10 – 20:20	***ΣΥΝΤΟΜΟ ΔΙΑΛΛΕΙΜΑ/SHORT BREAK***
Αίθουσα Α/Room A 20:20	ΛΗΞΗ ΣΥΝΕΔΡΙΟΥ/CONFERENCE CLOSING Προεδρείο/Chairs: Ε. Αγγελής, Α. Καραγρηγορίου, Γ. Παπαδόπουλος/ E. Aggelis, A. Karagrigoriou, G. Papadopoulos • Απονομή Βραβείου Καλύτερης Εργασίας Νέου Στατιστικού/Young Greek Statistician Award • Προτάσεις για το Συνέδριο Στατιστικής του 2023/Plans for Panhellenic Statistics Conference 2023 • Συζήτηση/Discussion and Concluding Remarks

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 34ου Πανελληνίου Συνεδρίου Στατιστικής (2022), σελ.21- 33

ΤΟ ΟΜΟΓΕΝΕΣ ΜΑΡΚΟΒΙΑΝΟ ΣΥΣΤΗΜΑ ΔΙΑΚΡΙΤΟΥ ΧΡΟΝΟΥ ΜΕ ΤΥΧΑΙΕΣ ΧΩΡΗΤΙΚΟΤΗΤΕΣ. ΕΦΑΡΜΟΓΗ ΣΤΗ ΜΕΛΕΤΗ ΟΥΡΑΣ ΜΕ ΤΥΧΑΙΟ ΑΡΙΘΜΟ ΕΞΥΠΗΡΕΤΗΤΩΝ

Γεώργιος Βασιλειάδης
Τμήμα Μαθηματικών,
Πανεπιστήμιο Δυτικής Μακεδονίας
gvasiliadis@uowm.gr

ΠΕΡΙΛΗΨΗ

Στην παρούσα εργασία παρουσιάζεται αρχικά το ομογενές Μαρκοβιανό σύστημα (ΟΜΣ) διακριτού χρόνου με τυχαίες χωρητικότητες στις καταστάσεις του. Πρόκειται για ένα κλειστό ΟΜΣ διακριτού χρόνου, μεγέθους $N \in \mathbb{N}$, στο οποίο θεωρούμε ότι κάθε κατάσταση i του χώρου καταστάσεων $S = \{1, 2, \ldots, k\}$, παρουσιάζει σε κάθε χρονική στιγμή $t \in \mathbb{N}$, μία πεπερασμένη χωρητικότητα $c_i(t)$, $i = 1, 2, \ldots, k$, η οποία είναι τυχαία μεταβλητή με τιμές στο σύνολο $\{0, 1, 2, \ldots, N\}$ και γνωστή κατανομή. Στη συνέχεια το γενικό αυτό μοντέλο χρησιμοποιείται για να μελετήσουμε ένα σύστημα αναμονής διακριτού χρόνου στο οποίο το πλήθος των διαθέσιμων εξυπηρετητών δεν είναι σταθερό σε κάθε χρονική στιγμή. Για το υπό μελέτη σύστημα αναμονής θεωρούμε ότι τόσο οι αφίξεις όσο και οι εξυπηρετήσεις πραγματοποιούνται σύμφωνα με τη γεωμετρική κατανομή και το πλήθος των πελατών σε κάθε χρονική στιγμή δεν μπορεί να ξεπεράσει ένα πεπερασμένο αριθμό πελατών N. Όσον αφορά το πλήθος των εξυπηρετητών του συστήματος, θεωρούμε ότι είναι $s \in \mathbb{N}, s \geq 2$, και σε κάθε χρονική στιγμή κάθε ένας από αυτούς έχει μία σταθερή πιθανότητα θ να μην είναι ενεργός (να είναι χαλασμένος) για τη συγκεκριμένη χρονική στιγμή.

Λέξεις Κλειδιά: Μαρκοβιανά συστήματα διακριτού χρόνου, Μαρκοβιανά συστήματα με χωρητικότητες, συστήματα αναμονής διακριτού χρόνου.

1. ΕΙΣΑΓΩΓΗ

Θεωρούμε ένα σύστημα τα μέλη του οποίου είναι ταξινομημένα σε k, $k \in$, καταστάσεις σύμφωνα με κάποια χαρακτηριστικά τους. Οι καταστάσεις του συστήματος ορίζονται με τέτοιο τρόπο ώστε σε κάθε χρονική στιγμή κάθε μέλος να ανήκει σε μία μόνο κατάσταση. Όλα τα μέλη του συστήματος έχουν τη δυνατότητα μετακίνησης από κατάσταση σε κατάσταση σύμφωνα με τη μαρκοβιανή ιδιότητα σε χρόνο διακριτό. Θεωρούμε επίσης ότι το πλήθος των μελών του συστήματος είναι

σταθερό. Δεν υπάρχει η δυνατότητα εισόδου νέων μελών ούτε επιτρέπεται η έξοδος στους ήδη υπάρχοντες.

Ένα σύστημα με τα παραπάνω χαρακτηριστικά ονομάζεται κλειστό Μαρκοβιανό σύστημα διακριτού χρόνου (Bartholomew, 1982) και για τη μελέτη του χρησιμοποιούμε τους παρακάτω συμβολισμούς:

- $t = 1, 2, \ldots$, η παράμετρος που δηλώνει τα βήματα (χρονικές στιγμές),
- $S = \{1, 2, \ldots, k\}$, ο χώρος καταστάσεων,
- N, $N \in \mathbb{N}$, το πλήθος των μελών του συστήματος.

Θεωρούμε επομένως ένα σύστημα τα N μέλη του οποίου είναι κατανεμημένα στις k σε πλήθος καταστάσεις και πραγματοποιούν μετακινήσεις μέσα στο σύστημα καθώς ο χρόνος ρέει διακριτά. Συμβολίζουμε με p_{ij}, $i, j = 1, 2, \ldots, k,$ την πιθανότητα μετάβασης από την κατάσταση i στη j σ' ένα βήμα. Οι πιθανότητες αυτές είναι σταθερές, ανεξάρτητες του χρόνου και δίνονται ως στοιχεία ενός πίνακα διάστασης $k \times k$, $\mathbf{P} = \left(p_{ij} \right)$, $i, j = 1, 2, \ldots, k$, ο οποίος ονομάζεται πίνακας μετάβασης του συστήματος. Επίσης συμβολίζουμε με $n_{ij}(t)$, $i, j = 1, 2, \ldots, k,$ το πλήθος των μελών του συστήματος που μετακινούνται από την κατάσταση i στη j ανάμεσα στις χρονικές στιγμές t και $t+1,$ και με $n_i(t)$, $i = 1, 2, \ldots, k,$ το πλήθος των μελών που βρίσκονται στην i κατάσταση τη χρονική στιγμή t. Η πληθυσμιακή δομή του συστήματος σε κάθε χρονική στιγμή t δίνεται από το διάνυσμα

$$\mathbf{n}(t) = \left(n_1(t), n_2(t), \ldots, n_k(t) \right),$$

το οποίο ονομάζεται διάνυσμα κατάστασης του συστήματος για τη χρονική στιγμή t. Για τις βασικές έννοιες στις μαρκοβιανές αλυσίδες και τη θεμελίωση του ομογενούς Μαρκοβιανού συστήματος (ΟΜΣ) αναφέρουμε ενδεικτικά τα Bartholomew (1982), Isaacson and Madsen (1976), Vassiliou (1997), ενώ σχετικά με εφαρμογές των ΟΜΣ αναφέρουμε τις εργασίες McClean et al (1998), Tsaklidis and Soldatos (2003), Dimitriou and Tsantas (2010), Odhner and Asada (2010), Lykou and Tsaklidis (2021).

Γενίκευση του κλασικού ΟΜΣ διακριτού χρόνου αποτελεί ένα σύστημα στο οποίο κάθε κατάσταση i του χώρου καταστάσεων S παρουσιάζει πεπερασμένη χωρητικότητα $c_i \in \mathbb{N}$ (ΟΜΣ/c). Το γεγονός ότι κάθε κατάσταση i του χώρου καταστάσεων S ενός ΟΜΣ διακριτού χρόνου παρουσιάζει πεπερασμένη χωρητικότητα $c_i \in \mathbb{N}$, έχει ως αποτέλεσμα το μέγεθος οποιασδήποτε κατάστασης i να μην μπορεί να υπερβεί την τιμή c_i, δηλαδή

$$n_i(t) \le c_i, \ i \in S,$$

για κάθε χρονική στιγμή t, $t = 1, 2, \ldots$. Αναλυτική περιγραφή και μελέτη συστημάτων με αυτό το χαρακτηριστικό παρουσιάζεται στην εργασία Vasiliadis and Tsaklidis (2011).

Κατά τη μελέτη αυτών των συστημάτων θεωρήθηκε ότι η χωρητικότητα των καταστάσεων είναι σταθερή σε κάθε χρονική στιγμή. Η χωρητικότητα όμως μιας

κατάστασης είναι ένα μέγεθος που μπορεί να μεταβάλλεται στην εξέλιξη του χρόνου. Στην παρούσα εργασία εξετάζουμε την περίπτωση όπου οι χωρητικότητες των καταστάσεων ενός ΟΜΣ διακριτού χρόνου είναι τυχαίες μεταβλητές που ακολουθούν μία γνωστή κατανομή. Η μελέτη αυτού του συστήματος γίνεται με τη βοήθεια επαναληπτικών σχέσεων για τις παραγοντικές ροπές των μεγεθών των καταστάσεων του. Ως εφαρμογή αυτού του μοντέλου στη συνέχεια παρουσιάζεται ένα σύστημα αναμονής διακριτού χρόνου το οποίο δεν έχει σταθερό αριθμό εξυπηρετητών. Το σύστημα αυτό μπορεί να αναπαρασταθεί με τη βοήθεια ενός ΟΜΣ διακριτού χρόνου με τυχαίες χωρητικότητες στις καταστάσεις του.

2. ΤΟ ΟΜΣ ΔΙΑΚΡΙΤΟΥ ΧΡΟΝΟΥ ΜΕ ΤΥΧΑΙΕΣ ΧΩΡΗΤΙΚΟΤΗΤΕΣ

Θεωρούμε ένα κλειστό ΟΜΣ διακριτού χρόνου μεγέθους N, $N \in \mathbb{N}$, με χώρο καταστάσεων $S = \{1, 2, \ldots, k\}$ και με πίνακα μετάβασης $\mathbf{P} = \left(p_{ij} \right)$, $i, j = 1, 2, \ldots, k$. Κάθε κατάσταση i, $i \in S$, παρουσιάζει μία χωρητικότητα η οποία δεν είναι σταθερή στο χρόνο αλλά είναι μία τυχαία μεταβλητή που παίρνει τιμές στο σύνολο $\{0, 1, 2, \ldots, N\}$. Συμβολίζουμε με $c_i(t)$ την τυχαία μεταβλητή που παριστάνει τη χωρητικότητα της κατάστασης i τη χρονική στιγμή t. Για κάθε χρονική στιγμή t, θεωρούμε ότι οι τ.μ. $c_i(t)$ είναι ανεξάρτητες και ισόνομες με γνωστή συνάρτηση πιθανότητας $P(c_i(t) = x)$, $x = 0, 1, 2, \ldots, N$.

Το γεγονός ότι κάθε κατάσταση i του χώρου καταστάσεων S του συστήματος παρουσιάζει χωρητικότητα $c_i(t)$, έχει ως αποτέλεσμα το μέγεθος οποιασδήποτε κατάστασης i να μην μπορεί να υπερβεί την τιμή αυτή, δηλαδή

$$n_i(t) \le c_i(t), \ i \in S,$$

για κάθε χρονική στιγμή t. Έτσι, αν το πλήθος των μελών του συστήματος που αποφασίσουν να μετακινηθούν προς μία κατάσταση i, $i \in S$, σε κάποια χρονική στιγμή t δεν υπερβαίνει τη χωρητικότητα $c_i(t)$, τότε όλα τα μέλη εισέρχονται σ' αυτήν. Αν όμως σε κάποια χρονική στιγμή t, το πλήθος αυτό υπερβεί τη χωρητικότητα, τότε στην κατάσταση i εισέρχονται μόνο $c_i(t)$ από αυτά τα μέλη και τα μέλη αυτά αποτελούν το μέγεθος της κατάστασης για το χρονικό διάστημα $(t, t+1]$. Στην περίπτωση αυτή λέμε ότι η κατάσταση i παρουσιάζει υπερχείλιση. Θεωρούμε ότι τα μέλη του συστήματος που υπερχειλίζουν από μία κατάσταση σε μία χρονική στιγμή εισέρχονται σε ένα εικονικό αποθηκευτικό χώρο της κατάστασης και δεν μετακινούνται πλέον σύμφωνα με τον πίνακα μετάβασης $\mathbf{P}$. Συμβολίζουμε με $e_i(t)$, $i = 1, 2, \ldots, k$, το πλήθος των μελών που βρίσκονται στον αποθηκευτικό χώρο της i κατάστασης τη χρονική στιγμή t. Τα μέλη του συστήματος που βρίσκονται στον αποθηκευτικό χώρο μιας κατάστασης i, $i \in S$, έχουν τη δυνατότητα να επιστρέψουν στο σύστημα (να εισέλθουν στην κατάσταση i) και να συνεχίσουν να μετακινούνται

σύμφωνα με τον πίνακα μετάβασης $\mathbf{P}$. Όταν οι μετακινήσεις από και προς την κατάσταση i και η χωρητικότητα αυτής στις επόμενες χρονικές στιγμές δημιουργήσουν κενές θέσεις (το μέγεθος της κατάστασης δεν υπερβαίνει τη χωρητικότητα αυτής) τότε οι θέσεις αυτές συμπληρώνονται με μέλη που τυχόν βρίσκονται στον αποθηκευτικό χώρο αυτής.

Αν συμβολίσουμε με $m_i(t),\ i=1,2,\ldots,k$, το πλήθος των μελών που βρίσκονται στην κατάσταση i και στον αποθηκευτικό χώρο αυτής τη χρονική στιγμή t, τότε

$$n_i(t)=\begin{cases} m_i(t), & m_i(t)<c_i(t) \\ c_i(t), & m_i(t)\geq c_i(t) \end{cases}$$

ή

$$m_i(t)=n_i(t)+e_i(t),\ i=1,2,\ldots,k.$$

Επομένως το μέγεθος μιας κατάστασης i, $i\in S$, σε μία χρονική στιγμή t, που δίνεται από την τ.μ. $n_i(t)$, μπορεί να προσδιοριστεί άμεσα αν γνωρίζουμε την τιμή της τ.μ. $m_i(t)$ και τη χωρητικότητα της κατάστασης τη χρονική στιγμή t. Έτσι, για τη μελέτη της συμπεριφοράς του συστήματος στην εξέλιξη του χρόνου μπορούμε να χρησιμοποιήσουμε τις τ.μ. $m_i(t),\ i\in S$.

Στην Πρόταση 1 που ακολουθεί δίνεται μία επαναληπτική σχέση για τις αναμενόμενες τιμές των τ.μ. $m_i(t),\ i\in S$.

Πρόταση 1. *Έστω ένα ΟΜΣ με τυχαίες χωρητικότητες στις καταστάσεις του και με πίνακα μετάβασης* $\mathbf{P}=\left(p_{ij}\right),\ i,j=1,2,\ldots,k$. *Τότε οι αναμενόμενες τιμές των τ.μ.* $m_i(t+1),\ i=1,2,\ldots,k$, *δίνονται από τη σχέση*

$$E[\mathbf{m}(t+1)]=E[\mathbf{m}(t)]\mathbf{P}+\sum_{\mathbf{c}}\big(E[\mathbf{e}(t)\,|\,\mathbf{c}(t)=\mathbf{c}]P(\mathbf{c}(t)=\mathbf{c})\big)(\mathbf{I}-\mathbf{P}), \qquad (1)$$

όπου $\mathbf{m}(t)=\left(m_1(t),m_2(t),\ldots,m_k(t)\right)$, $\mathbf{e}(t)=\left(e_1(t),e_2(t),\ldots,e_k(t)\right)$, $\mathbf{c}(t)=\left(c_1(t),c_2(t),\ldots,c_k(t)\right)$ και $\mathbf{I}$ ο $k\times k$ μοναδιαίος πίνακας.

Απόδειξη. Για ένα ΟΜΣ με σταθερές χωρητικότητες στις καταστάσεις του ισχύει (Βασιλειάδης, 2019)

$$E[\mathbf{m}(t+1)]=E[\mathbf{n}(t)]\mathbf{P}+E[\mathbf{e}(t)].$$

Έτσι, αν θεωρήσουμε τη δεσμευμένη τ.μ. $\mathbf{m}(t+1)\,|\,\mathbf{c}(t)$ θα έχουμε:

$$\begin{aligned} E[\mathbf{m}(t+1)\,|\,\mathbf{c}(t)=\mathbf{c}] &= E[\mathbf{n}(t)\,|\,\mathbf{c}(t)=\mathbf{c}]\mathbf{P}+E[\mathbf{e}(t)\,|\,\mathbf{c}(t)=\mathbf{c})] \\ &= E[\mathbf{m}(t)-\mathbf{e}(t)\,|\,\mathbf{c}(t)=\mathbf{c}]\mathbf{P}+E[\mathbf{e}(t)\,|\,\mathbf{c}(t)=\mathbf{c})] \\ &= E[\mathbf{m}(t)\,|\,\mathbf{c}(t)=\mathbf{c}]\mathbf{P}+E[\mathbf{e}(t)\,|\,\mathbf{c}(t)=\mathbf{c})](\mathbf{I}-\mathbf{P}) \\ &= E[\mathbf{m}(t)]\mathbf{P}+E[\mathbf{e}(t)\,|\,\mathbf{c}(t)=\mathbf{c})](\mathbf{I}-\mathbf{P}). \end{aligned}$$

Άρα

$$E[\mathbf{m}(t+1)] = E[E[\mathbf{m}(t+1) \mid \mathbf{c}(t)]]$$
$$= \sum_{\mathbf{c}} E[\mathbf{m}(t+1) \mid \mathbf{c}(t) = \mathbf{c}]P(\mathbf{c}(t) = \mathbf{c})$$
$$= E[\mathbf{m}(t)]\mathbf{P} + \sum_{\mathbf{c}} (E[\mathbf{e}(t) \mid \mathbf{c}(t) = \mathbf{c}]P(\mathbf{c}(t) = \mathbf{c}))(\mathbf{I} - \mathbf{P}).$$

□

Από τη σχέση (1) προκύπτει ότι για να μπορέσουμε να υπολογίσουμε τις αναμενόμενες τιμές των τ.μ. $m_i(t)$, $i = 1, 2, \ldots, k$, για μία χρονική στιγμή t, χρειαζόμαστε τις αναμενόμενες τιμές $E[\mathbf{e}(t-1) \mid \mathbf{c}(t-1) = \mathbf{c}]$ για όλες τις δυνατές τιμές που μπορούμε να έχουμε στις χωρητικότητες των καταστάσεων τη χρονική στιγμή $t-1$. Για τον υπολογισμό αυτών των αναμενόμενων τιμών χρειαζόμαστε την κατανομή των τ.μ. $m_i(t-1)$, η οποία μπορεί να προκύψει με τη βοήθεια των μεικτών παραγοντικών ροπών αυτών των τ.μ.

Για τον σκοπό αυτό διατυπώνουμε στη συνέχεια την Πρόταση 2, στην οποία δίνεται μια σχέση για τον υπολογισμό των μεικτών παραγοντικών ροπών των τ.μ. $m_i(t)$, $i = 1, 2, \ldots, k$. Για τη διατύπωση της πρότασης χρησιμοποιούμε ένα γινόμενο διανυσμάτων (χρησιμοποιούμε το σύμβολο $\times$), το οποίο μοιάζει με το γινόμενο Kronecker. Συγκεκριμένα, για δύο διανύσματα $\mathbf{x}_1$, $\mathbf{x}_2$ το αποτέλεσμα του γινομένου $\mathbf{x}_1 \times \mathbf{x}_2$ είναι ένα διάνυσμα γραμμή, τα στοιχεία του οποίου προκύπτουν από το Kronecker γινόμενο $\mathbf{x}_1 \otimes \mathbf{x}_2$, με τη διαφορά ότι οι δυνάμεις αντικαθίστανται με παραγοντικά. Για παράδειγμα, αν $\mathbf{x}_1 = (a, b)$, $\mathbf{x}_2 = (c, d)$, τότε έχουμε

$$\mathbf{x}_1 \times \mathbf{x}_2 = (ac, ad, bc, bd),$$

και

$$\mathbf{x}_1 \times \mathbf{x}_1 = (a(a-1), ab, ba, b(b-1)),$$
$$\mathbf{x}_2 \times \mathbf{x}_2 = (c(c-1), cd, dc, d(d-1)).$$

Επίσης, κατά αναλογία της r - οστής δύναμης Kronecker $\mathbf{x}_1^{\otimes r}$ που ορίζεται από τις σχέσεις

$$\mathbf{x}_1^{\otimes 1} = \mathbf{x}_1, \mathbf{x}_1^{\otimes r} = \mathbf{x}_1 \otimes \mathbf{x}_1^{\otimes (r-1)}, r = 2, 3, \ldots,$$

θα έχουμε

$$\mathbf{x}_1^{\times 1} = \mathbf{x}_1, \mathbf{x}_1^{\times r} = \mathbf{x}_1 \times \mathbf{x}_1^{\times (r-1)}, r = 2, 3, \ldots.$$

Πρόταση 2. *Έστω ένα ΟΜΣ με τυχαίες χωρητικότητες στις καταστάσεις του και με πίνακα μετάβασης* $\mathbf{P} = \left(p_{ij} \right)$, $i, j = 1, 2, \ldots, k$. *Τότε οι μεικτές παραγοντικές ροπές των τ.μ.* $m_i(t+1)$, $i = 1, 2, \ldots, k$, *δίνονται από τη σχέση*

$$E[\mathbf{m}(t+1)^{\times r}] = \sum_{\mathbf{c}} \left(E[\mathbf{n}(t)^{\times r}]\mathbf{P}^{\otimes r} \right.$$

$$
\begin{aligned}
&+E[\mathbf{n}(t)^{\times(r-1)} \times \mathbf{e}(t)](\mathbf{P}^{\otimes(r-1)} \otimes \mathbf{I}) \\
&+E[\mathbf{n}(t)^{\times(r-2)} \times \mathbf{e}(t) \times \mathbf{n}(t)](\mathbf{P}^{\otimes(r-2)} \otimes \mathbf{I} \otimes \mathbf{P}) \\
&+\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
&+E[\mathbf{e}(t)^{\times(r-1)} \times \mathbf{n}(t)](\mathbf{I}^{\otimes(r-1)} \otimes \mathbf{P}) \\
&+E[\mathbf{e}(t)^{\times(r-2)} \times \mathbf{n}(t) \times \mathbf{e}(t)](\mathbf{I}^{\otimes(r-2)} \otimes \mathbf{P} \otimes \mathbf{I}) \\
&+E[\mathbf{e}(t)^{\times r}]\mathbf{I}^{\otimes r}\big) P(\mathbf{c}(t) = \mathbf{c}), \qquad (2)
\end{aligned}
$$

όπου $r \in \square$.

Απόδειξη. Το αποτέλεσμα προκύπτει άμεσα χρησιμοποιώντας την αντίστοιχη επαναληπτική σχέση που ισχύει στην περίπτωση των σταθερών χωρητικοτήτων στις καταστάσεις ενός ΟΜΣ (Βασιλειάδης, 2019) και δεσμεύοντας ως προς την τ.μ. $\mathbf{c}(t)$ που παριστά τις χωρητικότητες που μπορούμε να έχουμε στις καταστάσεις τη χρονική στιγμή t .

□

Με τη βοήθεια των παραγοντικών και των μεικτών παραγοντικών ροπών των τ.μ. $m_i(t),\ i = 1, 2, \ldots, k$, μπορούμε να προσδιορίσουμε τις κατανομές αυτών των τ.μ. για κάθε χρονική στιγμή t (Vasiliadis and Tsaklidis, 2011). Έτσι καταλήγουμε στο επόμενο πόρισμα, όπου δίνεται η κατανομή των τ.μ. $m_i(t),\ i = 1, 2, \ldots, k$ και η από κοινού κατανομή αυτών των τ.μ. (Βασιλειάδης, 2019).

Πόρισμα 1. *Έστω ένα ΟΜΣ διακριτού χρόνου με χωρητικότητες στις καταστάσεις του και με πίνακα μετάβασης* $\mathbf{P} = \left(p_{ij} \right),\ i, j = 1, 2, \ldots, k$. *Οι κατανομές των τ.μ.* $m_i(t)$, $i = 1, 2, \ldots, k$, *για κάθε χρονική στιγμή* t , *δίνονται από τη σχέση*

$$
P\left[m_i(t) = n \right] = \frac{1}{n!} \sum_{j=0}^{N} \frac{(-1)^j}{j!} E[m_i^{(n+j)}(t)], \qquad (3)
$$

για $i = 1, 2, \ldots, k$, $t = 1, 2, \ldots,$ *και* $n = 0, 1, \ldots, N$.

Η από κοινού κατανομή των τ.μ. $m_i(t)$, $i = 1, 2, \ldots, k$, *για κάθε χρονική στιγμή* t , *δίνεται από τη σχέση*

$$
P\left[m_1(t) = r_1, m_2(t) = r_2, \ldots, m_k(t) = r_k \right] = \frac{1}{\prod_{i=1}^{k} r_i!} E\left[\prod_{i=1}^{k} m_i^{(r_i)}(t) \right]. \qquad (4)
$$

Χρησιμοποιώντας τη σχέση (3) μπορούμε να προσδιορίσουμε και την κατανομή των τ.μ. $n_i(t)$ και $e_i(t)$, $i = 1, 2, \ldots, k$, για μία χρονική στιγμή t , αφού η τιμή της τ.μ. μεταβλητής $m_i(t)$, αν γνωρίζουμε τη χωρητικότητα της κατάστασης i τη χρονική στιγμή t , καθορίζει άμεσα και τις τιμές των τ.μ. $n_i(t)$ και $e_i(t)$. Είναι

$$P\left[n_i(t)=n\right]=\sum_{c=0}^{N}P\left[n_i(t)=n\mid c_i(t)=c\right]P\left[c_i(t)=c\right],$$

όπου

$$P\left[n_i(t)=n\mid c_i(t)=c\right]=\begin{cases}P\left[m_i(t)=n\right], & n=0,1,\ldots,c-1,\\ \sum_{w=c_i}^{N}P\left[m_i(t)=w\right], & n=c,\end{cases}$$

και

$$P\left[e_i(t)=n\right]=\sum_{c=0}^{N}P\left[e_i(t)=n\mid c_i(t)=c\right]P\left[c_i(t)=c\right],$$

όπου

$$P\left[e_i(t)=n\mid c_i(t)=c\right]=\begin{cases}\sum_{w=0}^{c}P\left[m_i(t)=w\right], & n=0,\\ P\left[m_i(t)=n+c\right], & n=1,2,\ldots,N-c.\end{cases}$$

Θέλοντας να υπολογίσουμε τις παραγοντικές ροπές των τ.μ. $m_i(t)$, $i=1,2,\ldots,k$, σε μία χρονική στιγμή t χρησιμοποιώντας την Πρόταση 2, είναι απαραίτητο να γνωρίζουμε τις παραγοντικές ροπές και τις μεικτές παραγοντικές ροπές των τ.μ. $n_i(t-1)$, $e_i(t-1)$, $i=1,2,\ldots,k$, οι οποίες προκύπτουν αν γνωρίζουμε την κατανομή του διανύσματος $\mathbf{m}(t-1)$ και την κατανομή της χωρητικότητας κάθε κατάστασης για τη χρονική στιγμή $t-1$. Επομένως, θεωρώντας ότι η κατανομή του αρχικού διανύσματος $\mathbf{m}(0)$ και η κατανομή της χωρητικότητας κάθε κατάστασης σε κάθε χρονική στιγμή είναι γνωστά, μπορούμε να χρησιμοποιήσουμε επαναληπτικά την Πρόταση 2 και το Πόρισμα 1 έτσι ώστε να υπολογίσουμε τόσο τις παραγοντικές ροπές των τ.μ. $m_i(t)$, $i=1,2,\ldots,k$, όσο και την κατανομή του διανύσματος $\mathbf{m}(t)$ για οποιαδήποτε χρονική στιγμή t.

Ο αλγόριθμος που περιγράψαμε παραπάνω είναι επαναληπτικός και για τον υπολογισμό της κατανομής του διανύσματος $\mathbf{m}(t)$ για μία χρονική στιγμή t, απαιτείται ο υπολογισμός των κατανομών του ίδιου διανύσματος για όλες τις προηγούμενες χρονικές στιγμές $t-1, t-2, \ldots, 1$. Ένας εναλλακτικός τρόπος προσδιορισμού αυτής της κατανομής είναι να θεωρήσουμε μία νέα μαρκοβιανή αλυσίδα με χώρο καταστάσεων S' το σύνολο όλων των δυνατών δομών $\mathbf{m}(t)$ που μπορούμε να έχουμε οποιαδήποτε χρονική στιγμή t. Αφού το μέγεθος του συστήματος είναι πεπερασμένο και ίσο με N το πλήθος των δυνατών δομών (συμβολίζουμε με l) θα είναι πεπερασμένο και ίσο με

$$l=\binom{k+N-1}{N}.$$

Η πιθανότητα μετάβασης από μία δομή σε μία άλλη είναι σταθερή στον χρόνο αφού η συνάρτηση κατανομής για τις χωρητικότητες των καταστάσεων είναι η ίδια για κάθε χρονική στιγμή. Έτσι, αν θεωρήσουμε την ομογενή μαρκοβιανή αλυσίδα με χώρο καταστάσεων $S' = \{\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_l\}$, πίνακα μετάβασης $\mathbf{Q} = \left(q_{ij}\right),\ i, j \in S'$ και πιθανότητες μετάβασης

$$q_{ij} = P(\mathbf{m}(t+1) = \mathbf{m}_j \mid \mathbf{m}(t) = \mathbf{m}_i),\ i,\ j \in S',\ \forall t,$$

τότε

$$\mathbf{q}(t) = \mathbf{q}(0)\mathbf{Q}^t,$$

όπου

$$\mathbf{q}(t) = (P(\mathbf{m}(t) = \mathbf{m}_1), P(\mathbf{m}(t) = \mathbf{m}_2), \ldots, P(\mathbf{m}(t) = \mathbf{m}_l)).$$

Ο πίνακας μετάβασης αυτής της μαρκοβιανής αλυσίδας μπορεί να προσδιοριστεί με τη βοήθεια του προηγούμενου αλγορίθμου. Υποθέτοντας ότι $\mathbf{m}(t) = \mathbf{m}_i, i \in S'$, μπορούμε με τη βοήθεια της Πρότασης 2 και του Πορίσματος 1 να υπολογίσουμε την κατανομή του διανύσματος $\mathbf{m}(t+1)$ που είναι τα στοιχεία της i γραμμής του πίνακα $\mathbf{Q}$.

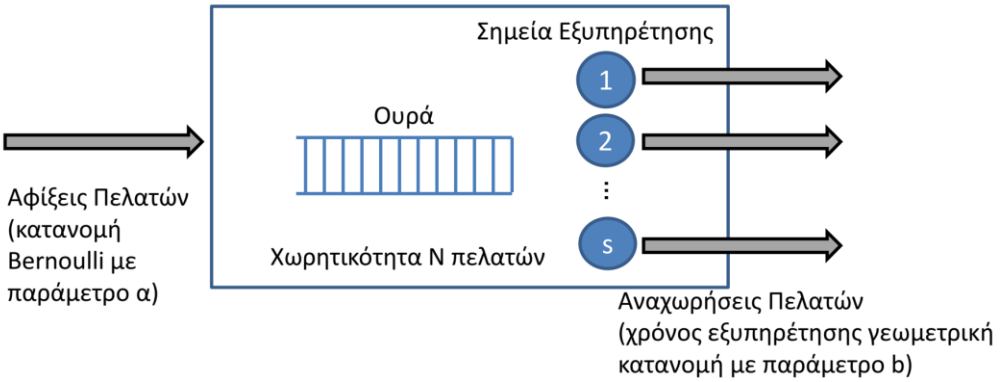
3. ΣΥΣΤΗΜΑ ΑΝΑΜΟΝΗΣ ΔΙΑΚΡΙΤΟΥ ΧΡΟΝΟΥ ΜΕ ΤΥΧΑΙΟ ΠΛΗΘΟΣ ΕΞΥΠΗΡΕΤΗΤΩΝ

Θεωρούμε ένα σύστημα αναμονής (ουρά) διακριτού χρόνου στο οποίο υπάρχουν $s \in \mathbb{N}, s \geq 2,$ παράλληλα σημεία εξυπηρέτησης και το σύστημα έχει πεπερασμένη χωρητικότητα N πελατών. Υποθέτουμε ότι οι αφίξεις των πελατών στο σύστημα πραγματοποιούνται σύμφωνα με την κατανομή Bernoulli με παράμετρο a και ο χρόνος που απαιτείται για την εξυπηρέτηση ενός πελάτη (σε οποιοδήποτε από τα σημεία εξυπηρέτησης) ακολουθεί τη γεωμετρική κατανομή με παράμετρο b. Θεωρούμε επίσης ότι σε κάθε χρονική στιγμή t $(t = 1, 2, \ldots)$, κάθε ένας από τους s εξυπηρετητές του συστήματος έχει μία σταθερή πιθανότητα θ να είναι χαλασμένος για τη χρονική περίοδο $(t, t+1]$ και να μην μπορεί να εξυπηρετήσει κάποιον πελάτη. Επομένως, το πλήθος από τα s σημεία εξυπηρέτησης που είναι ενεργά σε κάθε χρονική στιγμή και μπορούν να εξυπηρετήσουν ένα πελάτη δεν είναι σταθερό, αλλά είναι μία τυχαία μεταβλητή. Υποθέτουμε ότι τόσο ο καθορισμός της κατάστασης κάθε εξυπηρετητή (ενεργός ή χαλασμένος) όσο οι αφίξεις και οι αναχωρήσεις συμβαίνουν ταυτόχρονα στην αρχή κάθε χρονικής περιόδου t. Αν σε κάποια χρονική στιγμή ένας πελάτης φτάσει στο σύστημα για να εξυπηρετηθεί και όλοι οι ενεργοί εξυπηρετητές είναι κατειλημμένοι, τότε ο πελάτης εισέρχεται σε μία ουρά και περιμένει μέχρι να υπάρξει κάποιος διαθέσιμος ενεργός εξυπηρετητής. Στην ίδια ουρά εισέρχονται και οι πελάτες που δεν καταφέρνουν να ολοκληρώσουν την εξυπηρέτησή τους γιατί κατά τη διάρκεια της εξυπηρέτησης ο εξυπηρετητής χαλάει. Το σύστημα που περιγράψαμε παρουσιάζεται γραφικά στο Σχήμα 1.

Για να μελετήσουμε την εξέλιξη στον χρόνο της πληθυσμιακής δομής αυτού του συστήματος αναμονής, μπορούμε να χρησιμοποιήσουμε ένα Μαρκοβιανό σύστημα

διακριτού χρόνου με τυχαίες χωρητικότητες στις καταστάσεις του. Θεωρούμε ένα ΟΜΣ διακριτού χρόνου το οποίο αποτελείται από 2 καταστάσεις και $N+1$ μέλη, τα οποία αντιστοιχούν στους πελάτες του συστήματος αναμονής και μπορούν να βρίσκονται είτε στις καταστάσεις είτε στους αποθηκευτικούς χώρους αυτών.

Σχήμα 1. Σύστημα αναμονής διακριτού χρόνου



Η πρώτη κατάσταση αντιστοιχεί στην πηγή πελατών από την οποία πραγματοποιούνται οι αφίξεις των πελατών, ενώ η δεύτερη κατάσταση αντιστοιχεί στο σύστημα αναμονής. Σε κάθε χρονική στιγμή μπορεί να έχουμε το πολύ μία άφιξη στο σύστημα αναμονής, άρα από την πρώτη κατάσταση του ΟΜΣ μπορεί να μετακινηθεί μόνο ένα μέλος. Αυτό σημαίνει ότι, αν θεωρήσουμε ότι η χωρητικότητα της πρώτης κατάστασης είναι τ.μ., θα είναι $P(c_1(t)=1)=1, \forall t$. Από τη δεύτερη κατάσταση του ΟΜΣ, το πλήθος των μελών που μπορούν να μετακινηθούν μία χρονική στιγμή ισούται με το πλήθος των ενεργών εξυπηρετητών τη συγκεκριμένη χρονική στιγμή. Έτσι, η χωρητικότητα της δεύτερης κατάστασης σε μία χρονική στιγμή t θα ισούται με το πλήθος των ενεργών εξυπηρετητών τη χρονική στιγμή t . Το πλήθος των ενεργών εξυπηρετητών όμως είναι μία τυχαία μεταβλητή που ακολουθεί τη διωνυμική κατανομή με παραμέτρους s και $1-\theta$. Επομένως για την κατανομή της τ.μ. $c_2(t)$ (χωρητικότητα δεύτερης κατάστασης), θα έχουμε

$$P(c_2(t)=c)=\binom{s}{c}(1-\theta)^c\theta^{s-c},\ c=0,1,\ldots,s, \forall t.$$

Προφανώς ο αποθηκευτικός χώρος της δεύτερης κατάστασης $e_2(t)$, θα αντιστοιχεί στο μέγεθος της ουράς του συστήματος αναμονής τη χρονική στιγμή t .

Για να μπορέσουμε να μελετήσουμε τη συμπεριφορά του συστήματος χρειαζόμαστε τον πίνακα μετάβασης σύμφωνα με τον οποίο πραγματοποιούνται οι μετακινήσεις των μελών από κατάσταση σε κατάσταση. Ένα μέλος του συστήματος που σε κάποια χρονική στιγμή t είναι στην κατάσταση 1 (όχι στον αποθηκευτικό χώρο αυτής), μετακινείται στην κατάσταση 2 με πιθανότητα $p_{12}=a,$ αφού η πιθανότητα να

έχουμε μία άφιξη στο σύστημα αναμονής κατά τη χρονική στιγμή t είναι ίση με a. Επίσης, ένα μέλος του συστήματος που σε κάποια χρονική στιγμή t είναι στην κατάσταση 2 (όχι στον αποθηκευτικό χώρο αυτής), μετακινείται στην κατάσταση 1 με πιθανότητα $p_{21} = b,$ αφού η πιθανότητα να εξυπηρετηθεί ένα μέλος που βρίσκεται σε κάποιο ενεργό εξυπηρετητή του συστήματος αναμονής είναι ίση με b. Έτσι, οι μετακινήσεις των μελών του συστήματος από κατάσταση σε κατάσταση πραγματοποιούνται σύμφωνα με τον πίνακα μετάβασης

$$\mathbf{P} = \begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix}.$$

Η πληθυσμιακή δομή του συστήματος τη χρονική στιγμή t περιγράφεται από τα διανύσματα

$$\mathbf{m}(t) = \big(m_1(t), m_2(t)\big),\ \mathbf{n}(t) = \big(n_1(t), n_2(t)\big),\ \mathbf{e}(t) = \big(e_1(t), e_2(t)\big).$$

Εφαρμόζοντας τον επαναληπτικό αλγόριθμο που περιγράψαμε στην προηγούμενη ενότητα είμαστε σε θέση να υπολογίσουμε τόσο τις αναμενόμενες τιμές όσο και τις κατανομές των τ.μ. $m_i(t)$, $n_i(t)$, $e_i(t)$, $i = 1, 2,$ για οποιαδήποτε χρονική στιγμή t.

Θεωρώντας ότι στο ΟΜΣ υπάρχουν $N+1$ μέλη οι δυνατές τιμές για την τ.μ. $m_2(t)$ είναι $0, 1, 2, \ldots, N+1$. Στο σύστημα αναμονής που μελετάμε όμως, το πλήθος των πελατών δεν μπορεί να υπερβαίνει την τιμή N. Έτσι κατά την εφαρμογή του αλγορίθμου υπολογισμού της κατανομής του διανύσματος $\mathbf{m}(t)$, για να μπορέσουμε να υπολογίσουμε την κατανομή του μεγέθους του συστήματος αναμονής, θεωρούμε την τ.μ. $\tilde{\mathbf{m}}(t) = (\tilde{m}_1(t), \tilde{m}_2(t))$ για την οποία για κάθε χρονική στιγμή t ισχύει

$$P\big[\tilde{\mathbf{m}}(t) = (m_1, m_2)\big] = \begin{cases} P\big[\mathbf{m}(t) = (m_1, m_2)\big], (m_1, m_2) \neq (0, N+1), \\ P\big[\mathbf{m}(t) = (1, N)\big] + P\big[\mathbf{m}(t) = (0, N+1)\big],\ (m_1, m_2) = (1, N). \end{cases}$$

Έτσι η τ.μ. $\tilde{m}_1(t)$ παίρνει τιμές στο σύνολο $\{1, 2, \ldots, N+1\}$ και με αυτό τον τρόπο εξασφαλίζουμε ότι σε κάθε χρονική στιγμή t στην πρώτη κατάσταση του ΟΜΣ που αντιστοιχεί στην πηγή πελατών του συστήματος αναμονής, θα υπάρχει τουλάχιστον ένας πελάτης ο οποίος θα μπορεί να ζητήσει να εξυπηρετηθεί από το σύστημα την επόμενη χρονική στιγμή.

Αντίστοιχα μπορούμε να θεωρήσουμε τις τυχαίες μεταβλητές $\tilde{\mathbf{n}}(t) = (\tilde{n}_1(t), \tilde{n}_2(t))$, $\tilde{\mathbf{e}}(t) = (\tilde{e}_1(t), \tilde{e}_2(t))$, η κατανομή των οποίων για οποιαδήποτε χρονική στιγμή t προκύπτει άμεσα από την κατανομή της τ.μ. $\tilde{m}(t) = (\tilde{m}_1(t), \tilde{m}_2(t))$ και τις χωρητικότητες των καταστάσεων. Έτσι, για κάθε χρονική στιγμή t, το πλήθος των πελατών του συστήματος αναμονής θα δίνεται από την τ.μ. $\tilde{m}_2(t)$, το πλήθος των πελατών που εξυπηρετείται στο σημείο εξυπηρέτησης θα δίνεται από την τ.μ. $\tilde{n}_2(t)$, ενώ το πλήθος των πελατών που βρίσκονται στην ουρά του συστήματος αναμονής θα δίνεται από την τ.μ. $\tilde{e}_2(t)$. Τόσο οι αναμενόμενες τιμές όσο και η κατανομή αυτών των

τ.μ. μπορούν να υπολογιστούν για κάθε χρονική στιγμή t εφαρμόζοντας είτε τον επαναληπτικό αλγόριθμο που περιγράψαμε είτε χρησιμοποιώντας τη μαρκοβιανή αλυσίδα με χώρο καταστάσεων το σύνολο όλων των δυνατών δομών $\tilde{\mathbf{m}}(t)$ που μπορούμε να έχουμε οποιαδήποτε χρονική στιγμή t.

4. ΑΡΙΘΜΗΤΙΚΟ ΠΑΡΑΔΕΙΓΜΑ

Θεωρούμε ένα σύστημα αναμονής το οποίο έχει δύο εξυπηρετητές και μέγιστη χωρητικότητα 6 πελάτες. Σε κάθε χρονική στιγμή επομένως το πλήθος των πελατών που βρίσκονται στην ουρά και στα σημεία εξυπηρέτησης δεν μπορεί να ξεπερνά τους 6 πελάτες. Υποθέτουμε ότι οι πελάτες φθάνουν στο σύστημα σε χρόνο διακριτό σύμφωνα με μία κατανομή Bernoulli με παράμετρο $a = 0.8$, δηλαδή κάθε χρονική στιγμή η πιθανότητα να έρθει ένας πελάτης είναι 0.8. Επίσης, υποθέτουμε ότι ο χρόνος που απαιτείται για την ολοκλήρωση της εξυπηρέτησης των πελατών ακολουθεί γεωμετρική κατανομή με παράμετρο $b = 0.4$. Θεωρούμε ότι σε κάθε χρονική στιγμή t ($t = 1, 2, \ldots$), κάθε ένας από τους 2 εξυπηρετητές του συστήματος έχει πιθανότητα $\theta = 0.1$ να είναι χαλασμένος για τη χρονική περίοδο $(t, t+1]$ και να μην μπορεί να εξυπηρετήσει κάποιον πελάτη.

Το σύστημα αυτό μπορεί να μελετηθεί με τη βοήθεια ενός ΟΜΣ διακριτού χρόνου το οποίο αποτελείται από 2 καταστάσεις και τα μέλη του είναι οι πελάτες. Η πρώτη κατάσταση του συστήματος αντιπροσωπεύει την πηγή των πελατών. Σε κάθε χρονική στιγμή μπορεί να φθάσει στο σύστημα ένας μόνο πελάτης, δηλαδή μπορεί να πραγματοποιηθεί μετακίνηση ενός μόνο μέλους, επομένως η χωρητικότητα της πρώτης κατάστασης θα είναι $c_1(t) = 1$. Έτσι θα έχουμε $P(c_1(t) = 1) = 1, \forall t$. Η δεύτερη κατάσταση του συστήματος αντιστοιχεί στο σύστημα αναμονής. Αφού υπάρχουν 2 εξυπηρετητές και σε κάθε χρονική στιγμή κάποιος μπορεί να είναι ενεργός ή χαλασμένος, η χωρητικότητα της κατάστασης θα παίρνει τιμές στο σύνολο $\{0, 1, 2\}$ με πιθανότητες

$$P(c_2(t) = c) = \binom{2}{c} 0.9^c 0.1^{2-c}, \; c = 0, 1, 2, \forall t.$$

Αφού το σύστημα αναμονής έχει χωρητικότητα 6 πελατών, σε οποιαδήποτε χρονική στιγμή στην κατάσταση 2 και στον αποθηκευτικό χώρο αυτής δεν μπορούν να υπάρχουν πάνω από 6 μέλη. Επομένως, το μέγεθος του ΟΜΣ που θα χρησιμοποιήσουμε για την μελέτη του συστήματος αναμονής θα είναι $N = 7$ (ένα περισσότερο από τη χωρητικότητα του συστήματος αναμονής) και ο πίνακας μετάβασης $\mathbf{P}$ των μελών του συστήματος αυτού θα είναι ο εξής:

$$\mathbf{P} = \begin{pmatrix} 0.2 & 0.8 \\ 0.4 & 0.6 \end{pmatrix}.$$

Τη χρονική στιγμή $t = 0$ το διάνυσμα κατάστασης είναι $\mathbf{m}(0) = (7, 0)^T$. Χρησιμοποιώντας τα θεωρητικά αποτελέσματα (Πρόταση 2 και Πόρισμα 1) και με τη

βοήθεια ενός κώδικα που κατασκευάστηκε στην R, παρουσιάζουμε στον Πίνακα 1 τη συμπεριφορά (στην εξέλιξη του χρόνου) των αναμενόμενων τιμών για το πλήθος των πελατών που βρίσκεται στο σύστημα, το πλήθος των πελατών που εξυπηρετείται και το πλήθος των πελατών που βρίσκεται στην αναμονή (ουρά του συστήματος).

Πίνακας 1. *Κατανομή - Αναμενόμενες τιμές*

	$t=1$	$t=2$	$t=3$		$t=120$		$t=\infty$
$P[\tilde{\mathbf{m}}(t)=(7,0)]$	0.2	0.1034	0.07109		0.00871		0.00871
$P[\tilde{\mathbf{m}}(t)=(6,1)]$	0.8	0.5101	0.3816		0.05433		0.05433
$P[\tilde{\mathbf{m}}(t)=(5,2)]$	0	0.3865	0.42064		0.10294		0.10294
$P[\tilde{\mathbf{m}}(t)=(4,3)]$	0	0	0.12667		0.1322		0.1322
$P[\tilde{\mathbf{m}}(t)=(3,4)]$	0	0	0		0.17027		0.17027
$P[\tilde{\mathbf{m}}(t)=(2,5)]$	0	0	0		0.21454		0.21454
$P[\tilde{\mathbf{m}}(t)=(1,6)]$	0	0	0		0.31701		0.31701
$E[\tilde{m}_1(t)]$	6.2	5.72	5.397		2.68737		2.68737
$E[\tilde{m}_2(t)]$	0.8	1.28	1.603		4.31263		4.31263
$E[\tilde{n}_1(t)]$	1	1	1		1		1
$E[\tilde{n}_2(t)]$	0.792	1.2	1.363		1.7403		1.7403
$E[\tilde{e}_1(t)]$	5.2	4.72	4.397		1.68737		1.68737
$E[\tilde{e}_2(t)]$	0.008	0.08	0.24		2.57233		2.57233

ABSTRACT

In this paper we consider the discrete-time homogeneous Markov system (HMS) with random state capacities, which is a generalization of the classical HMS where the states' capacities are infinite. In order to examine the variability of the state sizes, recursive formulae for their factorial moments are derived. Next, the HMS with random state capacities is used in order to examine the behavior of a discrete-time queue, where the arrivals and the service time are taken place according to the geometric distribution,

there is a random number of available servers at any time t and the size of queue cannot exceed a finite number N .

ΑΝΑΦΟΡΕΣ

Bartholomew, D.J. (1982). *Stochastic Models for Social Processes*, 3rd edn., New York: John Wiley.

Βασιλειάδης, Γ. (2019). Μελέτη της ουράς Geo/Geo/1/N με τη βοήθεια ενός Μαρκοβιανού συστήματος διακριτού χρόνου. *Πρακτικά 32ου Πανελλήνιου Συνεδρίου Στατιστικής*, 77-89.

Dimitriou, V.A. and Tsantas, N. (2010). Evolution of a time dependent Markov model for training and recruitment decisions in manpower planning. *Linear Algebra Appl.*, **433**, 1950-1972.

Isaacson, D.L. and Madsen, R.W. (1976). *Markov Chains. Theory and Applications*, New York: John Wiley.

Lykou, R. and Tsaklidis G. (2021). Particle Filtering: A Priori Estimation of Observational Errors of a State-Space Model with Linear Observation Equation. *Mathematics* **9**, **no. 12,** 1445.

McClean, S.I., McAlea, B. and Millard, P. (1998). Using a Markov reward model to estimate spend-down costs for a geriatric department. *J. Operat. Res. Soc.*, **10**, 1021-1025.

Odhner, L. and Asada, H. (2010). Kalman filter for inhomogeneous population Markov chains with application to stochastic recruitment control of muscle actuators. *American Control Conference (ACC), 2010*, 4774-4781.

Tsaklidis G. and Soldatos K.P. (2003). Modelling of continuous time homogeneous Markov system with fixed size as elastic solid. *Appl. Math. Modell*, **27**, 877-887.

Vasiliadis, G. and Tsaklidis, G. (2011). On the distributions of the state sizes of the closed discrete-time Homogeneous Markov System with finite state capacities (HMS/c). *Markov Processes and Related Fields*, **17**, 91-118.

Vasiliadis, G. (2014). Transient analysis of the M/M/k/N/N queue using a continuous time homogeneous Markov system with finite state size capacity. *Communications in Statistics - Theory and Methods*, **43:7**, 1548-1562.

Vasiliadis, G. (2016). Transient analysis of a finite source discrete-time queueing system using homogeneous Markov system with state size capacities (HMS/c). *Communications in Statistics - Theory and Methods*, **43:5**, 1403-1423.

Vassiliou, P.-C. G. (1997). The evolution of the theory of non-homogeneous Markov systems. *Stochastic Models Data Anal.*, **13, no. 3-4**, 159-176.

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 34ου Πανελληνίου Συνεδρίου Στατιστικής (2022)(σελ.34-45)

Ανάλυση πατεντών μηχανικής μάθησης: αξιολόγηση τάσεων και ευρημάτων

Κ. Γεωργίου[1], Κ. Χαρμάνας[1], Ν. Μήττας[2], Ε. Αγγελής[1]
[1]Τμήμα Πληροφορικής, Σχολή Θετικών Επιστημών, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
{konsgeor, kcharman, lef}@csd.auth.gr
[2]Τμήμα Χημείας, Διεθνές Πανεπιστήμιο Ελλάδος
{nmittas}@chem.ihu.gr

ΠΕΡΙΛΗΨΗ

Τα τελευταία χρόνια παρατηρείται μία ραγδαία άνοδος του κλάδου της Μηχανικής Μάθησης (Machine Learning) και των εφαρμογών της σε πολλά πεδία όπως η Τεχνητή Νοημοσύνη, έχοντας ως αποτέλεσμα την αυξημένη ανάγκη των εταιρειών για την κατοχύρωσή σύνθετων ευρεσιτεχνιών του αντίστοιχου πεδίου. Στην παρούσα εργασία, με βάση την παρατήρηση αυτών των εξελίξεων, διεξάγουμε μία ανάλυση πατεντών (Patent Analysis) οι οποίες αναφέρονται σε μεθόδους Μηχανική Μάθησης, αντλώντας δεδομένα από το Γραφείο Κατοχύρωσης Πατεντών των Ηνωμένων Πολιτειών (USPTO). Αξιοποιώντας αυτή την πληροφορία, εφαρμόσαμε τεχνικές Περιγραφικής Στατιστικής για να παρουσιάσουμε τα κύρια δημογραφικά στοιχεία και τις αντίστοιχες τεχνολογίες. Επιπλέον, χρησιμοποιήσαμε Ανάλυση Δικτύων και Επεξεργασία Φυσικής Γλώσσας προκειμένου να μελετήσουμε την αλληλεπίδραση των πατεντών με βάση τις αναφορές τους και να εντοπίσουμε θεματικές περιοχές. Με τα ευρήματα της έρευνας μπορούμε να εξάγουμε συμπεράσματα για τις τάσεις, τις κύριες τεχνολογίες και τις προκλήσεις του πεδίου της Μηχανικής Μάθησης αλλά και το πως οι εφαρμογές της αντανακλώνται στις σχετικές πατέντες. Από την πλευρά των επιχειρήσεων τεχνολογίας, τα ευρήματα προσφέρουν κατευθυντήριες γραμμές σε σημαντικά στάδια του στρατηγικού πλάνου των εταιρειών, όπως η ανακάλυψη υποσχόμενων τεχνολογιών και η ανίχνευση των κύριων ανταγωνιστών.

Λέξεις Κλειδιά: Μηχανική Μάθηση, Ανάλυση Πατεντών, Ανάλυση Κειμένου, Ανάλυση Δικτύων

1. ΕΙΣΑΓΩΓΗ

Η κατοχύρωση μίας πατέντας ευρεσιτεχνίας αποτελεί σημαντικό κομμάτι της ιδιοκτησίας ενός προϊόντος, μίας μεθοδολογίας ή ενός λογισμικού. Πολλοί οργανισμοί, ερευνητές και εκπαιδευτικά ιδρύματα ασχολούνται ενεργά με την εξασφάλιση των δικαιωμάτων μίας πατέντας για να την αξιοποιήσουν εμπορικά ή ερευνητικά. Ως αποτέλεσμα, τα τελευταία 50 χρόνια, και ιδιαίτερα μετά την δεκαετία 2000-2010, οι κατοχυρώσεις πατεντών γνωρίζουν εντυπωσιακή άνοδο. Ως ερευνητικό πεδίο, η ανάλυση πατεντών είναι ιδιαίτερα χρήσιμη στην ανίχνευση νεών τεχνολογιών, τάσεων και την πρόβλεψη των μελλοντικών εξελίξεων όσον αφορά τεχνολογικούς τομείς. Αυτό συμβαίνει διότι οι πατέντες αποτελούν εξαιρετικά σημαντική πηγή δεδομένων για νέες εφευρέσεις, ιδέες και μεθοδολογίες καθώς εκφράζουν τους τεχνολογικούς σκοπούς πίσω από μία εφεύρεση. Φυσικά, καθώς οι κατοχυρώσεις πατεντών πληθαίνουν, είναι απαραίτητη η σωστή οργάνωση και αποθήκευσή τους. Αυτό το σκοπό εξυπηρετούν τα διάφορα γραφεία πατεντών στα οποία οι οργανισμοί μπορούν να κατοχυρώνουν τις πατέντες τους. Ένα από τα πιο γνωστά γραφεία πατεντών είναι το USPTO, με ανάλογα γραφεία στην Ευρώπη (EPO), την Κίνα (CNIPA) και την Ιαπωνία (JPO) να έχουν αξιόλογη δραστηριότητα.

Η Μηχανική Μάθηση αφορά αλγορίθμους που βελτιώνουν και εξυπηρετούν την «εμπειρία» της τεχνητής νοημοσύνης και η βασική της ιδέα είναι η εκπαίδευση ενός μοντέλου με κάποια δεδομένα που εξυπηρετούν ως είσοδος προκειμένου να προβλεφθεί μία έξοδος και το μοντέλο να μπορεί να μάθει από αυτά τα δεδομένα για να προβλέψει νέες περιπτώσεις. Οι εφαρμογές της μηχανικής μάθησης είναι ποικίλες και γνωρίζουν ολοένα και περισσότερη απήχηση τόσο στον βιομηχανικό και στον ακαδημαϊκό όσο και σε κρατικούς και κυβερνητικούς μηχανισμούς. Έτσι, με δεδομένο αυτή τη ραγδαία αύξηση των εφαρμογών της μηχανικής μάθησης, οι οργανισμοί και οι εταιρίες που αξιοποιούν τις τεχνικές και τους αλγορίθμους της για να εξυπηρετήσουν τις ανάγκες που προκύπτουν έχουν εισαχθεί σε έναν «αγώνα» κατοχύρωσης σχετικών πατεντών ώστε να εξασφαλίσουν την εμπορική εκμετάλλευση των μεθοδολογιών και των εφευρέσεων που έχουν σχέση με το πεδίο. Με αυτό τον τρόπο, τα δεδομένα πατεντών σχετικά με τη μηχανική μάθηση αυξάνονται καθημερινά και χρόνο με το χρόνο, αποτελώντας μία πολύτιμη πηγή πληροφορίας για τις τάσεις και τις εξελίξεις του κλάδου.

Με βάση τις παραπάνω καταστάσεις, και υποκινούμενοι από το μεγάλο ενδιαφέρον και τις εφαρμογές της Μηχανικής Μάθησης, στην παρούσα εργασία διεξάγουμε μια μελέτη σε δεδομένα πατεντών του πεδίου. Τα δεδομένα που χρησιμοποιούνται έχουν τη μορφή οργανωμένου κειμένου στο οποίο εμπεριέχεται χρήσιμη περιγραφική πληροφορία, όπως τα έτη καταχώρησης και παραχώρησης, οι εφευρέτες, ο κάτοχος κ.α. Το περιεχόμενο και τα πνευματικά δικαιώματα κάθε πατέντας αποτυπώνονται από τον τίτλο, την περίληψη καθώς και τους κωδικούς κατηγοριοποίησης της, οι οποίοι αναθέτονται από το USPTO. Τέλος, οι διασυνδέσεις της πατέντας με άλλες σχετικές

ευρεσιτεχνίες παρουσιάζονται μέσω των αναφορών της (citations) οι οποίες μπορούν να αποκαλύψουν σχέσεις μεταξύ τεχνολογιών και να αναδείξουν εκείνες με τη μεγαλύτερη επιρροή.

Τα ανακτημένα δεδομένα υποβλήθηκαν σε διαδικασίες καθαρισμού για απαλοιφή θορύβου, κενών πεδίων και μετατροπής του κειμένου σε επεξεργάσιμη δομή. Στα αποτελέσματα των παραπάνω τεχνικών ανάλυσης παρουσιάζουμε τις κύριες εταιρίες που κατέχουν πατέντες Μηχανικής Μάθησης και την χρονική εξέλιξη των πατεντών. Επίσης, χρησιμοποιώντας τις παραπομπές κάθε πατέντας, ανακαλύπτουμε τις τεχνολογίες και τις εταιρίες με τη μεγαλύτερη επιρροή. Τέλος, η Ανάλυση Δικτύων μας επιτρέπει να εξάγουμε τα τεχνολογικά πεδία που πραγματεύονται οι πατέντες και κατά πόσο συμφωνούν με τις τελευταίες εξελίξεις του πεδίου.

Τα ερευνητικά ερωτήματα που καλούμαστε να απαντήσουμε στην παρούσα εργασία μπορούν να εκφραστούν ως εξής:

[EE$_1$] Πως διαμορφώνονται τα δημογραφικά στοιχεία των πατεντών μηχανικής μάθησης;
Το παρόν ερευνητικό ερώτημα αποτελεί μία περιγραφική μελέτη των κυριότερων δημογραφικών στοιχείων (χώρες, οργανισμοί, υποομάδες) των πατεντών μηχανικής μάθησης με σκοπό να κατανοήσουμε τις βασικές τάσεις του κλάδου.

[E.E$_2$] Ποιοι είναι οι θεματικοί άξονες των πατεντών μηχανικής μάθησης;

Σε αυτό το ερευνητικό ερώτημα, χρησιμοποιώντας τεχνικές ανάλυσης δικτύων, κάνουμε μία απόπειρα ανίχνευσης των κύριων θεματικών περιοχών που πραγματεύονται οι πατέντες μηχανικής μάθησης, οι οποίες αφορούν διαφορετικές περιοχές του κλάδου.

[E.E$_3$] Ποια χαρακτηριστικά των πατεντών έχουν μεγαλύτερη επιρροή;

Στο τελευταίο ερευνητικό ερώτημα, με βάση τις αναφορές κάθε πατέντας, ανακαλύπτουμε ποια χαρακτηριστικά πατεντών (εταιρίες, κλάσεις, χώρες) έχουν τη μεγαλύτερη επιρροή και αναφέρονται περισσότερο από άλλες πατέντες κατά την κατοχύρωσή τους.

2. ΑΝΑΣΚΟΠΗΣΗ ΒΙΒΛΙΟΓΡΑΦΙΑΣ

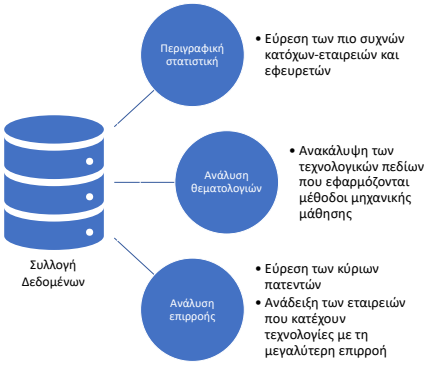
Οι εφαρμογές ανάλυσης πατεντών συχνά διεισδύουν στη φύση των εγγραφών ευρεσιτεχνίας με σκοπό την ανακάλυψη γνώσης με διαφορετικά κριτήρια και πεδία ενδιαφέροντος. Όπως προαναφέραμε, οι εγγραφές πατεντών περιέχουν και πληροφορίες κειμένου (τίτλος , περίληψη) που περιγράφουν τη φύση και τις αναλυτικές κατοχυρώσεις της κάθε ευρεσιτεχνίας, προσφέροντας σημαντικά στοιχεία για την εξόρυξη συλλογικής γνώσης χρησιμοποιώντας δεδομένα κειμένων και τεχνικές Επεξεργασία Φυσικής Γλώσσας (text mining). Προηγούμενες εφαρμογές συστήνουν τη χρήση των συγκεκριμένων πεδίων κειμένου καθώς περιέχουν αναλυτικότερη

πληροφορία από αυτήν της κωδικοποίησης CPC και μπορούν εξυπηρετούν σε εφαρμογές ομαδοποίησης πατεντών και εύρεση κύριων θεματολογιών [1-3].
Επιπλέον, προηγούμενες μελέτες που περιστρέφονται γύρω από εφαρμογές ανάλυσης πατεντών συνήθως χρησιμοποιούν τις παραπομπές μεταξύ των πατεντών για διάφορους σκοπούς που σχετίζονται κυρίως με την εύρεση των κύριων τεχνολογιών [6] και την ανάθεση στρατηγικών θέσεων στις εμπλεκόμενες εταιρείες [7]. Γενικότερα, οι παραπομπές πατεντών έχουν αποδεχθεί χρήσιμη πληροφορία η οποία αντικατοπτρίζει μέχρι ένα βαθμό την τεχνολογική και οικονομική αξία της κάθε ευρεσιτεχνίας [8,9]. Έχοντας αυτές τις πληροφορίες ως βάση για την παρούσα μελέτη, αποφασίσαμε να κατασκευάσουμε ένα δίκτυο παραπομπής πατεντών (Patent Citation Network – PCN) τόσο για την αξιολόγηση των πατεντών ως μονάδες όσο και για τη συλλογική ανάθεση επιρροής των εταιρειών και των κύριων θεματολογιών.

3. ΜΕΘΟΔΟΛΟΓΙΑ

Στην παρούσα εργασία εστιάζουμε στην εξαγωγή σημαντικής πληροφορίας από τεχνολογικές πατέντες που χρησιμοποιούν μεθόδους μηχανικής μάθησης, εφαρμόζοντας βασικές τεχνικές ανάλυσης πατεντών. Στο πρώτο στάδιο της έρευνας μας μελετήσαμε εις βάθος το συνεργατικό σχήμα κατηγοριοποίησης πατεντών (CPC), με σκοπό την επιλογή και την ανάκτηση των κατάλληλων εγγραφών από το USPTO (Συλλογή Δεδομένων). Στη συνέχεια, θα παρουσιάσουμε κάποια βασικά στατιστικά των πιο σημαντικών χαρακτηριστικών των συλλεγμένων εγγραφών (Περιγραφική στατιστική). Έπειτα εκμεταλλευόμενοι των αποτελεσμάτων που προκύπτουν από τα προηγούμενα βήματα θα εφαρμόσουμε δύο πιο σύνθετες μεθοδολογίες για την ανακάλυψη των βασικών θεματολογιών (Ανάλυση Θεματολογιών) και τη μελέτη συσχετίσεων επιρροής χρησιμοποιώντας πληροφορίες παραπομπών μεταξύ των πατεντών (Ανάλυση επιρροής) αντίστοιχα. Η συνολική ροή της μεθοδολογίας αυτής της εργασίας απεικονίζεται στην Εικόνα 1.

Εικόνα 1. *Σχήμα Μεθοδολογίας*



Συλλογή Δεδομένων

Έπειτα από τη χειροκίνητη μελέτη του CPC, καταλήξαμε στη συλλογή πατεντών που ανήκουν στην κατηγορία Υπολογιστικές διαδικασίες βασισμένες σε συγκεκριμένα υπολογιστικά μοντέλα –μηχανικής μάθηση, η οποία έχει κωδικοποιηθεί ως **G06N20**, ή στις δύο υποκατηγορίες της που ονομάζονται Μέθοδοι πυρήνων, π.χ. μηχανές υποστήριξης διανυσμάτων [SVM] και Συνθετική μάθηση αντίστοιχα. Ως αποτέλεσμα, η συλλογή που ανακτήθηκε περιέχει 17412 πατέντες που καλύπτουν όλο το εύρος των διαθέσιμων εγγραφών από το 1970 έως και το 2022. Η δομή της κάθε εγγραφής διαθέτει όλες τις απαραίτητες πληροφορίες που προαναφέραμε για την διεξαγωγή των απαραίτητων συμπερασμάτων που σχετίζονται με τους στόχους της συγκεκριμένης εργασίας.
Στη συνέχεια παρατηρήσαμε πως πολλές εγγραφές του συνόλου δεδομένου που συλλέξαμε κατέχουν σχεδόν ή καθολικά ισάξια χαρακτηριστικά καθώς κάποιες σημαντικές πατέντες ανανεώνονται μετά τη λήξη τους η επεκτείνονται με επιπρόσθετες λειτουργίες. Για να αντιμετωπίσουμε το συγκεκριμένο εμπόδιο, αναλύσαμε τις σχέσεις μεταξύ των πατεντών χρησιμοποιώντας επίσημες πηγές δεδομένων του UPSTO[1] με σκοπό την ομαδοποίηση της συνολικής πληροφορίας εκείνων που συσχετίζονται με τουλάχιστον έναν τρόπο σύνδεσης από αυτές που αναφέραμε. Οι σχέσεις μεταξύ πατεντών υπάρχουν διότι πολλές φορές μία πατέντα μπορεί να υποβληθεί ξανά στο γραφείο κατοχύρωσης ως συνέχεια ή τροποποίηση μιας υπάρχουσας πατέντας. Η ανίχνευση σχέσεων μεταξύ δύο ή περισσότερων πατεντών μας επέτρεψε να ενώσουμε κοινές πατέντες και να παράξουμε «ενώσεις» πατεντών (patent unions). Από το σημείο αυτό οι ομάδες πατεντών που προκύπτουν θα επεξεργάζονται ως μια ολότητα ώστε να αποφευχθούν μη έγκυρα αποτελέσματα και συμπεράσματα.

Περιγραφική στατιστική

Σε αυτό το στάδιο της μεθοδολογίας μας εφαρμόσαμε βασικές τεχνικές περιγραφικής στατιστικής με σκοπό να ανακαλύψουμε τα κύρια δημογραφικά στοιχεία που προκύπτουν από τις εγγραφές που έχουμε συλλέξει. Πιο αναλυτικά, μελετήσαμε αρχικά τη συνολική αύξηση των κατοχυρωμένων πατεντών ανά χρόνιά που αντιπροσωπεύει τόσο την εξέλιξη και την πληθώρα της μηχανικής μάθησης όσο και το ενδιαφέρον των αντίστοιχων εταιρειών για τις συγκεκριμένες πατέντες. Επιπλέον, παρουσιάζουμε στη συνέχεια τις εταιρείες που κατέχουν τις περισσότερες πατέντες, αναδεικνύοντας με αυτόν τον τρόπο τους κύριους συνεισφορείς και ενδιαφερόμενους του τεχνολογικού πεδίου από την οπτική του πλήθους διακριτών πατεντών. Αναλύσαμε τους κωδικούς CPC με τη μεγαλύτερη συνύπαρξη, καθώς κάθε πατέντα μπορεί να συνδέεται με παραπάνω από μία υποκλάσεις, με σκοπό να διασταυρώσουμε και να επαληθεύσουμε τα αποτελέσματα των τεχνικών Ανάλυσης θεματολογιών που

[1] https://patentsview.org/download/data-download-tables

θα αναπτύξουμε στην επόμενη υπό ενότητα. Οι συγκεκριμένες εφαρμογές που συζητήσαμε αποτελούν τα θεμέλια των μεθοδολογιών ανάλυσης πατεντών και μας βοηθούν στην περιγραφή της βασική πληροφορίας του συγκεκριμένου τεχνολογικού πεδίου.

Ανάλυση θεματολογιών

Στην παρούσα εργασία εφαρμόσαμε συνδυαστικές μεθόδους Επεξεργασίας Φυσικής Γλώσσας και Ανάλυσης Δικτύων για την εύρεση των κύριων θεματολογιών χρησιμοποιώντας τους περιγραφικούς τίτλους των συλλεγμένων πατεντών. Πιο αναλυτικά, το πρώτα βήμα αυτής της μεθοδολογίας είναι η μετατροπή της πληροφορίας κειμένου σε πίνακα πατέντας-όρος (Document Term Matrix) εφαρμόζοντας βασικές τεχνικές επεξεργασίας κειμένου όπως διαγραφή σημείων στίξης και κοινών συνδετικών λέξεων, μετατροπή κειμένου σε πεζούς χαρακτήρες και κλάδεμα (Stemming) κ.α. οι οποίες περιγράφονται αναλυτικά στην Ikonomakis et al. [4]. Σε αυτό το σημείο θα πρέπει να αναφέρουμε πως αφαιρέσαμε από την ανάλυση μας τους όρους που είχαν μικρότερη εμφάνιση από το 0.2% και μεγαλύτερη από 0.08% όλων των εγγραφών.
Έπειτα, έχοντας έτοιμο τον κατάλληλο αριθμητικό πίνακα εφαρμόσαμε τεχνικές ανάλυσης συνεμφάνισης λέξεων (co-word analysis), αξιοποιώντας τη μετρική Inclusion Index [5] (δείκτης συμπερίληψης) η οποία ορίζεται ως εξής (1):

$$I.I = \frac{C_{ij}}{\min\left(C_i, C_j\right)} \quad (1)$$

όπου C_{ij} είναι το πλήθος των εγγραφών στα οποία συνυπάρχουν οι όροι i και j και C_i (αντίστοιχα C_j) το πλήθος των εγγράφων στα οποία εμφανίζονται ο όρος i (αντίστοιχα ο όρος j). Η συγκεκριμένη μετρική εξυπηρετεί στην εύρεση ζευγαρωτών συσχετίσεων μεταξύ των διακριτών όρων που προέκυψαν προηγουμένως, με σκοπό την ομαδοποίηση εκείνων που παρουσίασαν υψηλό δείκτη συμπερίληψης. Ως αποτέλεσμα, στο επόμενο βήμα δημιουργήσαμε ένα δίκτυο μεταξύ των όρων, οι οποίοι αποτελούν τους κόμβους του, και στη συνέχεια θέσαμε την ύπαρξη μη κατευθυνόμενη ακμής μεταξύ δύο όρων με βάση το $I.I$ με βάρος μεγαλύτερη ή ίση του 0.2 εκφράζοντας την ύπαρξη υψηλής συσχέτισης μεταξύ των όρων. Το δίκτυο που προέκυψε υποβλήθηκε στη συνέχεια στον αλγόριθμο του Leiden [10] για ανίχνευση κοινοτήτων (communities), οι οποίες περιείχαν όρους που εκφράζουν κοινές τεχνολογικές πατέντες και μεθοδολογίες. Ο αλγόριθμος του Leiden αποτελεί έναν αρκετά δημοφιλή τρόπο ανίχνευσης κοινοτήτων, ο οποίος λαμβάνοντας υπόψιν την αρθρωτότητα (Modularity) [11] του δικτύου εντοπίζει τον βέλτιστο διαχωρισμό (partition) του δικτύου σε κοινότητες, μέσω της επαναλαμβανόμενης μετακίνησης κόμβων σε διαφορετικές ομάδες ώστε να καθοριστεί για κάθε κόμβο η ομάδα (κοινότητα) στην οποία έχει τις περισσότερες συνδέσεις.
Έχοντας κατασκευάσει το τελικό δίκτυο κοινοτήτων, στη συνέχεια ορίζουμε τον παρακάτω τύπο για την ανάθεση συσχέτισης κάθε όρου (*rel*) με την κάθε κοινότητα

(2) με σκοπό να ανατρέξουμε στην ταξινόμηση των εγγράφων στις κοινότητες που προκύπτουν με τη χρήση των βαρών rel και τον πίνακα πατέντα-όρος.

$$rel_{ij} = \frac{E_{ij}}{E_i} \quad (2)$$

Στον παραπάνω τύπο ο όρος E_{ij} δηλώνει τον αριθμό των ακμών που συνδέουν τον όρο i με όρους που ανήκουν στην κοινότητα j, ενώ παράλληλα ο όρος E_i δηλώνει το άθροισμα των ακμών που συμμετέχει ο όρος i. Τελικά, για την ταξινόμηση κάθε εγγραφής στις κοινότητες ($\pi\kappa$) χρησιμοποιούμε τον παρακάτω τύπο συνδυάζοντας τις δύο δομές που προαναφέραμε, όπου ο πίνακας πατέντας-όρος χαρακτηρίζεται ως πo (3). Επιπλέον, στον παρακάτω τύπο $o\ \acute{o}\rho o\varsigma\ \pi\kappa_{ij}$ αναφέρεται στο βάρος σύνδεσης (ταξινόμηση) της εγγραφής i στην κοινότητα j και ταυτόχρονα ο όρος πo_i τον αριθμό των όρων που υπάρχουν στην εγγραφή i. Γενικά, ο πίνακας $\pi\kappa$ μας βοηθάει στην ερμηνεία της κάθε κοινότητας και παράλληλα τον χρησιμοποιούμε για τον υπολογισμό της συνολικής επιρροής της κάθε κοινότητας, μέσω παραθέσεων, στις πατέντες μηχανικής μάθησης.

$$\pi\kappa = \pi o \cdot rel\ , \pi\kappa_{ij} = \frac{\pi\kappa_{ij}}{\pi o_i} \quad (3)$$

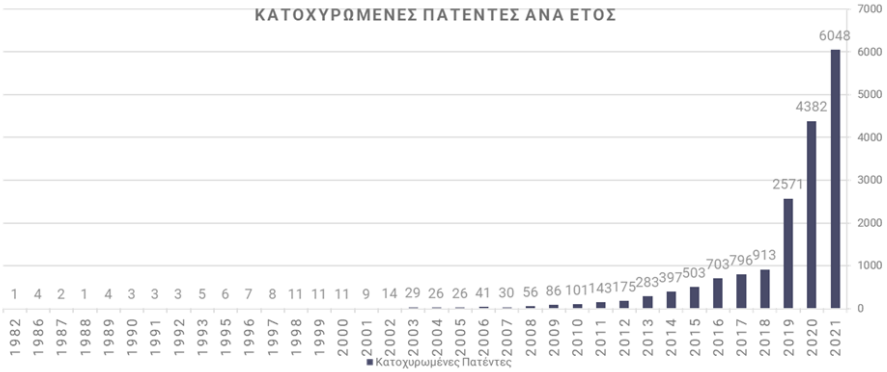
Ανάλυση επιρροής

Στο επόμενο στάδιο της εργασίας, κατασκευάσαμε ένα απλό PCN το οποίο μας βοήθησε στην ανακάλυψη της σχετικής τοποθέτησης κάθε εγγραφής στον ορίζοντα των ευρεσιτεχνιών μηχανικής. Στα πλαίσια της μεθοδολογίας μας συσχετίσαμε την αξία κάθε πατέντας με τον απόλυτο αριθμό των κατευθυνόμενων ακμών που καταλήγουν σε αυτήν και στη συνέχεια τη συλλογική επιρροή κάθε στοιχείου (εταιρεία, θεματολογίας κτλ.) χρησιμοποιώντας το άθροισμα των ατομικών αξιολογήσεων των πατεντών που σχετίζονται με το κάθε στοιχείο. Τελικά, χρησιμοποιώντας τα αποτελέσματα των συγκεκριμένων εφαρμογών παρουσιάζουμε τις εταιρείες και χώρες με το μεγαλύτερο στρατηγικό πλεονέκτημα στο τεχνολογικό πεδίο που εξετάζουμε καθώς και τις αντίστοιχες σημαντικότερες θεματολογίες τεχνολογιών που εφαρμόζουν μεθόδους μηχανικής μάθησης με βάση τις υποκλάσεις και τις κοινότητες που ανιχνεύονται στα δίκτυα.

4. ΑΠΟΤΕΛΕΣΜΑΤΑ

[ΕΕ$_1$] Πως διαμορφώνονται τα δημογραφικά στοιχεία των πατεντών μηχανικής μάθησης;

Εικόνα 2. *Χρονική εξέλιξη κατοχύρωσης πατεντών*



Πίνακας 1. *Κύρια Δημογραφικά Στοιχεία και ποσοστά κατοχής*

Κύριες χώρες	Κύριοι οργανισμοί	Κύριες CPC υποκλάσεις
USA (72%)	IBM (13%)	G06N7/005: Probabilistic networks (18%) (1)
Japan (6%)	Microsoft (5%)	G06N3/08: Learning Methods (17%) (2)
Korea (2%)	Google (3%)	G06N3/0454: Multiple Neural Networks (13%) (3)
China (2%)	Amazon (2%)	G06N5/003: Heuristics/Dynamic Trees (11%) (4)
Germany (2%)	Meta (Facebook) (2%)	G06N5/04: Inference Methods (11%) (5)
Ireland (1%)	Capital One Services (2%)	G06N5/022: Knowledge Engineering (8%) (6)
Canada (1%)	Accenture (2%)	G06N3/0445: Feedback Networks (6%) (7)
Israel (1%)	Cisco (2%)	G06N3/084: Back Propagation (6%) (8)
India (1%)	NEC Labs (1%)	G06N5/02: Knowledge representation (6%) (9)
Great Britain (1%)	Intel (1%)	G06F40/30: Semantic Analysis (5%) (10)

Από τα ευρήματα στην Εικόνα 1 και τον Πίνακα 1, παρατηρούμε πως αρχικά, οι πατέντες μηχανικής μάθησης πληθαίνουν σταθερά ανά τα χρόνια, με τις κατοχυρώσεις να εκτοξεύονται μετά το 2015. Αυτό συμπίπτει με την σταθερή άνοδο του ενδιαφέροντος για τον κλάδο της μηχανικής μάθησης τόσο από επιχειρήσεις και οργανισμούς όσο και από την ακαδημαϊκή κοινότητα.

Επιπλέον, στον Πίνακα 1 παρουσιάζονται οι κύριες χώρες, οργανισμοί και CPC υποκλάσεις που συνδέονται με τις ανακτημένες πατέντες. Σε επίπεδο χωρών, παρατηρούμε το συγκριτικό πλεονέκτημα των Ηνωμένων Πολιτειών, ένα εύρημα το οποίο ήταν αναμενόμενο καθώς οι πατέντες που συλλέξαμε προέρχονται αποκλειστικά

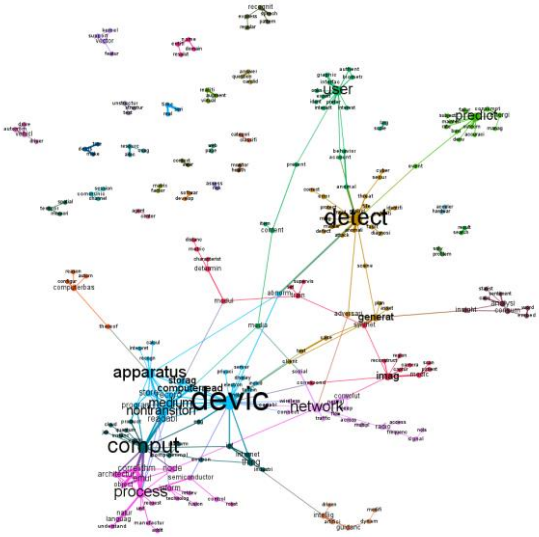
από το Αμερικανικό γραφείο πατεντών. Παρόλαυτα, οι ασιατικές χώρες (Ιαπωνία, Κορέα, Κίνα) φαίνεται να έχουν ένα αξιοσημείωτο μερίδιο στην κατοχή πατεντών με την Ευρώπη (Γερμανία, Ιρλανδία) να ακολουθεί.

Σε επίπεδο οργανισμών παρατηρούμε πως μεγάλες τεχνολογικές επιχειρήσεις επενδύουν ενεργά στη Μηχανική Μάθηση και κατοχυρώνουν πατέντες σχετικές με το πεδίο. Ανάμεσά τους βλέπουμε εταιρίες που παρέχουν καινοτόμα λογισμικά και υπηρεσίες νέφους για εκτέλεση μοντέλων (Microsoft, Google, Amazon), εταιρίες ασφάλειας λογισμικού (Cisco, NEC Labs), οργανισμούς ανάλυσης δεδομένων (Accenture), εταιρίες κοινωνικής δικτύωσης (Meta) και εταιρίες παραγωγής υλισμικού (IBM, Intel).

Τέλος, οι κύριες CPC υποκλάσεις χαρακτηρίζονται από υπολογιστικές μεθόδους και μαθηματικά μοντέλα καθώς και στατιστικές μεθοδολογίες που υποδηλώνουν την φύση των πατεντών μηχανικής μάθησης αλλά και τη σύνδεση που έχει το πεδίο με τη στατιστική.

[Ε.Ε$_2$] Ποιοι είναι οι θεματικοί άξονες των πατεντών μηχανικής μάθησης;

Εικόνα 3. *Δίκτυο κοινοτήτων κύριων όρων*



Αριθμός	Τίτλος Κοινότητας
1	Ασφάλεια Λογισμικού
2	Ψηφιακά Δεδομένα
3	Δομές για επεξεργασία Δεδομένων
4	Διαδραστική μάθηση
5	Ψηφιακή μουσική και εικόνα
6	Νευρωνικά δίκτυα – Βαθιά μάθηση
7	Δ.Τ.Π, κβαντική υπολογιστική, νέφη
8	Πρόβλεψη κατανάλωσης ενέργειας
9	Ανάλυση συναισθήματος αγοράς
10	Τεχνητή νοημοσύνη

Στην εικόνα 3 παρουσιάζουμε το δίκτυο κοινοτήτων που προκύπτει από την μεθοδολογία που περιγράφουμε στο κεφάλαιο 3. Οι κοινότητες έχουν ανιχνευθεί μέσω του αλγορίθμου του Leiden και αποτελούνται από όρους που συνεμφανίζονται στους τίτλους των πατεντών. Παράλληλα στο διπλανό πίνακα παραθέτουμε τις δέκα πιο μεγάλες κοινότητες μαζί με έναν ενδεικτικό τίτλο που τις αντιπροσωπεύει ως θεματικός άξονας.

Το παραπάνω δίκτυο περιέχει και υπολογιστικές μεθόδους που συμφωνούν με τις CPC κλάσεις που εξήχθησαν στο προηγούμενο ερευνητικό ερώτημα. Επιπλέον, βλέπουμε πως η Μηχανική Μάθηση γνωρίζει πολλές εφαρμογές σε διάφορα πεδία όπως η ασφάλεια λογισμικού, η ανάλυση ψηφιακών δεδομένων και συναισθήματος, τα νευρωνικά δίκτυα και η βαθιά μάθηση αλλά και σε μικρότερο βαθμό στα αυτόνομα οχήματα και στην εικονική πραγματικότητα.

[Ε.Ε3] Ποια χαρακτηριστικά των πατεντών έχουν μεγαλύτερη επιρροή;

Πίνακας 2. *Συγκεντρωτικές Παραθέσεις Χαρακτηριστικών*

Αθροιστικές Παραθέσεις			
Χώρες	Οργανισμοί	CPC υποκλάσεις	Κοινότητες
USA 5764	Google 672	(1) 1749	(1) 730
Japan 181	IBM 487	(5) 952	(8) 615
Canada 76	Microsoft 432	(8) 745	(2) 570
Ireland 74	Amazon 432	(3) 698	(4) 524
Germany 62	SAS 110	(6) 627	(3) 490

Σύμφωνα με τον Πίνακα 2, η συλλογική επιρροή των χαρακτηριστικών που μελετήσαμε δεν συνδέεται σε μεγάλο βαθμό με την συνολική κατοχή πατεντών του κάθε χαρακτηριστικού.

5. ΣΥΜΠΕΡΑΣΜΑΤΑ – ΜΕΛΛΟΝΤΙΚΗ ΕΡΕΥΝΑ

Τα συμπεράσματα που προκύπτουν από την ανάλυση είναι πως, σχετικά με το πρώτο ερευνητικό ερώτημα, οι κύριοι συντελεστές (χώρες, οργανισμοί) είναι αναμενόμενοι καθώς απαρτίζονται από κολοσσούς που καινοτομούν σε πολλά πεδία της Επιστήμης Δεδομένων. Επιπλέον, οι κύριες υποκλάσεις φαίνεται να αποτελούνται από μαθηματικές και στατιστικές μεθόδους ενώ οι κατοχυρώσεις πατεντών Μηχανικής Μάθησης ακολουθούν ανοδική πορεία ανά τα χρόνια, ιδίως από το 2015 και μετέπειτα. Αναφορικά με το δεύτερο ερώτημα, οι πιο μεγάλες και σημαντικές κοινότητες αφορούν την εφαρμογή της μηχανικής μάθησης στην ασφάλεια λογισμικού, στην ανάλυση ψηφιακών δεδομένων και σε υπολογιστικές διαδικασίες ενώ παράλληλα οι κοινότητες που ανιχνεύθηκαν συνάδουν με τις κύριες CPC υποκατηγορίες που συνυπάρχουν στα δεδομένα.
Τέλος, όσον αφορά τα χαρακτηριστικά με τη μεγαλύτερη επιρροή, οι οργανισμοί με τις περισσότερες αναφορές έχουν και μεγαλύτερη συμμετοχή και κατοχή πατεντών, χωρίς όμως αυτό να συνεπάγεται ότι ένας οργανισμός με πολλές πατέντες έχει απαραίτητα και πολλές αναφορές. Επιπλέον, βλέπουμε πως με βάση και τις κοινότητες αλλά και τις CPC κατηγορίες, επιβεβαιώνεται η παραπάνω δήλωση καθώς παρατηρήσαμε πως το περιεχόμενο των πατεντών και όχι η αντιστοίχιση των πατεντών

σε CPC υποκλάσεις και κοινότητες είναι αυτό που κατέχει σημαντικό ρόλο στην επιρροή τους.
Κάποιοι περιορισμοί στην παρούσα εργασία σχετίζονται τόσο με το ότι το σύνολο δεδομένων μας προερχόταν από μία πηγή δεδομένων όσο και με την επιλογή και το φιλτράρισμα των μεθοδολογιών και των δεδομένων. Για μελλοντική έρευνα ενθαρρύνουμε την επέκταση των μεθοδολογιών μας και σε άλλες πηγές δεδομένων (ακαδημαϊκή βιβλιογραφία, κοινωνικά δίκτυα, εφαρμογές) και την διερεύνηση πρόσθετων μεθόδων για εξαγωγή θεματολογιών και την ανάλυση επιρροής.

ABSTRACT

In recent years there has been a rapid growth of the Machine Learning industry and its applications in many fields such as Artificial Intelligence, resulting in the increased need of companies to patent complex patents in the field. In the present work, based on the observation of these developments, we conduct a Patent Analysis which refers to Machine Learning methods, drawing data from the United States Patent Office (USPTO). Utilizing this information, we applied Descriptive Statistics techniques to present the main demographics and the corresponding technologies. In addition, we used Network Analysis and Natural Language Processing to study the interaction of patents based on their reports and to identify subject areas. With the findings of the research, we can draw conclusions about the trends, the main technologies and the challenges in the field of Machine Learning but also how its applications are reflected in the relevant patents. From the technology companies 'point of view, the findings provide guidelines at important stages of companies' strategic planning, such as discovering promising technologies and identifying key competitors.

ΑΝΑΦΟΡΕΣ

[1]Yoon, B., & Park, Y. (2004). A text-mining-based patent network: Analytical tool for high-technology trend. The Journal of High Technology Management Research, 15(1), 37-50.
[2] Yoon, J., & Kim, K. (2011). Identifying rapidly evolving technological trends for R&D planning using SAO-based semantic patent networks. Scientometrics, 88(1), 213-228.
[3] Tseng, Y. H., Lin, C. J., & Lin, Y. I. (2007). Text mining techniques for patent analysis. Information processing & management, 43(5), 1216-1247.
[4] Ikonomakis, M., Kotsiantis, S., & Tampakas, V. (2005). Text classification using machine learning techniques. WSEAS transactions on computers, 4(8), 966-974.
[5] He, Q. (1999). Knowledge discovery through co-word analysis.
[6] Kim, G., & Bae, J. (2017). A novel approach to forecast promising technology through patent analysis. Technological Forecasting and Social Change, 117, 228-237.
[7] Daim, T., Lai, K. K., Yalcin, H., Alsoubie, F., & Kumar, V. (2020). Forecasting technological positioning through technology knowledge redundancy: Patent citation analysis of IoT, cybersecurity, and Blockchain. Technological Forecasting and Social Change, 161, 120329.

[8] Albert, M. B., Avery, D., Narin, F., & McAllister, P. (1991). Direct validation of citation counts as indicators of industrially important patents. Research policy, 20(3), 251-259.
[9] Sampat, B. N., & Ziedonis, A. A. (2004). Patent citations and the economic value of patents. In Handbook of quantitative science and technology research (pp. 277-298). Springer, Dordrecht.
[10] Traag, V. A., Waltman, L., & Van Eck, N. J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. Scientific reports, 9(1), 1-12.
[11] Newman, M. E. (2006). Modularity and community structure in networks. Proceedings of the national academy of sciences, 103(23), 8577-8582.

Ελληνικό Στατιστικό Ινστιτούτο

Πρακτικά 34ου Πανελλήνιου Στατιστικού Συνεδρίου (2022) (σελ. 46-55)

Η ΚΑΤΑΝΟΜΗ ΤΟΥ ΣΥΝΟΛΙΚΟΥ ΑΡΙΘΜΟΥ ΤΩΝ ΕΠΙΤΥΧΙΩΝ ΣΕ ΑΣΘΕΝΕΙΣ ΡΟΕΣ ΠΕΡΙΟΡΙΣΜΕΝΟΥ ΜΗΚΟΥΣ

Σ.Δ. Δαφνής[1], *Φ.Σ. Μακρή*[2]
[1] Τμήμα Επιστήμης Φυτικής Παραγωγής, Γεωπονικό Πανεπιστήμιο Αθηνών
sdafnis@aua.gr
[2] Τμήμα Μαθηματικών, Πανεπιστήμιο Πατρών
makri@math.upatras.gr

ΠΕΡΙΛΗΨΗ

Στην παρούσα εργασία επεκτείνουμε τον ορισμό της ασθενούς ροής σε ακολουθίες δίτιμων δοκιμών, ώστε μία ασθενής ροή να περιλαμβάνει έναν ελάχιστο και ένα μέγιστο αριθμό επιτυχιών. Θεωρώντας n δίτιμες δοκιμές, μελετάμε την κατανομή του συνολικού αριθμού των επιτυχιών σε όλες τις ασθενείς ροές του προαναφερόμενου τύπου που έχουν καταμετρηθεί. Η μελέτη μας προϋποθέτει κατάλληλη γενίκευση της μεθόδου εμφύτευσης σε Μαρκοβιανή αλυσίδα. Πιο συγκεκριμένα, απαιτεί την εισαγωγή και τη μελέτη της οικογένειας των εμφυτεύσιμων μεταβλητών επιστρέψιμου - πολυωνυμικού τύπου. Η νέα αυτή οικογένεια γενικεύει οικογένειες εμφυτεύσιμων μεταβλητών, οι οποίες έχουν μελετηθεί εκτεταμένα στη βιβλιογραφία.

Λέξεις Κλειδιά: Δίτιμες δοκιμές, Μαρκοβιανή αλυσίδα, εμφυτεύσιμες μεταβλητές επιστρέψιμου - πολυωνυμικού τύπου, Γεωπονία

1. ΕΙΣΑΓΩΓΗ

Στο επίκεντρο του ενδιαφέροντος της Θεωρίας Πιθανοτήτων τις τελευταίες δεκαετίες είναι η μελέτη διακριτών κατανομών που σχετίζονται με την εμφάνιση ροών, λόγω της ευρείας εφαρμοσιμότητάς τους σε διάφορες επιστημονικές περιοχές (βλέπετε Balakrishnan and Koutras (2002)).

Ας θεωρήσουμε μια ακολουθία n δοκιμών Bernoulli $Z_1, Z_2, \ldots, Z_n$ και $Z_i = 1$ (επιτυχία) ή $Z_i = 0$ (αποτυχία) με πιθανότητα p_i και $q_i = 1 - p_i$, αντίστοιχα, δηλ. $p_i = P(Z_i = 1) = 1 - P(Z_i = 0) = 1 - q_i$, $1 \leq i \leq n$.

Στην πρόσφατη δουλειά των Dafnis and Makri (2022) οι συγγραφείς χρησιμοποίησαν τις ροές επιτυχιών μήκους τουλάχιστον k για να εμπλουτίσουν την έρευνα στη Γεωπονία που σχετίζεται με μοντέλα μονάδας θερμότητας (heat unit models).

Το μοντέλο τους πηγάζει από την απλή ιδέα ότι μία συγκέντρωση 'ζεστών' ημερών μπορεί να θεωρηθεί μία μονάδα θερμότητας. Η προαναφερόμενη δουλειά μπορεί να ενισχυθεί αν επιβληθεί και ένα άνω όριο στον αριθμό επιτυχιών σε κάθε ροή που καταμετράται. Πράγματι, η σχετιζόμενη με τα μοντέλα αυτά βιβλιογραφία υποδεικνύει ότι η επιβολή ενός τέτοιου ορίου μπορεί να αποδειχθεί ιδιαίτερα χρήσιμη (βλέπετε π.χ. Bonhomme (2000)). Φυσικά, η ιδέα αυτή μοιάζει ιδιαίτερα λογική και σε άλλα προβλήματα σχετιζόμενα με τη Γεωπονία. Ας θεωρήσουμε το απλό παράδειγμα του αριθμού των συνεχόμενων βροχερών ημερών που απαιτούνται προκειμένου να ποτιστεί μία καλλιέργεια. Είναι λογικό κάποιος/α να καταμετρήσει όλες τις περιόδους από τουλάχιστον k συνεχόμενες βροχερές ημέρες (η εκτίμηση της παραμέτρου k θα διαφοροποιείται για διαφορετικές καλλιέργειες). Αλλά, είναι προφανές ότι η υπερβολικά μεγάλη ποσότητα νερού μπορεί να βλάψει την καλλιέργεια. Συνεπώς, θα μπορούσαν να είναι επιθυμητές περίοδοι από τουλάχιστον k και το πολύ m ($k \leq m \leq n$) συνεχόμενες βροχερές ημέρες (η εκτίμηση της παραμέτρου m θα εξαρτάται, επίσης, από τη συγκεκριμένη καλλιέργεια).

Η προσπάθεια να μοντελοποιηθούν τέτοιου είδους προβλήματα της πραγματικής ζωής γεννά μία γενικευμένη στατιστική συνάρτηση ροών, η οποία είναι ο αριθμός των ροών επιτυχιών μήκους τουλάχιστον k και το πολύ m σε μία ακολουθία δοκιμών Bernoulli. Είναι δε ιδιαίτερα σημαντικός ο συνολικός αριθμός των επιτυχιών σε όλες τις προαναφερόμενες ροές. Στην παρούσα εργασία θα επικεντρώσουμε την προσοχή μας σε αυτή τη στατιστική συνάρτηση (βλέπετε, τους Antzoulakos et al. (2003) και Makri et al. (2007)). Η μελέτη μας θα πραγματοποιηθεί μέσα στο γενικότερο πλαίσιο των r-ασθενών ροών (βλέπετε τους Dafnis and Makri (2022)), το οποίο προσφέρει μεγαλύτερη προσαρμοστική ικανότητα σε περιπτώσεις εφαρμογών με μεγαλύτερη πολυπλοκότητα (η επιπλέον παράμετρος r, $r \geq 0$, επιτρέπει μέγιστη απόσταση $r+1$ μεταξύ κάθε δύο διαδοχικών επιτυχιών στη ροή). Υπενθυμίζουμε ότι η 0-ασθενής ροή είναι η τυπική ροή. Συνεπώς, τα νέα αποτελέσματα αυτής της δουλειάς καλύπτουν και την περίπτωση τυπικών ροών.

Για τη μελέτη της νέας τυχαίας μεταβλητής (ο συνολικός αριθμός επιτυχιών σε όλες τις ασθενείς ροές μήκους τουλάχιστον k και το πολύ m) πρέπει να αναπτυχθούν νέα μεθοδολογικά εργαλεία. Η μέθοδος εμφύτευσης σε Μαρκοβιανή αλυσίδα εμπλουτίζεται και η οικογένεια των εμφυτεύσιμων μεταβλητών επιστρέψιμου - πολυωνυμικού τύπου σε Μαρκοβιανή αλυσίδα εισάγεται και μελετάται. Η νέα αυτή οικογένεια γενικεύει την οικογένεια εμφυτεύσιμων μεταβλητών διωνυμικού τύπου που έχει εισαχθεί από τους Koutras and Alexandrou (1995), την οικογένεια εμφυτεύσιμων μεταβλητών επιστρέψιμου τύπου που έχει εισαχθεί από τους Han and Aki (1999) και την οικογένεια εμφυτεύσιμων μεταβλητών πολυωνυμικού τύπου που έχει εισαχθεί από τους Antzoulakos et al. (2003).

Σε όλη την εργασία συμβολίζουμε με $[x]$ το μεγαλύτερο ακέραιο που είναι μικρότερος ή ίσος του x και με $\delta_{i,j}$ τη συνάρτηση Delta του Kronecker για τους ακεραίους i, j.

2. Γενικά Αποτελέσματα

Πρόσφατα, οι Dafnis and Makri (2022) εισήγαγαν τις r-ασθενείς ροές μήκους τουλάχιστον k. Θα ορίσουμε τώρα τις r-ασθενείς ροές μήκους τουλάχιστον k και το πολύ m.

Ορισμός 1. Έστω $Z_1, Z_2, \ldots, Z_n$ μία ακολουθία δίτιμων δοκιμών ($Z_i = 0$ ή $Z_i = 1$), αριθμημένων από το 1 ως το n. Τότε, για $k \geq 2$ και $m \geq k$, μία r-ασθενής 1-ροή μήκους τουλάχιστον k και το πολύ m είναι ένας σχηματισμός από τουλάχιστον k και το πολύ m μονάδες ($1s$), όταν οποιεσδήποτε δύο διαδοχικές μονάδες στις δοκιμές i, j $(1 \leq i < j \leq n)$ μπορούν να έχουν μέγιστη απόσταση $r+1$, δηλ. $\forall i, j$ με $Z_i = 1$, $Z_j = 1$ και $Z_{i+t} = 0$ για $t = 1, \ldots, j-i-1$, τότε $d(Z_i, Z_j) \leq r+1$.

Για τον ορισμό της μετρικής της απόστασης σε ακολουθία δίτιμων δοκιμών, παραπέμπουμε στους Dafnis and Makri (2022).

Ας συμβολίσουμε με $L_{n,k,m,r}$ το συνολικό αριθμό των $1s$ που περιέχονται σε όλες τις r-ασθενείς 1-ροές μήκους τουλάχιστον k και το πολύ m σε μία ακολουθία δοκιμών Bernoulli $Z_1, Z_2, \ldots, Z_n$ $(k \leq m)$ με πιθανότητα επιτυχίας (αποτυχίας) στην t δοκιμή $p_t = Pr(Z_t = 1)$ $(q_t = Pr(Z_t = 0) = 1 - p_t)$, $t \geq 1$.

Προκειμένου να διευκολύνουμε την κατανόηση του παραπάνω ορισμού και της στατιστικής συνάρτησης, δίνουμε ένα παράδειγμα. Ας θεωρήσουμε ένα πείραμα κατά το οποίο $n = 30$ δοκιμές Bernoulli, αριθμημένες από το 1 ως το 30, οδηγούν στην ακολουθία αποτελεσμάτων 101001011001110100101101011001. Τότε, οι 1-ασθενείς 1-ροές μήκους τουλάχιστον 2 και το πολύ 4 είναι οι: $|1,2,3|$, $|6,7,8,9|$, $|12,13,14,15,16|$ και παρατηρώντας τις τιμές της νέας τυχαίας μεταβλητής μπορούμε να γράψουμε $L_{2,2,4,1} = 0$, $L_{3,2,4,1} = 2$, $L_{6,2,4,1} = 2$, $L_{8,2,4,1} = 4$, $L_{9,2,4,1} = 5$, $L_{16,2,4,1} = 9$, $L_{24,2,4,1} = 13$, $X_{25,2,4,1} = 4$, $L_{25,2,4,1} = 13$, $X_{26,2,4,1} = 3$, $L_{26,2,4,1} = 9$, $X_{30,2,4,1} = 3$, $L_{30,2,4,1} = 9$.

Έχοντας ορίσει την τυχαία μεταβλητή (τ.μ.) του τρέχοντος ενδιαφέροντος, μπορούμε να αναπτύξουμε τα απαραίτητα μεθοδολογικά εργαλεία για τη μελέτη της κατανομής της. Κατά τη διάρκεια των τελευταίων δεκαετιών η μέθοδος εμφύτευσης σε Μαρκοβιανή αλυσίδα έχει αναπτυχθεί και χρησιμοποιηθεί εκτεταμένα για τη μελέτη κατανομών ροών και σχηματισμών (βλέπετε τους Antzoulakos et al. (2003), Fu and Lou (2003), Fu and Koutras (1994), Han and Aki (1999), Koutras and Alexandrou (1995)).

Στη συνέχεια θα δώσουμε την έννοια της εμφυτεύσιμης μεταβλητής σε Μαρκοβιανή αλυσίδα, η οποία εισήχθη από τους Fu and Koutras (1994). Έστω X_n (n ένας μη-αρνητικός ακέραιος) μία μη-αρνητική, πεπερασμένη τ.μ. που παίρνει ακέραιες τιμές και $\ell_n = \sup\{x : \Pr(X_n = x) > 0\}$ το ανώτερο σημείο της.

Ορισμός 2. Η τ.μ, X_n θα ονομάζεται εμφυτεύσιμη μεταβλητή σε Μαρκοβιανή αλυσίδα αν

(a) υπάρχει Μαρκοβιανή αλυσίδα $\{Y_t,\ t \geq 0\}$ ορισμένη σε ένα διακριτό χώρο

καταστάσεων Ω, ο οποίος μπορεί να διαμεριστεί ως εξής

$$\Omega = \bigcup_{x \geq 0} C_x,$$

(b) η συνάρτηση πιθανότητας της X_n μπορεί να υπολογισθεί θεωρώντας την προβολή του χώρου πιθανότητας της Y_n στο C_x, δηλ.

$$\Pr(X_n = x) = \Pr(Y_n \in C_x), \quad n \geq 0,\ x = 0, 1, \ldots, \ell_n.$$

Στη βιβλιογραφία έχουν μελετηθεί διαφορετικοί τύποι εμφυτεύσιμων τυχαίων μεταβλητών σε Μαρκοβιανή αλυσίδα. Οι Koutras and Alexandrou (1995), μελετώντας την κατανομή του αριθμού ροών και σαρώσεων, όρισαν τη μεταβλητή διωνυμικού τύπου εμφυτεύσιμη σε Μαρκοβιανή αλυσίδα (MVB). Οι Han and Aki (1999) όρισαν τη μεταβλητή επιστρέψιμου τύπου εμφυτεύσιμη σε Μαρκοβιανή αλυσίδα (MVR)για να επιτρέψουν τη μελέτη της κατανομής του αριθμού ροών με ακριβές μήκος. Οι Antzoulakos et al. (2003) μελέτησαν το συνολικό αριθμό επιτυχιών σε ροές μήκους τουλάχιστον k και εισήγαγαν τη μεταβλητή πολυωνυμικού τύπου εμφυτεύσιμη σε Μαρκοβιανή αλυσίδα (MVP). Η μελέτη της κατανομής της τ.μ. $L_{n,k,m,r}$ απαιτεί μία γενίκευση των προαναφερόμενων οικογενειών εμφυτεύσιμων μεταβλητών. Συνεπώς, στη συνέχεια θα ορίσουμε τη μεταβλητή επιστρέψιμου-πολυωνυμικού τύπου εμφυτεύσιμη σε Μαρκοβιανή αλυσίδα (MVRP).

Ορισμός 3. Η τ.μ. X_n θα καλείται μεταβλητή επιστρέψιμου-πολυωνυμικού τύπου εμφυτεύσιμη σε Μαρκοβιανή αλυσίδα (MVRP) αν

(a) υπάρχει Μαρκοβιανή αλυσίδα $\{Y_t,\ t \geq 0\}$ ορισμένη σε ένα διακριτό χώρο καταστάσεων Ω, ο οποίος μπορεί να διαμεριστεί ως εξής

$$\Omega = \bigcup_{x \geq 0} C_x, \quad C_x = \{c_{x0}, c_{x1}, ..., c_{x,s-1}\},$$

(b) υπάρχουν δύο θετικοί ακέραιοι m και k τέτοιοι ώστε για $t \geq 1$

$$\Pr(Y_t \in C_y | Y_{t-1} \in C_x) = 0, \quad \text{για κάθε} \quad y \neq x-m, x-m+1, \ldots, x-1, x, x+1, \ldots, x+k,$$

(c) η συνάρτηση πιθανότητας της X_n μπορεί να υπολογισθεί θεωρώντας την προβολή του χώρου πιθανότητας της Y_n στο C_x, δηλ.

$$\Pr(X_n = x) = \Pr(Y_n \in C_x), \quad n \geq 0,\ x \geq 0.$$

Στον Ορισμό 3 έχουμε θεωρήσει, χωρίς βλάβη της γενικότητας, ότι όλοι οι χώροι υποκαταστάσεων $C_x, x = 0, 1, \ldots$ έχουν τον ίδιο πεπερασμένο πληθάριθμο $s = |C_x|$.

Για $k = 1$ και $m = 0$, ο Ορισμός 3 ανάγεται στη μεταβλητή διωνυμικού τύπου εμφυτεύσιμη σε Μαρκοβιανή αλυσίδα. Για $k = 1$ και $m = 1$, ο Ορισμός 3 ανάγεται στη μεταβλητή επιστρέψιμου τύπου εμφυτεύσιμη σε Μαρκοβιανή αλυσίδα. Για

$m = 0$, ο Ορισμός 3 ανάγεται στη μεταβλητή πολυωνυμικού τύπου εμφυτεύσιμη σε Μαρκοβιανή αλυσίδα.

Από τη συνθήκη (b) του Ορισμού 1 προκύπτει ότι οι μόνες εφικτές μεταβάσεις για την $\{Y_t,\, t \geq 0\}$ είναι αυτές που διεξάγονται μέσα στον ίδιο χώρο υποκαταστάσεων C_x ή από την υποκατάσταση C_x σε μία από τις επόμενες υποκαταστάσεις C_{x+i} $(i = 1, \ldots, k)$, ή σε μία από τις προηγούμενες υποκαταστάσεις C_{x-i} $(i = 1, \ldots, m)$. Έτσι, γεννιούνται οι επόμενοι $(m + k + 1)$ $s \times s$ πίνακες πιθανοτήτων μετάβασης

$$A_{t,i}(x) = (\Pr(Y_t \in c_{x+i,j'} | Y_{t-1} \in c_{x,j})), \quad -m \leq i \leq k,\ t \geq 1,\ x \geq 0.$$

Συμβολίζουμε με $\mathbf{f}_t(x)$ τα διανύσματα πιθανοτήτων

$$\mathbf{f}_t(x) = (\Pr(Y_t \in c_{x0}), \Pr(Y_t \in c_{x1}), ..., \Pr(Y_t \in c_{x,s-1})), \quad t \geq 0,\ x \geq 0$$

και με

$$\boldsymbol{\pi}_x = (\Pr(Y_0 \in c_{x0}), \Pr(Y_0 \in c_{x1}), ..., \Pr(Y_0 \in c_{x,s-1})), \quad x \geq 0,$$

τις αρχικές πιθανότητες της Μαρκοβιανής αλυσίδας.

Είναι προφανές ότι η συνάρτηση πιθανότητας $f_n(x)$ της X_n μπορεί να εκφραστεί ως

$$f_n(x) = \mathbf{f}_n(x)\mathbf{1}', \quad n \geq 0,\ x \geq 0, \tag{1}$$

όπου $\mathbf{1} = (1, 1, ..., 1)$ είναι το διάνυσμα γραμμή του R^s το οποίο έχει όλα τα στοιχεία του ίσα με 1.

Η σύμβαση $\Pr(X_0 = 0) = 1$ συνεπάγεται ότι

$$\boldsymbol{\pi}_0\mathbf{1}' = \mathbf{f}_0(x)\mathbf{1}' = (\Pr(Y_0 \in c_{0,0}), \Pr(Y_0 \in c_{0,1}), ..., \Pr(Y_0 \in c_{0,s-1}))\mathbf{1}' = 1,$$

και $\boldsymbol{\pi}_x\mathbf{1}' = 0$ για $x \geq 1$.

Συμβολίζουμε με $\varphi_t(z)$ και $\phi(z, w)$ την απλή και τη διπλή γεννήτρια συνάρτηση, αντίστοιχα, που είναι

$$\varphi_t(z) = \sum_{x=0}^{\infty} Pr(X_t = x)z^x, \quad \phi(z, w) = \sum_{t=0}^{\infty} \varphi_t(z)w^t,$$

με $\boldsymbol{\varphi}_t(z)$ και $\boldsymbol{\phi}(z, w)$ την απλή και τη διπλή διανυσματική γεννήτρια συνάρτηση, αντίστοιχα, που είναι

$$\boldsymbol{\varphi}_t(z) = \sum_{x=0}^{\infty} \mathbf{f}_t(x)z^x, \quad \boldsymbol{\phi}(z, w) = \sum_{t=0}^{\infty} \boldsymbol{\varphi}_t(z)w^t$$

και με I τον ταυτοτικό πίνακα $s \times s$.

Συμβολίζουμε, τέλος, με $\mu_t = E(X_t)$, $t \geq 1$ τη μέση τιμή της X_t και με $M(w) = \sum_{t=1}^{\infty} \mu_t w^t$, τη γεννήτρια της.
Το ακόλουθο θεώρημα επιτρέπει τον υπολογισμό της συνάρτησης πιθανότητας μίας MVRP.

Θεώρημα 1. Η ακολουθία των διανυσμάτων $\mathbf{f}_t(x)$ ικανοποιεί την αναδρομική σχέση

$$\mathbf{f}_t(x) = \sum_{i=-m}^{k} \mathbf{f}_{t-1}(x-i) A_{t,i}(x-i), \quad t \geq 1,\ x \geq 0,$$

με $\mathbf{f}_t(x) = \mathbf{0}$, για $x < 0$, και $\mathbf{f}_0(x) = \boldsymbol{\pi}_x$, για $0 \leq x \leq \ell_n$.

Απόδειξη. Η αναδρομή προκύπτει άμεσα από το θεώρημα ολικής πιθανότητας και τον Ορισμό 3.

Στη συνέχεια, θα εστιάσουμε την προσοχή μας στην περίπτωση πινάκων που δεν εξαρτώνται από το x. Στην ειδική περίπτωση αυτή, τα αποτελεσματα των Koutras and Alexandrou (1995) και Antzoulakos et al. (2003) γενικεύονται με τη χρήση της αναδρομικής σχέσης του Θεωρήματος 1, και προκύπτουν τα ακόλουθα θεωρήματα.

Θεώρημα 2. Αν $A_{t,i}(x) = A_{t,i}$ για κάθε $x \geq 0$, $t \geq 1$ και $-m \leq i \leq k$, τότε η (απλή) διανυσματική γεννήτρια μιας $MVRP$ τ.μ. X_t μπορεί να εκφραστεί ως

$$\boldsymbol{\varphi}_t(z) = \boldsymbol{\pi}_0 \prod_{j=1}^{t} \left(\sum_{i=-m}^{k} A_{j,i} z^i \right), \quad t \geq 1.$$

Θεώρημα 3. Στην περίπτωση μίας ομογενούς $MVRP$ ($A_{t,i}(x) = A_i$ για κάθε $x \geq 0$, $t \geq 1$ και $-m \leq i \leq k$) η διπλή διανυσματική γεννήτρια της X_t μπορεί να εκφραστεί ως

$$\boldsymbol{\phi}(z,w) = \boldsymbol{\pi}_0 \left(I - w \sum_{i=-m}^{k} A_i z^i \right)^{-1}.$$

Θεώρημα 4. Στην περίπτωση μίας ομογενούς $MVRP$ ($A_{t,i}(x) = A_i$ για κάθε $x \geq 0$, $t \geq 1$ και $-m \leq i \leq k$) η μέση τιμή της X_t και η γεννήτρια των μέσων τιμών μπορούν να εκφραστούν ως

$$\mu_t = E(X_t) = \boldsymbol{\pi}_0 \left(\sum_{j=1}^{t} \left(\sum_{i=-m}^{k} A_i \right)^{j-1} \right) \left(\sum_{i=-m}^{k} i A_i \right) \mathbf{1}'.$$

$$M(w) = \sum_{t=1}^{\infty} \mu_t w^t = \frac{w}{1-w} \boldsymbol{\pi}_0 \left(I - w \sum_{i=-m}^{k} A_i \right)^{-1} \left(\sum_{i=-m}^{k} i A_i \right) \mathbf{1}'.$$

3. Η κατανομή του συνολικού αριθμού των επιτυχιών που περιέχονται σε όλες τις r-ασθενείς 1-ροές μήκους τουλάχιστον k και το πολύ m

Το επόμενο Θεώρημα δίνει τη συνάρτηση πιθανότητας της τ.μ. $L_{n,k,m,r}$.

Θεώρημα 5. Η συνάρτηση πιθανότητας $f_n(x) = P(L_{n,k,m,r} = x)$ της τ.μ. $L_{n,k,m,r}$ $(2 \leq k \leq m,\, r \geq 1)$ δίνεται από τη σχέση (1), όπου $\mathbf{f}_t(x)$ είναι τα διανύσματα πιθανότητας που ικανοποιούν την αναδρομή του Θεωρήματος 1,
$\ell_n = \left[\frac{n}{m+r+1}\right] m + \left(n - \left[\frac{n}{m+r+1}\right](m+r+1)\right) I_{[k,\infty)}\left(n - \left[\frac{n}{m+r+1}\right](m+r+1)\right)$,
$s = (m+1)r + m + 2$, ο $A_{t,0}$ είναι ένας πίνακας $s \times s$, ο οποίος έχει όλα του τα στοιχεία ίσα με 0 εκτός από τα στοιχεία:

- $(1,1)$, που είναι ίσο με q_t,
- $(m+2+jr, 1)$, $j = 1, ..., m+1$, που είναι όλα ίσα με q_t,
- $(i, i+1)$, $1 \leq i \leq k-1$, που είναι όλα ίσα με p_t,
- $(i, m+3+(i-2)r)$, $2 \leq i \leq m+2$, που είναι όλα ίσα με q_t,
- $(m+2, m+2)$, που είναι ίσο με p_t,
- $(m+2+jr+i, j+3)$, για $k \geq 3$, $1 \leq i \leq r$, $j = 0, \ldots, max\{0, k-3\}$, που είναι όλα ίσα με p_t,
- $(m+1+i, m+2+i)$, για $r \geq 2$, $(j-2)r+2 \leq i \leq (j-1)r$, $j = 2, \ldots, m+2$, που είναι όλα ίσα με q_t,
- $((r+1)m+2+i, m+2)$, $1 \leq i \leq r$, που είναι όλα ίσα με p_t,

ο $A_{t,1}$ είναι ένας πίνακας $s \times s$, ο οποίος έχει όλα του τα στοιχεία ίσα με 0 εκτός από τα στοιχεία $(i, i+1)$, $i = k+1, \ldots, m$ και τα στοιχεία $(m+2+(k-1)r+jr+i, k+2+j)$, $i = 1, \ldots r$, $j = 0, \ldots m-k-1$, που είναι όλα ίσα με p_t, ο $A_{t,k}$ είναι ένας πίνακας $s \times s$, ο οποίος έχει όλα του τα στοιχεία ίσα με 0 εκτός από το στοιχείο $(k, k+1)$ και τα στοιχεία $(m+2+(k-2)r+i, k+1)$, $i = 1, \ldots r$, που είναι όλα ίσα με p_t, ο $A_{t,-m}$ είναι ένας πίνακας $s \times s$, ο οποίος έχει όλα του τα στοιχεία ίσα με 0 εκτός από το στοιχείο $(m+1, m+2)$ και τα στοιχεία $(m+2+(m-1)r+i, m+2)$, $i = 1, \ldots r$, που είναι όλα ίσα με p_t και όλοι οι πίνακες $A_{t,i}$, για $i = -m+1, \ldots, -1$ και $i = 2, \ldots, k-1$, είναι $s \times s$ μηδενικοί πίνακες.

Απόδειξη. Θα αποδείξουμε πρώτα ότι η $L_{n,k,m,r}$ ανήκει στην οικογένεια των $MVRP$ τ.μ.

Θέτουμε
$\ell_n = \left[\frac{n}{m+r+1}\right] m + \left(n - \left[\frac{n}{m+r+1}\right](m+r+1)\right) I_{[k,\infty)}\left(n - \left[\frac{n}{m+r+1}\right](m+r+1)\right)$
και εισάγουμε το χώρο καταστάσεων $\Omega = \bigcup_{y=0}^{\ell_n} C_y$ όπου οι C_y, $y = 0, 1, \ldots, \ell_n$ είναι ξένοι μεταξύ τους υπόχωροι με $\mid C_y \mid = (m+1)r + m + 2$, $y = 0, 1, \ldots, \ell_n$, στοιχεία, που συμβολίζονται ως εξής:

$$C_y = \{c_{y,0}, c_{y,1}, \ldots, c_{y,m+1}, c_{y,1^{(1)}}, c_{y,1^{(2)}}, \ldots, c_{y,1^{(r)}},$$
$$c_{y,2^{(1)}}, c_{y,2^{(2)}}, \ldots, c_{y,2^{(r)}}, \ldots, c_{y,(m+1)^{(1)}}, c_{y,(m+1)^{(2)}}, \ldots, c_{y,(m+1)^{(r)}}\}.$$

Στη συνέχεια εισάγουμε μια Μαρκοβιανή αλυσίδα $\{Y_t,\, t \geq 0\}$ στον Ω ως εξής:

$Y_t \in c_{y,i} = \{(y,i)\}$, ή ισοδύναμα $Y_t = (y,i)$, αν στα πρώτα t αποτελέσματα $Z_1, Z_2, \ldots, Z_t$ ο συνολικός αριθμός των $1s$ που περιέχονται σε όλες τις r-ασθενείς 1-ροές μήκους τουλάχιστον k και το πολύ m είναι y και
(α) $i = 0$, αν
(1) $y = 0$ και $\prod_{j=1}^{t}(1 - Z_j) = 1$ (δεν έχει εμφανισθεί 1 μέχρι το αποτέλεσμα t) ή
(2) για $t \geq r+1$, $\prod_{i=0}^{r}(1 - Z_{t-i}) = 1$ (το μήκος της τρέχουσας ροής αποτυχιών είναι μεγαλύτερο από r).
(β) $i = j$, $j = 1, 2, \ldots, m$, αν το αποτέλεσμα t είναι το τελευταίο 1 ($Z_t = 1$) μιας r-ασθενούς 1-ροής μήκους j.
(γ) $i = m+1$, αν το τελευταίο αποτέλεσμα είναι το τελευταίο 1 ($Z_t = 1$) μιας r-ασθενούς 1-ροής μήκους τουλάχιστον $m+1$.
(δ) $i = j^{(d)}$, $1 \leq j \leq m$, $1 \leq d \leq r$, αν το αποτέλεσμα t είναι το d-οστό συνεχόμενο 0 μιας ροής αποτυχιών
($\prod_{i=0}^{d-1}(1 - Z_{t-i}) = 1$, $Z_{t-d} = 1$), η οποία ακολουθεί μια r-ασθενή 1-ροή μήκους j.
(ε) $i = (m+1)^{(k)}$, $1 \leq k \leq r$, αν το αποτέλεσμα t είναι το k-οστό συνεχόμενο 0 μιας ροής αποτυχιών
($\prod_{i=0}^{k-1}(1 - Z_{t-i}) = 1$, $Z_{t-k} = 1$), η οποία ακολουθεί μια r-ασθενή 1-ροή μήκους τουλάχιστον m.

Με αυτή την προεργασία η τ.μ. $L_{n,k,m,r}$ γίνεται $MVRP$ με διάνυσμα αρχικών πιθανοτήτων

$$\boldsymbol{\pi}_0 = (1,\, 0,\, 0,\, \ldots,\, 0)_{1\times((m+1)r+m+2)}\,,$$

και πίνακες $A_{t,i}$, $i = -k, \ldots, m$ οι οποίοι έχουν τα στοιχεία που περιγράφονται στο θεώρημα.

Το αποτέλεσμα προκύπτει άμεσα από το Θεώρημα 1 και τον Ορισμό 3.

Ας επιστρέψουμε στο παράδειγμα της Ενότητας 2. Για την ακολουθία των αποτελεσμάτων 10100101100111010010110101 1001 και την ειδική περίπτωση $k = 2$, $m = 4$, $r = 1$, οι καταστάσεις της Μαρκοβιανής αλυσίδας του Θεωρήματος 5 είναι $Y_1 = (0,1)$, $Y_2 = (0,1^{(1)})$, $Y_3 = (2,2)$, $Y_4 = (2,2^{(1)})$, $Y_5 = (2,0)$, $Y_6 = (2,1)$, $Y_7 = (2,1^{(1)})$, $Y_8 = (4,2)$, $Y_9 = (5,3)$, $Y_{10} = (5,3^{(1)})$, $Y_{11} = (5,0)$, $Y_{12} = (5,1)$, $Y_{13} = (7,2)$, $Y_{14} = (8,3)$, $Y_{15} = (8,3^{(1)})$, $Y_{16} = (9,4)$, $Y_{17} = (9,4^{(1)})$, $Y_{18} = (9,0)$, $Y_{19} = (9,1)$, $Y_{20} = (9,1^{(1)})$, $Y_{21} = (11,2)$, $Y_{22} = (12,3)$, $Y_{23} = (12,3^{(1)})$, $Y_{24} = (13,4)$, $Y_{25} = (13,4^{(1)})$, $Y_{26} = (9,5)$, $Y_{27} = (9,5)$, $Y_{28} = (9,5^{(1)})$, $Y_{29} = (9,0)$, $Y_{30} = (9,1)$. Σε αυτή την περίπτωση οι μη-μηδενικοί πίνακες πιθανοτήτων μετάβασης

είναι οι εξής:

$$A_{t,0} = \begin{pmatrix} (\cdot,0) & (\cdot,1) & (\cdot,2) & (\cdot,3) & (\cdot,4) & (\cdot,5) & (\cdot,1^{(1)}) & (\cdot,2^{(1)}) & (\cdot,3^{(1)}) & (\cdot,4^{(1)}) & (\cdot,5^{(1)}) \\ q_t & p_t & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & q_t & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & q_t & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & q_t & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & q_t & 0 \\ 0 & 0 & 0 & 0 & 0 & p_t & 0 & 0 & 0 & 0 & q_t \\ q_t & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ q_t & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ q_t & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ q_t & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ q_t & 0 & 0 & 0 & 0 & p_t & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$A_{t,1} = \begin{pmatrix} (\cdot,0) & (\cdot,1) & (\cdot,2) & (\cdot,3) & (\cdot,4) & (\cdot,5) & (\cdot,1^{(1)}) & (\cdot,2^{(1)}) & (\cdot,3^{(1)}) & (\cdot,4^{(1)}) & (\cdot,5^{(1)}) \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & p_t & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & p_t & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & p_t & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & p_t & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$A_{t,2} = \begin{pmatrix} (\cdot,0) & (\cdot,1) & (\cdot,2) & (\cdot,3) & (\cdot,4) & (\cdot,5) & (\cdot,1^{(1)}) & (\cdot,2^{(1)}) & (\cdot,3^{(1)}) & (\cdot,4^{(1)}) & (\cdot,5^{(1)}) \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & p_t & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & p_t & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$A_{t,-4} = \begin{pmatrix} (\cdot,0) & (\cdot,1) & (\cdot,2) & (\cdot,3) & (\cdot,4) & (\cdot,5) & (\cdot,1^{(1)}) & (\cdot,2^{(1)}) & (\cdot,3^{(1)}) & (\cdot,4^{(1)}) & (\cdot,5^{(1)}) \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & p_t & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & p_t & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

ΑΝΑΦΟΡΕΣ

Antzoulakos, D.L., Bersimis, S. and Koutras M.V. (2003). On the distribution of the total number of run lengths. *Annals of the Institute of Statistical Mathematics*, **55**, 865-884.

Balakrishnan, N. and Koutras, M.V. (2002). *Runs and Scans with Applications,* New York: John Wiley.

Bonhomme R. (2000). Bases and limits to using "degree.day" units. *European journal of agronomy*, **13**, 1-10.

Dafnis S.D. and Makri F.S. (2022). Weak runs in sequences of binary trials. *Metrika*, **85**, 573-603.

Fu J.C. and Koutras M.V. (1994). Distribution theory of runs: a Markov chain approach. *Journal of the American Statistical Association*, **89**, 1050- 1058.

Fu J.C. and Lou W.Y.W. (2003). *Distribution theory of runs and patterns and its applications: a finite Markov imbedding approach.* Singapore: World Scientific Publishing.

Han Q. and Aki S. (1999) Joint distributions of runs in a sequence of multi-state trials. *Annals of the Institute of Statistical Mathematics*, **51**, 419-447.

Koutras M.V. and Alexandrou V.A. (1995). Runs, scans and urn model distributions: a unified Markov chain approach. *Annals of the Institute of Statistical Mathematics*, **47**, 743-766.

Makri F.S., Philippou A.N. and Psillakis Z.M. (2007). Success run statistics defined on an urn model. *Advances in Applied Probability*, **39**, 991-1019.

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 34ου Πανελληνίου Συνεδρίου Στατιστικής (2022), (σελ. 56-69)

ΣΥΜΠΕΡΑΣΜΑΤΟΛΟΓΙΑ ΑΞΙΟΠΙΣΤΙΑΣ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΑ ΑΝΑΛΟΓΙΣΤΙΚΑ - ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΑ ΜΑΘΗΜΑΤΙΚΑ

Α. Μακρίδης, Χ. Μεσελίδης, Α. Καραγρηγορίου
Εργαστήριο Στατιστικής και Ανάλυσης Δεδομένων
Πανεπιστήμιο Αιγαίου
{amakridis, meselidis, alex.karagrigoriou}@aegean.gr

ΠΕΡΙΛΗΨΗ

Στην παρούσα εργασία αναπτύσσεται μια καινοτόμος προσέγγιση πρόβλεψης δείκτη ζημίας χρησιμοποιώντας μια ειδική μορφή ημιμαρκοβιανών διαδικασιών. Ορίζονται τρία επίπεδα του (μηνιαίου) δείκτη ζημίας τα οποία θεωρούνται ως οι καταστάσεις μιας ημιμαρκοβιανής διαδικασίας και χρησιμοποιείται η μεθοδολογία ημιμαρκοβιανών διαδικασιών για την εκτίμηση των πιθανοτήτων μετάβασης μεταξύ των επιπέδων/καταστάσεων του δείκτη ζημίας και κατ΄ επέκταση για την αξιολόγηση των αποτελεσμάτων των ασφαλιστικών συμβούλων insurance advisors) και την έγκαιρη εξυγίανση των χαρτοφυλακίων τους (insurance portfolio).

Λέξεις κλειδιά: Αναλογισμός, Αξιοπιστία, Ημιμαρκοβιανές διαδικασίες, Γενικές Ασφαλίσεις, Δείκτης ζημίας.

1. ΕΙΣΑΓΩΓΗ

Ένα σύστημα που μπορεί να έχει έναν πεπερασμένο αριθμό ρυθμών απόδοσης μπορεί να μοντελοποιηθεί και να αξιολογηθεί ως προς την αξιοπιστία του, μέσω ενός συστήματος πολλών καταστάσεων (multi state system, MSS). Οι βασικές έννοιες της αξιοπιστίας MSS προτάθηκαν τη δεκαετία του ΄70 ενώ γενικεύσεις και προεκτάσεις τους παρουσιάσθηκαν τη δεκαετία του ΄80 (Natvig 1982; Murchland 1975; El-Neveihi and Proschan 1984). Για πρόσφατες θεωρητικές μελέτες και εφαρμογές στη θεωρία αξιοπιστίας των MSS ο αναγνώστης μπορεί να ανατρέξει στους Lisnianski et al. (2010), Lisnianski and Levitin (2003) και Natvig (2011). Όσον αφορά συστήματα ημιμαρκοβιανών διαδικασιών συνεχούς χρόνου και σχετικά θέματα αξιοπιστίας ενδεικτικές αναφορές αποτελούν οι μελέτες των Limnios and Oprişan (2001) και Limnios and Ouhbi (2003 & 2006).

Όσον αφορά ασφαλιστικές εταιρείες που δραστηριοποιούνται στον χώρο των γενικών ασφαλίσεων, ο έλεγχος της επάρκειας του καθαρού ασφαλίστρου μπορεί να διερευνηθεί μέσω του (μηνιαίου) δείκτη ζημίας ο οποίος δύναται να θεωρηθεί ως σύστημα πολλαπλών καταστάσεων. Ως καθαρό ασφάλιστρο (ή απλά, ασφάλιστρο) νοείται εκείνο το οποίο είναι σε θέση να καλύψει το ύψος των απαιτήσεων που δημιουργούνται απο τις ασφαλιστικές εργασίες της εταιρείας. Το ασφάλιστρο προσαυξημένο ώστε να απορροφά τα έξοδα και να λαμβάνει υπόψη του το περιθώριο κέρδους, δημιουργεί το εμπορικό ασφάλιστρο. Ο (μηνιαίος) δείκτης ζημίας επομένως μπορεί να οριστεί ως ο λόγος του συνολικού ύψους των απαιτήσεων προς το καθαρό ασφάλιστρο. Έτσι ένα ορθώς (δίκαιο από την σκοπιά της στατιστικής και των πιθανοτήτων) ορισμένο ασφάλιστρο οδηγεί τον δείκτη ζημίας σε μια 'τυπική' κατάσταση (ούτε ζημιογόνο ούτε κερδοφόρο) που κινείται γύρω από τη μονάδα (1). Για περισσότερες πληροφορίες σχετικά με αναλογιστικές τεχνικές γενικών ασφαλίσεων παραπέμπουμε στους Parodi (2014) και Klugman et al. (2019). Ένα κανάλι διανομής ασφαλιστικών προϊόντων στο κοινό, αποτελούν οι ασφαλιστικοί σύμβουλοι. Οι ασφαλιστικοί σύμβουλοι που εκπροσωπούν μία εταιρεία αποτελούν καίριο συστατικό στοιχείο της επιτυχίας της και αυτό καθίσταται εμφανές αν αναλογιστεί κανείς το γεγονός ότι οι εργασίες τους είναι άμεσα συνυφασμένες με τα έσοδα και κατ' επέκταση, τα κέρδη της εταιρείας. Υψίστης σημασίας αποτελεί επομένως η αξιολόγηση των αποτελεσμάτων τους και υπό το πρίσμα αυτό, ο δείκτης ζημίας δύναται να αποτελέσει ένα εργαλείο ελέγχου – αξιολόγησης των αποτελεσμάτων των ασφαλιστικών συμβούλων, ενώ στο πλαίσιο των συστημάτων πολλαπλών καταστάσεων (MSS) μπορεί να χρησιμοποιηθεί για την πρόβλεψη της κατάστασης στην οποία μπορεί να περιέλθει ένας ασφαλιστικός σύμβουλος (στην παρούσα μελέτη θεωρούμε τρεις καταστάσεις: κερδοφόρο, τυπική και ζημιογόνο), επιτρέποντας στην εταιρεία να λαμβάνει έγκαιρα τα ενδεδειγμένα μέτρα και αποφάσεις για τη διατήρηση ή την εξυγίανση του χαρτοφυλακίου εκάστου συμβούλου.

Έστω μια διαδικασία ορισμένη σε ένα χώρο πιθανοτήτων (Ω, F, P) και χώρο καταστάσεων $E = 1, 2, \ldots, N$ όπου για παράδειγμα, η κατάσταση N αφορά στην πλήρη απόδοση (λειτουργία) του συστήματος ενώ η κατάσταση 1 στην πλήρη αποτυχία (μη λειτουργία). Οι μαρκοβιανές διαδικασίες αποτελούν τυπικά εργαλεία για τη μοντελοποίηση ενός τέτοιου συστήματος. Σε αυτή την εργασία εστιάζουμε σε συστήματα πολλών καταστάσεων τα οποία μοντελοποιούμε μέσω ημιμαρκοβιανών διαδικασιών, οι οποίες γενικεύουν τυπικές μαρκοβιανές διαδικασίες επιτρέποντας γενικές κατανομές για τους χρόνους παραμονής (sojourn times) στις καταστάσεις του συστήματος (Limnios and Oprişan 2001). Η ευελιξία αυτή καθιστά τις ημιμαρκοβιανές διαδικασίες προτιμητέες όχι μόνο για μελέτες αξιοπιστίας συστημάτων αλλά και για εφαρμογές, γενικότερα. Για τους σκοπούς της παρούσας μελέτης οι χρόνοι παραμονής σε μια δεδομένη κατάσταση θεωρείται ότι ακολουθούν κατανομές που ανήκουν σε μια γενικευμένη κλάση κατανομών. Ένα από τα σημαντικότερα όσο και ενδιαφέροντα θεωρητικά χαρακτηριστικά της προτεινόμενης κλάσης κατανομών είναι

ότι αυτή είναι κλειστή ως προς την ελάχιστη και τη μέγιστη διατεταγμένη παρατήρηση (βλ. Balasubramanian et al. 1991; Barbu et. al 2017). Αξίζει να σημειωθεί ότι αρκετές δημοφιλείς κατανομές αξιοπιστίας όπως η εκθετική, η Weibull και η Pareto αποτελούν μέλη της συγκεκριμένης κλάσης κατανομών.

Το περίγραμμα της εργασίας είναι το εξής: Στην ενότητα 2 παρουσιάζεται η θεωρία των ημιμαρκοβιανών διαδικασιών και ορίζεται η γενικευμένη κλάση κατανομών. Στην ενότητα 3 παρουσιάζεται η διαδικασία εκτίμησης και στην ενότητα 4 οι πιθανότητες μετάβασης και οι εκτιμητές των εμπλεκομένων παραμέτρων. Η εφαρμογή της μεθοδολογίας στο μηνιαίο δείκτη ζημίας και οι πιθανότητες μετάβασης για διάφορα χρονικά διαστήματα παρουσιάζονται στην ενότητα 5.

2. ΗΜΙΜΑΡΚΟΒΙΑΝΕΣ ΔΙΑΔΙΚΑΣΙΕΣ ΚΑΙ ΣΥΣΤΗΜΑΤΑ ΠΟΛΛΩΝ ΚΑΤΑΣΤΑΣΕΩΝ

Οι ημιμαρκοβιανές διαδικασίες είναι τυπικά εργαλεία για τη μοντελοποίηση τεχνικών συστημάτων. Μια τέτοια κατηγορία στοχαστικών διαδικασιών γενικεύει τις τυπικές μαρκοβιανές διαδικασίες άλματος επιτρέποντας γενικές κατανομές για χρόνους παραμονής (Limnios and Oprişan 2001).

Ας υποθέσουμε ότι η χρονική εξέλιξη ενός συστήματος διέπεται από μια στοχαστική διαδικασία $Z = (Z_t)_{t\in\mathbb{R}_+}$. Ορίζουμε με $S = (S_n)_{n\in\mathbb{N}}$ τις διαδοχικές χρονικές στιγμές κατά τις οποίες παρατηρείται αλλαγή κατάστασης (μετάβαση από μια κατάσταση σε άλλη) της $(Z_t)_{t\in\mathbb{R}_+}$ και με $J = (J_n)_{n\in\mathbb{N}}$ τις διαδοχικές καταστάσεις κατά τις χρονικές αυτές στιγμές. Ορίζουμε επίσης $X = (X_n)_{n\in\mathbb{N}}$ να είναι οι διαδοχικοί χρόνοι παραμονής (sojourn times) στις εκάστοτε καταστάσεις. Έτσι,

$$X_n = S_n - S_{n-1},\ n \in \mathbb{N}^*,$$

και, κατά σύμβαση, θεωρούμε $X_0 = S_0 = 0$.

Ας θυμηθούμε εδώ τους ορισμούς μιας ανανεωτικής διαδικασίας Markov και μιας ημιμαρκοβιανής διαδικασίας (βλ. Limnios and Oprişan 2001). Έστω ότι $(J, S) = (J_n, S_n)_{n\in\mathbb{N}}$ ικανοποιεί τη σχέση

$$\begin{aligned}&\mathbb{P}(J_{n+1} = j, S_{n+1} - S_n \le t | J_0, \cdots, J_n; S_1, \cdots, S_n)\\ &= \mathbb{P}(J_{n+1} = j, S_{n+1} - S_n \le t | J_n), j \in E, t \in \mathbb{R}_+,\end{aligned}$$

τότε

- το ζεύγος (J, S) αποτελεί την ανανεωτική διαδικασία Markov
- η $(J_n)_{n\in\mathbb{N}}$ καλείται εμβαπτισμένη μαρκοβιανή αλυσίδα (embedded Markov chain) και

- $Z = (Z_t)_{t\in\mathbb{R}_+}$ ονομάζεται ημιμαρκοβιανή διαδικασία που σχετίζεται με το ζεύγος (J, S), όπου

$$Z_t := J_{N(t)} \quad \Leftrightarrow \quad J_n = Z_{S_n},$$

ενώ

$$N(t) := \max\{n \in \mathbb{N} \mid S_n \leq t\},\ t \in \mathbb{R}_+, \tag{1}$$

είναι η απαριθμητή διαδικασία του πλήθους των μεταβάσεων στο χρονικό διάστημα $(0, t]$. Έτσι, Z_t ορίζεται ως η κατάσταση του συστήματος τη χρονική στιγμή t.

Μια ημιμαρκοβιανή διαδικασία χαρακτηρίζεται από

1. την αρχική κατανομή (αρχικές πιθανότητες) $\alpha = (\alpha_1, \ldots, \alpha_N)$ όπου

$$\alpha_j := \mathbb{P}(J_0 = j),\ j \in E,$$

και

2. από τον πυρήνα που δίνεται από τον τύπο

$$Q_{ij}(t) := \mathbb{P}(J_n = j, X_n \leq t | J_{n-1} = i).$$

Ορίζουμε επίσης

1. τις πιθανότητες μετάβασης της εμβαπτισμένης μαρκοβιανής αλυσίδας $(J_n)_{n\in\mathbb{N}}$,

$$p_{ij} := \mathbb{P}(J_n = j | J_{n-1} = i) = \lim_{t\to\infty} Q_{ij}(t),$$

καθώς και

2. τις συναρτήσεις κατανομής του χρόνου παραμονής

$$\begin{aligned} W_{ij}(t) &:= \mathbb{P}(S_n - S_{n-1} \leq t | J_{n-1} = i, J_n = j) \\ &= \mathbb{P}(X_n \leq t | J_{n-1} = i, J_n = j). \end{aligned}$$

Δεν είναι δύσκολο να δειχθεί ότι $Q_{ij}(t) = p_{ij} W_{ij}(t)$.

Έστω T_{ij} ο χρόνος που δαπανάται (χρόνος παραμονής) στην κατάσταση i πριν το σύστημα μεταβεί (απευθείας) στην κατάσταση j. Συμβολίζουμε με $F_{ij}(t; \theta_{ij})$ τη συνάρτηση κατανομής του χρόνου παραμονής, όπου θ_{ij} η m-διάσταση εμπλεκόμενη παράμετρος. Υποθέτουμε ότι η κατανομή του χρόνου T_{ij} είναι απόλυτα συνεχής παντού ως προς το μέτρο Lebesgue με συνάρτηση πυκνότητας πιθανότητας $f_{ij}(t; \theta_{ij})$.

Η δυναμική του συστήματος που προτείνεται είναι η εξής: η επόμενη κατάσταση που θα μεταβεί το σύστημα, μετά την κατάσταση i, είναι αυτή για την οποία ο

χρόνος T_{ij} είναι ο ελάχιστος (μεταξύ των χρόνων $T_{i1}, \ldots, T_{i,i-1}, T_{i,i+1}, \ldots, T_{iN}$). Έτσι, για το ημιμαρκοβιανό αυτό μοντέλο, ο πυρήνας παίρνει τη μορφή

$$\begin{aligned} Q_{ij}(t) &= \mathbb{P}(\min_l T_{il} \leq t, T_{ij} \leq T_{il}, \forall l | J_{n-1} = i) \\ &= \mathbb{P}(\min_l T_{il} \leq t | J_{n-1} = i, J_n = j) \times \mathbb{P}(T_{ij} \leq T_{il}, \forall l | J_{n-1} = i) \\ &= p_{ij} W_i(t), \end{aligned}$$

όπου

$$p_{ij} = \mathbb{P}(J_n = j | J_{n-1} = i) = \mathbb{P}(T_{ij} \leq T_{il}, \forall l | J_{n-1} = i)$$

και

$$\begin{aligned} W_{ij}(t) &= \mathbb{P}(X_n \leq t | J_{n-1} = i, J_n = j) \\ &= \mathbb{P}(\min_l T_{il} \leq t | J_{n-1} = i, J_n = j) \\ &= \mathbb{P}(\min_l T_{il} \leq t | J_{n-1} = i) =: W_i(t), \text{ ανεξάρτητο από το } j, \end{aligned}$$

που αντιπροσωπεύει τη συνάρτηση κατανομής του χρόνου παραμονής στην κατάσταση i ανεξάρτητα από το πού θα γίνει η επόμενη μετάβαση. Είναι προφανές ότι

$$\sum_j Q_{ij}(t) = W_i(t).$$

Έστω ότι η $W_i(t)$ είναι απόλυτα συνεχής ως προς το μέτρο Lebesgue με πυκνότητα πιθανότητας $f_i(t)$.

2.1 Γενικευμένη Κλάση Κατανομών

Η μορφή των κατανομών που εξετάζονται στην εργασία αυτή παρουσιάζεται σε αυτήν την ενότητα. Πιο συγκεκριμένα, εξετάζουμε την περίπτωση όπου οι κατανομές $F_{ij}(\cdot; a_{ij})$, $i, j = 1, \ldots, N$, κάποιων τυχαίων μεταβλητών X_{ij} είναι της ίδιας συναρτησιακής μορφής αλλά με διαφορετικές παραμέτρους a_{ij}. Με άλλα λόγια, εστιάζουμε σε ανεξάρτητες αλλά όχι απαραίτητα ισόνομες τυχαίες μεταβλητές. Το τυπικό μέλος αυτής της κλάσης κατανομών έχει παράμετρο ίση με τη μονάδα (1) και θεωρείται ότι επαληθεύει την πιο κάτω σχέση.

$$F_{ij}(t; a_{ij}) := 1 - (1 - F(t; 1))^{a_{ij}}. \tag{2}$$

Υποθέσουμε ότι η $F_{ij}(t; a_{ij})$ είναι συνεχής ως προς το μέτρο Lebesgue και συμβολίσουμε την πυκνότητά της με $f_{ij}(t; a_{ij})$.

Θεώρημα 1. Έστω $X_1, \ldots, X_N$ ανεξάρτητες αλλά όχι αναγκαστικά ισόνομες τυχαίες μεταβλητές έτσι ώστε $X_i \sim F(x; a_i)$ που ανήκει στην κλάση (2). Τότε, η

συνάρτηση κατανομής $F^{(1)}$ της ελάχιστης διατεταγμένης τυχαίας μεταβλητής $X_{(1)}$ ανήκει επίσης στην (2) (βλ. Barbu et. al 2017).

Η απλούστερη διακριτή κατανομή που ανήκει στην κλάση (2) είναι η Γεωμετρική κατανομή. Στις συνεχείς κατανομές περιλαμβάνονται οι Pareto, Kumaraswamy, Weibull καθώς και ειδικές περιπτώσεις της τελευταίας, όπως η Εκθετική, η Rayleigh και η Erlang truncated exponential.

Για την πιο πάνω κλάση κατανομών, μπορεί εύκολα να δειχθεί το ακόλουθο αποτέλεσμα σχετικά με τα κύρια χαρακτηριστικά της ημιμαρκοβιανής διαδικασίας. Για ευκολία χρησιμοποιείται ο ακόλουθος συμβολισμός $F(t) := F(t;1)$, $f(t) := f(t;1)$ και $Q_{ij}\left(t; a_{ik}; k = 1, \ldots, N\right) := Q_{ij}(t)$.

Πρόταση 1. Σύμφωνα με το προτεινόμενο μοντέλο, ισχύουν τα ακόλουθα (βλ. Barbu et. al 2017):

$$Q_{ij}(t) = \frac{a_{ij}}{\sum\limits_{k\in E} a_{ik}} \left[1 - (1 - F(t))^{\sum\limits_{k\in E} a_{ik}}\right], \tag{3}$$

$$p_{ij} = \frac{a_{ij}}{\sum\limits_{k\in E} a_{ik}}, \tag{4}$$

$$W_i(t) = 1 - [1 - F(t)]^{\sum\limits_{j=1}^{N} a_{ij}} \tag{5}$$

και

$$f_i(t) = \sum_{j=1}^{N} a_{ij} \left(1 - F(t)\right)^{\sum\limits_{j=1}^{N} a_{ij}} \frac{f(t)}{1 - F(t)}. \tag{6}$$

3. ΕΚΤΙΜΗΣΗ ΒΑΣΙΚΩΝ ΠΑΡΑΜΕΤΡΩΝ

Για την εκτίμηση παραμέτρων (τόσο εκείνων που αφορούν την κατανομή όσο και των αρχικών πιθανότητων) εφαρμόζεται η μέθοδος μέγιστης πιθανοφάνειας. Μπορούν να ληφθούν υπόψη διάφορα σενάρια ενός ή πολλών μονοπατιών (pathways), τόσο με όσο και χωρίς λογοκρισία (censoring). Στην παρούσα εργασία το ενδιαφέρον εστιάζεται στην γενική περίπτωση $L, L \geq 1$, μονοπατιών χωρίς λογοκρισία.

Θεωρώντας L τυχαία δείγματα μονοπατιών μιας ημιμαρκοβιανής διαδικασίας έστω $\left\{j_0^{(l)}, x_1^{(l)}, j_1^{(l)}, x_2^{(l)}, \ldots, j_{N^l(M)}^{(l)}\right\}$, $l = 1, \ldots, L$, για το χρονικό διάστημα $(0, M)$, η αντίστοιχη συνάρτηση πιθανοφάνειας γράφεται στη μορφή

$$\begin{aligned} \mathcal{L} &= \prod_{l=1}^{L} \alpha_{j_0}^{(l)} p_{j_0^{(l)} j_1^{(l)}} f_{j_0^{(l)}}(x_1^{(l)}) \ldots f_{j_{N^l(M)-1}^{(l)}}(x_{N^l(M)}^{(l)}) \qquad (7) \\ &= \left(\prod_{i \in E} \alpha_i^{N_{i,0}^{(L)}}\right) \left(\prod_{i,j \in E} p_{ij}^{\sum_{l=1}^{L} N_{ij}^{(l)}(M)}\right) \times \left(\prod_{l=1}^{L} \prod_{i \in E} \prod_{k=1}^{N_i^{(l)}(M)} f_i(x_i^{(l,k)})\right), \end{aligned}$$

όπου

- $N_{i,0}^{(L)} := \sum_{l=1}^{L} \mathbb{1}_{\{J_0^{(l)}=i\}}$,
- $N_i^{(l)}(M)$: το πλήθος μεταβάσεων στην κατάσταση i μέχρι τη χρονική στιγμή M, για το l μονοπάτι, $l = 1, \ldots, L$,
- $N_{ij}^{(l)}(M)$: το πλήθος μεταβάσεων από την κατάσταση i στην κατάσταση j μέχρι τη χρονική στιγμή M, για το μονοπάτι l, $l = 1, \ldots, L$,
- $N_{ij}(L, M) := \sum_{l=1}^{L} N_{ij}^{(l)}(M)$,
- $x_i^{(l,k)}$: ο χρόνος παραμονής στην κατάσταση i κατά την k επίσκεψη, $k = 1, \ldots, N_i^{(l)}(M)$ για το μονοπάτι l, $l = 1, \ldots, L$.

Οι εκτιμητές των παραμέτρων a_{ij} δίνονται από τη σχέση

$$\widehat{a}_{ij}(L, M) = -\frac{N_{ij}(L, M)}{\sum_{l=1}^{L} \sum_{k=1}^{N_i^{(l)}(M)} \log\left(1 - F\left(x_i^{(l,k)}\right)\right)}, \qquad (8)$$

ενώ οι εκτιμητές των αρχικών πιθανοτήτων δίνονται από τη σχέση

$$\widehat{\alpha}_i(L, M) = \frac{N_{i,0}^{(L)}}{L}, \qquad (9)$$

όπου $F(\cdot) \equiv F(\cdot, 1)$ το τυπικό μέλος της γενικευμένης κλάσης κατανομών (2).

4. ΑΝΑΝΕΩΤΙΚΗ ΣΥΝΑΡΤΗΣΗ ΜΑΡΚΟΒ ΚΑΙ ΠΙΘΑΝΟΤΗΤΕΣ ΜΕΤΑΒΑΣΗΣ ΗΜΙΜΑΡΚΟΒΙΑΝΗΣ ΔΙΑΔΙΚΑΣΙΑΣ - ΕΚΤΙΜΗΣΗ ΒΑΣΙΚΩΝ ΠΑΡΑΜΕΤΡΩΝ

Η ενότητα αυτή πραγματεύεται δύο σημαντικές για τη μελέτη της συμπεριφοράς μιας ημιμαρκοβιανής διαδικασίας ποσότητες, ήτοι την ανανεωτική συνάρτηση

Markov, (βλ. σχέση (10)) και τη συνάρτηση μετάβασης της ημιμαρκοβιανής διαδικασίας (βλ. σχέση (12)). Αυτές οι δύο ποσότητες είναι σημαντικές τόσο αυτόνομα, όσο και στο ρόλο που διαδραματίζουν σε πιο σύνθετες ποσότητες όπως για παράδειγμα είναι διάφοροι δείκτες αξιοπιστίας (που όμως δεν θα μας απασχολήσουν στην παρούσα εργασία), (βλ. Barbu and Limnios 2008 & Ouhbi and Limnios 1996).

Η ανανεωτική συνάρτηση Markov, συμβολίζεται με $\Psi_{ij}(t)$, $i, j \in E$, $t \geq 0$, και ορίζεται ως η αναμενόμενη τιμή του αριθμού των επισκέψεων $N_j(t)$ στην κατάσταση j μέχρι κάποια χρονική στιγμή t γνωρίζοντας ότι η διαδικασία ξεκίνησε στην κατάσταση i στην αρχική χρονική στιγμή $t = 0$. Ως εκ τούτου, η ανανεωτική συνάρτηση Markov δίνεται από τη σχέση (βλ. Limnios and Ouhbi, 2006):

$$\begin{aligned} \Psi_{ij}(t) &:= \mathbb{E}_i[N_j(t)] = \sum_{n=1}^{\infty} Q_{ij}^{(n)}(t) \\ &= \sum_{n=1}^{\infty} \sum_{k \in E} \int_0^t Q_{ik}(s) Q_{kj}^{(n-1)}(t-s) ds \end{aligned} \tag{10}$$

$$\tag{11}$$

Ο ημιμαρκοβιανός πίνακας μεταβάσεων ορίζεται ως

$$P_{ij}(t) := \mathbb{P}(Z_t = j | Z_0 = i), i, j \in E. \tag{12}$$

και μπορεί να υπολογιστεί από τη σχέση

$$P(t) = \left((I_N - Q)^{(-1)} \star (I_N - W) \right)(t) = (\Psi \star (I_N - W))(t), \tag{13}$$

όπου $P(t) = (P_{ij}(t))_{i,j \in E}$, $Q(t) = (Q_{ij}(t))_{i,j \in E}$, $\Psi(t) = (\Psi_{ij}(t))_{i,j \in E}$, I_N είναι ο $N \times N$ ταυτοτικός πίνακας. Αποδεικνύεται εύκολα ότι $(I_N - Q)^{(-1)}(t) = \Psi(t)$.

5. ΕΦΑΡΜΟΓΗ ΣΤΟΝ ΑΝΑΛΟΓΙΣΜΟ

Η προτεινόμενη μεθοδολογία έχει μια ευρεία γκάμα εφαρμογών που ξεφεύγουν από τα συνηθισμένα τεχνικά ή γεωφυσικά φαινόμενα, αλλά μπορεί να βρει εφαρμογή στην Οικονομική και Αναλογιστική Επιστήμη. Αυτή η οπτική πλευρά σκιαγραφείται στην ενότητα αυτή με σκοπό να αναδείξει και να αξιολογήσει προβλήματα αξιοπιστίας που συναντώνται στον Αναλογισμό με χρήση του προτεινόμενου ημιμαρκοβιανού μοντέλου. Πιο συγκεκριμένα ο (μηνιαίος) δείκτης ζημίας θα αξιοποιηθεί ως εργαλείο ελέγχου και αξιολόγησης της αποτελεσματικότητας των ασφαλιστικών συμβούλων (insurance advisors) από τη σκοπιά της κερδοφορίας της ασφαλιστικής εταιρείας. Κατ΄ ουσίαν εντάσσοντας στο πλαίσιο των συστημάτων πολλών καταστάσεων (MSS) το δείκτη ζημίας κάθε ασφαλιστικού συμβούλου, καθίσταται δυνατή μέσω των πιθανοτήτων μετάβασης, η πρόβλεψη της κατάστασης στην οποία μπορεί να περιέλθει το

χαρτοφυλάκιο του συμβούλου. Στην παρούσα μελέτη θεωρούμε τρεις καταστάσεις, ήτοι την κερδοφόρο, την τυπική (ούτε κερδοφόρος, ούτε ζημιογόνος) και την ζημιογόνο και δίνουμε ένα εύχρηστο εργαλείο για την έγκαιρη αναδιάρθρωση ασφαλιστικών χαρτοφυλακίων (και κατ΄ επέκταση των συμβούλων που τα κατέχουν) τα οποία είναι ή προβλέπεται (δηλ. εκτιμάται μέσω των πιθανοτήτων μετάβασης) ότι θα καταστούν ζημιογόνα, για την εταιρεία.

Έστω τυχαίο δείγμα 9 ασφαλιστικών συμβούλων, που αντιπροσωπεύει τα $L = 9$ τυχαία μονοπάτια μιας ημιμαρκοβιανής διαδικασίας. Έστω επίσης ότι τρεις (3) εκ των εννέα είναι τυπικοί ασφαλιστικοί σύμβουλοι (δηλαδή σύμβουλοι με χαρτοφυλάκια ούτε κερδοφόρα, ούτε ζημιογόνα), τρεις (3) είναι κερδοφόροι και οι υπόλοιποι τρεις (3) είναι ζημιογόνοι ασφαλιστικοί σύμβουλοι.
Το ενδιαφέρον εστιάζεται στον υπολογισμό του μηνιαίου δείκτη ζημίας (ΜΔΖ) για περίοδο 24 μηνών που ορίζεται ως

$$M\Delta Z = \frac{\text{Μηνιαίο Ύψος Απαιτήσεων}}{\text{Μηνιαίο Ασφάλιστρο}} \tag{14}$$

και θεωρητικά λαμβάνει τιμές από το μηδέν (0) όταν δεν υπάρχουν απαιτήσεις μέχρι μια υψηλή θετική (πεπερασμένη) τιμή που όμως θεωρητικά μπορεί να τεθεί ίση με το άπειρο. Για σκοπούς εφαρμογής της προτεινόμενης μεθοδολογίας, οι χρόνοι μεταξύ των μεταβάσεων (δηλ. οι χρόνοι παραμονής σε κάθε μια κατάσταση) θεωρείται ότι ακολουθούν την κατανομή **Weibull** με σταθερή παράμετρο μορφής (**shape**) που εδώ λαμβάνεται ίση με 2 και μεταβαλλόμενη (από κατάσταση σε κατάσταση) άγνωστη παράμετρο θέσης (scale).

Μία στοχευμένη διαμέριση (κατηγοριοποίηση) του συνόλου των πιθανών μηνιαίων δεικτών ζημίας θα δώσει ως αποτέλεσμα τις καταστάσεις του ημιμαρκοβιανού μοντέλου. Στην κατάσταση 1 βρίσκονται όσες τιμές απέχουν πάνω από μία τυπική απόκλιση (τ.α.) αριστερά (κάτω) του μέσου μηνιαίου δείκτη ζημίας (πράσινη -κερδοφόρος κατάσταση). Η κατάσταση 2 (κίτρινη - τυπική κατάσταση) αποτελείται από τις τιμές που βρίσκονται εντός ακτίνας το πολύ μιας τυπικής απόκλισης εκατέρωθεν του μέσου, ενώ η κατάσταση 3 (κόκκινη - ζημιογόνος κατάσταση) αποτελείται από τις τιμές που απέχουν τουλάχιστον μία τυπική απόκλιση δεξιά (άνω) του μέσου. Να σημειωθεί ότι τα τρία αυτά όρια έχουν υπολογιστεί με βάση την απόδοση ενός τυπικού ασφαλιστικού συμβούλου. Η διαμέριση είναι δυνατόν να γίνει όχι με βάση την μια τυπική απόκλιση αλλά τις δύο ή και περισσότερες.

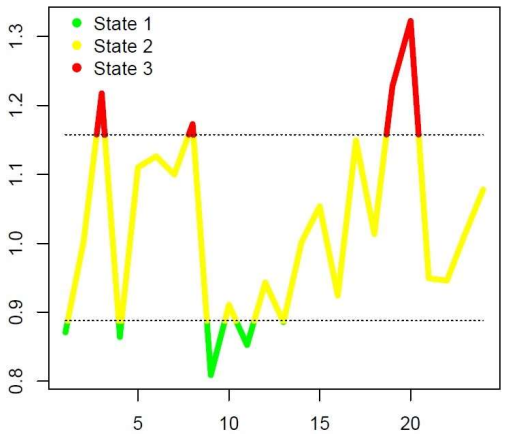
Τα Σχήματα 1 και 2 παρουσιάζουν ένα τυχαίο μονοπάτι της 24-μηνης πορείας του δείκτη ενός τυπικού ασφαλιστικού συμβούλου με όρια υπολογισμένα ώστε να απέχουν μία (Σχήμα 1) ή δύο (Σχήμα 2) αντίστοιχα τυπικές αποκλίσεις εκατέρωθεν του μέσου δείκτη ζημίας. Είναι φανερό ότι η δεύτερη προσέγγιση δεν φαίνεται τόσο δόκιμη μιας και λόγω της μεγάλης απόστασης μεταξύ των ορίων ενδέχεται να μην αποτυπώνουν σωστά την πραγματική εικόνα. Σε οποιαδήποτε περίπτωση, ο ερευνητής μπορεί να επιλέξει τα όρια με όποιον τρόπο θεωρεί ότι αποτυπώνει καλύτερα

το στόχο του ασφαλιστικού οργανισμού. Γενικά τα όρια των τριών καταστάσεων καθορίζονται από τις σχέσεις

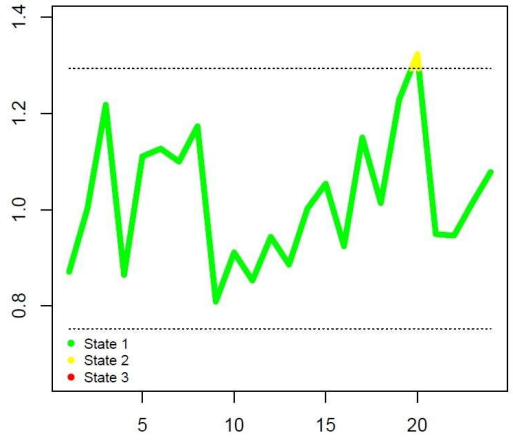
$$1 \pm k \times std \tag{15}$$

όπου std η τυπική απόκλιση των μηναιών δεικτών ζημίας και 1 ο μέσος μηναίος τυπικός δείκτης ζημίας.

Σχήμα 1: ΜΔΖ ενός τυπικού ασφαλιστικού συμβούλου με όρια $\pm 1 \cdot$ τ.α.

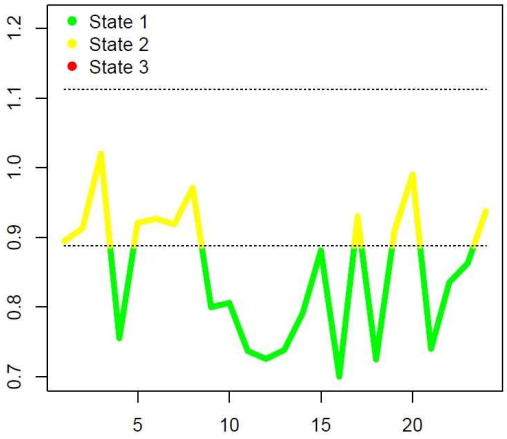


Σχήμα 2: ΜΔΖ ενός τυπικού ασφαλιστικού συμβούλου με όρια $\pm 2 \cdot$ τ.α.

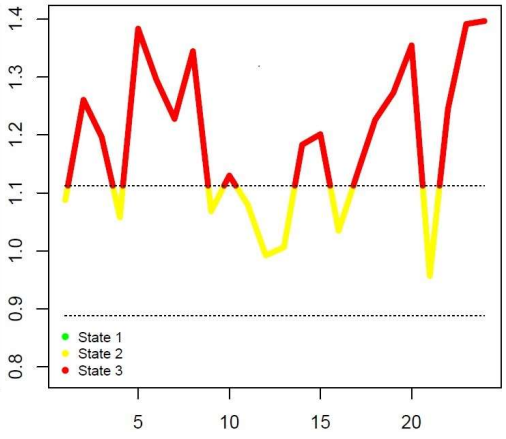


Τυχαία μονοπάτια ενός κερδοφόρου και ενός ζημιογόνου ασφαλιστικού συμβούλου παρουσιάζονται στα Σχήματα 3 και 4 αντίστοιχα με όρια όμως βασισμένα στην απόδοση τυπικού ασφαλιστικού συμβούλου. Παρατηρείται ότι κατά την 24-μηνη πορεία του κερδοφόρου ασφαλιστικού συμβούλου όχι μόνο δεν έχει μεταβεί ποτέ στην ΄κόκκινη΄ κατάσταση αλλά καταγράφει τιμές μικρότερες της μονάδας καθ΄ όλη σχεδόν την πορεία του. Αντίθετα, η πορεία του ζημιογόνου ασφαλιστικού συμβούλου καταγράφει τιμές μεγαλύτερες της μονάδας σε όλο σχεδόν το 24-μηνο με το χειρότερο σενάριο να παρατηρείται στις πλείστες των περιπτώσεων.

Σχήμα 3: ΜΔΖ ενός κερδοφόρου ασφαλιστικού συμβούλου



Σχήμα 4: ΜΔΖ ενός ζημιογόνου ασφαλιστικού συμβούλου



Στην προσπάθεια προσέγγισης του πιο πάνω σεναρίου στα πλαίσια του προτεινόμενου ημιμαρκοβιανού μοντέλου, θα πρέπει αρχικά να εκτιμηθούν οι τιμές των scale παραμέτρων, a_{ij}, της Weibull κατανομής των χρόνων παραμονής στην κατάσταση i πριν μεταβεί το σύστημα στην κατάσταση j, $i, j = 1, 2, 3, i \neq j$ (βλ. Πίνακα 1) καθώς και οι πιθανότητες μετάβασης, p_{ij}, της εμβαπτισμένης Μαρκοβιανής αλυσίδας (χρησιμοποιώντας την σχέση (4)) όπως φαίνονται στον Πίνακα 2.

Πίνακας 1: Πίνακας εκτιμητριών $\hat{a}_{ij}$

$\hat{a}_{ij}$	1	2	3
1	0	0.8809	2.1005
2	1.5406	0	0.2972
3	1.2267	1.8389	0

Πίνακας 2: Πίνακας εκτιμητριών $\hat{p}_{ij}$

$\hat{p}_{ij}$	1	2	3
1	0	0.2955	0.7045
2	0.8383	0	0.1617
3	0.4002	0.5998	0

Έχοντας εκτιμήσει όλες τις απαραίτητες ποσότητες του ημιμαρκοβιανού υποδείγματος, μπορεί κανείς να προχωρήσει σε εκτίμηση-πρόβλεψη των πιθανοτήτων μετάβασης της ημιμαρκοβιανής διαδικασίας, $P_{ij}(t)$, για t χρονικά βήματα μπροστά. Η πρόβλεψη για χρονική περίοδο ενός μήνα ($t = 1$) υποδηλώνει ότι είναι πιο πιθανό η διαδικασία να παραμείνει στην ίδια κατάσταση από την οποία έχει ξεκινήσει

(Πίνακας 3). Η πρόβλεψη για πάροδο 2 μηνών δηλώνει ότι αν έχει ξεκινήσει από την κατάσταση 1 ή 2 η πιο δημοφιλής είναι η κατάσταση 2, ενώ αν έχει ξεκινήσει από την κατάσταση 3 τότε με πιθανότητα 45.84% είναι πιθανότερο να παραμείνει στην ίδια κατάσταση με δεύτερη εναλλακτική την κατάσταση 2 με αρκετά κοντινή πιθανότητα, της τάξης του 42% (Πίνακας 4).

Πίνακας 3: Πίνακας εκτιμητριών $\hat{P}_{ij}$ για περίοδο 1 μήνα

$\hat{P}_{ij}(t=1)$	1	2	3
1	0.6681	0.3024	0.0295
2	0.1058	0.7312	0.1630
3	0.0402	0.2161	0.7437

Πίνακας 4: Πίνακας εκτιμητριών $\hat{P}_{ij}$ για περίοδο 2 μηνών

$\hat{P}_{ij}(t=2)$	1	2	3
1	0.3828	0.4736	0.1435
2	0.1940	0.5214	0.2845
3	0.1213	0.4203	0.4584

Αν το παράθυρο της πρόβλεψης μεγαλώσει, και συγκεκριμένα γίνει τουλάχιστον 4 μήνες ($t = 4$), η στάσιμη κατανομή φαίνεται να επιτυγχάνεται (Πίνακες 5 και 6).

Πίνακας 5: Πίνακας εκτιμητριών $\hat{P}_{ij}$ για περίοδο 4 μηνών

$\hat{P}_{ij}(t=4)$	1	2	3
1	0.1973	0.4835	0.3192
2	0.1823	0.4801	0.3376
3	0.1729	0.4749	0.3522

Πίνακας 6: Πίνακας εκτιμητριών $\hat{P}_{ij}$ για περίοδο 6 μηνών

$\hat{P}_{ij}(t=6)$	1	2	3
1	0.1377	0.4688	0.3935
2	0.1352	0.4672	0.3976
3	0.1325	0.4649	0.4026

6. ΣΥΜΠΕΡΑΣΜΑΤΑ

Η παρούσα εργασία αξιοποιεί μια ειδική μορφή ημιμαρκοβιανών διαδικασιών και τη θεωρία αξιοπιστίας για να αξιολογήσει την αποτελεσματικότητα ασφαλιστικών συμβούλων ενός ασφαλιστικού οργανισμού με βάση το μηνιαίο δείκτη ζημίας. Πιο συγκεκριμένα, ο δείκτης ζημίας κάθε συμβούλου αντιμετωπίζεται ως ένα μονοπάτι

(path) μιας ημιμαρκοβιανής διαδικασίας με κερδοφόρες καθώς και ζημιογόνες καταστάσεις.

Αν και στην παρούσα εργασία ο δείκτης ζημίας έχει βασιστεί στο καθαρό ασφάλιστρο (συνολικό ύψος απαιτήσεων προς καθαρό ασφάλιστρο) με μέση τιμή του δείκτη, την μονάδα (1), η αναπροσαρμογή και μετάβαση στον εμπορικό δείκτη μπορεί εύκολα να γίνει ενσωματώνοντας στον δείκτη, το περιθώριο κέρδους του ασφαλιστικού οργανισμού. Επίσης είναι προφανές ότι είναι δυνατόν με τη βοήθεια της σχέσης (15) να καθοριστούν όρια με τρόπο ώστε να υπάρχει και μια warning (τέταρτη) κατάσταση ώστε ο σύμβουλος να έχει τη δυνατότητα να προβεί από μόνος του, σε παρεμβάσεις, ώστε να εξυγιάνει και να επαναφέρει στην κεδροφόρο κατάσταση, το χαρτοφυλάκιο που διαχειρίζεται. Επιπρόσθετα, ο ερευνητής μπορεί να παρακολουθεί την εξέλιξη του δείκτη στο χρόνο, ταυτόχρονα με τις εκτιμήσεις των πιθανοτήτων μετάβασης από κατάσταση σε κατάσταση ώστε να εντοπίσει χρονικές στιγμές όπου διακρίνεται τάση υπέρβασης του τυπικού ορίου και με τον τρόπο αυτό η διαδικασία να λειτουργήσει για σκοπούς πρόληψης και τελικά με κατάλληλες ενέργειες να αποφευχθεί η ανεπιθύμητη μετάβαση.

Τέλος η κατανομή του χρόνου κατανομής είναι δυνατόν να είναι οποιαδήποτε κατανομή που ανήκει στη γενικευμένη κλάση κατανομών που ορίσθηκε στην εργασία αυτή. Η κλάση κατανομών περιλαμβάνει τόσο την εκθετική κατανομή, οπότε και αναφερόμαστε σε μαρκοβιανές διαδικασίες όσο και κατανομές με ουρές που είναι είτε ελαφρότερες είτε βαρύτερες της εκθετικής που αφορούν ημιμαρκοβιανές διαδικασίες, όπως για παράδειγμα, η Weibull (που χρησιμοποιήθηκε στην εφαρμογή), η Pareto και άλλες.

Abstract

An innovative approach of (monthly) loss ratio forecasting is developed using a special type of semi-Markov processes. Three levels of loss ratio are considered as the states of a semi-Markov process, and semi-Markov process methodology is employed for estimating transition probabilities of loss ratio levels transit from a predefined level to another one.

REFERENCES

Balasubramanian, K., Beg, M. I. and Bapat, R. B. (1991). On families of distributions closed under extrema, *Sankhya: The Indian Journal of Statistics A*, **53**, 375-388.

Barbu, V. S., Karagrigoriou, A., Makrides, A. (2017). Semi-Markov Modelling for Multi-State Systems, *Meth. & Comput. Appl. Prob.*, **19**, 1011-1028.

Barbu, V. S., Limnios, N. (2008). *Semi-Markov Chains and Hidden Semi-Markov Models toward Applications*, Springer, London.

El-Neveihi, E. and Proschan, F. (1984). Degradable systems: a survey of multistate system theory, *Comm. Statist. Theory Methods*, **13**, 405-432.

Klugman, S. A., Panjer, H. H, and Willmot, G. E. (2019). *Loss Models: From Data to Decisions*, 5th ed., Wiley, New York.

Limnios, N. and Oprişan, G. (2001). *Semi-Markov Processes and Reliability*, Birkhäuser, Boston.

Limnios, N. and Ouhbi, B. (2003). Empirical estimators of reliability and related functions for semi-Markov systems, In: Lindqvist, B. H., Doksum, K. A. (eds), *Mathematical and Statistical Methods in Reliability*, **7**, World Scientific, Singapore, 469-484.

Limnios, N. and Ouhbi, B. (2006). Nonparametric estimation of some important indicators in reliability for semi-Markov processes, *Stat. Methodol.*, **3**, 341-350.

Lisnianski, A., Frenkel, I. and Ding, Y. (2010). *Multi-state System Reliability Analysis and Optimization for Engineers and Industrial Managers*, Springer, London.

Lisnianski, A. and Levitin, G. (2003). *Multi-state System Reliability: Assessment, Optimization and Applications*, World Scientific, Singapore.

Murchland, J. (1975). Fundamental concepts and relations for reliability analysis of Multistate systems, In: Barlow, R. E., Fussell, J. B. and Singpurwalla N. (eds), *Reliability and Fault Tree Analysis: Theoretical and Applied Aspects of System Reliability*, SIAM, Philadelphia, 581-618.

Natvig, B. (1982). Two suggestions of how to define a multistate coherent system, *Adv. in Appl. Probab.*, **14**, 434-455.

Natvig, B. (2011). *Multistate Systems Reliability. Theory with Applications*, Wiley, New York.

Ouhbi, B. and Limnios, N. (1996). Non-parametric estimation for semi-Markov kernels with application to reliability analysis, *Appl. Stoch. Models Data Anal.*, **12**, 209-220.

Ouhbi, B. and Limnios, N. (1999). Non-parametric estimation for semi-Markov processes based on its hazard rate functions, *Statist. Infer. Stoch. Processes*, **2**(2), 151-173.

Parodi, P. (2014). *Pricing in General Insurance*, 1st ed, Chapman and Hall/CRC.

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 34ου Πανελληνίου Συνεδρίου Στατιστικής (2022), σελ.70-82

ΠΑΡΑΓΟΝΤΕΣ ΠΟΥ ΕΠΗΡΕΑΖΟΥΝ ΤΙΣ ΕΚΠΟΜΠΕΣ ΤΟΥ ΔΙΟΞΕΙΔΙΟΥ ΤΟΥ ΑΝΘΡΑΚΑ ΚΑΙ ΤΗΝ ΠΕΡΙΒΑΛΛΟΝΤΙΚΗ ΠΟΛΙΤΙΚΗ ΤΗΣ ΕΥΡΩΠΑΙΚΗΣ ΕΝΩΣΗΣ

Μελπομένη Μασούρα[1], Ειρήνη Νικολοπούλου[2] Σόνια Μαλεφάκη[3]
[1]Σχολή Θετικών Επιστημών και Τεχνολογίας, Ελληνικό Ανοιχτό Πανεπιστήμιο
std138386@eap.gr
[2]Τμήμα Μαθηματικών, Πανεπιστήμιο Πατρών
nikirene@upatras.gr
[3]Τμήμα Μηχανολόγων & Αεροναυπηγών Μηχανικών, Πανεπιστήμιο Πατρών
smalefaki@upatras.gr

ΠΕΡΙΛΗΨΗ

Είναι κοινά αποδεκτό, ότι οι ανθρώπινες δραστηριότητες (βιομηχανία, κτηνοτροφία, γεωργία κ.α.) τα τελευταία χρόνια έχουν ως αποτέλεσμα την υποβάθμιση του περιβάλλοντος, η οποία πλέον μοιάζει απειλητική για το μέλλον του πλανήτη και τη βιοποικιλότητά του. Ένα από τα σημαντικότερα περιβαλλοντικά προβλήματα της εποχής μας είναι το φαινόμενο του θερμοκηπίου καθώς διαταράσσονται οι θερμοκρασιακές συνθήκες που επικρατούν στην επιφάνεια της γης με σοβαρό αντίκτυπο στη χλωρίδα και στην πανίδα της. Τα αέρια του θερμοκηπίου είναι περίπου 20, αλλά σύμφωνα με τη Διακυβερνητική Επιτροπή για την Κλιματική Αλλαγή, αποτελούνται κατά κύριο λόγο (περίπου 76.7%) από διοξείδιο του άνθρακα (CO_2). Η Ευρωπαϊκή Ένωση (ΕΕ) αναγνωρίζοντας από νωρίς το πρόβλημα προχώρησε σε διεθνείς νομοθετικές πράξεις (πρωτόκολλα) για τον περιορισμό των κλιματικών μεταβολών. Η παρούσα εργασία εξετάζει τις σχέσεις που συνδέουν την οικονομική δραστηριότητα των χωρών μελών της ΕΕ με το περιβαλλοντικό τους αποτύπωμα το χρονικό διάστημα 1995-2018 καθώς και την επίδραση των πρωτοκόλλων του Κιότο (1997) και της Ντόχα (2012) σε αυτό. Σκοπός μας είναι η αποτίμηση της επίδρασης του κατά κεφαλήν ΑΕΠ, της γεωργικής προστιθέμενης αξίας, της έκτασης καλλιεργήσιμης γης και της κατανάλωσης ανανεώσιμων και μη πηγών ενέργειας στις κατά κεφαλήν εκπομπές CO_2. Μέσω μικτών μοντέλων επιδράσεων (Mixed Effect Models) επιβεβαιώνεται η επίδραση του πρωτοκόλλου της Ντόχα στις εκπομπές CO_2, ενώ παρουσιάζεται αρνητική επίδραση του κατά κεφαλήν ΑΕΠ, της κατανάλωσης μη ανανεώσιμων πηγών ενέργειας και της προστιθέμενης γεωργικής αξίας στις εκπομπές CO_2. Αντίθετα, η επίδραση της κατανάλωσης ανανεώσιμων πηγών ενέργειας και της καλλιεργήσιμης γης στις εκπομπές CO_2 είναι θετική. Επίσης επιβεβαιώνεται η περιβαλλοντική καμπύλη του Kuznet στο σύνολο δεδομένων μας.

Λέξεις Κλειδιά: εκπομπές CO_2, μοντέλα μικτών επιδράσεων, περιβαλλοντική καμπύλη Kuznet

1. ΕΙΣΑΓΩΓΗ

Ένα από τα σημαντικότερα σύγχρονα προβλήματα είναι η υποβάθμιση του περιβάλλοντος που είναι ιδιαίτερα έντονη τις τελευταίες δεκαετίες και οφείλεται σε πολλούς παράγοντες. Ένας από τους βασικότερους είναι η υψηλή συγκέντρωση CO_2 στην ατμόσφαιρα η οποία έχει αυξηθεί από την αρχή της βιομηχανικής επανάστασης έως σήμερα κατά 35%. Αυτή η αύξηση οφείλεται κατά κύριο λόγο στη χρήση των ορυκτών καυσίμων και στην αποψίλωση των δασών είτε για να δημιουργηθούν νέες εκτάσεις για γεωργικές καλλιέργειες, είτε για την εκμετάλλευση της ξυλείας, είτε λόγω των καταστροφικών πυρκαγιών. Ακολουθούν η παραγωγή ενέργειας από μη ανανεώσιμες πηγές (καύση λιγνίτη), η βιομηχανική δραστηριότητα και οι μεταφορές (IPCC 2013).

Η Ευρωπαϊκή Ένωση (ΕΕ) εντόπισε από πολύ νωρίς το πρόβλημα και ξεκινώντας από τη δεκαετία του 70 διαμόρφωσε σταδιακά την πολιτική της για το περιβάλλον. Προχώρησε στην ανάπτυξη και εφαρμογή περιβαλλοντικών προγραμμάτων καθώς και σε διεθνείς νομοθετικές πράξεις (πρωτόκολλα) με σημαντικότερους σταθμούς τα πρωτόκολλα του Κιότο (1997) και της Ντόχα (2012). Το πρωτόκολλο του Κιότο υπογράφηκε το 1997 ενώ τέθηκε σε ισχύ το 2005 και μέχρι σήμερα έχει υπογραφεί από 192 κράτη (UNFCCC 1997). Περιλαμβάνει τις δεσμεύσεις που έχουν αναλάβει οι εκβιομηχανισμένες χώρες για τον περιορισμό των εκπομπών ορισμένων αερίων με στόχο τη μείωση τους τουλάχιστον κατά 5% την πενταετία 2008-2012 σε σύγκριση με τα επίπεδα του 1990. Τα υπογράφοντα κράτη-μέλη καλούνται να εξασφαλίσουν ότι οι εκπομπές για έξι (6) συνολικά αέρια, δεν θα υπερβούν τα όρια που τίθενται από το πρωτόκολλο ενώ θα υφίστανται κυρώσεις σε περίπτωση μη επίτευξης των στόχων για τους οποίους δεσμευτήκαν. Το πρωτόκολλο της Ντόχα υπογράφηκε το 2012 και αποτελεί μια τροποποίηση του πρωτοκόλλου του Κιότο (United Nations 2012). Σύμφωνα με το πρωτόκολλο της Ντόχα, θεσπίστηκε μια δεύτερη περίοδος ανάληψης υποχρεώσεων, από την 1η Ιανουαρίου 2012 έως την 31η Δεκεμβρίου 2020 και αφορά στη μείωση των εκπομπών αερίων του θερμοκηπίου για την περίοδο 2013-2020 κατά 20% σε σχέση με τα επίπεδα του 1990.

Η εφαρμογή περιβαλλοντικών πολιτικών οδήγησε στην ευαισθητοποίηση κράτους και πολιτών στα περιβαλλοντικά προβλήματα, με αποτέλεσμα τα τελευταία χρόνια να γίνεται μία συστηματική προσπάθεια αποσύνδεσης της οικονομικής ανάπτυξης των χωρών και της περιβαλλοντικής υποβάθμισης τους. Για τη μελέτη της περιβαλλοντικής υποβάθμισης μιας χώρας σε σχέση με την αντίστοιχη οικονομική της ανάπτυξης συχνά χρησιμοποιείται το μοντέλο της καμπύλης Kuznets (Kuznet 1955), ελαφρά τροποποιημένο και προσαρμοσμένο στα περιβαλλοντικά δεδομένα (Grossman and Krueger 1991). Η καμπύλη προτείνει και πάλι μια σχέση μορφής ανεστραμμένου U μεταξύ της περιβαλλοντικής υποβάθμισης και της οικονομικής ανάπτυξης της χώρας. Η περιβαλλοντική καμπύλη Kuznets, διερευνά τη σχέση μεταξύ του βαθμού οικονομικής ανάπτυξης μιας χώρας – που μπορεί να εκτιμηθεί μέσω του κατά κεφαλήν ΑΕΠ της (ή του συνολικού ΑΕΠ της) και της αντίστοιχης περιβαλλοντικής υποβάθμισής της.

Εικόνα 1.1. *Περιβαλλοντική καμπύλη του Kuznet*

Αρχικά όσο μεγαλώνει το ΑΕΠ αυξάνει και η περιβαλλοντική υποβάθμιση, αλλά από ένα σημείο και μετά, με αύξηση του κατά κεφαλήν ΑΕΠ παρατηρείται μείωση της περιβαλλοντικής υποβάθμισης κάτι που υποδηλώνει μια στροφή των πιο αναπτυγμένων οικονομικά κοινωνιών, σε ένα πιο οικολογικό και «βιώσιμο» τρόπο ζωής (Εικόνα 1.1).

Οι Waheed et al. (2017) μελέτησαν την επίδραση της δασικής έκτασης, της κατά κεφαλήν γεωργικής προστιθέμενης αξίας και των ανανεώσιμων πηγών ενέργειας στις εκπομπές CO_2 στο Πακιστάν την περίοδο 1990-2014. Εφαρμόζοντας οικονομετρικές μεθόδους (Autoregressive distributed lag models) κατέληξαν πως μακροπρόθεσμα υπάρχει στατιστικά σημαντική θετική επίδραση των ανανεώσιμων πηγών ενέργειας και της δασικής έκτασης στις εκπομπές CO_2 με αποτελέσματα οι εκπομπές να μειώνονται. Αντίθετα υπάρχει αρνητική επίδραση της γεωργικής προστιθέμενης αξίας που οδηγεί στην αύξηση των εκπομπών CO_2. Η μελέτη του Pata (2021) αφορά στη σύνδεση των ανανεώσιμων πηγών ενέργειας, της παγκοσμιοποίησης και της γεωργίας με τις εκπομπές CO_2 και το οικολογικό αποτύπωμα των χωρών Βραζιλία-Ρωσία-Ινδία-Κίνα (BRIC) κατά την περίοδο 1971-2016. Διαπίστωσε πως για τη Ρωσία και την Ινδία οι ανανεώσιμες πηγές ενέργειας φαίνεται να μην έχουν επίδραση στις εκπομπές CO_2. Τέλος, κατέληξε πως η γεωργία είναι σημαντικός παράγοντας για τη ρύπανση του περιβάλλοντος, ωστόσο η πιο αποδοτική χρήση των γεωργικών εκτάσεων με χρήση τεχνολογιών φιλικότερων προς το περιβάλλον οδηγούν στη μείωση των εκπομπών CO_2.

Οι Mahmood et. al (2019), μελέτησαν τη σχέση της ανάπτυξης γεωργικών δραστηριοτήτων και εκπομπών CO_2 στη Σαουδική Αραβία την περίοδο 1971-2014. Διαπίστωσαν πως η αγροτική προστιθέμενη αξία και οι ανανεώσιμες πηγές ενέργειας επηρεάζουν θετικά, μειώνοντας τις κατά κεφαλήν εκπομπές CO_2. Επίσης στην εργασία τους επιβεβαιώνουν την περιβαλλοντική καμπύλη του Kuznet (Environmental Kuznet Curve-EKC) καθώς στο μοντέλο, το πρόσημο των συντελεστών του ΑΕΠ και του $ΑΕΠ^2$ είναι θετικό και αρνητικό αντίστοιχα. Στην εργασία των Aydoğan and Vardar (2019), μελετήθηκαν παράγοντες που επηρεάζουν τις κατά κεφαλήν εκπομπές CO_2 σε

7 χώρες (Βραζιλία, Κίνα, Ινδία, Ινδονησία, Μεξικό, Ρωσία και Τουρκία) την περίοδο 1990-2014. Διαπίστωσαν πως η αγροτική προστιθέμενη αξία, το ΑΕΠ και η κατανάλωση μη ανανεώσιμων πηγών ενέργειας επηρεάζουν αρνητικά, αυξάνοντας τις κατά κεφαλήν εκπομπές CO_2 σε αντίθεση με την κατανάλωση ανανεώσιμων πηγών ενέργειας που τις επηρεάζουν θετικά. Ομοίως και σε αυτή τη μελέτη επιβεβαιώνεται η περιβαλλοντική καμπύλη του Kuznet (EKC) για τις παραπάνω χώρες. Στην παρούσα εργασία μελετάμε την επίδραση που έχουν στις κατά κεφαλήν εκπομπές CO_2, τα πρωτόκολλα του Κιότο (1997) και τις Ντόχα (2012), το κατά κεφαλήν ΑΕΠ, η γεωργική προστιθέμενη αξία, η έκταση καλλιεργήσιμης γης και η κατανάλωση ανανεώσιμων και μη πηγών ενέργειας.

Η λοιπή εργασία έχει οργανωθεί ως εξής: Στη δεύτερη ενότητα γίνεται μία σύντομη παρουσίαση των δεδομένων μας ενώ στην τρίτη ενότητα μελετάμε την εξέλιξη στο χρόνο των κατά κεφαλήν εκπομπών διοξειδίου του άνθρακα. Στην τέταρτη ενότητα γίνεται διερεύνηση της επίδρασης της υπογραφής των πρωτοκόλλων του Κιότο και της Ντόχα καθώς επίσης και παραγόντων όπως το κατά κεφαλήν Ακαθάριστο Εγχώριο Προϊόν (ΑΕΠ), η γεωργική προστιθέμενη αξία, η έκταση καλλιεργήσιμης γης, η κατανάλωση ανανεώσιμων και μη πηγών ενέργειας στις κατά κεφαλήν εκπομπές CO_2. Στην τελευταία ενότητα παρουσιάζονται τα σημαντικότερα συμπεράσματα της παρούσας εργασίας και επισημαίνονται θέματα για περαιτέρω έρευνα.

2. ΠΕΡΙΓΡΑΦΗ ΔΕΔΟΜΕΝΩΝ

Με στόχο τον προσδιορισμό των παραγόντων που επηρεάζουν τις εκπομπές CO_2 στην ΕΕ για το διάστημα 1995-2018, χρησιμοποιήθηκαν δεδομένα που προέρχονται από την Ευρωπαϊκή Στατιστική Υπηρεσία (Eurostat) και την Παγκόσμια Τράπεζα Ανοιχτών Δεδομένων (Word Data Bank). Τα δεδομένα αφορούν 28 χώρες που ανήκαν στην ΕΕ το 2018 ανεξάρτητα από το πότε εισήλθαν σε αυτήν (Αυστρία, Βέλγιο, Βουλγαρία, Κύπρο, Τσεχία, Γερμανία, Δανία, Εσθονία, Ελλάδα, Ισπανία, Φιλανδία, Γαλλία, Κροατία, Ουγγαρία, Ιρλανδία, Ιταλία, Λιθουανία, Λουξεμβούργο, Λετονία, Μάλτα, Ολλανδία, Πολωνία, Πορτογαλία, Ρουμανία, Σουηδία, Σλοβενία, Σλοβακία και Ηνωμένο Βασίλειο). Το σύνολο των δεδομένων αποτελείται 28 (χώρες) x 24 (έτη) = 672 παρατηρήσεις.

2.1 Κατά κεφαλήν εκπομπές διοξειδίου του άνθρακα

Ως εκπομπές διοξειδίου του άνθρακα (CO_2) θεωρούμε αυτές που προέρχονται από την καύση ορυκτών καυσίμων και την παραγωγή τσιμέντου ενώ εξαιρούνται οι εκπομπές από τη χρήση γης, όπως η αποψίλωση των δασών. Επίσης, περιλαμβάνουν το διοξείδιο του άνθρακα που παράγεται κατά την κατανάλωση στερεών, υγρών και αερίων καυσίμων. Στις παραπάνω μετρήσεις δεν συμπεριλαμβάνεται το διοξείδιο του άνθρακα που παράγεται κατά τις μετακινήσεις πλοίων και αεροσκαφών για διεθνείς μεταφορές λόγω της δυσκολίας κατανομής των καυσίμων μεταξύ των δικαιούχων χωρών. Τα δεδομένα λήφθηκαν από την Παγκόσμια Τράπεζα Ανοιχτών Δεδομένων

(https://databank.worldbank.org/source/environment-social-and-governance-(esg)-data) και ως μονάδα μέτρησης χρησιμοποιούνται οι κατά κεφαλήν τόνοι εκπομπών CO_2 (tons per capita).

2.2 Κατά κεφαλήν ακαθάριστο εγχώριο προϊόν

Το κατά κεφαλήν ακαθάριστο εγχώριο προϊόν (GDP) είναι μια μέτρηση που αποτυπώνει την οικονομική δραστηριότητα μιας χώρας ανά άτομο και υπολογίζεται διαιρώντας το ΑΕΠ της χώρας με τον πληθυσμό της. Είναι το τυπικό μέτρο της προστιθέμενης αξίας που δημιουργείται μέσω της παραγωγής αγαθών και υπηρεσιών σε μια χώρα κατά τη διάρκεια μιας ορισμένης περιόδου. Μετρά επίσης το εισόδημα που αποκτάται από αυτή την παραγωγή ή το συνολικό ποσό που δαπανάται για τελικά αγαθά και υπηρεσίες (μείον τις εισαγωγές). Είναι ένας παγκόσμιος δείκτης για την άνθηση ή ύφεση της οικονομικής δραστηριότητας των χωρών και αποτελεί ένα από τα δημοφιλέστερα κριτήρια για το χαρακτηρισμό μιας χώρας ως οικονομικά ανεπτυγμένης. Οι μικρές, πλούσιες χώρες και οι πιο ανεπτυγμένες βιομηχανικά χώρες τείνουν να έχουν το υψηλότερο κατά κεφαλήν ΑΕΠ. Τα δεδομένα λήφθηκαν από την Παγκόσμια Τράπεζα Ανοιχτών Δεδομένων (https://ec.europa.eu/eurostat /databrowser/view/NAMA_10_GDP__custom_5092506/default/table?lang=en) και ως μονάδα μέτρησης χρησιμοποιούνται τα εκατομμύρια ευρώ ανά κάτοικο (million per capita).

2.3 Γεωργική προστιθέμενη αξία

Η γεωργική προστιθέμενη αξία (Agriculture) εκφράζει το ποσοστό του ΑΕΠ που οφείλεται σε τρεις (3) βασικές παραγωγικές δραστηριότητες βάσει της Διεθνούς Πρότυπης Βιομηχανικής Ταξινόμησης Όλων των Οικονομικών Δραστηριοτήτων (International Standard Industrial Classification of All Economic Activities - ISIC).

Η πρώτη δραστηριότητα περιλαμβάνει δύο βασικούς τομείς, την παραγωγή γεωργικών και ζωικών προϊόντων. Καλύπτει επίσης τις μορφές βιολογικής γεωργίας, τις γενετικά τροποποιημένες καλλιέργειες και την εκτροφή γενετικά τροποποιημένων ζώων. Τέλος, περιλαμβάνονται δραστηριότητες συναφείς με τη γεωργική και τη ζωική παραγωγή με στόχο την προετοιμασία προϊόντων για την πρωτογενή αγορά.

Η δεύτερη δραστηριότητα περιλαμβάνει την παραγωγή ξυλείας για μεταποιητικές βιομηχανίες καθώς και τη συλλογή δασικών προϊόντων εκτός ξυλείας (πχ. μανιτάρια, μούρα, καρποί). Εκτός από την παραγωγή ξυλείας, περιλαμβάνει την παραγωγή προϊόντων που υφίστανται ελάχιστη επεξεργασία, όπως καυσόξυλα, κάρβουνο και ροκανίδια που χρησιμοποιούνται σε μη επεξεργασμένη μορφή. Αυτές οι δραστηριότητες μπορούν να πραγματοποιηθούν σε φυσικά ή φυτεμένα δάση.

Η τρίτη δραστηριότητα περιλαμβάνει την αλιεία και την υδατοκαλλιέργεια, και αφορά στη χρήση αλιευτικών πόρων από θαλάσσια, υφάλμυρα ή γλυκά νερά με στόχο τη συλλογή ζωντανών υδρόβιων οργανισμών (κυρίως ψάρια, μαλάκια και μαλακόστρακα) συμπεριλαμβανομένων φυτών από τα ωκεάνια, παράκτια ή εσωτερικά ύδατα για ανθρώπινη κατανάλωση και άλλους σκοπούς. Η συλλογή μπορεί να γίνεται

με το χέρι ή συνηθέστερα από διάφορους τύπους αλιευτικών εργαλείων όπως δίχτυα, πετονιές και σταθερές παγίδες. Τα παραπάνω δεδομένα προέρχονται από την Παγκόσμια Τράπεζα Ανοιχτών Δεδομένων (https://databank.worldbank.org/source/2?series=NV.AGR.TOTL.ZS&country=&l=en).

2.4 Έκταση καλλιεργήσιμης γης

Η έκταση καλλιεργήσιμης γης (Crop) περιλαμβάνει την έκταση παραγωγικής γης που διατίθεται για καλλιέργεια και συγκομιδή ενώ δεν περιλαμβάνει εκτάσεις με νέες φυτείες που δεν είναι ακόμη παραγωγικές. Στην έκταση αυτή συνυπολογίζεται η έκταση πολλαπλής συγκομιδής (συγκομιδή προϊόντων πάνω από μία φορά το χρόνο). Τα δεδομένα προέρχονται από την Ευρωπαϊκή Στατιστική Υπηρεσία (https://ec.europa.eu/eurostat/databrowser/view/APRO_CPNH1_H__custom_2335105/default/table και https://ec.europa.eu/eurostat/databrowser/view/APRO_CPNH1__custom_2335120/default/table) και αφορούν σε εκτάσεις καλλιέργειας/συγκομιδής δημητριακών για την παραγωγή σιτηρών με μονάδα μέτρησης τα 1000 εκτάρια (1000 ha). Στη μελέτη μας χρησιμοποιήθηκε το ποσοστό της έκτασης καλλιεργήσιμης γης επί της συνολικής έκτασης της χώρας, καθώς υπάρχει μεγάλη ανομοιογένεια ως προς το μέγεθος των χωρών.

2.5 Κατανάλωση ανανεώσιμών και μη πηγών ενέργειας

Η κατανάλωση ανανεώσιμων πηγών ενέργειας (RNRG_consumption) αφορά τη συνολική ενέργεια που προέρχεται από ανανεώσιμες πηγές ενέργειας όπως αιολικά πάρκα, φωτοβολταϊκά συστήματα κ.α.. Αντίστοιχα η κατανάλωση μη ανανεώσιμων πηγών ενέργειας (CNRG_consumption) αφορά τη συνολική ενέργεια που προέρχεται από μη ανανεώσιμους φυσικούς πόρους (ορυκτά καύσιμα, φυσικό αέριο). Τα δεδομένα μας προέρχονται από την Παγκόσμια Τράπεζα Ανοιχτών Δεδομένων και την Ευρωπαϊκή Στατιστική Υπηρεσία (https://ec.europa.eu/eurostat/databrowser/view/NRG_BAL_C__custom_2335765/default/table, https://databank.worldbank.org/source/environment -social-and-governance-(esg)-data#). Ως μονάδα μέτρησης χρησιμοποιούνται οι χιλιάδες τόνοι ισοδύναμου πετρελαίου (thousand tons of oil equivalent).

2.6 Υπογραφή πρωτοκόλλων Κιότο και Ντόχα

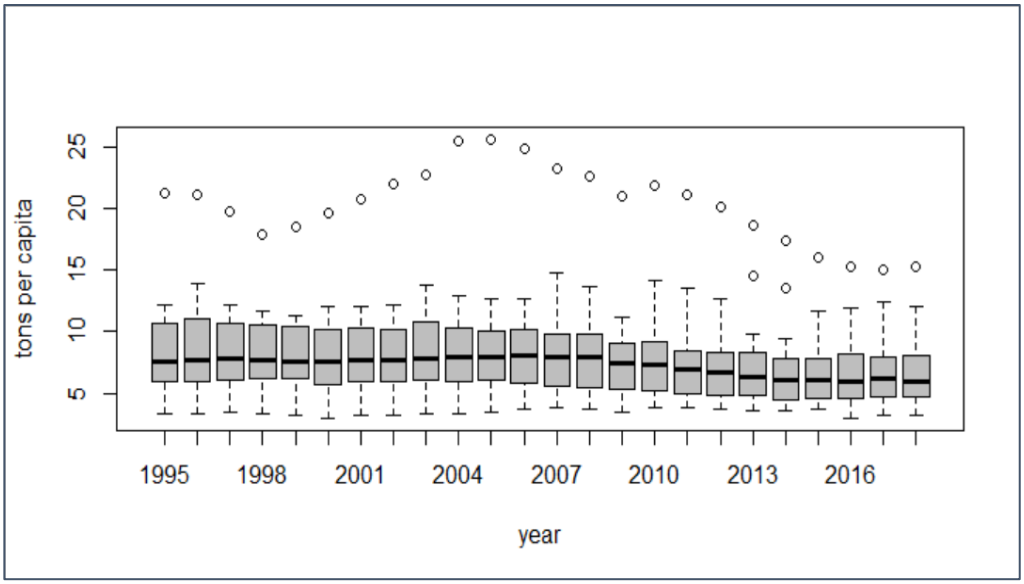
Χρησιμοποιώντας δύο δείκτριες μεταβλητές επισημαίνουμε τις χρονολογίες υπογραφής των πρωτοκόλλων του Κιότο και της Ντόχα 1997 και 2012 αντίστοιχα. Τα πρωτόκολλα αυτά είναι οι σημαντικοί σταθμοί στην Ευρωπαϊκή περιβαλλοντική πολιτική και μας ενδιαφέρει να μελετήσουμε την επίδραση τους στις εκπομπές του διοξειδίου του άνθρακα.

Στο σύνολο των δεδομένων δεν υπάρχουν ελλιπείς τιμές εκτός από τη μεταβλητή έκταση καλλιεργήσιμης γης (Crop) που υπάρχουν 12 ελλιπείς τιμές και αφορούν στη Μάλτα και στην Κροατία για τα έτη 1995-1999, στη Λετονία για το έτος 2010 και στο Ηνωμένο Βασίλειο για το έτος 2015.

3. ΕΞΕΛΙΞΗ ΔΕΙΚΤΗ ΕΚΠΟΜΠΩΝ CO_2 ΩΣ ΠΡΟΣ ΤΟ ΧΡΟΝΟ

Στη συνέχεια θα μελετήσουμε την εξέλιξη των κατά κεφαλήν εκπομπών διοξειδίου του άνθρακα το χρονικό διάστημα 1995 – 2018. Στην Εικόνα 3.1 παρουσιάζονται τα θηκογράμματα των εκπομπών διοξειδίου του άνθρακα ανά έτος. Παρατηρούμε μια ελαφριά ανοδική τάση της διαμέσου των κατά κεφαλήν εκπομπών CO_2 με την πάροδο του χρόνου μέχρι το 2008 ενώ έπειτα παρατηρείται μια ελαφριά πτώση. Επίσης παρατηρούνται εξωκείμενα σημεία (outliers), που αφορούν στις κατά κεφαλήν εκπομπές CO_2 του Λουξεμβούργου, με αποτέλεσμα το Λουξεμβούργο να εξαιρεθεί από την περαιτέρω ανάλυση μας.

Εικόνα 3.1. *Θηκογράμματα των κατά κεφαλήν εκπομπών CO_2 για το χρονικό διάστημα 1995-2018 για τις 28 χώρες που το 2018 άνηκαν στην ΕΕ*



Στον Πίνακα 3.1 παρουσιάζονται για κάθε έτος η μέγιστη (max), η ελάχιστη (min), η μέση τιμή (mean), η διάμεσος (median) καθώς επίσης το πρώτο (Q_1) και το τρίτο (Q_3) τεταρτημόριο, το διατεταρτημοριακό εύρος (IQR) και το εύρος (range) των κατά κεφαλήν εκπομπών CO_2. ανά έτος.

Παρατηρούμε ότι οι μέσοι κατά κεφαλήν τόνοι εκπομπών CO_2 παραμένουν σχεδόν σταθεροί μέχρι 2008, ενώ στη συνέχεια μειώνονται σταθερά μέχρι το 2018 όπου έχουμε τις μικρότερες μέσες κατά κεφαλήν εκπομπές CO_2 (6.5529). Επίσης, παρατηρούμε μια σημαντική μείωση στη μέγιστη τιμή των κατά κεφαλήν τόνων εκπομπών CO_2 από 25.4969 το 2004 σε 15.3302 το 2018, ενώ η ελάχιστη τιμή παραμένει σχεδόν σταθερή κατά την πάροδο των ετών που εξετάζουμε. Παρά το μεγάλο εύρος των τιμών (max-min), παρατηρούμε ότι το κεντρικό 50% των

παρατηρήσεων συγκεντρώνονται σε ένα μικρό εύρος τιμών (το IQR κυμαίνεται από 3.1077 μονάδες το 2013 μέχρι 5.0429 το 1996), υποδηλώνοντας ότι τουλάχιστον οι μισές χώρες έχουν σχετικά σταθερές κατά κεφαλήν εκπομπές διοξειδίου του CO_2.

Πίνακας 3.1. *Περιγραφικά στατιστικά των κατά κεφαλήν εκπομπών CO_2 για το χρονικό διάστημα 1995-2018 για τις 28 χώρες που το 2018 άνηκαν στην ΕΕ*

CO2_PC	min	Q1	median	mean	Q3	max	IQR	range
1995	3.3333	6.0213	7.5114	8.2046	10.6454	21.2420	4.6240	17.9087
1996	3.3662	6.0222	7.7449	8.5170	11.0651	21.1479	5.0429	17.7817
1997	3.5062	6.1521	7.8543	8.3028	10.6283	19.7163	4.4762	16.2101
1998	3.3734	6.1452	7.7510	8.1817	10.5468	17.8714	4.4016	14.4980
1999	3.1584	6.1711	7.5715	8.0250	10.3427	18.4680	4.1716	15.3096
2000	2.9271	5.8593	7.5786	8.0023	10.1205	19.6654	4.2612	16.7383
2001	3.1491	5.9178	7.6694	8.2601	10.3267	20.7010	4.4089	17.5519
2002	3.1816	5.9689	7.6604	8.2558	10.1685	22.0541	4.1996	18.8725
2003	3.2882	6.1330	7.8063	8.6023	10.6314	22.7177	4.4984	19.4295
2004	3.3670	6.0103	7.8857	8.6293	10.1219	25.4969	4.1116	22.1299
2005	3.4483	6.1818	7.9962	8.5358	9.8682	25.6687	3.6864	22.2204
2006	3.6964	5.8168	8.0764	8.5638	10.0866	24.8394	4.2697	21.1429
2007	3.8767	5.6461	7.9688	8.5296	9.7940	23.2087	4.1479	19.3320
2008	3.7248	5.5192	7.9267	8.2617	9.7078	22.5929	4.1887	18.8681
2009	3.4366	5.3241	7.4590	7.5754	8.9886	21.0132	3.6645	17.5766
2010	3.8278	5.2960	7.3522	7.8823	9.0452	21.8166	3.7492	17.9889
2011	3.8355	5.0091	6.8924	7.5486	8.2844	21.0863	3.2753	17.2508
2012	3.7211	4.9882	6.6427	7.2633	8.1869	20.1339	3.1987	16.4127
2013	3.6034	4.9385	6.3475	7.0641	8.0462	18.6801	3.1077	15.0766
2014	3.5868	4.5483	6.0972	6.7164	7.7352	17.3462	3.1870	13.7593
2015	3.6789	4.6014	6.0302	6.6089	7.8276	16.0287	3.2263	12.3498
2016	2.9647	4.6480	5.9830	6.5952	8.1848	15.2230	3.5367	12.2583
2017	3.2479	4.7488	6.1809	6.6462	7.8951	15.0922	3.1464	11.8443
2018	3.1983	4.7139	5.9567	6.5529	8.0770	15.3302	3.3631	12.1319

4. ΠΑΡΑΓΟΝΤΕΣ ΠΟΥ ΕΠΗΡΕΑΖΟΥΝ ΤΙΣ ΕΚΠΟΜΠΕΣ CO_2

Θα θέλαμε να δούμε αν και πως η υπογραφή των πρωτοκόλλων του Κιότο και της Ντόχα σε συνδυασμό με μεταβλητές όπως το κατά κεφαλήν Ακαθάριστο Εγχώριο Προϊόν – ΑΕΠ), η γεωργική προστιθέμενη αξία, η έκταση της καλλιεργήσιμης γης, η κατανάλωση ανανεώσιμων και μη πηγών ενέργειας επηρεάζουν τις κατά κεφαλήν εκπομπές διοξειδίου του άνθρακα των χωρών μελών της ΕΕ την περίοδο 1995 – 2018. Επίσης μελετάμε την εγκυρότητα της περιβαλλοντικής καμπύλης του Kuznet στο συγκεκριμένο σύνολο δεδομένων.

Χρησιμοποιώντας τα δεδομένα που περιγράψαμε στην Ενότητα 2 και το πακέτο `lme4` της `R` (Bates, et al. 2014), κατασκευάζουμε μοντέλα μεικτών επιδράσεων (Mixed Effect Models), τα οποία θα περιγράφουν την επίδραση των παραπάνω παραγόντων στις κατά κεφαλήν εκπομπές CO_2 (CO_2_PC). Αρχικά κατασκευάσαμε το μοντέλο 1, με τυχαίους σταθερούς όρους ως προς τις χώρες και σταθερές κλίσεις ως προς το χρόνο, το οποίο δίνεται από τη σχέση 1.

$$y_i = X_i\beta + Z_i b_i + \varepsilon_i, i = 1, \ldots, 27 \quad (1)$$
$$b_i \sim N(0, {\sigma_b}^2), \varepsilon_i \sim N(0, \sigma^2 I)$$

όπου

$$X_1 = \cdots = X_{27} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ \vdots & \vdots \\ 1 & 24 \end{bmatrix}, \qquad Z_1 = \cdots = Z_{27} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

Το δισδιάστατο διάνυσμα σταθερών επιδράσεων β αποτελείται από τη μέση τιμή του σταθερού όρου β_1 και την κοινή κλίση β_2. Τα μονοδιάστατα διανύσματα τυχαίων επιδράσεων $b_i,\ i = 1 \ldots .27$, περιγράφουν μια μετατόπιση του σταθερού όρου για κάθε χώρα. Επειδή υπάρχει κοινή κλίση, αυτές οι μετατοπίσεις διατηρούνται για όλες τα έτη. Στη συνέχεια κατασκευάσαμε το μοντέλο 2, με τυχαίους σταθερούς όρους και τυχαίες κλίσεις ως προς τον χρόνο ανά χώρα. Τα δύο παραπάνω μοντέλα συγκρίθηκαν χρησιμοποιώντας τη συνάρτηση `anova` της R (Chambers and Hastie 1992) και επιλέχθηκε το μοντέλο 2 ως βέλτιστο με βάση το κριτήριο Akaike (AIC) (Burnham and Anderson 2004). Στο μοντέλο αυτό προσθέτουμε τους υπόλοιπους παράγοντες, το ΑΕΠ (GDP_PC), τη γεωργική προστιθέμενη αξία (Agriculture), το ποσοστό της έκτασης καλλιεργήσιμης γης (Crop), τις ανανεώσιμες πηγές ενέργειας (RNRG_consumption), τις μη ανανεώσιμες πηγές ενέργειας (CNRG_consumption) καθώς και τις μεταβλητές Doha και Kyoto, ενώ για τον έλεγχο εγκυρότητας της περιβαλλοντικής καμπύλης του Kuznet, συμπεριλάβαμε και τον παράγοντα κατά κεφαλήν ΑΕΠ στο τετράγωνο (GDP_PC2) (μοντέλο 3).

Με την εφαρμογή του μοντέλου 3, παρατηρούμε πως ο παράγοντας Kyoto δεν είναι στατιστικά σημαντικός σε αντίθεση με τον παράγοντα Doha. Αυτό δείχνει πως η υπογραφή του πρωτοκόλλου της Ντόχα (2012) φαίνεται να αποδίδει άμεσα καρπούς προς την κατεύθυνση της μείωσης των κατά κεφαλήν εκπομπών CO_2. Ωστόσο, μέρος των καρπών αυτών μπορεί να αποδοθεί και στο πρωτόκολλο του Κιότο καθώς απαιτείται χρόνος για να υλοποιηθούν και να αποδώσουν οι δράσεις που έχουν ξεκινήσει με την υπογραφή του το 1997 και την ενεργοποίηση του το 2005. Παραλείποντας τον παράγοντα Kyoto οδηγούμαστε στο τελικό μοντέλο (μοντέλο 4) (Πίνακας 4.1).

Στο μοντέλο 4 (Πίνακας 4.1) παρατηρούμε πως όλοι οι παράγοντες είναι στατιστικά σημαντικοί. Οι μέσες κατά κεφαλήν εκπομπές διοξειδίου του άνθρακαμειώνονται

κατά 0.11 ανά έτος για σταθερές τιμές των μεταβλητών κατά κεφαλήν ΑΕΠ, γεωργική προστιθέμενη αξία, έκταση καλλιεργήσιμης γης και κατανάλωσης ανανεώσιμων και μη πηγών ενέργειας. Επίσης, μειώνονται κατά 0.69 για κάθε μονάδα αύξησης του ποσοστού της έκτασης της καλλιεργήσιμης γης και κατά 0.85 για κάθε μονάδα αύξησης της κατανάλωσης ανανεώσιμων πηγών ενέργειας. Αντίθετα, διαπιστώνουμε αύξηση κατά 0.18 για κάθε μονάδα αύξησης της γεωργικής προστιθέμενης αξίας και κατά 2.88 για κάθε μονάδα αύξησης του κατά κεφαλήν ΑΕΠ.

Πίνακας 4.1. *Μοντέλο 4: Τελικό μοντέλο*

Model 4	Est	S.E.	t val.	p	Random Effect Std. Dev
Intercept	8.86	0.64	13.81	0.00	0.46
Year	-0.11	0.02	-5.05	0.00	
GDP_PC	2.88	0.32	8.88	0.00	
GDP_PC2	-1.34	0.20	-6.67	0.00	
Agriculture	0.18	0.07	2.69	0.00	
Crop	-0.69	0.18	-3.76	0.00	
RNRG_consumption	-0.85	0.15	-5.55	0.00	
CNRG_consumption	1.38	0.28	4.87	0.00	
Doha	-0.29	0.08	-3.67	0.00	

Επιπλέον, επιβεβαιώνεται και η περιβαλλοντική καμπύλη του Kuznets (EKC) σύμφωνα με την οποία με αύξηση του κατά κεφαλήν ΑΕΠ αυξάνονται οι κατά κεφαλήν εκπομπές διοξειδίου του άνθρακα μέχρι κάποιο όριο και στη συνέχεια παρατηρείται μείωση των κατά κεφαλήν εκπομπών CO_2. Κάτι τέτοιο υποδηλώνει μια στροφή των πιο αναπτυγμένων οικονομικά κοινωνιών, σε ένα πιο οικολογικό και «βιώσιμο» τρόπο ζωής.

5. ΣΥΜΠΕΡΑΣΜΑΤΑ

Στην παρούσα εργασία μελετήθηκε η εξέλιξη των κατά κεφαλήν εκπομπών CO_2 τα έτη 1995-2018 χρησιμοποιώντας δεδομένα από την Παγκόσμια Τράπεζα Ανοιχτών δεδομένων και την Ευρωπαϊκή Στατιστική Υπηρεσία. Επίσης μελετήθηκε η σχέση που συνδέει τις κατά κεφαλήν εκπομπές CO_2 με παράγοντες όπως το κατά κεφαλήν Ακαθάριστο Εγχώριο Προϊόν (ΑΕΠ), η γεωργική προστιθέμενη αξία, το ποσοστό της έκτασης της καλλιεργήσιμης γης, η κατανάλωση ανανεώσιμων και μη πηγών ενέργειας σε συνδυασμό με την υπογραφή των πρωτοκόλλων του Κιότο (1997) και της Ντόχα (2012). Διαπιστώθηκε ότι το κατά κεφαλήν ΑΕΠ, η κατανάλωση μη ανανεώσιμων πηγών ενέργειας και της γεωργικής προστιθέμενης αξίας, επιδρούν αρνητικά αυξάνοντας τις κατά κεφαλήν εκπομπές CO_2. Αντίθετα με την αύξηση του ποσοστού της έκτασης καλλιεργήσιμης γης και την αύξηση χρήσης ανανεώσιμων πηγών ενέργειας οι κατά κεφαλήν εκπομπές CO_2 μειώνονται. Η υπογραφή του πρωτοκόλλου του Κιότο φαίνεται να μην αποδίδει άμεσα καρπούς καθώς απαιτείται χρόνος για την

υλοποίηση των πολυετών δράσεων που έχουν ξεκινήσει με την υπογραφή του. Από την άλλη, η υπογραφή του πρωτοκόλλου της Ντόχα φαίνεται να αποδίδει άμεσα καρπούς, μέρος των οποίων μπορεί ωστόσο να αποδοθεί και στο πρωτόκολλο του Κιότο.

Τέλος, επιβεβαιώνεται η περιβαλλοντική καμπύλη Kuznet (EKC) στο σύνολο των 27 χωρών-μελών της ΕΕ καθώς στο τελικό μοντέλο η σχέση των κατά κεφαλήν εκπομπών CO_2 αναφορικά με το ΑΕΠ και το ΑΕΠ^2 είναι θετική και αρνητική αντίστοιχα. Συνεπώς όσο μεγαλώνει το κατά κεφαλήν ΑΕΠ αυξάνει και η περιβαλλοντική υποβάθμιση, αλλά από ένα σημείο και μετά, με αύξηση του κατά κεφαλήν ΑΕΠ παρατηρείται μείωση της περιβαλλοντικής υποβάθμισης. Αυτή η σχέση υποδηλώνει μια στροφή των πιο αναπτυγμένων οικονομικά χωρών, σε ένα πιο οικολογικό και πιο «βιώσιμο» τρόπο ζωής αποσυνδέοντας την περιβαλλοντική επιβάρυνση από την οικονομική ανάπτυξη.

Τα αποτελέσματα της παρούσας εργασίας έρχονται σε πλήρη συμφωνία και με τα αποτελέσματα των Aydoğan and Vardar (2019) και Waheed et al. (2017) ως προς την αρνητική επίδραση του ΑΕΠ, των μη ανανεώσιμων πηγών ενέργειας και της γεωργικής προστιθέμενης αξίας στις εκπομπές CO_2. Μερικώς συμφωνούν και με τους Mahmood et. al (2019) και Pata (2019) ως προς τον παράγοντα της γεωργικής προστιθέμενης αξίας καθώς οι μελέτες τους καταλήγουν πως η αύξηση του συγκεκριμένου παράγοντα μπορεί να οδηγήσει σε μείωση των εκπομπών διοξειδίου του άνθρακα στην ατμόσφαιρα. Τέλος, σε ότι αφορά στην περιβαλλοντική καμπύλη του Kuznet η παρούσα εργασία έρχεται σε πλήρη συμφωνία με τις μελέτες των Aydoğan and Vardar (2019) και Mahmood et. al (2019) καθώς και οι τρεις μελέτες την επιβεβαιώνουν.

Ευελπιστούμε η παρούσα εργασία να αποτελέσει ερέθισμα για τη συστηματικότερη μελέτη των παραγόντων που επηρεάζουν τις εκπομπές του διοξειδίου του άνθρακα και την περιβαλλοντική πολιτική όχι μόνο σε επίπεδο ΕΕ αλλά και σε παγκόσμιο επίπεδο. Επέκταση της μελέτης θα μπορούσε να γίνει στο σύνολο χωρών εντός και εκτός του Οργανισμού Οικονομικής Συνεργασίας και Ανάπτυξης (ΟΟΣΑ). Για μία διεξοδικότερη μελέτη των εκπομπών CO_2, θα μπορούσαν στο προτεινόμενο μοντέλο να ενσωματωθούν και άλλοι παράγοντες όπως η παγκοσμιοποίηση, οι μετακινήσεις και να μελετηθούν πιο σύνθετα μοντέλα με αλληλεπιδράσεις. Τέλος θα είχε ενδιαφέρον και η ταξινόμηση των χωρών ως προς τις εκπομπές CO_2.

Τα περιβαλλοντικά προβλήματα δε γνωρίζουν σύνορα, αλλά επηρεάζουν όλο τον κόσμο, απαιτώντας ουσιαστικά κοινές δράσεις και συνευθύνη για την αντιμετώπισή τους. Είναι πλέον κοινά αποδεκτό ότι οικονομική ανάπτυξη που δε λαμβάνει υπόψη την προστασία του περιβάλλοντος και τη διατήρηση του οικοσυστήματος δεν είναι αποδεκτή στην Ευρωπαϊκή κοινότητα. Προς την κατεύθυνση αυτή, η ΕΕ λαμβάνει μια σειρά μέτρων και πρωτοβουλιών για τον περιορισμό των κλιματικών μεταβολών με στόχο την επίτευξη της αειφόρου ανάπτυξης.

ABSTRACT

It is commonly accepted that human activities (industry, livestock, agriculture, etc.) in recent years have resulted in the degradation of the environment, which now seems threatening to the future of the planet and its biodiversity. One of the most important environmental problems of our time is the greenhouse effect as the temperature conditions prevailing on the earth's surface are disturbed with a serious impact on its flora and fauna. Greenhouse gases are about 20, but according to the Intergovernmental Panel on Climate Change, they consist mainly (about 76.7%) of carbon dioxide (CO_2). Recognizing the problem early on, the European Union (EU) has adopted international legislation (protocols) to reduce climate change. This paper examines the links between the economic activity of EU Member States and their environmental footprint from 1995 to 2018 and the impact of the Kyoto (1997) and Doha (2012) protocols. Our main aim is to assess the impact of per capita GDP, agricultural added value, percentage of arable land and the consumption of renewable and non-renewable energy sources on per capita CO_2 emissions. Mixed Effect Models confirm the impact of the Doha protocol, renewable energy consumption and arable land in reducing CO_2 emissions. On the other hand, the impact of GDP per capita, non-renewable energy consumption and agricultural added value is negative as CO2 emissions are increasing. The Kuznet environmental curve is also confirmed.

ΑΝΑΦΟΡΕΣ

Aydoğan, B, and G Vardar. 2019. "Evaluating the role of renewable energy, economic growth and agriculture on CO2 emission in E7 countries." *International Journal of Sustainable Energy* 335-348.

Bates, D, B Bolker, M Mächler, and S Walker. 2014. "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software.*

Burnham, K P, and D R Anderson. 2004. "Multimodel Inference: Understanding AIC and BIC in Model Selection." *Sociological Methods & Research* 261–304.

Chambers, J M, and T J Hastie. 1992. *Statistical Models in S.* Wadsworth & Brooks/Cole.

Grossman, G, and A B Krueger. 1991. "Environmental Impacts of a North American Free Trade Agreement." *US-Mexico Free Trade Agreement.*

IPCC. 2013. *Climate Change 2013: The physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change.* Cambridge University Press.

Kuznet, S. 1955. "Economic Growth and Income Inequality." *American Economic Review* 45: 1-28.

Mahmood, H, T T Y Alkhateeb, M M Z Al-QahtanI, Z Allam, N Ahmad, and M Furqan. 2019. "Agriculture development and CO2 emissions nexus in Saudi Arabia." *PLOS ONE.*

Pata, U K. 2021. "Linking renewable energy, globalization, agriculture, CO2 emissions and ecological footprint in BRIC countries: A sustainability perspective." *Renewable Energy* 197-208.

UNFCCC. 1997. *Kyoto Protocol to the United Nations frameworkconvention on climate change.* Kyoto: FCCC/CP/1997/L.7/Add.1.

United Nations. 2012. *Doha Amendment to the Kyoto protocol.* Doha: C.N.718.2012.TREATIES-XXVII.7.c.

Waheed, R, D Chang, S Sarwar, and W Chen. 2017. "Forest, agriculture, renewable energy, and CO2 emission." *Journal of Cleaner Production* 4231-4238.

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 34ου Πανελληνίου Συνεδρίου Στατιστικής (2016), σελ.83-98

ΕΝΑ ΜΕΙΚΤΟ SEIHCRDV-UKF ΜΟΝΤΕΛΟ ΓΙΑ ΤΗΝ ΠΡΟΒΛΕΨΗ ΤΟΥ COVID-19. ΕΦΑΡΜΟΓΗ ΣΤΙΣ ΗΜΕΡΗΣΙΕΣ ΚΑΤΑΓΡΑΦΕΣ ΠΑΝΔΗΜΙΑΣ ΣΤΗ ΓΑΛΛΙΑ

Βασίλειος Ε. Παπαγεωργίου[1], Γεώργιος Τσακλίδης[2]
Τμήμα Μαθηματικών, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης, Θεσσαλονίκη 54124, Ελλάδα
vpapageor@math.auth.gr[1]; tsaklidi@math.auth.gr[2]

ΠΕΡΙΛΗΨΗ

Στο παρόν άρθρο, προτείνεται για τη μελέτη της εξέλιξης της πανδημίας COVID-19 ένα διαμερισματικό SEIHCRDV μοντέλο – μια επέκταση του κλασικού μοντέλου SIR – το οποίο λαμβάνει υπόψη τους πληθυσμούς των ευάλωτων (susceptible) και των εκτεθειμένων (exposed) ασθενών, των μολυσματικών (infectious), νοσηλευόμενων σε νοσοκομεία (hospitalized) και μονάδες εντατικής θεραπείας (icu admitted), αναρρωμένων (recovered), νεκρών (deceased) και εμβολιασμένων (vaccinated), σε συνδυασμό με ένα unscented φίλτρο Kalman (UKF), παρέχοντας μια δυναμική εκτίμηση των εξαρτώμενων από το χρόνο παραμέτρων του συστήματος. Εξετάζεται η αξιοπιστία του μοντέλου για μια μακρά περίοδο 265 ημερών, όπου παρατηρούνται δύο σημαντικά κύματα κρουσμάτων, ξεκινώντας τον Ιανουάριο του 2021 που σηματοδοτεί την έναρξη των εμβολιασμών στην Ευρώπη, παρέχοντας ενθαρρυντικές προβλέψεις. Ιδιαίτερη έμφαση δίνεται στον υπολογισμό ενός πιο αντιπροσωπευτικού βασικού αναπαραγωγικού αριθμού (basic reproduction number) R_0 και στη διερεύνηση της ασυμπτωτικής ευστάθειας των ισοζυγίων του μοντέλου συναρτήσει του δείκτη R_0. Η μεθοδολογία που παρουσιάζεται μπορεί εύκολα να εφαρμοστεί σε άλλες επιδημίες, καθώς οι προτεινόμενες καταστάσεις και μεταβάσεις είναι αντιπροσωπευτικές για τις περισσότερες από αυτές.

Λέξεις Κλειδιά: Unscented Kalman Filter, Επιδημιολογία, COVID-19, Διαμερισματικά Μοντέλα, Μοντέλα Χώρου Κατάστασης, Δυναμική Εκτίμηση Παραμέτρων

1. ΕΙΣΑΓΩΓΗ

Ο νέος κορονοϊός SARS-CoV-2 (COVID-19) ξέσπασε στην πόλη Wuhan της Κίνας τον Δεκέμβριο του 2019 (Muralidar et al. 2020). Αυτή η εξαιρετικά μεταδοτική νόσος ανακηρύχτηκε ως πανδημία από τον Παγκόσμιο Οργανισμό Υγείας τον Ιανουάριο του 2020, ενώ ένα χρόνο αργότερα, ο ιός είχε μολύνει περισσότερους από 100 εκατομμύρια ανθρώπους (Coronavirus Research Center of John Hopkins University 2021), παρά τα ποικίλα υγειονομικά μέτρα που θεσπίστηκαν από τις εθνικές κυβερνήσεις.

Η σοβαρότητα της κατάστασης μαζί με την ανάγκη για έγκαιρη πρόληψη, οδήγησαν σε διερεύνηση της φύσης του COVID-19, κατά κύριο λόγο μέσω μαθηματικής μοντελοποίησης. Συνήθως, η εξάπλωση μολυσματικών ασθενειών περιγράφεται με τη χρήση διαμερισματικών μοντέλων, όπου το SIR (Susceptible-Infected-Recovered) αντιπροσωπεύει το πιο ευρέως γνωστό μοντέλο (Brauer et al. 2019). Ως αποτέλεσμα, πολλά άρθρα βασίζουν την εξερεύνηση της δυναμικής του COVID στο μοντέλο SIR (Cooper I. et al 2020) ή σε ορισμένες από τις επεκτάσεις αυτού, όπως το SIRS (Susceptible-Infected-Recovered-Susceptible) (Salman A.M. et al 2021), τα μοντέλα SEIR (susceptible-exposed-infectious-recovered) (He et al. 2020) ή SEIRD (Susceptible-Exposed-Infectious-Recovered-Deceased) (Rajagopal K. et al 2020).

Οι ανωτέρω προσεγγίσεις δεν είναι σε θέση να περιγράψουν τη σύνθετη δυναμική της πανδημίας, ειδικά για μεγάλα χρονικά διαστήματα, λόγω διαφόρων διακυμάνσεων στις παραμέτρους της νόσου. Για παράδειγμα, η θέσπιση περιοριστικών μέτρων όπως η χρήση μασκών σε κλειστούς και δημόσιους χώρους

ή η θέσπιση των lockdowns σε πολλές χώρες ανά τον κόσμο, μείωσε τα καθημερινά ποσοστά μολύνσεων (Atalan 2020), ενώ η εμφάνιση παραλλαγών όπως η B.1.1.7 (άλφα) και η B.1.617.2 (δέλτα) σχετίζονται με αυξημένη μεταδοτικότητα (Kidd et al. 2021), αυξημένο κίνδυνο νοσηλείας και εισαγωγής στη ΜΕΘ, καθώς και αυξημένη θνησιμότητα (Tuite et al. 2021, Veneti et al. al. 2021. Δεδομένου ότι οι ασυμπτωματικές μολύνσεις είναι δύσκολο να ανιχνευθούν ενώ πάντα είναι υπαρκτές οι περιπτώσεις ψευδώς θετικών-αρνητικών PCR και rapid τεστ, αντιλαμβανόμαστε την ύπαρξη πιθανών αβεβαιοτήτων στα αναφερόμενα δεδομένα (Hu et al. 2020, Keeling et al. 2020). Επιπλέον, οι καταστάσεις και οι αντίστοιχες μεταβάσεις της πλειονότητας των διαμερισματικών μοντέλων δεν λαμβάνουν υπόψη την πλήρη δυναμική της νόσου. Επομένως, η μετάβαση από τη ντετερμινιστική σε μια στοχαστική προσέγγιση διαφαίνεται απαραίτητη.

Οι Singh et al. (2021) χρησιμοποιούν το βασικό φίλτρο Kalman για να λάβουν εκτιμήσεις για την εξέλιξη της πανδημίας στην Ινδία, αλλά όπως αναφέρουν, αυτές οι εκτιμήσεις είναι αξιόπιστες μόνο για σύντομο χρονικό διάστημα. Οι Ndanguza et al. (2016) συνδυάζουν ένα μοντέλο SEIR με ένα εκτεταμένο φίλτρο Kalman με δυναμική εκτίμηση παραμέτρων του επιδημιολογικού μοντέλου. Οι Zhu et al. (2021) προτείνουν ένα μοντέλο SEIRD-EKF με δυναμική εκτίμηση παραμέτρων, ενώ οι Song et al. (2021) χρησιμοποιούν το ίδιο μοντέλο, με τη διαφορά ότι η εκτίμηση παραμέτρων του SEIRD πραγματοποιείται μέσω επαναληπτικής μεθόδου βελτιστοποίησης που βασίζεται στη μέγιστη πιθανότητα.

Στο παρόν άρθρο, προτείνουμε ένα νέο μεικτό SEIHCRDV (Susceptible-Exposed-Infectious-Hospitalized-ICU admitted-Recovered-Deceased-Vaccinated) μοντέλο σε συνδυασμό με ένα unscented φίλτρο Kalman (UKF) με δυναμική εκτίμηση παραμέτρων, που μπορεί να αντιμετωπίσει αποτελεσματικά τις έντονες διακυμάνσεις που συνοδεύουν την εξάπλωση του COVID-19, μετά το ξεκίνημα της περιόδου των εμβολιασμών, παρέχοντας μια πολύ πιο αντιπροσωπευτική εικόνα της καθημερινής εξέλιξης της πανδημίας, λόγω του αυξημένου αριθμού κατάλληλων καταστάσεων και μεταβάσεων.

Η αύξηση του αριθμού των πεπλεγμένων διαφορικών εξισώσεων, όπου η πλειονότητα των καταστάσεων (6 από της 8) είναι παρατηρήσιμες, σε συνδυασμό με τη δυναμική εκτίμηση παραμέτρων που προκύπτει μέσω της ανατροφοδότησης από τις ημερήσιες καταγραφές, αποτρέπει ακραίες εκτιμήσεις παραμέτρων, δίνοντας αξιόπιστες προβλέψεις ακόμη και για τις κρυφές καταστάσεις του μοντέλου. Η συμπερίληψη των καταστάσεων των νοσηλευόμενων σε νοσοκομεία και ΜΕΘ παρέχει ένα ακόμη σημαντικό πλεονέκτημα στην παρούσα ανάλυση. Ο COVID-19 χαρακτηρίζεται από υψηλό ποσοστό ασυμπτωματικών φορέων, γεγονός που οδηγεί στο συμπέρασμα πως τα καθημερινά επίπεδα ενεργών κρουσμάτων και αναρρωμένων θα μπορούσαν να θεωρηθούν ως δείκτες χαμηλής αξιοπιστίας για την εξέλιξη της πανδημίας. Αντίθετα, οι πληθυσμοί των νοσηλευόμενων σε νοσοκομεία και ΜΕΘ ελέγχονται διεξοδικά για την επιβεβαίωση της λοίμωξης από COVID, καθιστώντας τις ημερήσιες καταγραφές αυτών των δύο καταστάσεων ως τους πιο ακριβείς δείκτες για την αξιολόγηση της προσαρμοστικής-προγνωστικής ικανότητας του προτεινόμενου μοντέλου. Θα πρέπει να σημειωθεί ότι η προσθήκη των προαναφερθέντων καταστάσεων δεν δημιουργεί σημαντική υπολογιστική επιβάρυνση στο μοντέλο.

Επιπλέον, μέσω μαθηματικής ανάλυσης προτείνεται ένας εναλλακτικός-βελτιωμένος τύπος για τον δείκτη R_0 (βασικός αναπαραγωγικός ρυθμός) με βάση το προτεινόμενο διαμερισματικό μοντέλο, εξάγοντας συμπεράσματα για την εξέλιξη και το μέλλον της πανδημίας. Ακόμη, διερευνάτε η ύπαρξη ενδημικού ισοζυγίου και η ευστάθεια του ισοζυγίου έλλειψης της νόσου από τον πληθυσμό συναρτήσει των τιμών του R_0. Τέλος, ελέγχουμε την αξιοπιστία του μοντέλου στις ημερήσιες καταγραφές στη Γαλλία, καλύπτοντας μια μακρά περίοδο 265 ημερών, παρέχοντας εκτιμήσεις για τους πληθυσμούς των μολυσματικών, νοσηλευόμενων σε νοσοκομεία και ΜΕΘ, αναρρωμένων, αποθανόντων και εμβολιασμένων.

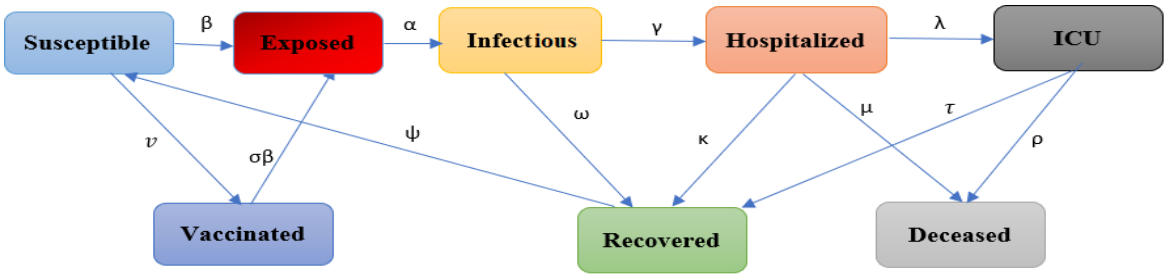
2. ΜΕΘΟΔΟΙ ΚΑΙ ΜΑΘΗΜΑΤΙΚΗ ΜΟΝΤΕΛΟΠΟΙΗΣΗ

2.1. Το προτεινόμενο μοντέλο SEIHCRDV

Στην παρούσα παράγραφο παρουσιάζουμε το διαμερισματικό μοντέλο SEIHCRDV, μέσω του οποίου εξετάζουμε την εξάπλωση του κορονοϊού στη Γαλλία βασιζόμενοι στις ημερήσιες καταγραφές. Όπως αναφέρθηκε προηγουμένως, προτείνεται μια επέκταση του κλασικού μοντέλου SIR, εισάγοντας πέντε

επιπλέον καταστάσεις/διαφορικές εξισώσεις (δ.ε.), προκειμένου να ενσωματώσουμε στο μοντέλο τους πληθυσμούς των εκτεθειμένων, νοσηλευόμενων σε νοσοκομεία, νοσηλευόμενων σε ΜΕΘ, αποθανόντων εξαιτίας του COVID-19 και εμβολιασμένων.

Στο Σχήμα 1, παρουσιάζονται οι μεταβάσεις μεταξύ των καταστάσεων του προτεινόμενου επιδημιολογικού μοντέλου. Η επιλογή των παρουσιαζόμενων καταστάσεων και μεταβάσεων εξετάστηκε με μεγάλη προσοχή, με στόχο την ακριβέστερη περιγραφή της εξάπλωσης της πανδημίας χωρίς να αυξηθεί σημαντικά η πολυπλοκότητα του μοντέλου.



Σχήμα 1. *Διαγραμματική αναπαράσταση του προτεινόμενου SEIHCRDV μοντέλου*

Το σύστημα (2) παρέχει ένα νέο ντετερμινιστικό μοντέλο της μορφής

$$\dot{\boldsymbol{X}}(t) = f(t, \boldsymbol{X}) \tag{1}$$

που περιγράφει την εξέλιξη της εξάπλωσης του COVID-19 στον πληθυσμό. Στον πίνακα 1, περιέχονται οι συμβολισμοί των καταστάσεων και των μεταβάσεων του μοντέλου συνοδευόμενοι από τη διαισθητική τους ερμηνεία.

Πίνακας 1. *Ορισμός παραμέτρων και καταστάσεων του προτεινόμενου SEIHCRDV μοντέλου*

Συμβολισμός	Ορισμός Παραμέτρων/Καταστάσεων
S	Ευάλωτοι
E	Εκτεθειμένοι
I	Μολυσματικοί
H	Νοσηλευόμενοι σε Νοσοκομεία
C	Νοσηλευόμενοι σε ΜΕΘ
R	Αναρρωμένοι
D	Αποθανόντες
V	Εμβολιασμένοι
Λ	Ρυθμός Γεννήσεων
α	Ρυθμός Επώασης
β	Ρυθμός Μόλυνσης
γ	Ρυθμός Εισαγωγών σε Νοσοκομεία
δ	Ρυθμός Θνησιμότητας από άλλα αίτια
λ	Ρυθμός Μετάβασης από Νοσοκομείο σε ΜΕΘ
κ	Ρυθμός Ανάρρωσης μετά από Νοσηλεία σε Νοσοκομείο
μ	Ρυθμός Θνησιμότητας μετά από Νοσηλεία σε Νοσοκομείο
ν	Ρυθμός Πλήρους Εμβολιασμού
ρ	Ρυθμός Θνησιμότητας μετά από Νοσηλεία σε ΜΕΘ
σ	Ρυθμός Μόλυνσης Εμβολιασμένων
τ	Ρυθμός Ανάρρωσης μετά από Νοσηλεία σε ΜΕΘ
ψ	Ρυθμός Μετάβασης από Αναρρωμένος σε Ευάλωτος
ω	Ρυθμός Ανάρρωσης Κρουσμάτων με Ήπια Συμπτώματα

Το σύστημα δ.ε. (2) που ακολουθεί παρουσιάζει αναλυτικά τη δομή και την εξέλιξη του μοντέλου,

$$\begin{aligned}
\frac{dS(t)}{dt} &= \Lambda - \frac{\beta S(t) I(t)}{N} - \nu S(t) + \psi R(t) - \delta S(t) \\
\frac{dE(t)}{dt} &= \frac{\beta S(t) I(t)}{N} + \frac{\sigma \beta V(t) I(t)}{N} - aE(t) - \delta E(t) \\
\frac{dI(t)}{dt} &= aE(t) - \gamma I(t) - \omega I(t) - \delta I(t) \\
\frac{dH(t)}{dt} &= \gamma I(t) - \lambda H(t) - \kappa H(t) - \mu H(t) - \delta H(t) \\
\frac{dC(t)}{dt} &= \lambda H(t) - \tau C(t) - \rho C(t) - \delta C(t) \\
\frac{dR(t)}{dt} &= \omega I(t) + \kappa H(t) + \tau C(t) - \psi R(t) - \delta R(t) \\
\frac{dD(t)}{dt} &= \mu H(t) + \rho C(t) \\
\frac{dV(t)}{dt} &= \nu S(t) - \frac{\sigma \beta V(t) I(t)}{N} - \delta V(t)
\end{aligned} \tag{2}$$

που αξιοποιείται κατά το θεωρητικό μέρος της μελέτης μας. Οι ρυθμοί μεταβάσεων γ, ψ, ω και σ μπορούν να θεωρηθούν σταθεροί, ενώ η ποσότητα $1 - \sigma$ αναπαριστά τη μέση προστασία που παρέχεται από τον πλήρη εμβολιασμό (Papageorgiou και Tsaklidis 2023).

Στο μοντέλο (2), παρατηρούμε ότι δεν υπάρχει μετάβαση μεταξύ μολυσματικών και αποθανόντων. Επιπλέον, περιέχεται μια διαφορική εξίσωση που αντιστοιχεί στον πληθυσμό των πλήρως εμβολιασμένων ατόμων, αναπαριστώντας ένα μοντέλο που ταιριάζει καλύτερα στα μεταγενέστερα στάδια της πανδημίας, δηλαδή από τον Ιανουάριο του 2021 και έπειτα. Η περίοδος μεταξύ της εμφάνισης του COVID-19 (8 Δεκεμβρίου 2019) (Muralidar et al. 2020) και της έναρξης της περιόδου εμβολιασμού, ήταν επαρκής για την κατανόηση των κινδύνων της νόσου και για την ενημέρωση των πολιτών σχετικά με τη σοβαρότητα της κατάσταση. Ως εκ τούτου, θεωρούμε το ποσοστό των κρουσμάτων που απεβίωσαν εξαιτίας του COVID-19 χωρίς να εισαχθούν σε νοσοκομείο ή ΜΕΘ αμελητέο. Ακόμη, η αμελητέα επίδραση αυτής της μετάβασης, θα μπορούσε να θεωρηθεί ως μέρος του γκαουσιανού θορύβου που προστίθεται στις εξισώσεις κατάστασης, οδηγώντας σε ένα στοχαστικό ισοδύναμο του ντετερμινιστικού SEIHCRDV.

Πρόταση 1. Ο βασικός αναπαραγωγικός ρυθμός (BAP) του προτεινόμενου SEIHCRDV μοντέλου είναι

$$R_0 = \frac{\beta\alpha(S^0 + \sigma V^0)}{N(\alpha+\delta)(\gamma+\omega+\delta)} = \frac{\beta\alpha\Lambda(\delta+\sigma\nu)}{N\delta(\alpha+\delta)(\gamma+\omega+\delta)(\nu+\delta)}.$$

Απόδειξη. Η διαδικασία υπολογισμού βασίζεται στη χρήση του πίνακα επόμενης γενιάς (next-generation matrix) που προτείνεται από τους van den Driessche και Watmough (2002) για τον ορισμό του βασικού αναπαραγωγικού ρυθμού R_0. Σημειώνουμε ότι ο συντελεστής R_0 αποτελεί τον αριθμό των δευτερογενών μολύνσεων που προκύπτουν από ένα ήδη μολυσμένο άτομο του πληθυσμού.

Έστω $\boldsymbol{X} = (E, I, H, C)^T$ το διάνυσμα που περιέχει τις 4 καταστάσεις που δηλώνουν την ύπαρξη της νόσου στον ασθενή και $\boldsymbol{Y}^0 = (S^0, 0,0,0,0,0, V^0) = \left(\frac{\Lambda}{\nu+\delta}, 0,0,0,0,0, \frac{\nu\Lambda}{\delta(\nu+\delta)}\right)$, το ισοζύγιο απουσίας της νόσου από τον πληθυσμό, όπου όλος ο πληθυσμός είναι συγκεντρωμένος στις καταστάσεις ευάλωτος και εμβολιασμένος $\boldsymbol{X}^0 = \mathbf{0}^T$, μιας και δεν θα πρέπει να υπάρχουν περιστατικά μολύνσεων κατά τη διάρκεια περιόδων που ισχύει το ισοζύγιο. Έστω $\mathcal{F}(\boldsymbol{X})$ και $\mathrm{V}(\boldsymbol{X})$ διανύσματα διάστασης 4×1, των οποίων η i συνιστώσα περιέχει το ρυθμό εμφάνισης νέων μολύνσεων που εισέρχονται στην κατάσταση i καθώς και το ρυθμό εξόδου ενός ατόμου από την κατάσταση i για μη μολυσματικούς λόγους, αντιστοίχως. Οι μετακινήσεις μεταξύ των

καταστάσεων του $\boldsymbol{X}$ δεν αποτελούν νέες μολύνσεις, παρά μετακινήσεις ενός ήδη μολυσμένου ατόμου εντός του συστήματος. Έτσι, $\dot{\boldsymbol{X}}^T = \mathcal{F}(\boldsymbol{X}) - \mathrm{V}(\boldsymbol{X})$, όπου

$$\mathcal{F}(\boldsymbol{X}) = \left(\frac{\beta SI}{N} + \frac{\sigma\beta VI}{N}, 0,0,0\right)^T$$

και

$$\mathrm{V}(\boldsymbol{X}) = ((a+\delta)E, (\gamma+\omega+\delta)I - aE, (\lambda+\kappa+\mu+\delta)H - \gamma I, (\tau+\rho)C - \lambda H)^T.$$

Ο πίνακας επόμενης γενιάς ορίζεται ως το γινόμενο $\boldsymbol{FV}^{-1}$ όπου $\boldsymbol{F}, \boldsymbol{V}$ είναι οι Ιακωβιανοί πίνακες των $\mathcal{F}(\boldsymbol{X}), \mathrm{V}(\boldsymbol{X})$ υπολογισμένοι στο ισοζύγιο $\boldsymbol{Y}^0$. Έτσι, οι $\boldsymbol{F}$ και $\boldsymbol{V}$ είναι πίνακες διάστασης 4×4 και μορφής

$$\boldsymbol{F}_{|Y^0} = \begin{pmatrix} 0 & \frac{\beta S^0 + \sigma\beta V^0}{N} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \tag{3}$$

και

$$\boldsymbol{V}_{|Y^0} = \begin{pmatrix} \delta + a & 0 & 0 & 0 \\ -a & (\gamma+\omega+\delta) & 0 & 0 \\ 0 & -\gamma & (\lambda+\kappa+\mu+\delta) & 0 \\ 0 & 0 & -\lambda & (\tau+\rho+\delta) \end{pmatrix}, \tag{4}$$

αντίστοιχα, ενώ ο $\boldsymbol{V}$ είναι ένας αντιστρέψιμος κάτω τριγωνικός πίνακας. Έτσι έχουμε

$$\boldsymbol{V}^{-1} = \begin{pmatrix} \frac{1}{\alpha+\delta} & 0 & 0 & 0 \\ \frac{\alpha}{(\alpha+\delta)(\gamma+\omega+\delta)} & \frac{1}{\gamma+\omega+\delta} & 0 & 0 \\ \frac{\alpha\gamma}{(\alpha+\delta)(\kappa+\lambda+\mu+\delta)(\gamma+\omega+\delta)} & \frac{\gamma}{(\kappa+\lambda+\mu+\delta)(\gamma+\omega+\delta)} & \frac{1}{\kappa+\lambda+\mu+\delta} & 0 \\ \frac{\alpha\gamma\lambda}{(\alpha+\delta)(\kappa+\lambda+\mu+\delta)(\gamma+\omega+\delta)(\tau+\rho+\delta)} & \frac{\gamma\lambda}{(\kappa+\lambda+\mu+\delta)(\gamma+\omega+\delta)(\tau+\rho+\delta)} & \frac{\lambda}{(\kappa+\lambda+\mu+\delta)(\tau+\rho+\delta)} & \frac{1}{(\tau+\rho+\delta)} \end{pmatrix},$$

ενώ ο δείκτης R_0 είναι η φασματική ακτίνα του πίνακα επόμενης γενιάς $\boldsymbol{FV}^{-1}$, και δίνεται ως

$$R_0 = \rho(\boldsymbol{FV}^{-1}) = \frac{\beta\alpha(S^0 + \sigma V^0)}{N(\alpha+\delta)(\gamma+\omega+\delta)} \tag{5}$$

όπου με $\rho(\boldsymbol{A})$ συμβολίζουμε τη φασματική ακτίνα του πίνακα $\boldsymbol{A}$.

Αντικαθιστώντας τις τιμές του ισοζυγίου $\boldsymbol{Y}^0$ στη σχέση (5) έχουμε

$$R_0 = \frac{\alpha\beta\Lambda(\delta+\sigma\nu)}{N\delta(\alpha+\delta)(\gamma+\omega+\delta)(\nu+\delta)}. \tag{6}$$

Με παρόμοιο τρόπο, μπορούμε να υπολογίσουμε τον (BAP) για το σύστημα (2) αφαιρώντας την επίδραση εξωγενών παραγόντων, όπως γεννήσεις και θνησιμότητα από άλλα αίτια, δηλαδή

$$R_0 = \frac{\beta(S^0 + \sigma V^0)}{N(\gamma+\omega)}. \tag{7}$$

Ένα δυναμικό ισοδύναμο του (BAP) R_0 που μπορεί να υπολογιστεί δυναμικά κατά τη διάρκεια της πανδημίας είναι ο αποτελεσματικός αναπαραγωγικός ρυθμός R_t, που περιγράφεται από τον τύπο

$$R_t = \frac{\beta_t(S_t + \sigma V_t)}{N(\gamma_t+\omega)}. \tag{8}$$

Πρόταση 2. Για το προτεινόμενο μοντέλο SEIHCRDV υπάρχει μοναδικό, μη-τετριμμένο ενδημικό ισοζύγιο όταν $R_0 > 1$.

Απόδειξη. Έστω $\boldsymbol{Y}^* = (S^*, E^*, I^*, H^*, C^*, R^*, D^*, V^*)^T$ το ενδημικό ισοζύγιο του μοντέλου SEIHCRDV, που προκύπτει εξισώνοντας όλες τις δ.ε. του συστήματος με το 0. Μπορούμε να περιορίσουμε την ανάλυση στις 7 εξισώσεις, εφόσον το $D(t)$ μπορεί να περιγραφεί ως γραμμικός συνδυασμός των άλλων 7 καταστάσεων. Ερευνούμε την ύπαρξη και τη μοναδικότητα του ισοζυγίου αυτού συναρτήσει των τιμών του βασικού αναπαραγωγικού ρυθμού R_0.

Για λόγους συντομίας, θέτουμε $\beta_1 = \frac{\beta}{N} > 0$. Τότε, από το σύστημα (2) έχουμε,

$$\begin{gathered}
\Lambda - \beta_1 S^* I^* - \upsilon S^* + \psi R^* - \delta S^* = 0 \\
\beta_1 S^* I^* + \sigma \beta_1 V^* I^* - a E^* - \delta E^* = 0 \\
\alpha E^* - \gamma I^* - \omega I^* - \delta I^* = 0 \\
\gamma I^* - \lambda H^* - \kappa H^* - \mu H^* - \delta H^* = 0 \\
\lambda H^* - \tau C^* - \rho C^* - \delta C^* = 0 \\
\omega I^* + \kappa H^* + \tau C^* - \psi R^* - \delta R^* = 0 \\
\upsilon S^* - \sigma \beta_1 V^* I^* - \delta V^* = 0 \,.
\end{gathered}$$

Λαμβάνοντας υπόψη το παραπάνω μη γραμμικό σύστημα, στοχεύουμε στο να περιγράψουμε τις καταστάσεις του συστήματος συναρτήσει της ποσότητας I^*, καταλήγοντας στις ακόλουθες σχέσεις:

$$E^* = \frac{\gamma + \omega + \delta}{\alpha} I^* \tag{9}$$

$$H^* = \frac{\gamma}{\lambda + \kappa + \mu + \delta} I^* \tag{10}$$

$$C^* = \frac{\lambda \gamma}{(\lambda + \kappa + \mu + \delta)(\tau + \rho + \delta)} I^* \tag{11}$$

$$R^* = \frac{\omega(\lambda + \kappa + \mu + \delta)(\tau + \rho + \delta) + \kappa\gamma(\tau + \rho + \delta) + \tau\lambda\gamma}{(\psi + \delta)(\lambda + \kappa + \mu + \delta)(\tau + \rho + \delta)} I^* = c_1 I^* \tag{12}$$

$$S^* = \frac{\Lambda + \psi c_1 I^*}{\beta_1 I^* + \upsilon + \delta} \tag{13}$$

$$V^* = \frac{\upsilon\Lambda + \upsilon\psi c_1 I^*}{(\sigma\beta_1 I^* + \delta)(\beta_1 I^* + \upsilon + \delta)}. \tag{14}$$

Θα πρέπει να τονιστεί ότι $c_1 \in \left(0, \frac{\gamma+\omega}{\psi+\delta}\right)$. Για το c_1 της σχέσης (12) έχουμε

$$c_1 = \frac{1}{\psi + \delta}\left[\omega + \frac{\gamma\,(\lambda\tau + \kappa(\tau + \rho + \delta))}{(\lambda + \kappa + \mu + \delta)(\tau + \rho + \delta)}\right] < \frac{\gamma + \omega}{\psi + \delta}\ , \tag{15}$$

καθώς $\lambda\tau + \kappa(\tau + \rho + \delta) < (\lambda + \kappa + \mu + \delta)(\tau + \rho + \delta)$.

Αντικαθιστώντας τις σχέσεις (9), (13) και (14) στη 2η εξίσωση του μη γραμμικού συστήματος δ.ε., παίρνουμε την κυβική εξίσωση χωρίς σταθερό όρο,

$$a\beta_1(\Lambda + \psi c_1 I^*)(\sigma\beta_1 I^* + \delta)I^* + \sigma\beta_1 a\upsilon(\Lambda + \psi c_1 I^*)I^* - (a + \delta)(\gamma + \omega + \delta)(\sigma\beta_1 I^* + \delta)(\beta_1 I^* + \delta + \upsilon)I^* = 0\,. \tag{16}$$

Η τετραγωνική εξίσωση που προκύπτει από τη σχέση (16) είναι η

$$\begin{aligned}
&\sigma\beta_1^2[\alpha\psi c_1 - (a + \delta)(\gamma + \omega + \delta)]I^{*2} \\
&\quad + [\alpha\beta_1\big(\Lambda\beta_1\sigma + \psi c_1(\delta + \sigma\upsilon)\big) - \beta_1(\alpha + \delta)(\gamma + \omega + \delta)(\sigma\delta + \delta + \sigma\upsilon)]I^* \\
&\quad + \alpha\beta_1\Lambda(\delta + \sigma\upsilon) - \delta(\alpha + \delta)(\gamma + \omega + \delta)(\delta + \upsilon) = 0\ .
\end{aligned} \tag{17}$$

Κάνοντας χρήση του κανόνα του Vieta, μπορούμε να προσδιορίσουμε το πρόσημο των λύσεων της εξίσωσης (17). Έστω I_1^* και I_2^* οι λύσεις της (17), όπου

$$I_1^* I_2^* = \frac{\alpha\beta_1 \Lambda(\delta+\sigma\nu) - \delta(\alpha+\delta)(\gamma+\omega+\delta)(\delta+\nu)}{\sigma\beta_1^2[\alpha\psi c_1 - (a+\delta)(\gamma+\omega+\delta)]}. \quad (18)$$

Ο παρονομαστής της (18) είναι αυστηρά αρνητικός, εφόσον $\alpha\psi c_1 < a\frac{\psi}{\psi+\delta}(\omega+\gamma) < \alpha(\omega+\gamma) < (\alpha+\delta)(\gamma+\omega+\delta)$. Ακόμη, ο αριθμητής της (18) είναι θετικός όταν

$$a\beta_1\Lambda(\delta+\sigma\nu) - \delta(\alpha+\delta)(\gamma+\omega+\delta)(\delta+\nu) > 0\,,$$

οπότε, με βάση την σχέση (6) έχουμε $R_0 > 1$. Επομένως, αν $R_0 > 1$, καταλήγουμε σε ετερόσημες ρίζες οδηγώντας σε μοναδικό ενδημικό ισοζύγιο, μιας και η αρνητική ρίζα θα πρέπει να απορριφθεί καθώς δεν είναι δυνατόν να έχουμε αρνητικό αριθμό κρουσμάτων. Μέσω των εξισώσεων (9) – (14) υπολογίζονται οι λοιπές ποσότητες του ισοζυγίου.

Πρόταση 3. Το ισοζύγιο απουσίας της νόσου $\boldsymbol{Y}^0$ είναι τοπικά ασυμπτωτικά ευσταθές όταν $R_0 < 1$, οριακά ευσταθές όταν $R_0 = 1$ και ασταθές όταν $R_0 > 1$.

Απόδειξη. Αρχικά, περιορίζουμε την ανάλυσή μας στις 7 δ.ε., εφόσον το $D(t)$ μπορεί να περιγραφεί ως γραμμικός συνδυασμός των υπόλοιπων 7 καταστάσεων. Γραμμικοποιούμε το σύστημα (2) παίρνοντας τον αντίστοιχο Ιακωβιανό πίνακα υπολογισμένο στο ισοζύγιο απουσίας της νόσου από τον πληθυσμό $\boldsymbol{Y}^0$,

$$\boldsymbol{J}_{DFE} = \begin{pmatrix} -(v+\delta) & 0 & -\beta_1 S^0 & 0 & 0 & \psi & 0 \\ 0 & -(\alpha+\delta) & \beta_1(S^0+\sigma V^0) & 0 & 0 & 0 & 0 \\ 0 & \alpha & -(\gamma+\omega+\delta) & 0 & 0 & 0 & 0 \\ 0 & 0 & \gamma & -(\lambda+\kappa+\mu+\delta) & 0 & 0 & 0 \\ 0 & 0 & 0 & \lambda & -(\tau+\rho+\delta) & 0 & 0 \\ 0 & 0 & \omega & \kappa & \tau & -(\psi+\delta) & 0 \\ v & 0 & -\sigma\beta_1 V^0 & 0 & 0 & 0 & -\delta \end{pmatrix}. \quad (19)$$

Το σύστημα (2) είναι τοπικά ασυμπτωτικά ευσταθές όταν όλες οι ιδιοτιμές του πίνακα $\boldsymbol{J}_{DFE}$ λαμβάνουν χώρα στο αριστερό μιγαδικό ημιεπίπεδο. Το χαρακτηριστικό πολυώνυμο του πίνακα $\boldsymbol{J}_{DFE}$ είναι το

$$\begin{aligned} p(s) &= \det(\boldsymbol{J}_{DFE} - s\boldsymbol{I}_{7x7}) \\ &= (s+\delta)(s+\lambda+\kappa+\mu+\delta)(s+v+\delta)(s+\tau+\rho+\delta)(s+\delta+\psi) \\ &\quad (s^2 + (a+\gamma+\omega+2\delta)s + \delta(\alpha+\gamma+\delta+\omega) + \alpha(\gamma+\omega-\beta_1 S^0 - \sigma\beta_1 V^0) = 0, \quad (20) \end{aligned}$$

οδηγώντας σε 5 αρνητικές ιδιοτιμές, συγκεκριμένα

$$s_1 = -\delta, s_2 = -(\lambda+\kappa+\mu+\delta), s_3 = -(v+\delta), s_4 = -(\tau+\rho+\delta) \text{ και } s_5 = -(\delta+\psi).$$

Για το τετραγωνικό πολυώνυμο της σχέσης (20), εφαρμόζουμε το κριτήριο των Routh-Hurwitz, σύμφωνα με το οποίο οι ρίζες του πολυωνύμου θα βρίσκονται στο αριστερό μιγαδικό ημιεπίπεδο όταν οι συντελεστές $(a+\gamma+\omega+2\delta)$ και $\delta(\alpha+\gamma+\delta+\omega)+\alpha(\gamma+\omega-\beta_1 S^0-\sigma\beta_1 V^0)$ είναι ταυτόχρονα θετικοί. Καθώς $(a+\gamma+\omega+2\delta) > 0$, το κριτήριο των Routh-Hurwitz ικανοποιείται αν και μόνο αν

$$\delta(\alpha+\gamma+\delta+\omega)+\alpha(\gamma+\omega-\beta_1 S^0-\sigma\beta_1 V^0) > 0$$

ή

$$\beta_1 a(S^0+\sigma V^0) < a(\gamma+\omega+\delta)+\delta(\gamma+\omega+\delta)$$

ή

$$\beta_1 a(S^0 + \sigma V^0) < (a + \delta)(\gamma + \omega + \delta)$$

καθιστώντας το σύστημα (2) τοπικά ασυμπτωτικά ευσταθές αν και μόνο αν $R_0 < 1$.

Όταν $R_0 = 1$, το τετραγωνικό πολυώνυμο της σχέσης (21) παίρνει τη μορφή

$$s^2 + (a + \gamma + \omega + 2\delta)s = 0,$$

δίνοντας τις λύσεις

$$s_1 = 0\,, \qquad s_2 = -(a + \gamma + \omega + 2\delta) < 0. \tag{21}$$

Η s_1 βρίσκεται πάνω στον φανταστικό άξονα $y'y$, καθιστώντας το σύστημα οριακά ευσταθές, εφόσον οι υπόλοιπες 6 ιδιοτιμές λαμβάνουν χώρα στο αριστερό μιγαδικό ημιεπίπεδο. Όταν $R_0 > 1$, η τετραγωνική εξίσωση έχει λύσεις με θετικό πραγματικό μέρος καθιστώντας το σύστημα (2) ασταθές, δηλαδή η ύπαρξη έστω και ενός μικρού αριθμού κρουσμάτων μπορεί να οδηγήσει σε ραγδαία εξάπλωση της πανδημίας.

2.2 Unscented φίλτρο Kalman

Το unscented φίλτρο Kalman (UKF) είναι μια παραλλαγή του γραμμικού φίλτρου Kalman που στοχεύει στην μείωση/εξάλειψη των προβλημάτων που προκαλούνται από τις μη γραμμικότητες του συστήματος, χρησιμοποιώντας ένα σύνολο σ-σημείων (sigma points), με στόχο την ακριβή περιγραφή της κατανομής καταστάσεων και παρατηρήσεων. Σκοπός του UKF είναι η αποτελεσματική εξάλειψη του θορύβου που συνοδεύει τις παρατηρήσεις με στόχο να αποκαλυφθεί η πραγματική εξέλιξη τόσο των καταστάσεων όσο των παρατηρήσεων ενός μοντέλου χώρου κατάστασης. Οι σχέσεις (22) και (23) αντιπροσωπεύουν τις εξισώσεις κατάστασης και παρατήρησης, όπου $\boldsymbol{x}_k^T = [S_k \;\; E_k \;\; I_k \;\; H_k \;\; C_k \;\; D_k \;\; V_k]$ είναι το διάνυσμα καταστάσεων του φαινομένου. Η εξίσωση (22) προκύπτει από τη διακριτοποίηση του συστήματος δ.ε. (2) και την προσθήκη γκαουσιανού θορύβου, ενώ η (23) αντιπροσωπεύει την εξίσωση παρατήρησης

$$\boldsymbol{x}_k = f(\boldsymbol{x}_{k-1}) + \boldsymbol{v}_k\,, \tag{22}$$

$$\boldsymbol{y}_k = h(\boldsymbol{x}_k) + \boldsymbol{w}_k\,. \tag{23}$$

Η προσθήκη των θορύβων $\boldsymbol{v}_k$ και $\boldsymbol{w}_k$ στις (22) και (23), σηματοδοτούν τη μετάβαση από την ντετερμινιστική στη στοχαστική σκοπιά. Τα τυχαία διανύσματα $\boldsymbol{v}_k$ και $\boldsymbol{w}_k$ αντιπροσωπεύουν λευκούς θορύβους με μηδενική μέση τιμή και πίνακες διακύμανσης-συνδιακύμανσης $\boldsymbol{Q}$ και $\boldsymbol{R}$, αντίστοιχα.

Αρχικά, θεωρούμε μια αρχικοποίηση του διανύσματος καταστάσεων και του αντίστοιχου πίνακα διακύμανσης-συνδιακύμανσης, $\hat{\boldsymbol{x}}_0 = E[\boldsymbol{x}_0]$ και $\boldsymbol{P}_0 = E[(\boldsymbol{x}_0 - \hat{\boldsymbol{x}}_0)(\boldsymbol{x} - \hat{\boldsymbol{x}}_0)^T]$. Για την εφαρμογή του αλγορίθμου UKF, κατασκευάζεται μια σειρά από σ-σημεία $\boldsymbol{s}_i$ με αντίστοιχα βάρη 1ης τάξης w_i^a, όπου

$$E[\boldsymbol{x_j}] = \sum_{i=0}^{N} w_i^a s_{i,j}\,, \qquad j = 1, \dots, L \tag{24}$$

και βάρη 2ης τάξης w_i^c, με

$$E[\boldsymbol{x_j x_l}] = \sum_{i=0}^{N} w_i^c s_{i,j} s_{i,l}, \qquad j, l = 1, \dots, L. \tag{25}$$

Τα $2L + 1$ σ-σημεία υπολογίζονται ως

$$\boldsymbol{s}_0 = \overline{\boldsymbol{x}}_{k-1|k-1} \tag{26}$$

$$w_0^a = \frac{a^2 - 1}{a^2} \tag{27}$$

$$w_0^c = w_0^a + 1 - a^2 + \beta \tag{28}$$

$$\boldsymbol{s}_i = \bar{\boldsymbol{x}}_{k-1|k-1} + \alpha\sqrt{L}\boldsymbol{A}_i, \quad i = 1, \dots, L \tag{29}$$

$$\boldsymbol{s}_{L+i} = \bar{\boldsymbol{x}}_{k-1|k-1} - \alpha\sqrt{L}\boldsymbol{A}_i, \quad i = 1, \dots, L \tag{30}$$

$$w_i^a = w_i^c = \frac{1}{2a^2L}, \quad i = 1, \dots, 2L \tag{31}$$

(Wan και Van der Merwe 2000). Το διάνυσμα $\boldsymbol{A}_i$ είναι η i στήλη του πίνακα $\boldsymbol{A}$ όπου $\boldsymbol{P}_{k-1|k-1} = \boldsymbol{A}\boldsymbol{A}^T$, ενώ ο $\boldsymbol{A}$ μπορεί να υπολογιστεί μέσω της παραγοντοποίησης Cholesky. Η παράμετρος α καθορίζει τη διασπορά των σ-σημείων γύρω από το μέσο $\bar{\boldsymbol{x}}$ και συνήθως παίρνει τιμές $0 \leq \alpha \leq 1$. Η παράμετρος β χρησιμοποιείται για την περιγραφή της εκ των προτέρων κατανομής του $\boldsymbol{x}$, και στην περίπτωση κανονικής κατανομής η βέλτιστη τιμή είναι η $\beta = 2$ (Wan και Van der Merwe 2000).

Ομοίως, όπως σε άλλες μεθόδους φίλτρου Kalman, το UKF χρησιμοποιεί έναν επαναληπτικό αλγόριθμο πρόβλεψης-διόρθωσης με στόχο να παρέχει τις καλύτερες δυνατές εκτιμήσεις. Κατά τη διαδικασία πρόβλεψης, ο αλγόριθμος εκμεταλλεύεται τα σ-σημεία διαδίδοντάς τα μέσω της μη γραμμικής συνάρτησης $\boldsymbol{x}_i = f(\boldsymbol{s}_i), i = 0, \dots, 2L$ μέσω των οποίων υπολογίζονται η σταθμισμένη μέση τιμή και ο πίνακας διακύμανσης-συνδιακύμανσης για το k-οστό βήμα πρόβλεψης,

$$\bar{\boldsymbol{x}}_{k|k-1} = \sum_{i=0}^{2L} w_i^a \boldsymbol{x}_i \tag{32}$$

$$\boldsymbol{P}_{k|k-1} = \sum_{i=0}^{2L} w_i^c (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{k|k-1})(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{k|k-1})^T + \boldsymbol{Q} \ . \tag{33}$$

Δεδομένων των μέσων τιμών και των πινάκων διακύμανσης-συνδιακύμανσης για το k-οστό βήμα πρόβλεψης, $\bar{\boldsymbol{x}}_{k|k-1}$ και $\boldsymbol{P}_{k|k-1}$, νέα σ-σημεία διαδίδονται μέσω της συνάρτησης παρατήρησης κατά το k-οστό βήμα διόρθωσης μέσω της συνάρτησης $\boldsymbol{y}_i = h(\boldsymbol{s}_i), i = 0, \dots, 2L$. Τότε, η σταθμισμένη μέση τιμή και ο πίνακας διακύμανσης-συνδιακύμανσης των παραγόμενων σημείων $\boldsymbol{y}_i$ υπολογίζεται ως (Sarkka 2007)

$$\bar{\boldsymbol{y}} = \sum_{i=0}^{2L} w_i^a \boldsymbol{y}_i \tag{34}$$

$$\boldsymbol{S}_k = \sum_{i=0}^{2L} w_i^c (\boldsymbol{y}_i - \bar{\boldsymbol{y}})(\boldsymbol{y}_i - \bar{\boldsymbol{y}})^T + \boldsymbol{R}. \tag{35}$$

Οι πίνακες διασυνδιασποράς και κέρδους Kalman μπορούν να υπολογιστούν,

$$\boldsymbol{C}_{sz} = \sum_{i=0}^{2L} w_i^c (\boldsymbol{s}_i - \bar{\boldsymbol{x}}_{k|k-1})(\boldsymbol{y}_i - \bar{\boldsymbol{y}})^T \ , \tag{36}$$

$$\boldsymbol{K}_k = \boldsymbol{C}_{sz}\, \boldsymbol{S}_k^{-1} \ . \tag{37}$$

Τέλος, η μέση τιμή και ο πίνακας διακύμανσης-συνδιακύμανσης για το k-οστό βήμα διόρθωσης, δίνονται ως

$$\bar{\boldsymbol{x}}_{k|k} = \bar{\boldsymbol{x}}_{k|k-1} + \boldsymbol{K}_k(\boldsymbol{y}_k - \bar{\boldsymbol{y}}) \, , \tag{38}$$

$$\boldsymbol{P}_{k|k} = \boldsymbol{P}_{k|k-1} - \boldsymbol{K}_k \boldsymbol{S}_k \boldsymbol{K}_k^T. \tag{39}$$

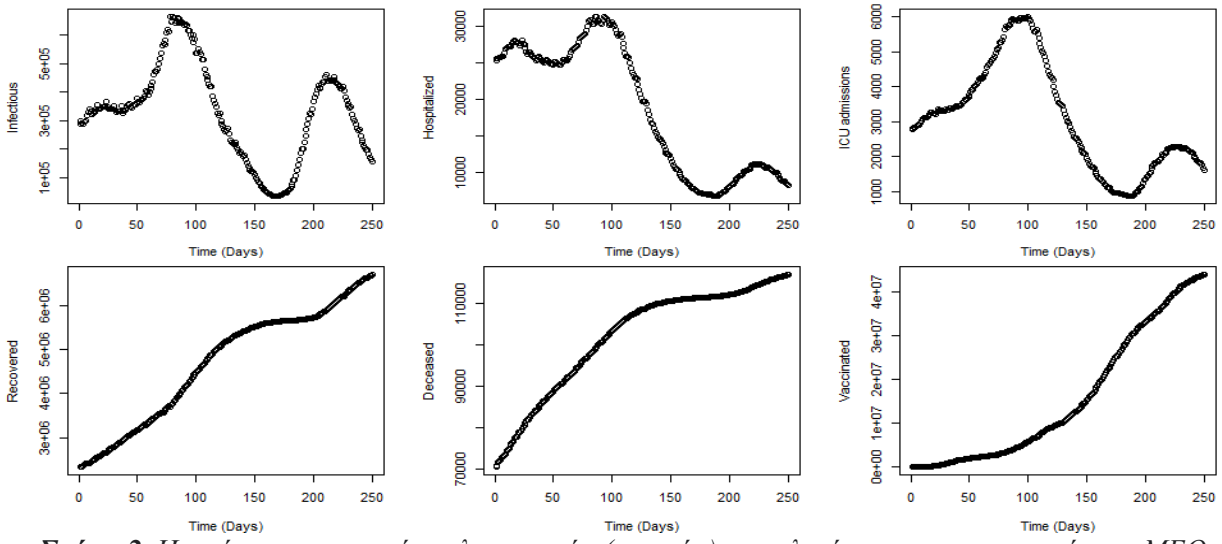
3. ΑΠΟΤΕΛΕΣΜΑΤΑ

Στην παρούσα ενότητα διερευνάται η προσαρμοστική και προγνωστική ικανότητα του μοντέλου SEIHCRDV-UKF με δυναμική εκτίμηση παραμέτρων στις ημερήσιες παρατηρήσεις COVID-19 στη Γαλλία. Τονίζουμε ότι ο πληθυσμός θεωρήθηκε κλειστός, δηλαδή οι ρυθμοί Λ και δ δε λήφθηκαν υπόψιν λόγω της αμελητέας επίδρασης τους. Εξετάζουμε την εξέλιξη της πανδημίας από τις 16 Ιανουαρίου έως τις

7 Οκτωβρίου 2021 (διάστημα 265 ημερών). Τα πλήρως εμβολιασμένα άτομα είτε έχουν λάβει δύο δόσεις από τα εμβόλια Pfizer ή Astra Zeneca ή μία δόση από τα εμβόλια Johnson & Johnson ή Moderna. Τα δεδομένα που χρησιμοποιούνται για την ανάλυση, περιλαμβάνουν καταγραφές για τους ενεργούς μολυσματικούς φορείς, νοσηλευόμενους σε νοσοκομεία και ΜΕΘ, αναρρωμένους, αποθανόντες και εμβολιασμένους. Οι παρατηρήσεις αυτών των 6 χρονοσειρών μπορούν να συλλεγούν μέσω της ιστοσελίδας data.europa.eu. που περιλαμβάνει τις επίσημες, ημερήσιες καταγραφές για τον COVID-19 για όλα τα κράτη-μέλη της Ευρωπαϊκής Ένωσης.

Η εξεταζόμενη χρονική περίοδος περιλαμβάνει δύο κύματα κρουσμάτων που κορυφώνονται περίπου την 80η (5 Απριλίου) και την 210η (15 Αυγούστου) ημέρα ανάλυσης, ενώ συνοδεύονται από αντίστοιχα κύματα νοσηλευόμενων σε νοσοκομείο και ΜΕΘ. Η πτωτική τάση του πρώτου κύματος πιθανότατα επηρεάστηκε έντονα από τα lockdowns που επιβλήθηκαν σε 16 γαλλικές περιφέρειες, ενώ θεσπίστηκε απαγόρευση κυκλοφορίας μεταξύ 7 μ.μ. και 6 π.μ. (20 Μαρτίου) σε εθνικό επίπεδο.

Η εμφάνιση του δεύτερου κύματος περίπου την 1η εβδομάδα του Ιουλίου (ημέρα 170) σχετίζεται σε μεγάλο βαθμό με τη χαλάρωση των περιοριστικών μέτρων, όπως την επαναλειτουργία εστιατορίων/μπαρ/καφέ με 50% πληρότητα (9 Ιουνίου), την άρση της υποχρεωτικότητας στη χρήση μάσκας σε εξωτερικούς χώρους (17 Ιουνίου) και την άρση της νυχτερινής απαγόρευσης κυκλοφορίας (20 Ιουνίου). Ένας άλλος σημαντικός παράγοντας για τη δεύτερη εδραίωση του COVID-19 στη Γαλλία είναι η διασπορά της μετάλλαξης δέλτα στη χώρα, όπου σύμφωνα με τον G. Attal (France 24, 2021), η εξαιρετικά μεταδοτική αυτή μετάλλαξη αντιπροσώπευε το 40% των νέων μολύνσεων από COVID-19.

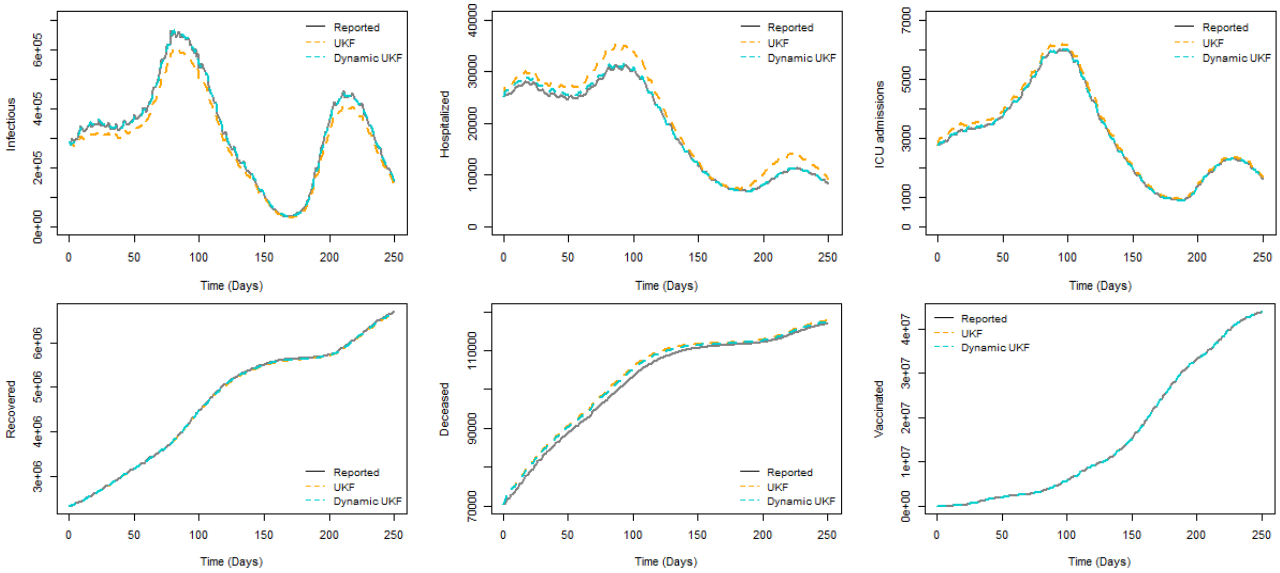


Σχήμα 2. *Ημερήσιες καταγραφές μολυσματικών (ενεργών), νοσηλευόμενων σε νοσοκομεία και ΜΕΘ, αναρρωμένων, αποθανόντων και εμβολιασμένων στη Γαλλία για διάστημα 250 ημερών*

Με βάση το σχήμα 2, παρατηρούμε την εμφάνιση 2 εξίσου ισχυρών κυμάτων μόλυνσης. Ωστόσο, κατά τη διάρκεια του 2ου κύματος τα αντίστοιχα κύματα νοσηλευόμενων σε νοσοκομεία και ΜΕΘ είναι σημαντικά εξασθενημένα συγκριτικά με τα πρώτα. Το φαινόμενο αυτό, τονίζει τη βελτιωμένη αντιμετώπιση του COVID-19 από τους υγειονομικούς, καθώς η εμπειρία που αποκτήθηκε από προηγούμενα κύματα μόλυνσης συνέβαλε σημαντικά στην κατανόηση των χαρακτηριστικών της νόσου, οδηγώντας σε πιο στοχευμένες και αποτελεσματικές τεχνικές θεραπείας. Επιπλέον, ο διαρκώς αυξανόμενος αριθμός των πλήρως εμβολιασμένων ατόμων παίζει καθοριστικό ρόλο στο προαναφερθέν φαινόμενο, καθώς φαίνεται να μειώνει (σχήμα 2) την πιθανότητα σοβαρών λοιμώξεων από τον COVID-19, οδηγώντας σε λιγότερες νοσηλείες σε νοσοκομεία και ΜΕΘ. Παράλληλα, παρατηρούμε αύξηση του συνολικού αριθμού των νεκρών αλλά με έντονα πτωτική τάση, ενώ η καμπύλη των αναρρωμένων παρουσιάζει την ακριβώς αντίθετη συμπεριφορά, υποδηλώνοντας την αποτελεσματικότητα των εμβολιασμών.

Μετά τα πρώτα συμπεράσματα που προέκυψαν από την παρατήρηση του Σχήματος 2, διερευνούμε την αποτελεσματικότητα της προσαρμογής – πρόβλεψης του μοντέλου SEIHCRDV-UKF στις παρατηρήσεις

COVID-19 στη Γαλλία. Επιπλέον, πραγματοποιείται σύγκριση της αποτελεσματικότητας του προτεινόμενου μοντέλου με τον αλγόριθμο SEIHCRDV-UKF με σταθερές παραμέτρους, τον SEIHCRDV-EKF με δυναμική εκτίμηση παραμέτρων, τον SEIRD-EKF με δυναμική εκτίμηση παραμέτρων (Zhu et al. 2021) και το ντετερμινιστικό SEIHCRDV. Στο Σχήμα 3, παρουσιάζεται η προσαρμοστική ικανότητα του SEIHCRDV-UKF με και χωρίς δυναμική εκτίμηση παραμέτρων χρησιμοποιώντας τις χρονοσειρές των μολυσματικών (ενεργών), νοσηλευόμενων σε νοσοκομεία και ΜΕΘ, αναρρωμένων, αποθανόντων και εμβολιασμένων στη Γαλλία. Το μοντέλο με τις σταθερές παραμέτρους εμφανίζει υποεκτίμηση των ενεργών κρουσμάτων ειδικά κατά τη διάρκεια των δύο κυμάτων μόλυνσης (Απρίλιο και Αύγουστο 2021) και υπερεκτίμηση στον αντίστοιχο αριθμό εισαγωγών σε νοσοκομεία και ΜΕΘ, σε σύγκριση με τα αποτελέσματα που προέκυψαν από το δυναμικό μοντέλο UKF – SEIHCRDV. Η διαφορά αυτή οφείλεται στην αδυναμία του μοντέλου αυτού να αναπροσαρμόσει τις παραμέτρους του ώστε κατά τη διάρκεια της πανδημίας να μειωθούν οι ρυθμοί μετάβασης σε νοσοκομεία και ΜΕΘ.



Σχήμα 3. *Αναπαράσταση της προσαρμοστικής ικανότητας των μοντέλων SEIHCRDV-UKF με και χωρίς δυναμική εκτίμηση παραμέτρων στις ημερήσιες παρατηρήσεις στη Γαλλία*

Πίνακας 2. *Τιμές NRMSE του ντετερμινιστικού SEIHCRDV μοντέλου, του UKF-SEIHCRDV με σταθερές παραμέτρους, του EKF-SEIRD με δυναμική εκτίμηση παραμέτρων και των EKF και UKF-SEIHCRDV με δυναμική εκτίμηση παραμέτρων*

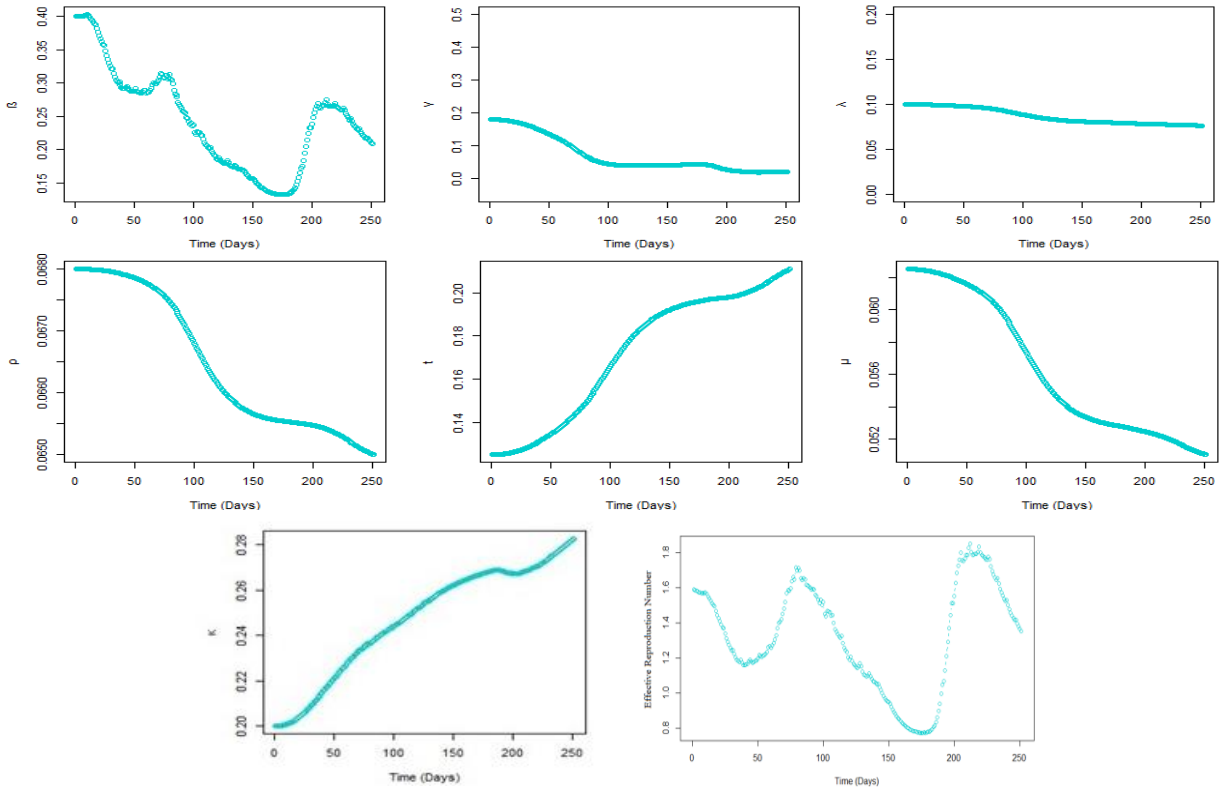
	Infectious	Hospitalized	ICU	Recovered	Deceased	Vaccinated
Deterministic	0.868560	0.960792	0.997148	0.967693	0.998600	0.846578
UKF-SEIHCRDV	0.214439	0.256795	0.098419	0.026397	0.092463	0.015045
Dynamic EKF-SEIRD	0.131866	-	-	0.024213	0.246692	-
Dynamic EKF-SEIHCRDV	0.108814	0.227158	0.057104	0.024754	0.083265	0.015059
Dynamic UKF-SEIHCRDV	**0.087466**	**0.058741**	**0.039427**	**0.020921**	**0.069456**	**0.015074**

Παράλληλα με το σχήμα 3, παρουσιάζεται ο Πίνακας 2 με τις τιμές NRMSE για το ντετερμινιστικό SEIHCRDV και τα τρία προαναφερθέντα στοχαστικά ισοδύναμα των UKF, EKF και UKF με δυναμική εκτίμηση παραμέτρων. Οι παρουσιαζόμενες τιμές NRMSE υπολογίζονται σύμφωνα με τον τύπο που παρουσιάζεται στο (Papageorgiou και Tsaklidis 2021). Τόσο το SEIHCRDV-UKF χωρίς την εκτίμηση

παραμέτρων όσο και τα δυναμικά SEIHCRDV-EKF και το SEIHCRDV-UKF ξεπερνούν σημαντικά την ακρίβεια του ντετερμινιστικού SEIHCRDV.

Επίσης, το προτεινόμενο SEIHCRDV-UKF με δυναμική εκτίμηση παραμέτρων παρέχει την πιο ακριβή περιγραφή της εξέλιξης της πανδημίας στη Γαλλία, σχεδόν και για τις 6 παρατηρήσιμες καταστάσεις. Ειδικότερα, το μοντέλο αυτό παράγει μειωμένες τιμές NRMSE κατά 145,17%, 337,16%, 149,62%, 26,17% και 33,12% σε σχέση με το SEIHCRDV-UKF και 24,41%, 286,71%, 44,83%, 83% και 18% σε σύγκριση με το SEIHCRDV-EKF για τους μολυσματικούς, τους νοσηλευόμενους σε νοσοκομεία και ΜΕΘ, τους αναρρωμένους και τους αποθανόντες αντίστοιχα.

Ένα σημαντικό χαρακτηριστικό που επαληθεύει την αξιοπιστία της προτεινόμενης δυναμικής των μοντέλων UKF – SEIHCRDV είναι η παρακολούθηση των μεταβαλλόμενων παραμέτρων (Σχήμα 4). Πρώτον, ο ρυθμός μόλυνσης β, ακολουθεί την εμφάνιση των δύο κυμάτων μόλυνσης που παρατηρούμε εντός της περιόδου των 250 ημερών, ενώ κατά τη διάρκεια του 2ου κύματος, η αύξηση του ρυθμού β και ο αριθμός των μολυσματικών-ενεργών κρουσμάτων ξεκινούν ταυτόχρονα. Οι παράμετροι μετάβασης γ και λ εμφανίζουν φθίνουσα συμπεριφορά κατά την εξέλιξη της πανδημίας, η οποία επίσης συμφωνεί με τις παρατηρήσιμες χρονοσειρές. Αυτό το φαινόμενο υπογραμμίζει τον μειωμένο κίνδυνο μόλυνσης από COVID, καθώς προχωράμε βαθύτερα στην περίοδο του εμβολιασμού.



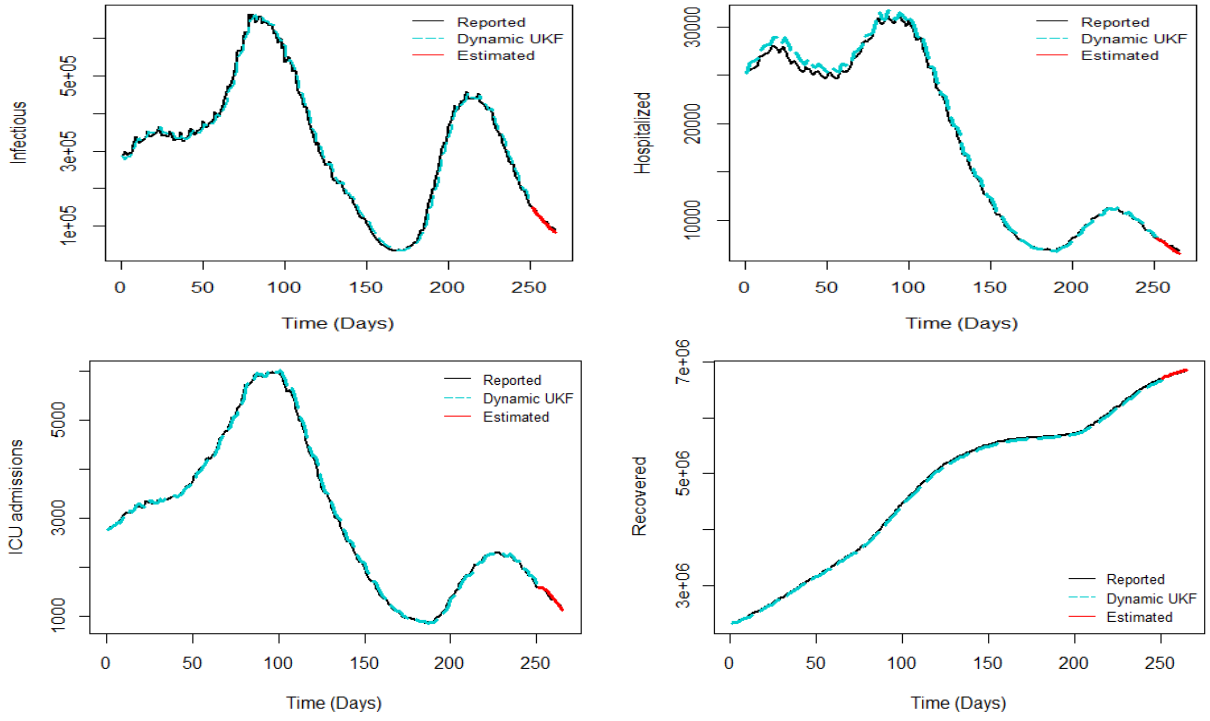
Σχήμα 4. *Η εξέλιξη των 7 μεταβαλλόμενων παραμέτρων που εκτιμώνται μέσω του δυναμικού SEIHCRDV-UKF και ο δείκτης R_t κατά τη διάρκεια 250 ημερών πανδημίας*

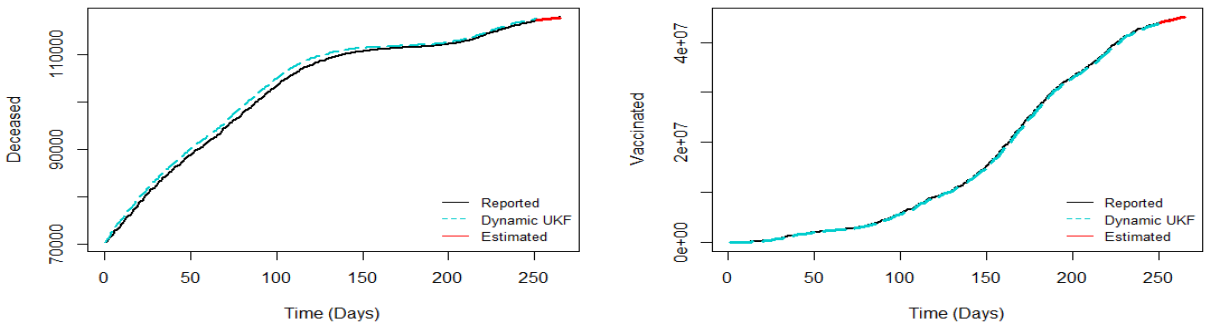
Μια άλλη αξιοσημείωτη παρατήρηση είναι η φθίνουσα τάση των ποσοστών θνησιμότητας μ και ρ κατά τη διάρκεια των 250 ημερών, όπου ο ρυθμός θνησιμότητας των νοσηλευόμενων περιστατικών μ γίνεται χαμηλότερος από τον ρυθμό θνησιμότητας των εισαγωγών σε ΜΕΘ ρ, γεγονός αρκετά αναμενόμενο. Καθώς ο συνολικός αριθμός θανάτων αυξάνεται με φθίνουσα τάση (Σχήμα 2), η πιθανότητα θνησιμότητας των εισαχθέντων σε ΜΕΘ γίνεται συγκριτικά υψηλότερη από την αντίστοιχη θνησιμότητα των περιστατικών που νοσηλεύονται απλά σε κάποια νοσοκομειακή μονάδα. Η συμπεριφορά των ποσοστών ανάρρωσης κ και τ επικυρώνει την αξιοπιστία του μοντέλου μας, καθώς και τα δύο αυξάνονται σημαντικά με την πάροδο του χρόνου, υπογραμμίζοντας τον αντίκτυπο του εμβολιασμού και την προσαρμοστικότητα

του εθνικού συστήματος υγείας της Γαλλίας στις προκλήσεις που θέτει ο νέος ιός. Ως εκ τούτου, το ποσοστό ανάρρωσης κ των εισαχθέντων σε νοσοκομεία παρουσιάζει ανοδική τάση ενώ οι τιμές αυτού είναι διαρκώς σε υψηλότερα επίπεδα κατά τη διάρκεια των 250 ημερών συγκριτικά με το ποσοστό ανάρρωσης των εισαχθέντων σε ΜΕΘ, τ. Στο Σχήμα 4, παρουσιάζουμε την εξέλιξη του αποτελεσματικού αναπαραγωγικού ρυθμού R_t, σύμφωνα με τη σχέση (8) κατά την εξεταζόμενη περίοδο των 250 ημερών. Σύμφωνα με αυτόν τον τύπο και τη δυνατότητα δυναμικού υπολογισμού των παραμέτρων του μοντέλου (2), είμαστε σε θέση να υπολογίζουμε το R_t αρκετά αποτελεσματικά σε καθημερινή βάση παρέχοντας έναν αντιπροσωπευτικό δείκτη της εξάπλωσης του COVID-19.

Τέλος, εξετάζεται η προβλεπτική ακρίβεια του προτεινόμενου μοντέλου σε νέες παρατηρήσεις. Γίνεται χρήση του NRMSE, καθώς για τιμές μικρότερες από τη μονάδα το προτεινόμενο μοντέλο παρέχει καλύτερες εκτιμήσεις σε σύγκριση με το μέσο όρο των χρονοσειρών. Στο σχήμα 5, παρουσιάζονται οι προβλέψεις του δυναμικού SEIHCRDV-UKF για 15 ημέρες μπροστά. Όλες οι τιμές των NRMSE για τις 6 παρατηρήσιμες καταστάσεις παρουσιάζονται ως εξαιρετικά υποσχόμενες – οι αυτές τιμές είναι αρκετά κοντά στο μηδέν – υπογραμμίζοντας την καταλληλότητα του μοντέλου στην περιγραφή και τη μοντελοποίηση πραγματικών επιδημιολογικών δεδομένων.

Πιο συγκεκριμένα, λαμβάνουμε τιμές NRMSE της τάξεως του 0,24756, 0,44588, 0,32585, 0,45927, 0,39823 και 0,18285 για τους ενεργούς μολυσματικούς, νοσηλευόμενους σε νοσοκομεία και ΜΕΘ, αναρρωμένους, αποθανόντες και εμβολιασμένους αντίστοιχα. Η υψηλότερη τιμή NRMSE αντιστοιχεί στην περίπτωση των αναρρωμένων όπου παρατηρούμε μια μικρή υπερεκτίμηση καθώς προχωράμε σε μεταγενέστερες χρονικές στιγμές. Αυτή η υπερεκτίμηση επηρεάζεται από τις ήπιες υποεκτιμήσεις στα περιστατικά των νοσηλευόμενων σε νοσοκομεία/ΜΕΘ, καθώς αυτές οι δύο καταστάσεις συσχετίζονται γραμμικά στο μοντέλο SEIHCRDV (2). Ως εκ τούτου, και οι 6 τιμές NRMSE είναι πολύ μικρότερες της μονάδας, οδηγώντας στο συμπέρασμα ότι το μοντέλο μας μπορεί να χειριστεί αποτελεσματικά την πρόβλεψη μελλοντικών καταστάσεων –ακόμα και για μισό μήνα μπροστά– ειδικά για τους πληθυσμούς των ενεργών κρουσμάτων, νοσηλευόμενων σε νοσοκομεία/ΜΕΘ και αποθανόντων που αποτελούν τις σημαντικότερες μεταβλητές για την αξιολόγηση της σοβαρότητας της πανδημίας.





Σχήμα 5. *Προβλεπτική ικανότητα του SEIHCRDV-UKF με δυναμική εκτίμηση παραμέτρων για μισό μήνα μπροστά*

4. ΣΥΖΗΤΗΣΗ ΚΑΙ ΣΥΜΠΕΡΑΣΜΑΤΑ

Συνήθως, η μαθηματική μοντελοποίηση μεταδοτικών ασθενειών βασίζεται σε ντετερμινιστικές προσεγγίσεις, οι οποίες συχνά χρησιμοποιούν μια επέκταση του τυπικού μοντέλου SIR, προσθέτοντας διαφορικές εξισώσεις για τις περιπτώσεις των εκτεθειμένων και των αποθανόντων. Ωστόσο, αυτή η ντετερμινιστική προσέγγιση αποτυγχάνει να περιγράψει με ακρίβεια τη σύνθετη δυναμική του COVID-19, καθώς οι παράμετροι των μοντέλων μεταβάλλονται έντονα με την εισαγωγή περιοριστικών μέτρων για την υγεία, όπως lockdowns και μάσκες, ή τη συνεχή εμφάνιση νέων μεταλλάξεων. Ακόμη, η ύπαρξη θορύβου στις ημερήσιες καταγραφές έχει τη βάση της στις ασυμπτωματικές μολύνσεις, οι οποίες είναι δύσκολο να ανιχνευθούν, και στα ποσοστά ψευδώς θετικών-αρνητικών rapid και PCR τεστ.

Στο παρόν άρθρο, προτείνεται ένα νέο, μεικτό μοντέλο SEIHCRDV-UKF με δυναμική εκτίμηση παραμέτρων, με στόχο την αντιμετώπιση των προαναφερθέντων περιορισμών των ντετερμινιστικών διαμερισμάτων μοντέλων, ενισχύοντας σημαντικά την προβλεπτική τους ικανότητα. Επεκτείνουμε το τυπικό μοντέλο SIR αυξάνοντας επαρκώς το πλήθος των διαφορικών εξισώσεων του συστήματος σε οκτώ, λαμβάνοντας επιπλέον υπόψη τους πληθυσμούς των εκτεθειμένων, νοσηλευόμενων σε νοσοκομεία και ΜΕΘ, αποθανόντων και εμβολιασμένων. Η συμπερίληψη των νοσηλευόμενων σε νοσοκομεία και ΜΕΘ προσφέρει ένα ακόμη σημαντικό πλεονέκτημα. Στην περίπτωση του COVID-19, που χαρακτηρίζεται από υψηλά ποσοστά ασυμπτωματικών φορέων, οι ημερήσιες μολύνσεις και αναρρώσεις μπορούν να θεωρηθούν ως δείκτες χαμηλής ακρίβειας για την εξέλιξη της πανδημίας. Από την άλλη πλευρά, τα άτομα που εισάγονται σε νοσοκομείο και ΜΕΘ ελέγχονται διεξοδικά για λοίμωξη από COVID, καθιστώντας τις παρατηρήσεις αυτών των δύο καταστάσεων ως τα πιο αξιόπιστα δεδομένα για την επικύρωση της ικανότητας προσαρμογής-πρόβλεψης του προτεινόμενου μοντέλου.

Επιπλέον, η μεθοδολογία του UKF αξιοποιεί τις ημερήσιες παρατηρήσεις διευκολύνοντας την αποτελεσματική εκτίμηση των τιμών των μεταβαλλόμενων παραμέτρων, οδηγώντας σε μια στοχαστική προσέγγιση. Η συμπερίληψη των παραμέτρων του μοντέλου στην διαδικασία «φιλτραρίσματος» βοηθά την παρακολούθηση των μεταβολών της πανδημίας. Τα παρουσιαζόμενα αποτελέσματα της προσαρμογής και της πρόβλεψης της εξάπλωσης της νόσου στη Γαλλία – ακόμη και για μισό μήνα μπροστά – επιβεβαιώνουν την αξιοπιστία του μοντέλου. Αυτή η στοχαστική προσέγγιση είναι απαραίτητη, καθώς τα σφάλματα στις παρατηρήσεις είναι γνωστά και η ενσωμάτωση όλων των πιθανών μεταβάσεων της πανδημίας θα οδηγούσε σε ένα ιδιαίτερα περίπλοκο μοντέλο, καθιστώντας τη διαδικασία υπολογιστικά δαπανηρή και την προβλεπτική απόδοση αναξιόπιστη. Για παράδειγμα, οι μεταβάσεις που χαρακτηρίζονται από αμελητέα ποσοστά, όπως η μετάβαση από τη μόλυνση απευθείας στην εισαγωγή στη ΜΕΘ ή από τη μόλυνση απευθείας στο θάνατο, περιλαμβάνονται στον πρόσθετο θόρυβο του συστήματος. Η διαδικασία αρχικοποίησης παραμέτρων σε μη γραμμικές μεθοδολογίες θα πρέπει να εκτελείται με μεγάλη προσοχή. Η δυναμική εκτίμηση παραμέτρων αντιμετωπίζει αυτή την αναγκαιότητα αποτελεσματικά, ενισχύοντας την αξιοπιστία του μοντέλου μας.

Η θεωρητική ανάλυση που βασίζεται στο προτεινόμενο μοντέλο SEIHCRDV αποκαλύπτει πολύτιμα συμπεράσματα σχετικά με την εξάπλωση της πανδημίας. Οι παραγόμενες σχέσεις για τα R_0 και R_t

παρέχουν μια πιο αντιπροσωπευτική εικόνα για το μέλλον της πανδημίας, ενώ λαμβάνουμε σημαντικά συμπεράσματα για την εξέλιξή της με βάση την τιμή του R_0. Σύμφωνα με προαναφερθείσες έρευνες, ο δείκτης R_0 για τον COVID είναι πολύ υψηλότερος της μονάδας, ενώ η τιμή του μπορεί να μειωθεί μόνο μέσω αποφασιστικών παρεμβάσεων, όπως είναι τα lockdowns. Δεδομένων αυτών –τουλάχιστον για το άμεσο μέλλον– θα πρέπει να δοθεί μεγαλύτερη έμφαση στον περιορισμό των σοβαρών λοιμώξεων και των θανάτων, καθώς ο περιορισμός της μεταδοτικότητας της νόσου φαίνεται αρκετά δύσκολος, υποστηρίζοντας τη σημασία του πλήρους εμβολιασμού στον πληθυσμό.

Η μεθοδολογία που παρουσιάζεται μπορεί εύκολα να εφαρμοστεί σε άλλες επιδημίες, καθώς οι προτεινόμενες καταστάσεις και μεταβάσεις είναι αντιπροσωπευτικές για τις περισσότερες από αυτές. Ταυτόχρονα, αυτή η μεθοδολογία είναι ιδανική σε περιπτώσεις όπου τόσο η δυναμική του συστήματος όσο και οι αντίστοιχες παρατηρήσεις σε πραγματικό χρόνο δεν μπορούν να αποτυπώσουν με ακρίβεια την εξάπλωση της επιδημίας λόγω αβεβαιοτήτων, με στόχο την εξάλειψη του θορύβου που επηρεάζει την εξέλιξη καταστάσεων και παρατηρήσεων.

ABSTRACT

In this paper, for the examination of the evolution of COVID-19 in the population, we propose a SEIHCRDV model – an extension of the classic SIR compartmental model – which also takes into consideration the populations of exposed, hospitalized, admitted in intensive care units (ICU), deceased and vaccinated cases, in combination with an unscented Kalman filter (UKF), providing a dynamic estimation of the time dependent parameters of the system. Apparently, this new consideration could be useful for examining also other pandemics. We examine the reliability of our model over a long period of 265 days, where we observe two major infection waves, starting from the January of 2021 which signified the initialization of vaccinations in Europe, providing quite encouraging predictive performance. Finally, special emphasis is given to the proposal of a representative basic reproductive number R_0 and the investigation of the stability of the model's equilibriums in accordance with the formula produced to estimate R_0. The presented methodology can be easily implemented in other epidemics, as the states and parameters of the model are representative for the majority of them.

ΑΝΑΦΟΡΕΣ

Atalan A. (2020). Is the lockdown important to prevent the COVID-19 pandemic? Effects on psychology, environment and economy-perspective. *Annals of Medicine and Surgery*, **56**, 38-42. https://doi.org/10.1016/j.amsu.2020.06.010

Brauer F., Castillo-Chavez C. and Feng Z. (2019). *Endemic Disease Models. In: Mathematical Models in Epidemiology. Texts in Applied Mathematics 69*, Springer, New York, NY. https://doi.org/10.1007/978-1-4939-9828-9_3

Cooper I., Mondal A. and Antonopoulos C.G. (2020). A SIR model assumption for the spread of COVID-19 in different communities. *Chaos, Solitons and Fractals*, **139**, 110057. https://doi.org/10.1016/j.chaos.2020.110057

France 24 (2021). Highly contagious Delta variant could ruin France's summer, warns government, https://www.france24.com/en/france/20210707-highly-contagious-delta-variant-could-ruin-france-s-summer-warns-government Accessed 7 July 2021.

He S., Peng Y. and Sun K. (2020). SEIR modeling of the COVID-19 and its dynamics. *Nonlinear Dyn*, **101**, 1667-1680. https://doi.org/10.1007/s11071-020-05743-y

Hu Z., Song C., Xu C., Jin G., Chen Y., et al. (2020). Clinical characteristics of 24 asymptomatic infections with COVID-19 screened among close contact in Nanjing, China. *Sci. China Life Sci.*, **63**, 706-711. https://doi.org/10.1007/s11427-020-1661-4

Keeling M.J., Hollingsworth T.D. and Read J.M. (2020). Efficacy of contact tracing for the containment of

the 2019 novel coronavirus (COVID-19). *J. Epidemiol. Community*, **74**(10), 861-866. https://doi.org/10.1136/jech-2020-214051

Kidd M., Richter A., Best A., Cumley N., et al. (2021). S-Variant SARS-CoV-2 Lineage B1.1.7 Is Associated With Significantly Higher Viral Load in Samples Tested by TaqPath Polymerase Chain Reaction. *The Journal of infectious diseases*, **223**(10), 1666–1670. https://doi.org/10.1093/infdis/jiab082

Muralidar S., Ambi S.V., Sekaran S. and Krishnan U.M. (2020). The emergence of COVID-19 as a global pandemic: Understanding the epidemiology, immune response and potential therapeutic targets of SARS-CoV-2. *Biochimie*, **179**, 85-100. https://doi.org/10.1016/j.biochi.2020.09.018

Ndanguza D., Mbalawata I.S. and Nsabimana J.P. (2016). Analysis of SDEs Applied to SEIR Epidemic Models by Extended Kalman Filter Method. *Applied Mathematics*, **7**, 2195-2211. http://dx.doi.org/10.4236/am.2016.717175

Papageorgiou V. and Tsaklidis G. (2021). Modeling of Premature Mortality Rates from Chronic Diseases in Europe, Investigation of Correlations, Clustering and Granger Causality. *Communication in Mathematical Biology and Neuroscience*, **2021**(67). https://doi.org/10.28919/cmbn/5926

Papageorgiou V.E. and Tsaklidis G. (2023). An improved epidemiological-unscented Kalman filter (hybrid SEIHCRDV-UKF) model for the prediction of COVID-19. Application on real-time data. *Chaos, Solitons & Fractals*, **166**, 112914. https://doi.org/10.1016/j.chaos.2022.112914

Rajagopal K., Hasanzadeh N., et al. (2020). A fractional-order model for the novel coronavirus (COVID-19) outbreak. *Nonlinear Dyn*, **101**, 711-718. https://doi.org/10.1007/s11071-020-05757-6

Salman A.M., Ahmed I., et al. (2021). Scenario analysis of COVID-19 transmission dynamics in Malaysia with the possibility of reinfection and limited medical resources scenarios. *Computers in Biology and Medicine* **133**, 104372. https://doi.org/10.1016/j.compbiomed.2021.104372

Sarkka S. (2007). On Unscented Kalman Filtering for State Estimation of Continuous-Time Nonlinear Systems. *IEEE Transactions on Automatic Control*, **52**(9), 1631-1641. https://doi.org/10.1109/TAC.2007.904453

Singh K.K., Kumar S., et al. (2021). Kalman filter based short term prediction model for COVID-19 spread. *Applied Intelligence*, **51**, 2714-2726. https://doi.org/10.1007/s10489-020-01948-1

Song J., Xie H., Gao B., Zhong Y., Gu C. and Choi K.S. (2021). Maximum likelihood-based extended Kalman filter for COVID-19 prediction. *Chaos, Solitons and Fractals*, **146**, 110922. https://doi.org/10.1016/j.chaos.2021.110922

Tuite A.R., Fisman D.N., Odutayo A., Bobos P., Allen V., et al. (2021). COVID-19 hospitalizations, ICU admissions and deaths associated with the new variants of concern. *Science Briefs of the Ontario COVID-19 Science Advisory Table*, **1**(18). https://doi.org/10.47326/ocsat.2021.02.18.1.0

Van den Driessche P. and Watmough J. (2002). Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission. *Mathematical Biosciences*, **180**, 29-48. https://doi.org/10.1016/S0025-5564(02)00108-6

Veneti L., Seppälä E., et al. (2021). Increased risk of hospitalisation and intensive care admission associated with reported cases of SARS-CoV-2 variants B.1.1.7 and B.1.351 in Norway, December 2020–May 2021. *PLos ONE* **16**(10), e0258513. https://doi.org/10.1371/journal.pone.0258513

Wan E.A. and Van Der Merwe R. (2000). The unscented Kalman filter for nonlinear estimation. *Proceedings of the IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium (Cat. No.00EX373)*, 153-158. https://doi.org/10.1109/ASSPCC.2000.882463

Zhu X., Gao B., Zhong Y., Gu C. and Choi K.S. (2021). Extended Kalman filter based on stochastic epidemiological model for COVID-19 modelling. *Computers in Biology and Medicine*, **137**, 104810. https://doi.org/10.1016/j.compbiomed.2021.104810

Ελληνικό Στατιστικό Ινστιτούτο

Πρακτικά 34ου Πανελληνίου Συνεδρίου Στατιστικής (2022) (σελ. 99-112)

ΑΝΑΛΥΣΗ ΣΥΣΤΑΔΩΝ ΣΤΙΣ ΕΤΗΣΙΕΣ ΔΑΠΑΝΕΣ ΤΟΥ ΣΥΣΤΗΜΑΤΟΣ ΥΓΕΙΑΣ ΕΠΙΛΕΓΜΕΝΩΝ ΧΩΡΩΝ ΤΗΣ ΕΥΡΩΠΑΙΚΗΣ ΕΝΩΣΗΣ ΑΠΟ ΤΟ 2004 ΩΣ ΚΑΙ ΤΟ 2018

***Χατζημιχαήλ Θ. Χριστίνα*[1], *Καραγιάννης Βασίλης*[2]**

[1,2] Τμήμα Μαθηματικών ΑΠΘ

[1]xristina.k.xatzimixail@gmail.com

[2]vkdstat@math.auth.gr

ΠΕΡΙΛΗΨΗ

Αδιαμφισβήτητα προτεραιότητα κάθε κράτους θα πρέπει να αποτελεί η οργάνωση του Συστήματος Υγείας και η δυνατότητα καθολικής πρόσβασης των πολιτών σε κάθε είδος υγειονομικής περίθαλψης. Στο πλαίσιο αυτό, στην παρούσα εργασία παρουσιάζονται και μελετώνται διαχρονικά από το 2004 έως το 2018, οι ετήσιες δαπάνες Ταμείων Κοινωνικής Ασφάλισης αλλά και οι συνολικές κρατικές δαπάνες, ανά χίλιους κατοίκους και ως ποσοστό του ετήσιου ΑΕΠ των χωρών Ελλάδας, Γερμανίας, Ισπανίας, Ολλανδίας, Εσθονίας, Βελγίου και Πορτογαλίας για υπηρεσίες περίθαλψης και αποκατάστασης εσωτερικών και εξωτερικών ασθενών, για εργαστηριακές εξετάσεις, διαγνωστικές απεικονίσεις και υπηρεσίες μεταφοράς ασθενών, για υπηρεσίες μακροχρόνιας περίθαλψης ασθενών τόσο στο σπίτι, όσο και εντός των νοσοκομείων, για φαρμακευτικά αγαθά και άλλες θεραπευτικές συσκευές μακροχρόνιας χρήσης. Οι δαπάνες υγείας, το ετήσιο ΑΕΠ και ο πληθυσμός κάθε χώρας την 1 Ιανουαρίου κάθε έτους, ως στοιχεία επίσημων στατιστικών, ανακτήθηκαν από τη βάση δεδομένων της EUROSTAT και η ανάλυση συστάδων σε διαχρονικά δεδομένα έγινε με εφαρμογή του αλγορίθμου KmL. Από τη διαχρονική ταξινόμηση των χωρών επαληθεύτηκαν παρατηρήσεις οικονομικών μελετών της Eurostat, του Παγκόσμιου Οργανισμού Υγείας (Π.Ο.Υ) και του Οργανισμού Οικονομικής Συνεργασίας και Ανάπτυξης (ΟΟΣΑ) σχετικά με την οργάνωση των Συστημάτων Υγείας και την εξέλιξη των δαπανών των χωρών που κατατάσσονται στην ίδια κλάση.

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: Δαπάνες υγείας, Επίσημες Στατιστικές, Ανάλυση Συστάδων, KmL

1. Εισαγωγή

Είναι κοινά αποδεκτό, ότι κάθε κράτος είναι αναγκαίο να μεριμνά για την καθολική δυνατότητα πρόσβασης των πολιτών σε οποιουδήποτε είδους ιατρικής περίθαλψης και κάθε Σύστημα Υγείας οφείλει να εξασφαλίσει το δικαίωμα σε κάθε πολίτη να απολαμβάνει το υψηλότερο εφικτό επίπεδο υγειονομικών παροχών. Επιπρόσθετα, οι διάφορες Ευρωπαϊκές χώρες οργανώνουν με διαφορετικό τρόπο το Σύστημα Υγείας τους και το Σύστημα Κοινωνικής Ασφάλισης. Συγχρόνως, η Ευρωπαϊκή Ένωση ως οικονομική ένωση παρατηρεί την μετακίνηση εργατικού προσωπικού και ενδιαφέρεται για τη θέσπιση νομοθεσίας η οποία στοχεύει στην ασφαλιστική κάλυψη, όσον αφορά την υγειονομική περίθαλψη, κάθε εργαζόμενου σε όποια Ευρωπαϊκή χώρα και αν βρεθεί. Στο πλαίσιο αυτό λοιπόν, ενδιαφέρον παρουσιάζει η κατεύθυνση της κρατικής δαπάνης κάθε χώρας ανά τομέα του υγειονομικού

συστήματος ανεξάρτητα από τον προμηθευτή υγείας. Σύμφωνα με το Σύστημα Λογαριασμών Υγείας (2018), στους προμηθευτές υγείας εντάσσονται: νοσοκομεία, δομές νοσηλευτικής φροντίδας, δομές αντιμετώπισης κινητικών προβλημάτων, δομές επανένταξης, ΚΑΠΗ, οίκοι ευγηρίας, φορείς παροχής εξωνοσοκομειακής φροντίδας, προμηθευτές βοηθητικών υπηρεσιών υγείας (κλινικά και διαγνωστικά εργαστήρια, ΕΚΑΒ, κ.λπ.), έμποροι λιανικής και άλλοι φορείς παροχής ιατρικών προϊόντων μακροχρόνιου ή μη ιατρικού εξοπλισμού και τέλος φορείς που ασχολούνται με την διοίκηση του τομέα της υγείας και την ασφάλιση υγείας (ΕΟΦ, ΕΦΕΤ). Οι κρατικοί χρηματοδοτικοί φορείς του τομέα υγείας είναι οι φορείς της κεντρικής κυβέρνησης του κάθε κράτους, δηλαδή το Υπουργείο Υγείας και Κοινωνικής Αλληλεγγύης, το Υπουργείο Οικονομικών, το Υπουργείο Παιδείας το Υπουργείο Εθνικής Άμυνας, άλλοι φορείς της τοπικής αυτοδιοίκησης και τα Ταμεία Κοινωνικής Ασφάλισης (ΤΚΑ). Οι φορείς της κεντρικής κυβέρνησης θεσπίζουν τις δαπάνες του κρατικού προϋπολογισμού για την υγεία. Αξίζει να σημειωθεί πως, η συνολική κρατική δαπάνη για κάποια παροχή του Συστήματος Υγείας είναι το άθροισμα της δαπάνης ΤΚΑ και της δαπάνης που ορίζεται από τον κρατικό προϋπολογισμό για την συγκεκριμένη υγειονομική παροχή.

2. Στατιστικοί δείκτες (statistical indexes) μελέτης

Στην παρούσα εργασία κεντρικό άξονα ενδιαφέροντος αποτελεί ο ορισμός και η σύγκριση με χρήση διερευνητικού αλγορίθμου ταξινόμησης 5 στατιστικών δεικτών για τις χώρες Γερμανία, Ολλανδία, Βέλγιο, Ελλάδα, Ισπανία, Πορτογαλία και Εσθονία. Σύμφωνα με την EUROSTAT (https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Statistical_indicator) ως στατιστικός δείκτης (statistical index) ορίζεται η διορθωμένη τιμή ενός στατιστικού δεδομένου για τουλάχιστον μια διάστασή του (συνήθως αφορά το σχετικό μέγεθος), σε συγκεκριμένη χρονική στιγμή, συγκεκριμένο χώρο ή άλλο σχετικό χαρακτηριστικό, ώστε να είναι δυνατή η χρήση του για συγκρίσεις. Αναλυτικότερα, στην παρούσα εργασία μελετήθηκαν για τις 7 ανωτέρω χώρες της Ευρωπαϊκής Ένωσης οι ακόλουθοι στατιστικοί δείκτες:

a) **δαπάνη ΤΚΑ προς συνολικές κρατικές δαπάνες**. Η δαπάνη ΤΚΑ για κάθε παροχή του Συστήματος Υγείας εκφράζεται ως ποσοστό επί της συνολικής κρατικής δαπάνης για τη συγκεκριμένη παροχή του Συστήματος Υγείας, με στόχο την ταξινόμηση των χωρών ως προς τον τρόπο κρατικής χρηματοδότησης των υγειονομικών παροχών. Με άλλα λόγια, πραγματοποιήθηκε ταξινόμηση των 7 συγκεκριμένων Ευρωπαϊκών χωρών ανάλογα με το αν το μεγαλύτερο μέρος των συνολικών κρατικών δαπανών υγείας αποτελεί δαπάνες του κρατικού προϋπολογισμού ή δαπάνες ΤΚΑ.
b) **δαπάνη προς ετήσιο ΑΕΠ**. Τόσο οι υγειονομικές δαπάνες ΤΚΑ όσο και τα συνολικά κρατικά έξοδα υγείας εκφράστηκαν ως ποσοστό του ΑΕΠ και οι χώρες ταξινομήθηκαν ως προς το ποσοστό του ετήσιου ΑΕΠ τους που αντιστοιχεί στην χρηματοδότηση μιας παροχής του Συστήματος Υγείας.
c) **δαπάνη προς μέγεθος του πληθυσμού** ή αλλιώς δαπάνη ανά χίλιους κατοίκους. Κάθε υγειονομική δαπάνη ΤΚΑ, αλλά και τα συνολικά κρατικά έξοδα υγείας σε κάθε περίπτωση κανονικοποιήθηκαν ανά χίλιους κατοίκους με σκοπό οι χώρες να ταξινομηθούν σύμφωνα με την κατά κεφαλήν χρηματοδότηση καθεμιάς παροχής του Συστήματος Υγείας.

3. Ταξινόμηση Υγειονομικών Παροχών

Οι παραπάνω στατιστικοί δείκτες μελετώνται, όπως έχει ήδη αναφερθεί για 7 Ευρωπαϊκές χώρες, για 15 διαδοχικά έτη από το 2004 έως το 2018 και για τις ακόλουθες υγειονομικές παροχές:

a) **ενδονοσοκομειακή θεραπευτική περίθαλψη και αποκατάσταση**, που περιλαμβάνει δραστηριότητες ενδονοσοκομειακής περίθαλψης που λαμβάνουν χώρα σε δημόσια και ιδιωτικά γενικά, ψυχιατρικά και ειδικά νοσοκομεία και ιατρικές και παραϊατρικές υπηρεσίες που παρέχονται στο πλαίσιο φροντίδας ασθενών που έχουν εισαχθεί σε κέντρα κλειστής αποκατάστασης (κωδικός ΕΛΣΤΑΤ/EUROSTAT: HC.1.1 + HC.2.1).
b) **εξωνοσοκομειακές θεραπευτικές υπηρεσίες**. Τα νοσοκομεία μέσω των εξωτερικών τους ιατρείων προσφέρουν και εξωνοσοκομειακές υπηρεσίες. Κατά συνέπεια στην κατηγορία αυτή ανήκουν ιατρικές και παραϊατρικές υπηρεσίες που παρέχονται σε εξωτερικούς ασθενείς και υπηρεσίες από κινητές μονάδες περίθαλψης, από ιδιωτικές κλινικές και διαγνωστικά κέντρα (κωδικός ΕΛΣΤΑΤ/EUROSTAT: HC.1.3 + HC.2.3).
c) **ενδονοσοκομειακή μακροχρόνια νοσηλευτική φροντίδα**. Στην κατηγορία αυτή ανήκουν υπηρεσίες νοσηλευτικής περίθαλψης σε εσωτερικούς ασθενείς που χρειάζονται βοήθεια σε συνεχή βάση (κωδικός ΕΛΣΤΑΤ/EUROSTAT: HC.3.1).
d) **μακροχρόνια νοσηλευτική φροντίδα κατ'οίκον**, που περιλαμβάνει κατ'οίκον ιατρική και παραϊατρική φροντίδα σε ασθενείς που χρειάζονται βοήθεια σε συνεχή βάση (κωδικός ΕΛΣΤΑΤ/EUROSTAT: HC.3.4).
e) **υπηρεσίες κλινικών εργαστηρίων, διαγνωστικής απεικόνισης και μεταφοράς ασθενών**. Στην συγκεκριμένη υγειονομική δραστηριότητα εντάσσονται υπηρεσίες σε μικροβιολογικά εργαστήρια όπως ιατρικές εξετάσεις, διαγνωστικά τεστ, διαγνωστικές υπηρεσίες απεικόνισης που παρέχονται σε εξωτερικούς ασθενείς όπως ακτινογραφίες, αξονικές τομογραφίες, μελέτες οστικής πυκνότητας και υπηρεσίες που αφορούν την μεταφορά από και προς τις μονάδες υγείας για σκοπούς ιατρικής φροντίδας αλλά και την μεταφορά με συμβατικά οχήματα όταν ο ασθενής αποζημιώνεται για τα σχετικά έξοδα (κωδικός ΕΛΣΤΑΤ/EUROSTAT: HC.4).
f) **μη διαρκή φαρμακευτικά και άλλα υγειονομικά αναλώσιμα**, όπως διάφορα φαρμακευτικά προϊόντα, φάρμακα, οροί, εμβόλια, επίδεσμοι, κ.λπ. (κωδικός ΕΛΣΤΑΤ/EUROSTAT: HC.5.1)
g) **θεραπευτικές συσκευές και άλλα ιατρικά αγαθά διαρκείας**, π.χ. γυαλιά οράσεως, ακουστικά βαρηκοΐας, ορθοπεδικές συσκευές κ.λπ. (κωδικός ΕΛΣΤΑΤ/EUROSTAT: HC.5.2)

4. Δεδομένα Μελέτης

Ο λόγος που στη συγκεκριμένη εργασία παρουσιάζονται μόνο κρατικά υγειονομικά έξοδα είναι γιατί για τις συγκεκριμένες Ευρωπαϊκές χώρες, με εξαίρεση την Ισπανία και την Πορτογαλία, αποτελούν περίπου το 70 με 95% των συνολικών δαπανών υγείας. Επιπρόσθετα, κρίνεται απαραίτητο να επισημανθεί ότι προκειμένου να υπάρξει η δυνατότητα σύγκρισης των δαπανών υγείας μεταξύ των ετών, κάθε στατιστικός δείκτης υπολογίστηκε με έτος βάσης το 2009. Συνεπώς, η αξία κάθε υγειονομικής δαπάνης, είτε αυτή αποτελεί δαπάνη ΤΚΑ είτε συνολικά κρατικά έξοδα, όπως επίσης και το ετήσιο ΑΕΠ κάθε χώρας υπολογίστηκαν με έτος βάσης το 2009 λαμβάνοντας υπόψιν το μέσο ετήσιο πληθωρισμό. Για την συγκεκριμένη εργασία χρησιμοποιήθηκαν δεδομένα σχετικά με τις κρατικές δαπάνες, το ετήσιο ΑΕΠ και τον πληθυσμό κάθε χώρας που δημοσιεύονται από την ΕΛΣΤΑΤ και την EUROSTAT. Μονάδα μέτρησης κάθε υγειονομικής δαπάνης, αλλά και του ΑΕΠ είναι το 1 εκατομμύριο ευρώ (εκατ. ευρώ) και ως ετήσιος πληθυσμός χρησιμοποιήθηκε η εκτίμηση που δίνεται από την EUROSTAT 1 Ιανουαρίου κάθε έτους για τον πληθυσμό κάθε χώρας. Αξίζει να αναφερθεί, ότι τα δεδομένα αυτά είναι απογραφικά, δεν έχουν προκύψει δηλαδή από κάποια δειγματοληψία.

5. Ο αλγόριθμος KmL για διαχρονικά δεδομένα

5.1 Η έννοια της τροχιάς

Ο αλγόριθμος ταξινόμησης KmL αποτελεί παραλλαγή του αλγορίθμου k-means και χρησιμοποιείται για την ομαδοποίηση διαχρονικών δεδομένων. Βασίζεται στο κριτήριο Expectation-Maximization επί των διασπορών «μέσα» στις ομάδες (Within Clusters Variance) και «μεταξύ» των ομάδων (Between Clusters Variance). Για την ανάπτυξη θεωρίας σχετικά με τον συγκεκριμένο αλγόριθμο, σε πρώτο στάδιο κρίνεται απαραίτητο να οριστεί η έννοια της τροχιάς.

Έστω S ένα σύνολο *n* αντικειμένων και για κάθε αντικείμενο μία ποσοτική μεταβλητή *X* μετριέται σε *v* διαφορετικές χρονικές στιγμές. Στην παρούσα εργασία το σύνολο S είναι ο πληθυσμός των 7 Ευρωπαϊκών χωρών, Ελλάδα, Γερμανία, Ισπανία, Ολλανδία, Βέλγιο, Εσθονία και Πορτογαλία και καταγράφονται για τα έτη 2004 έως 2018 οι στατιστικοί δείκτες που διατυπώθηκαν στην παράγραφο 2. Η τιμή της *X* για το αντικείμενο *j* στο χρόνο *i* συμβολίζεται με x_{ij}. Για την συγκεκριμένη εργασία ως αντικείμενο *j* αναφέρεται κάποια από τις παραπάνω χώρες της Ευρώπης και στο χρόνο *i* αντιστοιχεί κάποια χρονιά μεταξύ των ετών 2004 και 2018. Οι μετρήσεις της ίδιας ποσοτικής μεταβλητής *X* για τα ίδια αντικείμενα για v διαφορετικές χρονικές στιγμές αποτελούν επαναλαμβανόμενες μετρήσεις. Για κάθε αντικείμενο j η ακολουθία των x_{ij} καλείται τροχιά και παρουσιάζεται ως: $\mathbf{X}_j = \left(\mathrm{x}_{1j}, \mathrm{x}_{2j}, \mathrm{x}_{3j}, \dots, \mathrm{x}_{vj}\right)$ (Genolini and Falissard, 2010)

5.2 Ορισμός απόστασης μεταξύ των τροχιών και εύρεση βέλτιστου αριθμού κλάσεων

Όπως και στην περίπτωση του αλγορίθμου k-means, η ταξινόμηση των τροχιών σύμφωνα με τον αλγόριθμο KmL στηρίζεται στην έννοια της απόστασης. Στην παρούσα εργασία χρησιμοποιείται η ευκλείδεια απόσταση για τον υπολογισμό της απόστασης μεταξύ των τροχιών. Αναλυτικότερα και σύμφωνα με τους Calinski and Harabasz (1974), έστω $\mathbf{P}_1, \dots, \mathbf{P}_n$ τροχιές ή αλλιώς σημεία του v-διάστατου χώρου και εφόσον έχει οριστεί ένα μέτρο απόστασης μεταξύ των τροχιών ορίζεται ο $n \times n$ πίνακας $\mathbf{Q} = \left\{\mathrm{q}_{ij}\right\}_{i,j=1,\cdots,n}$ όπου q_{ij}: είναι η απόσταση μεταξύ των τροχιών $\mathbf{P}_i$ και $\mathbf{P}_j$. Επίσης ορίζεται ο $v \times v$ πίνακας συνδιασπορών $\mathbf{R} = \left\{\mathrm{r}_{ij}\right\}_{i,j=1,\dots,v}$ όπου r_{ij} η συνδιασπορά των τροχιών $\mathbf{P}_i$ και $\mathbf{P}_j$ και τέλος ορίζεται ο $v \times n$ πίνακας $\mathbf{X} = \left\{\mathrm{x}_{ij}\right\}$, όπου x_{ij} είναι η τιμή της ποσοτικής μεταβλητής του προβλήματος για το αντικείμενο *j* την χρονική στιγμή i. Έτσι, ο πίνακας $\mathbf{X}$ ισούται με $\mathbf{X} = (\mathbf{X}_1\ \mathbf{X}_2 \dots \mathbf{X}_n)$, όπου η στήλη $\mathbf{X}_j$ περιλαμβάνει την ακολουθία των παρατηρήσεων της τροχιάς $\mathbf{P}_j$. Για παράδειγμα $\mathbf{X}_1 = [\mathrm{x}_{11}\, \mathrm{x}_{21} \cdots \mathrm{x}_{v1}]^{\mathrm{T}}$. Στην παρούσα εργασία το πλήθος των τροχιών είναι 7, όσο και το πλήθος των Ευρωπαϊκών χωρών για τις οποίες μελετήθηκαν και καταγράφηκαν οι στατιστικοί δείκτες που αναφέρθηκαν παραπάνω. Ακόμα, οι τροχιές ανήκουν σε χώρο διάστασης 15, καθώς η ποσοτική μεταβλητή, που στην συγκεκριμένη περίπτωση είναι κάποιος από τους στατιστικούς δείκτες που διατυπώθηκαν στην παράγραφο 2, καταγράφεται για 15 διαδοχικά έτη από το 2004 έως το 2018. Επιπρόσθετα, δεν πρέπει να παραληφθεί ότι στην παρούσα εργασία η ποσοτική μεταβλητή *X* δεν είναι τυχαία μεταβλητή, διότι λαμβάνει σταθερές τιμές και όχι τιμές με κάποια πιθανότητα. Για κάθε μία από τις 7 υγειονομικές παροχές που δόθηκαν στην παράγραφο 3 αντιστοιχούν οι 5 στατιστικοί δείκτες της παραγράφου 2 με 15 διαχρονικές παρατηρήσεις.

Στον ευκλείδειο χώρο η απόσταση μεταξύ των τροχιών $\mathbf{P}_i$ και $\mathbf{P}_j$ συμβολίζεται με d_{ij} και το τετράγωνο της απόστασης αυτής δίνεται από (Calinski and Harabasz 1974):

$$\mathrm{d}_{ij}^2 = \left(\mathbf{X}_i - \mathbf{X}_j\right)^{\mathrm{T}}\left(\mathbf{X}_i - \mathbf{X}_j\right) \quad (1)$$

όπου όπως έχει ήδη αναφερθεί $\mathbf{X}_i$: $v \times 1$ πίνακας των συντεταγμένων της τροχιάς $\mathbf{P}_i$.

Στόχος κάθε ταξινόμησης είναι ο σχηματισμός ομοιογενών και καλά διαχωρισμένων κλάσεων. Η απόσταση μεταξύ των τροχιών κάθε κλάσης εκφράζεται από το άθροισμα τετραγώνων «μέσα» στις κλάσεις (Within Sum Squares), ενώ η απόσταση μεταξύ των κλάσεων, με άλλα λόγια η ανομοιογένεια

μεταξύ των κλάσεων, εκφράζεται από το άθροισμα τετραγώνων «μεταξύ» των κλάσεων (Between Cluster Sum Squares). Οι επόμενες προτάσεις βασίζονται στο κείμενο της εργασίας των Calinski and Harabasz, 1974.

Πρόταση 1: Έστω κλάση m με n_m τροχιές και κέντρο κλάσης την τροχιά $\overline{\mathbf{X}}_m$, όπου $\overline{\mathbf{X}}_m = \left[\frac{\sum_{j=1}^{n_m} \mathrm{x}_{1j}}{n_m} \frac{\sum_{j=1}^{\mathrm{n_m}} \mathrm{x}_{2j}}{n_m} \ldots \frac{\sum_{j=1}^{n_m} \mathrm{x}_{\mathrm{v}j}}{n_m}\right]^{\mathrm{T}}$. Το άθροισμα τετραγώνων «μέσα» στη κλάση m συμβολίζεται με WSS_m και ισούται με:

$$\mathrm{WSS}_m = \sum_{l=1}^{\mathrm{n_m}} (\mathbf{X}_l - \overline{\mathbf{X}}_m)^{\mathrm{T}} (\mathbf{X}_l - \overline{\mathbf{X}}_m) \quad (2)$$

Όπως έχει αναφερθεί η διασπορά εντός της κλάσης εκφράζεται από την απόσταση μεταξύ των τροχιών της κλάσης. Αν η κλάση m αποτελείται από 2 τροχιές $\mathbf{X}_1$ και $\mathbf{X}_2$, οι οποίες ανήκουν σε χώρο διάστασης 2, στην περίπτωση δηλαδή όπου για 2 χώρες και για κάποια από τις υγειονομικές παροχές καταγράφεται ένας στατιστικός δείκτης για 2 χρονικές στιγμές, τότε εφαρμόζοντας την σχέση (2) προκύπτει ότι $\mathrm{WSS}_m = \frac{1}{2}(\mathbf{X}_1 - \mathbf{X}_2)^{\mathrm{T}}(\mathbf{X}_1 - \mathbf{X}_2) = \frac{1}{2}\mathrm{d}_{12}^2$. Γενικά αποδεικνύεται η επόμενη πρόταση.

Πρόταση 2: Το άθροισμα τετραγώνων WSS_m στην κλάση m με πλήθος τροχιών n_m, οι οποίες ανήκουν στον v-διάστατο ευκλείδειο χώρο ισούται με:

$$\mathrm{WSS}_m = \mathrm{n}_m^{-1}(\, \mathrm{d}_{12}^2(m) + \mathrm{d}_{13}^2(m) + \ldots + \mathrm{d}_{\mathrm{n}_m-1\mathrm{n}_m}^2(m)) \quad (3)$$

όπου $\mathrm{d}_{\mathrm{ij}}^2(m)$ είναι το τετράγωνο της ευκλείδειας απόστασης μεταξύ των τροχιών $\mathbf{P}_i$ και $\mathbf{P}_j$ της κλάσης m.

Πρόταση 3: Αν με $\bar{\mathrm{d}}_m^2$ συμβολίζεται ο μέσος όρος των $\frac{\mathrm{n}_m(\mathrm{n}_m-1)}{2}$ τετραγώνων των αποστάσεων μεταξύ των τροχιών της κλάσης m τότε από την σχέση (3) προκύπτει ότι:

$$\mathrm{WSS}_m = \frac{\mathrm{n}_m-1}{2}\bar{\mathrm{d}}_m^2 \quad (4)$$

Πρόταση 4: Το άθροισμα τετραγώνων ανάμεσα (within) στις κλάσεις WCSS για την περίπτωση k κλάσεων δίνεται από:

$$\mathrm{WCSS} = \frac{1}{2}\left((\mathrm{n}_1 - 1)\bar{\mathrm{d}}_1^2 + \cdots + (\mathrm{n}_k - 1)\bar{\mathrm{d}}_k^2\right) \quad (5)$$

Πρόταση 5: Αν $\overline{\mathbf{X}} = \left[\frac{\sum_{j=1}^{n} \mathrm{x}_{1j}}{n} \frac{\sum_{j=1}^{n} \mathrm{x}_{2j}}{n} \ldots \frac{\sum_{j=1}^{n} \mathrm{x}_{\mathrm{v}j}}{n}\right]^{\mathrm{T}}$ είναι η μέση τροχιά των n τροχιών τότε το άθροισμα τετραγώνων μεταξύ (between) των κλάσεων για k κλάσεις, συμβολίζεται με BCSS και είναι ίσο με:

$$\mathrm{BCSS} = \sum_{m=1}^{k} \mathrm{n}_m(\overline{\mathbf{X}}_m - \overline{\mathbf{X}})^{\mathrm{T}}\,(\overline{\mathbf{X}}_m - \overline{\mathbf{X}}) \quad (6)$$

Πρόταση 6: Για το άθροισμα τετραγώνων BCSS αποδεικνύεται ότι:

$$\mathrm{BCSS} = \frac{1}{2}\left((k-1)\overline{\mathrm{d}^2} + (n-k)\mathrm{A}_k\right) \quad (7)$$

όπου $\mathrm{A}_k = \frac{1}{n-k}(\,(n_1 - 1)\left(\overline{\mathrm{d}^2} - \bar{\mathrm{d}}_1^2\right) + (n_2 - 1)\left(\overline{\mathrm{d}^2} - \bar{\mathrm{d}}_2^2\right) + \ldots + (n_k - 1)\left(\overline{\mathrm{d}^2} - \bar{\mathrm{d}}_k^2\right))$

και $\overline{\mathrm{d}^2}$ είναι ο μέσος όρος των τετραγώνων των ευκλείδειων αποστάσεων των $\frac{n(n-1)}{2}$ συνδυασμών ανά δυο που δημιουργούνται από n τροχιές.

Πρόταση 7: Για το άθροισμα τετραγώνων WCSS αποδεικνύεται ότι:

$$\mathrm{WCSS} = \frac{1}{2}(n-k)\left(\overline{\mathrm{d}^2} - \mathrm{A}_k\right) \quad (8)$$

Πρόταση 8: Το συνολικό άθροισμα τετραγώνων (Total Sum Squares) ισούται με το άθροισμα των WCSS και BCSS και αποδεικνύεται ότι είναι ίσο με:

$$\mathrm{TSS} = \frac{1}{2}(n-1)\overline{\mathrm{d}^2} \tag{9}$$

Ο βέλτιστος αριθμός κλάσεων για την ταξινόμηση τροχιών με τον αλγόριθμο KmL επιλέγεται με χρήση του κριτηρίου Calinsk και Harabasz

Κριτήριο Calinsk και Harabasz (1974): Το κριτήριο Calinsk και Harabasz είναι βασισμένο στην αρχή της ελάχιστης διασποράς. Συμβολίζοντας με k τον αριθμό κλάσεων στις οποίες ταξινομούνται οι τροχιές, επιλέγεται εκείνη η τιμή του k για την οποία μεγιστοποιείται ο λόγος των διασπορών:

$$\mathrm{C}(k) = \frac{\mathrm{BCSS}}{\mathrm{WCSS}}\frac{n-k}{k-1} \tag{10}$$

Αν στην σχέση (10) τα αθροίσματα τετραγώνων BCSS και WCSS αντικατασταθούν με τις σχέσεις (7) και (8) αντίστοιχα, αποδεικνύεται ότι:

$$\mathrm{C}(k) = \frac{\overline{\mathrm{d}^2} + \frac{n-k}{k-1}\mathrm{A}_k}{\overline{\mathrm{d}^2} - \mathrm{A}_k} \tag{11}$$

Πρόταση 9: Αν οι αποστάσεις μεταξύ των n τροχιών ανά δυο είναι ίσες τότε $\mathrm{C}(k) = 1$.

5.3 Αξιολόγηση της ομαδοποίησης

Στην παρούσα εργασία για την αξιολόγηση της ομαδοποίησης χρησιμοποιήθηκε ο συντελεστής silhouette (Rousseeuw, 1986) ο οποίος προκύπτει ως εξής: για μια τροχιά *i* μιας κλάσης υπολογίζεται η ευκλείδεια απόσταση από κάθε άλλη τροχιά της κλάσης αυτής και υπολογίζεται ο μέσος όρος αυτών των αποστάσεων ο οποίος συμβολίζεται με D_i. Στην συνέχεια, υπολογίζεται η απόσταση της συγκεκριμένης τροχιάς από κάθε τροχιά που ανήκει στην κοντινότερη κλάση και λαμβάνεται και πάλι ο μέσος όρος των ευκλείδειων αποστάσεων που προκύπτουν, ο οποίος συμβολίζεται με C_i. Έτσι, ο συντελεστής silhouette για την *i* τροχιά δίνεται από:

$$\mathrm{S}_i = \frac{\mathrm{C}_i - \mathrm{D}_i}{\max(\mathrm{C}_i, \mathrm{D}_i)} \tag{12}$$

Γίνεται αντιληπτό ότι, για κάθε τροχιά i ισχύει $-1 \leq \mathrm{S}_i \leq 1$. Θετικές τιμές του συντελεστή silhouette και ιδιαίτερα τιμές κοντά στην μονάδα φανερώνουν πολύ ικανοποιητική ταξινόμηση. Αντιθέτως, αν προκύψει αρνητικός συντελεστής silhouette για κάποια τροχιά τότε η τροχιά έχει ταξινομηθεί σε λανθασμένη κλάση και αν ο συντελεστής silhouette ισούται με μηδέν τότε η τροχιά βρίσκεται μεταξύ δύο κλάσεων. Στην περίπτωση που η κλάση αποτελείται από μία μόνο τροχιά, στην βιβλιογραφία δεν αναφέρεται ο τρόπος υπολογισμού του συντελεστή silhouette γι' αυτή την τροχιά και συμβατικά θεωρείται μηδενικός. Η επιλογή του μηδενός είναι τυχαία και φυσικά δεν σημαίνει ότι η τροχιά βρίσκεται μεταξύ δύο κλάσεων.

5.4 Πλεονεκτήματα και Μειονεκτήματα του αλγορίθμου KmL

Ο αλγόριθμος ταξινόμησης KmL, παρόλο που ως παραλλαγή του αλγορίθμου k-means εξαρτάται από την αρχική επιλογή κέντρων, είναι εύκολος στην κατανόηση και έχει χρησιμοποιηθεί με αρκετή επιτυχία σε μελέτες διερευνητικής ταξινόμησης, με αποτέλεσμα να είναι κατάλληλος για διερευνητική ταξινόμηση ατόμων ενός πληθυσμού. Συνεπώς, κρίνεται κατάλληλος για την ταξινόμηση των 7 χωρών της παρούσας μελέτης, όπου τα δεδομένα αποτελούν παραμέτρους απογραφής. Επιπρόσθετα, συγκριτικά με τις μεθόδους που βασίζονται σε πιθανοθεωρητικά μοντέλα και ελέγχους υποθέσεων, ο αλγόριθμος KmL δεν απαιτεί γνώση της κατανομής των δεδομένων, έλεγχο κανονικότητας ή

οποιαδήποτε άλλη παραμετρική παραδοχή και επομένως δεν είναι απαραίτητη οποιαδήποτε υπόθεση σχετικά με το σχήμα της τροχιάς. Τέλος, σε σχέση με άλλες παραλλαγές του αλγορίθμου k-means για την περίπτωση ταξινόμησης διαχρονικών δεδομένων, ο αλγόριθμος KmL διαχειρίζεται καλύτερα το πρόβλημα των χαμένων τιμών (Sentenac et. al., 2015).

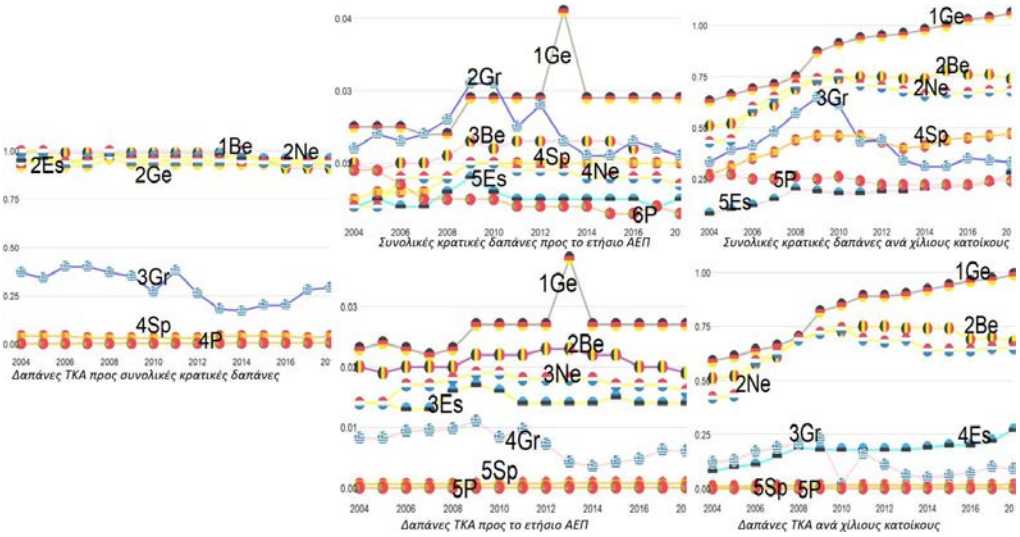
6. Αποτελέσματα

Τα αποτελέσματα ταξινόμησης των χωρών ως προς την εξέλιξη των 5 στατιστικών δεικτών, σύμφωνα με τον αλγόριθμο ταξινόμησης KmL, δίνονται με τη μορφή ομαδοποιημένων χρονοδιαγραμμάτων και κειμένου μόνο στην παράγραφο 6.1, ενώ λόγω έλλειψης χώρου στις παραγράφους 6.2 ως και 6.7 δίνονται με απόλυτα ανάλογο τρόπο μόνο με την μορφή κειμένου (τα ομαδοποιημένα χρονοδιαγράμματα για τις παραγράφους αυτές μπορούν να σταλούν σε οποιονδήποτε ενδιαφέρεται μετά από επικοινωνία με τους συγγραφείς).

6.1 Κρατική χρηματοδότηση ενδονοσοκομειακών υπηρεσιών περίθαλψης και αποκατάστασης ασθενών

Στο γράφημα που ακολουθεί συμβολίζουμε ως Be: Βέλγιο (Belgium), Ne: Ολλανδία (Netherlands), Ge: Γερμανία (German), Es: Εσθονία (Estonia), Gr: Ελλάδα (Greece), Sp: Ισπανία (Spain) και P: Πορτογαλία (Portugal). Ο αριθμός που προηγείται της συντομογραφίας της χώρας, φανερώνει την κλάση στην οποία ταξινομείται η χώρα.

Εικόνα 1. Κρατική χρηματοδότηση ενδονοσοκομειακών υπηρεσιών περίθαλψης και αποκατάστασης.



Όσον αφορά την κρατική χρηματοδότηση υπηρεσιών ενδονοσοκομειακής περίθαλψης και αποκατάστασης, το Βέλγιο δεν ταξινομείται στην ίδια κλάση με κάποια άλλη χώρα. Στο Βέλγιο έως το 2014 το 99% των συνολικών κρατικών δαπανών για την παραπάνω υγειονομική παροχή αποτελεί δαπάνες ΤΚΑ, ενώ από το 2015 το συγκεκριμένο ποσοστό φθίνει και σταθεροποιείται περίπου στο 95% για τα έτη 2016 μέχρι 2018. Ολλανδία, Γερμανία και Εσθονία ταξινομούνται στην ίδια κλάση στην οποία παραπάνω από το 95% της συνολικής κρατικής χρηματοδότησης προέρχεται από ΤΚΑ.

Στην Ελλάδα κύριος κρατικός χρηματοδοτικός φορέας υπηρεσιών περίθαλψης και αποκατάστασης ασθενών εντός των νοσοκομειακών μονάδων είναι οι φορείς της κεντρικής κυβέρνησης και αναφορικά με τον τρόπο χρηματοδότησης δεν ταξινομείται στην ίδια κλάση με κάποια άλλη χώρα. Το ποσοστό των συνολικών κρατικών δαπανών που αντιστοιχεί σε δαπάνες ΤΚΑ παρουσιάζει μεγαλύτερες ετήσιες διακυμάνσεις συγκριτικά με το ποσοστό των υπολοίπων χωρών. Η Πορτογαλία και η Ισπανία ταξινομούνται στην ίδια κλάση αναφορικά με τον τρόπο κρατικής χρηματοδότησης, τις δαπάνες ΤΚΑ ως ποσοστό του ΑΕΠ και τις κατά κεφαλή δαπάνες ΤΚΑ. Για τις χώρες αυτές, σχεδόν εξ' ολοκλήρου οι συνολικές κρατικές δαπάνες αποτελούν δαπάνες του κρατικού προϋπολογισμού. Αν και η Ολλανδία ταξινομείται στην ίδια κλάση με την Εσθονία όσον αφορά το ποσοστό του ΑΕΠ που αντιστοιχεί σε δαπάνες ΤΚΑ, οι αντίστοιχες κατά κεφαλήν δαπάνες της Εσθονίας διαφέρουν κατά πολύ από τις κατά κεφαλήν δαπάνες της Ολλανδίας, η οποία ταξινομείται στην ίδια κλάση με το Βέλγιο. Το γεγονός αυτό οφείλεται στο υψηλό ποσοστό ΑΕΠ της Ολλανδίας. Γενικότερα, παρατηρείται ότι οι δαπάνες ΤΚΑ ανά χίλιους κατοίκους της Γερμανίας, Ολλανδίας και Βελγίου είναι αρκετά αυξημένες συγκριτικά με τις αντίστοιχες δαπάνες των χωρών Εσθονίας, Ελλάδας, Πορτογαλίας και Ισπανίας. Διαπιστώνεται ότι οι συνολικές κρατικές δαπάνες ως ποσοστό του ΑΕΠ και ανά χίλιους κατοίκους των χωρών Γερμανίας, Ολλανδίας, Εσθονίας και Βελγίου δεν διαφέρουν από τις αντίστοιχες δαπάνες ΤΚΑ, καθώς κύριος κρατικός χρηματοδοτικός φορέας των χωρών αυτών είναι τα ΤΚΑ. Αντιθέτως, οι συνολικές κρατικές δαπάνες ως ποσοστό του ετήσιου ΑΕΠ και οι δαπάνες ανά χίλιους κατοίκους της Ελλάδας, Ισπανίας και Πορτογαλίας παρουσιάζονται αυξημένες και μάλιστα από το 2009 έως το 2011 το ποσοστό του ΑΕΠ της Ελλάδας που αντιστοιχεί σε συνολικές κρατικές δαπάνες για την συγκεκριμένη υγειονομική παροχή είναι το υψηλότερο μεταξύ των αντίστοιχων ποσοστών των υπολοίπων χωρών. Η Εσθονία διαθέτει τις χαμηλότερες κατά κεφαλήν συνολικές δαπάνες και ταξινομείται στην ίδια κλάση με την Πορτογαλία. Είναι φανερό ότι το ποσοστό του ετήσιου ΑΕΠ της Γερμανίας που αντιστοιχεί σε συνολικές κρατικές δαπάνες και το ποσοστό του ΑΕΠ που αντιστοιχεί σε δαπάνες ΤΚΑ για την συγκεκριμένη υγειονομική παροχή παρουσιάζει ραγδαία αύξηση το 2013, καθώς το ΑΕΠ της Γερμανίας εκείνης της χρονιάς μειώνεται έντονα και επανέρχεται σε σταθερά επίπεδα κατά τα έτη 2014 μέχρι 2018. Ωστόσο, διαπιστώνεται ότι το Γερμανικό κράτος αναπληρώνει το συγκεκριμένο έλλειμμα, αφού τόσο οι συνολικές δαπάνες όσο και οι δαπάνες ΤΚΑ για ενδονοσοκομειακές παροχές περίθαλψης και αποκατάστασης όχι μόνο δεν μειώνονται αλλά παρουσιάζουν συνεχώς αυξητική τάση από το 2009 έως το 2018. Το ίδιο παρατηρείται και κατά την χρηματοδότηση κάθε μίας από τις παρακάτω υγειονομικές δραστηριότητες του Γερμανικού Συστήματος Υγείας.

6.2 Κρατική χρηματοδότηση εξωνοσοκομειακών υπηρεσιών περίθαλψης και αποκατάστασης

Αναφορικά με τον τρόπο χρηματοδότησης, παρατηρήθηκε ότι το Βέλγιο και η Ολλανδία ταξινομούνται στην ίδια κλάση. Από το 2004 έως το 2015 σχεδόν το 100% των συνολικών κρατικών δαπανών των παραπάνω χωρών αποτελεί δαπάνες ΤΚΑ, ποσοστό το οποίο παρουσιάζει μικρή πτώση το 2016 και σταθεροποιείται έως το 2012. Εσθονία και Γερμανία ταξινομούνται στην ίδια κλάση, ενώ η Ελλάδα δεν κατηγοριοποιείται σε κοινή κλάση με κάποια άλλη χώρα, διότι για τα περισσότερα έτη κύριος κρατικός χρηματοδοτικός φορέας είναι οι φορείς της κεντρικής κυβέρνησης. Για τα έτη 2004, 2011 οι συνολικές κρατικές δαπάνες της Ελλάδας για την συγκεκριμένη υγειονομική παροχή είναι κατά το ήμισυ δαπάνες ΤΚΑ και για το 2005 και 2006 οι δαπάνες ΤΚΑ είναι λίγο υψηλότερες από τις αντίστοιχες δαπάνες του κρατικού προϋπολογισμού. Πορτογαλία και Ισπανία ταξινομούνται στην ίδια κλάση, καθώς το μεγαλύτερο μέρος της χρηματοδότησης υπηρεσιών περίθαλψης εξωτερικών ασθενών πραγματοποιείται από φορείς της κεντρικής κυβέρνησης. Συνεπώς, όσον αφορά τις δαπάνες ΤΚΑ ως

ποσοστό του ΑΕΠ και ανά χίλιους κατοίκους για την συγκεκριμένη υγειονομική δραστηριότητα, Ελλάδα, Πορτογαλία και Ισπανία ταξινομούνται στην ίδια κλάση. Οι δαπάνες ΤΚΑ ανά χίλιους κατοίκους του Βελγίου είναι αρκετά υψηλότερες από τις αντίστοιχες δαπάνες ΤΚΑ της Εσθονίας, διαφορά που δεν παρατηρήθηκε όταν οι συγκεκριμένες δαπάνες εκφράστηκαν ως ποσοστό του ετήσιου ΑΕΠ. Τόσο οι συνολικές, όσο και οι δαπάνες ΤΚΑ της Ολλανδίας και της Γερμανίας είναι υψηλότερες συγκριτικά με τις ανάλογες δαπάνες των υπολοίπων χωρών, αυξάνονται συνεχώς κατά έτος και έτσι οι χώρες αυτές ταξινομούνται στην ίδια κλάση. Εφόσον, κύριος κρατικός χρηματοδοτικός φορέας εξωνοσοκομειακών υπηρεσιών περίθαλψης και αποκατάστασης του Βελγίου, Εσθονίας, Ολλανδίας και Γερμανίας είναι τα ΤΚΑ δεν παρουσιάστηκαν σημαντικές διαφορές στην εξέλιξη των συνολικών κρατικών δαπανών και των αντίστοιχων δαπανών ΤΚΑ των χωρών αυτών. Αντιθέτως, αυξημένες παρουσιάστηκαν οι συνολικές δαπάνες της Πορτογαλίας και της Ισπανίας. Μάλιστα το ποσοστό του ετήσιου ΑΕΠ της Πορτογαλίας που αντιστοιχεί σε συνολικές κρατικές δαπάνες για παροχές περίθαλψης εξωτερικών ασθενών είναι υψηλότερο από το 2004 έως το 2012 συγκριτικά με το αντίστοιχο ποσοστό ΑΕΠ των υπολοίπων χωρών. Μεταξύ των χωρών, η Ελλάδα διαθέτει τις χαμηλότερες συνολικές κρατικές δαπάνες τόσο κατά κεφαλήν όσο και ως ποσοστό του ΑΕΠ.

6.3 Κρατική χρηματοδότηση υπηρεσιών κλινικών εργαστηρίων διαγνωστικής απεικόνισης και μεταφοράς ασθενών

Ολλανδία, Βέλγιο και Γερμανία ταξινομούνται στην ίδια κλάση όσον αφορά το ποσοστό των συνολικών κρατικών δαπανών που αποτελούν ετησίως δαπάνες ΤΚΑ για υπηρεσίες κλινικών εργαστηρίων, διαγνωστικής απεικόνισης και μεταφοράς ασθενών, καθώς το ποσοστό αυτό ετησίως είναι μεγαλύτερο του 95%. Σε αντίθεση με τις παροχές περίθαλψης εσωτερικών και εξωτερικών ασθενών, το μεγαλύτερο μέρος της κρατικής χρηματοδότησης υπηρεσιών κλινικών εργαστηρίων, διαγνωστικής απεικόνισης και μεταφοράς ασθενών στην Ελλάδα πραγματοποιείται από ΤΚΑ. Αναλυτικότερα, το ποσοστό των συνολικών κρατικών δαπανών της Ελλάδας που αντιστοιχεί σε δαπάνες ΤΚΑ για την συγκεκριμένη υγειονομική παροχή ενώ είναι σταθερό γύρω στο 90% κατά τα έτη 2004 έως 2008, φθίνει συνεχώς από το 2009 έως το 2012, αλλά δεν παρατηρήθηκε μικρότερο του 73% και αυξάνει εκ νέου το 2013. Για την Ελλάδα, που δεν ταξινομείται στην ίδια κλάση με κάποια άλλη χώρα, παρατηρήθηκαν έντονες ετήσιες μεταβολές αναφορικά με τον τρόπο χρηματοδότησης της συγκεκριμένης παροχής. Για την Εσθονία, που επίσης αποτελεί μια ξεχωριστή κλάση, παρατηρήθηκε ότι η κρατική χρηματοδότηση πραγματοποιείται περίπου κατά 75% ετησίως από ΤΚΑ. Το μεγαλύτερο μέρος των συνολικών κρατικών δαπανών της Ισπανίας και της Πορτογαλίας για την συγκεκριμένη υγειονομική δραστηριότητα είναι δαπάνες του κρατικού προϋπολογισμού και οι δύο αυτές χώρες ταξινομούνται στην ίδια κλάση σχετικά με τον τρόπο κρατικής χρηματοδότησης της συγκεκριμένης υγειονομικής παροχής. Ισπανία και Πορτογαλία είναι οι μοναδικές χώρες που ταξινομούνται σε κοινή κλάση όσον αφορά τις δαπάνες ΤΚΑ ως ποσοστό του ΑΕΠ και ανά χίλιους κατοίκους για κλινικές εξετάσεις και διαγνωστικούς ελέγχους. Διαπιστώνεται ότι με εξαίρεση τα έτη 2012 και 2016 έως 2018, λόγω του υψηλού ΑΕΠ, το ποσοστό του ετήσιου ΑΕΠ της Ολλανδίας που αντιστοιχεί σε δαπάνες ΤΚΑ είναι χαμηλότερο από το αντίστοιχο ποσοστό ΑΕΠ της Ελλάδας. Ωστόσο, οι κατά κεφαλήν δαπάνες ΤΚΑ της Ολλανδίας είναι υψηλότερες από τις αντίστοιχες κατά κεφαλήν δαπάνες της Ελλάδας. Γερμανία και Βέλγιο διαθέτουν τις υψηλότερες κατά κεφαλήν συνολικές κρατικές δαπάνες και δαπάνες ΤΚΑ, αλλά δεν ταξινομούνται στην ίδια κλάση. Παρατηρήθηκε ότι οι κατά κεφαλήν δαπάνες του Βελγίου σταθεροποιούνται από το 2012 έως το 2018, ενώ οι αντίστοιχες κατά κεφαλήν δαπάνες της Γερμανίας κατά τα παραπάνω έτη αυξάνουν συνεχώς. Το ποσοστό του ετήσιου ΑΕΠ της Εσθονίας που

αντιστοιχεί τόσο σε δαπάνες ΤΚΑ όσο και σε συνολικές κρατικές δαπάνες είναι μεγαλύτερο από αυτό των υπολοίπων χωρών κατά τα έτη 2008 έως 2018 με εξαίρεση το 2013. Μεταξύ των χωρών, οι συνολικές κρατικές δαπάνες ως ποσοστό του ΑΕΠ της Ολλανδίας για την παραπάνω υγειονομική παροχή είναι οι χαμηλότερες, ενώ οι μικρότερες κατά κεφαλήν συνολικές κρατικές δαπάνες σημειώνονται στην Ελλάδα.

6.4 Κρατική χρηματοδότηση υπηρεσιών μακροχρόνιας ενδονοσοκομειακής περίθαλψης ασθενών

Σε ότι αφορά τον τρόπο κρατικής χρηματοδότησης υπηρεσιών μακροχρόνιας ενδονοσοκομειακής φροντίδας, Ολλανδία και Ελλάδα ταξινομούνται στην ίδια κλάση. Οι συνολικές κρατικές δαπάνες των χωρών αυτών για την συγκεκριμένη υγειονομική παροχή αποτελούν σχεδόν εξ' ολοκλήρου δαπάνες ΤΚΑ. Κάθε μία από τις υπόλοιπες χώρες, αναφορικά με το ποσοστό των συνολικών κρατικών δαπανών που αντιστοιχεί σε δαπάνες ΤΚΑ, σχηματίζει ξεχωριστή κλάση. Σε αντίθεση με τις προηγούμενες υγειονομικές παροχές, κύριος κρατικός χρηματοδοτικός φορέας υπηρεσιών μακροχρόνιας ενδονοσοκομειακής περίθαλψης ασθενών της Πορτογαλίας είναι τα ΤΚΑ. Με εξαίρεση το 2004 και τα έτη 2014 έως 2016, όπου για την τριετία αυτή το 50% των συνολικών κρατικών εξόδων αποτελεί δαπάνες ΤΚΑ, κύριος κρατικός χρηματοδοτικός φορέας της παραπάνω υγειονομικής παροχής του Εσθονικού Συστήματος Υγείας είναι οι φορείς της κεντρικής κυβέρνησης. Στο Βέλγιο, ενώ κατά τα έτη 2004 και 2005 το ποσοστό των συνολικών κρατικών εξόδων για την παραπάνω υγειονομική παροχή που αντιστοιχεί σε δαπάνες ΤΚΑ είναι μικρότερο του 50%, από το 2006 έως το 2015 το ποσοστό αυτό αυξάνεται και σταθεροποιείται περίπου στο 52%. Το συγκεκριμένο ποσοστό φθίνει ραγδαία το 2014 και γίνεται σχεδόν μηδενικό έως το 2018. Η χρηματοδότηση υπηρεσιών μακροχρόνιας ενδονοσοκομειακής περίθαλψης ασθενών της Ισπανίας πραγματοποιείται αποκλειστικά από φορείς της κεντρικής κυβέρνησης. Ακόμα, διαπιστώθηκε ότι τόσο οι συνολικές κρατικές δαπάνες όσο και οι δαπάνες ΤΚΑ ως ποσοστό του ΑΕΠ και ανά χίλιους κατοίκους της Ολλανδίας είναι ιδιαίτερα αυξημένες συγκριτικά με τις αντίστοιχες δαπάνες των υπολοίπων χωρών. Επιπλέον, παρατηρήθηκε ότι οι δαπάνες ΤΚΑ του Βελγίου για παροχές μακροχρόνιας περίθαλψης εσωτερικών ασθενών ως ποσοστό του ΑΕΠ και ανά χίλιους κατοίκους φθίνουν ραγδαία το 2015 και είναι ιδιαίτερα χαμηλές από το 2015 έως το 2018. Ωστόσο το έλλειμμα που φάνηκε να παρουσιάζεται στην χρηματοδότηση από ΤΚΑ κατά τα έτη αυτά καλύπτεται από φορείς της κεντρικής κυβέρνησης, με αποτέλεσμα οι συνολικές κρατικές δαπάνες του Βελγίου ως ποσοστό του ΑΕΠ και ανά χίλιους κατοίκους να σημειώνουν τελικά ελάχιστη πτώση το 2015 και γενικότερα να είναι οι δεύτερες υψηλότερες δαπάνες μετά τις δαπάνες του Ολλανδικού κράτους για την συγκεκριμένη υγειονομική παροχή. Οι κατά κεφαλήν δαπάνες ΤΚΑ της Εσθονίας, Πορτογαλίας, Ελλάδας και Ισπανίας για υπηρεσίες μακροχρόνιας ενδονοσοκομειακής φροντίδας ασθενών είναι αρκετά χαμηλές, οι χαμηλότερες μεταξύ των υπολοίπων χωρών. Οι τέσσερις αυτές χώρες ταξινομούνται στην ίδια κλάση αναφορικά με τις κατά κεφαλήν δαπάνες ΤΚΑ. Όμως όσον αφορά τις κατά κεφαλήν συνολικές κρατικές δαπάνες η Ισπανία αποτελεί ξεχωριστή κλάση. Όπως έχει αναφερθεί οι δαπάνες ΤΚΑ της Ισπανίας για την προαναφερθείσα υγειονομική παροχή είναι μηδενικές, επομένως σε αυτή την περίπτωση οι συνολικές δαπάνες του ισπανικού κράτους είναι δαπάνες του κρατικού προϋπολογισμού. Μεταξύ των χωρών, η Ελλάδα διαθέτει τις χαμηλότερες δαπάνες τόσο ως ποσοστό του ΑΕΠ όσο και ανά χίλιους κατοίκους για υπηρεσίες μακροχρόνιας ενδονοσοκομειακής φροντίδας.

6.5 Κρατική χρηματοδότηση φαρμακευτικών αγαθών βραχυπρόθεσμης χρήσης

Όσον αφορά τον τρόπο κρατικής χρηματοδότησης φαρμακευτικών αγαθών βραχυπρόθεσμης χρήσης διαπιστώθηκε ότι το Βέλγιο, η Ολλανδία, η Εσθονία, η Ελλάδα και η Γερμανία ταξινομούνται στην ίδια κλάση, όπως βρέθηκε στα δεδομένα, κύριος κρατικός χρηματοδοτικός φορέας των χωρών αυτών για την συγκεκριμένη υγειονομική δραστηριότητα είναι τα ΤΚΑ. Αναλυτικότερα, για τις παραπάνω χώρες το ποσοστό των συνολικών κρατικών δαπανών που αποτελούν δαπάνες ΤΚΑ δεν είναι μικρότερο του 90%. Αντιθέτως, το ποσοστό των συνολικών κρατικών δαπανών της Ισπανίας και της Πορτογαλίας που αντιστοιχεί σε δαπάνες ΤΚΑ για φαρμακευτικά αγαθά βραχυπρόθεσμης χρήσης είναι για κάθε έτος αρκετά χαμηλό σχεδόν μηδενικό, με αποτέλεσμα οι δύο αυτές χώρες να ταξινομούνται στην ίδια κλάση σχετικά με τον τρόπο κρατικής χρηματοδότησης. Από το 2004 έως το 2012, τα ποσοστά του ετήσιου ΑΕΠ της Ελλάδας που αντιστοιχούν σε δαπάνες ΤΚΑ και σε συνολικές κρατικές δαπάνες είναι αρκετά υψηλά συγκριτικά με τα αντίστοιχα ετήσια ποσοστά ΑΕΠ των υπολοίπων χωρών, ενώ παρουσιάζουν πτωτική τάση από το 2011. Ιδιαίτερα αυξημένες είναι και οι κατά κεφαλήν δαπάνες της Ελλάδας για φαρμακευτικά αγαθά μη μακροχρόνιας χρήσης, οι οποίες φθίνουν συνεχώς από το 2010 έως το 2014. Η Γερμανία διαθέτει επίσης υψηλές φαρμακευτικές δαπάνες και αξίζει να αναφερθεί ότι από το 2013 έως το 2018 ενώ οι συνολικές κρατικές δαπάνες και οι δαπάνες ΤΚΑ της Γερμανίας ως ποσοστό του ΑΕΠ δεν διαφέρουν κατά πολύ από τις ανάλογες δαπάνες του Ελληνικού κράτους, οι αντίστοιχες κατά κεφαλήν δαπάνες της Γερμανίας είναι αρκετά υψηλότερες από εκείνες της Ελλάδας. Η Ελλάδα και η Γερμανία σε κάθε χρονοδιάγραμμα σχηματίζουν η κάθε μια τη δική της κλάση. Ολλανδία και Εσθονία ταξινομούνται στην ίδια κλάση αναφορικά με τις δαπάνες ΤΚΑ ως ποσοστό του ΑΕΠ για φαρμακευτικά αγαθά βραχυπρόθεσμης χρήσης, όπως παρατηρήθηκε δεν παρουσιάζουν έντονες ετήσιες αποκλίσεις. Ωστόσο, οι αντίστοιχες κατά κεφαλήν δαπάνες ΤΚΑ της Ολλανδίας είναι αρκετά υψηλότερες από τις ανάλογες δαπάνες της Εσθονίας και η κάθε μια ταξινομείται σε δική της κλάση. Δεν πρέπει να παραληφθεί, ότι μεταξύ των χωρών οι συνολικές κατά κεφαλήν δαπάνες του Εσθονικού κράτους για φαρμακευτικά αγαθά βραχυπρόθεσμης χρήσης είναι οι χαμηλότερες.

6.6 Κρατική χρηματοδότηση θεραπευτικών συσκευών μακροχρόνιας χρήσης

Αναφορικά με τον τρόπο κρατικής χρηματοδότησης θεραπευτικών συσκευών μακροχρόνιας χρήσης Ελλάδα και Βέλγιο ταξινομούνται στην ίδια κλάση, όπως παρατηρήθηκε, οι συνολικές κρατικές δαπάνες των δύο αυτών χωρών για την συγκεκριμένη υγειονομική παροχή είναι σχεδόν εξ' ολοκλήρου δαπάνες ΤΚΑ. Γερμανία και Ολλανδία επίσης ταξινομούνται στην ίδια κλάση, με κύριο κρατικό χρηματοδοτικό φορέα των χωρών αυτών για θεραπευτικές συσκευές διαρκείας τα ΤΚΑ. Αντιθέτως το μεγαλύτερο μέρος των συνολικών κρατικών δαπανών της Ισπανίας, Εσθονίας και Πορτογαλίας αναφέρεται σε δαπάνες του κρατικού προϋπολογισμού. Ωστόσο, κάθε μία από τις τρεις αυτές χώρες, όσον αφορά τον τρόπο κρατικής χρηματοδότησης, σχηματίζει ξεχωριστή κλάση. Αξίζει να αναφερθεί ότι, για την Ισπανία από τις υγειονομικές παροχές που μελετήθηκαν μέχρι στιγμής, μόνο στην χρηματοδότηση θεραπευτικών συσκευών διαρκείας η συνεισφορά των ΤΚΑ δεν είναι σχεδόν μηδενική, αλλά κυμαίνεται μεταξύ του 20 έως 25% περίπου. Η ταξινόμηση των χωρών που προκύπτει σχετικά με τις δαπάνες ΤΚΑ ως ποσοστό του ετήσιου ΑΕΠ ταυτίζεται με την ταξινόμηση που προκύπτει όταν οι δαπάνες αυτές εκφράζονται ανά χίλιους κατοίκους. Όσον αφορά λοιπόν τους συγκεκριμένους στατιστικούς δείκτες Εσθονία, Πορτογαλία και Ισπανία είναι οι μοναδικές χώρες που ταξινομούνται σε κοινή κλάση, όπως παρατηρήθηκε, διαθέτουν τις χαμηλότερες δαπάνες ΤΚΑ ως ποσοστό του ΑΕΠ

και ανά χίλιους κατοίκους και παρουσιάζουν παρόμοια χρονική εξέλιξη. Γερμανία και Ολλανδία καταγράφουν τις υψηλότερες συνολικές κρατικές δαπάνες και δαπάνες ΤΚΑ τόσο ως ποσοστό του ΑΕΠ όσο και κατά κεφαλήν για θεραπευτικές συσκευές διαρκείας. Ωστόσο, Γερμανία και Ολλανδία δεν ταξινομούνται στην ίδια κλάση. Μεταξύ των δυο αυτών χωρών, παρατηρήθηκαν ετήσιες διαφορές στις συγκεκριμένες δαπάνες. Συγκεκριμένα, παρατηρήθηκε ότι ενώ οι κατά κεφαλήν δαπάνες Ολλανδίας και Γερμανίας σχεδόν ταυτίζονται από το 2004 έως το 2013, το 2014 οι κατά κεφαλήν δαπάνες της Γερμανίας αυξάνουν και τελικά σταθεροποιούνται έως το 2018 και συγχρόνως οι αντίστοιχες κατά κεφαλήν δαπάνες της Ολλανδίας παρουσιάζουν μείωση και η ετήσια διαφορά μεταξύ των δαπανών των δύο αυτών χωρών διευρύνεται. Ιδιαίτερα αυξημένες ως ποσοστό του ετήσιου ΑΕΠ είναι και οι κρατικές δαπάνες της Ελλάδας για θεραπευτικές συσκευές διαρκείας και μάλιστα το 2016 είναι υψηλότερες μεταξύ των υπολοίπων χωρών. Ωστόσο, οι αντίστοιχες κατά κεφαλήν δαπάνες της Ελλάδας είναι μικρότερες σε σχέση με τις κατά κεφαλήν δαπάνες της Γερμανίας και της Ολλανδίας και φθίνουν συνεχώς από το 2009 έως το 2013. Το ποσοστό του ετήσιου ΑΕΠ που αποτελεί συνολικές κρατικές δαπάνες και δαπάνες ΤΚΑ της Ελλάδας για θεραπευτικές συσκευές μακροχρόνιας χρήσης, αλλά και οι αντίστοιχες κατά κεφαλήν δαπάνες παρουσιάζουν συνεχή άνοδο από το 2013 έως το 2016. Στην Ισπανία καταγράφονται οι χαμηλότερες συνολικές κρατικές δαπάνες κατά κεφαλήν και ως ποσοστό του ΑΕΠ για θεραπευτικές συσκευές διαρκείας.

6.7 Κρατική χρηματοδότηση υπηρεσιών μακροχρόνιας περίθαλψης ασθενών στο σπίτι

Το μεγαλύτερο μέρος, ποσοστό μεγαλύτερο του 90%, της κρατικής χρηματοδότησης υπηρεσιών μακροχρόνιας φροντίδας ασθενών στο σπίτι για την Ολλανδία, Πορτογαλία και Γερμανία πραγματοποιείται από ΤΚΑ και έτσι αναφορικά με τον τρόπο κρατικής χρηματοδότησης οι χώρες αυτές ταξινομούνται σε κοινή κλάση. Το ποσοστό των συνολικών κρατικών δαπανών της Εσθονίας που αντιστοιχεί σε δαπάνες ΤΚΑ έως το 2015 είναι μεγαλύτερο του 90%, αλλά φθίνει ραγδαία, κάτω του 50%, το 2017 και διατηρείται σταθερό και κατά το επόμενο έτος. Όπως παρατηρήθηκε, η Εσθονία, όσον αφορά τις δαπάνες ΤΚΑ προς συνολικές κρατικές δαπάνες, σχηματίζει ξεχωριστή κλάση. Αναφορικά με τον ίδιο στατιστικό δείκτη, κάθε μία από τις χώρες Βέλγιο και Ισπανία ταξινομούνται σε ξεχωριστή κλάση. Από το 2004 έως το 2015 στο Βέλγιο κατά το ήμισυ κύριος κρατικός χρηματοδοτικός φορέας υπηρεσιών μακροχρόνιας περίθαλψης ασθενών στο σπίτι είναι τα ΤΚΑ, ενώ για την συγκεκριμένη υγειονομική παροχή από το 2016 έως το 2018 κύριος κρατικός χρηματοδοτικός φορέας είναι οι φορείς της κεντρικής κυβέρνησης. Η συγκεκριμένη υγειονομική παροχή είναι η μοναδική, όπου για κάποια έτη παρατηρείται ότι το ποσοστό των συνολικών κρατικών δαπανών της Ισπανίας που αντιστοιχεί σε δαπάνες ΤΚΑ είναι μεγαλύτερο του 40%. Στην Ελλάδα από το 2004 έως το 2008 δεν υφίσταται κρατική χρηματοδότηση υπηρεσιών μακροχρόνιας φροντίδας ασθενών στο σπίτι και γενικότερα, με εξαίρεση τα έτη 2009 και 2010, οι δαπάνες ΤΚΑ για την συγκεκριμένη υγειονομική παροχή είναι μηδενικές, με αποτέλεσμα να είναι αναμενόμενο η Ελλάδα να διαθέτει μεταξύ των υπολοίπων χωρών, τις χαμηλότερες δαπάνες ως ποσοστό του ΑΕΠ και ανά χίλιους κατοίκους. Με μία πρώτη ματιά διαπιστώθηκε ότι όσο αναφορά τις συνολικές κρατικές δαπάνες και τις δαπάνες ΤΚΑ τόσο ως ποσοστό του ΑΕΠ όσο και ανά χίλιους κατοίκους οι χώρες διαιρούνται σε δύο κατηγορίες. Γερμανία, Ολλανδία και Βέλγιο διαθέτουν υψηλές δαπάνες, ενώ αρκετά χαμηλότερες είναι οι αντίστοιχες δαπάνες της Ισπανίας, Πορτογαλίας, Ελλάδας και Εσθονίας. Η μοναδική διαφορά μεταξύ της ταξινόμησης αναφορικά με τις δαπάνες ΤΚΑ ως ποσοστό του ΑΕΠ και της ταξινόμησης σε σχέση με τις αντίστοιχες δαπάνες ΤΚΑ ανά χίλιους κατοίκους είναι ότι στην περίπτωση των κατά κεφαλήν δαπανών η Ολλανδία ταξινομείται σε κοινή κλάση με την Γερμανία και όχι με το Βέλγιο. Για την

Ολλανδία και την Γερμανία, από το 2004 έως το 2016 οι δαπάνες ΤΚΑ ανά χίλιους κατοίκους σχεδόν ταυτίζονται. Το Βέλγιο σημειώνει τις υψηλότερες συνολικές κρατικές δαπάνες ως ποσοστό του ΑΕΠ, αλλά και κατά κεφαλήν για την συγκεκριμένη υγειονομική παροχή, οι οποίες μάλιστα αυξάνουν ραγδαία το 2016. Οι κλάσεις που προέκυψαν όσον αφορά τις συνολικές κρατικές κατά κεφαλήν δαπάνες είναι αρκετά καλά διαχωρισμένες με το Βέλγιο να μην ταξινομείται σε κοινή κλάση με άλλη χώρα. Γερμανία και Ολλανδία ταξινομούνται στην ίδια κλάση, ενώ οι χώρες με τις χαμηλότερες δαπάνες, δηλαδή Πορτογαλία, Ισπανία, Εσθονία και Ελλάδα σχηματίζουν ενιαία κλάση.

7. Σύγκριση αποτελεσμάτων με βιβλιογραφικές πηγές και συμπεράσματα για τον KmL

Εκπρόσωποι του Οργανισμού Οικονομικής Συνεργασίας και Ανάπτυξης (ΟΟΣΑ) σε αναφορά τους που εκδόθηκε το 2018 (Health at a glance, 2018) μελέτησαν δαπάνες υγείας των Ευρωπαϊκών χωρών. Αναλυτικότερα, μελέτησαν δαπάνες υγείας ως ποσοστό του ΑΕΠ και κατά κεφαλήν δαπάνες υγείας και παρατήρησαν γραφικά την εξέλιξη τους. Έπειτα, κατέληξαν σε παρατηρήσεις σχετικά με τις ετήσιες μεταβολές των δαπανών υγείας της κάθε χώρας, αλλά και σχετικά με το ποιες χώρες παρουσιάζουν παρόμοια εξέλιξη δαπανών υγείας διαχρονικά. Τα αποτελέσματα ταξινόμησης στην παρούσα εργασία επιβεβαιώνονται από τα συμπεράσματα της παραπάνω οικονομικής μελέτης. Διαπιστώνεται λοιπόν ότι ο αλγόριθμος KmL, αποτελεί ένα χρήσιμο εργαλείο για την διερεύνηση των σχέσεων μεταξύ των χωρών, ως προς τον τρόπο και το ύψος της κρατικής χρηματοδότησης των διαφόρων παροχών των Συστημάτων Υγείας και κατ' επέκταση της σχετικής νομοθεσίας για την ασφαλιστική κάλυψη των εργαζομένων και την δυνατότητα πρόσβασης τους σε κάθε είδος ιατροφαρμακευτικής περίθαλψης.

ABSTRACT

Undoubtedly, the organization of a Health Care System as well as the access of its citizens to all kinds of health services should be a priority of every state. Thus, in the present paper we present the expenses of Social Healthcare Funds, but also total state expenses per thousand inhabitants and as a percentage of the annual GDP of the countries of Greece, Germany, Netherlands, Estonia, Belgium and Portugal for inpatient and outpatient curative and rehabilitative care, laboratory services, imaging services, patient transport, long-term care, both at home as well as in the hospital, pharmaceuticals and other medical non-durable goods and therapeutic appliances and other medical durable goods. In addition, illustrated the evolution of the percentage of total state expenses which is expenses of Social Security Funds of the above countries for each of the above health services. Health care expenses, the annual GDP and the population of every country as of Jan. 1st of every year, are published by EUROSTAT. In the sequel, categorized health expenses longitudinal data of each country according to the KmL algorithm. From the longitudinal classification of countries, verified observations of economic studies of Eurostat, the World Health Organization (WHO) and the Organization for Economic Cooperation and Development (OECD) on the organization of Health Systems and the evolution of the expenses of the countries in the same class.

KEY WORDS: Health-Care Expenses, Official Statistics, Cluster Analysis, KmL

ΑΝΑΦΟΡΕΣ

Calinski T. and Harabasz (1974). A dendrite method for cluster analysis. *Communications in statistics*,311,1-27.

Rousseeuw P. (1986). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*. Elsevier Science Publishers B.V.

Kaushik M. and Mathur B. (2014). Comparative Study of k-means and Hierarchical Clustering Techniques, *International Journal of Software and Hardware Research in Engineering*, ISSN No: 2347-4890, Volume 2, Issue 6.

Νιάκας Δ. (2014). Η οικονομική κρίση και οι επιπτώσεις στο ελληνικό σύστημα υγείας, *Ελληνική Επιθεώρηση Διαιτολογίας- Διατροφής* 5(1): 3-7.

Maresso A., Mladovsky P., Thomson S., Sagan A., Karanikolos M., Richardson E., Cylus J., Eretovits T., Jowett M., Figueraw J., Kluge H. (2015). *Economic crisis, health systems and health in Europe: Country experience*, WHO.

Sentenac M., Genolini C., Alacoque X. and Arnaud C. (2015). KmL and KmL3d: R Packages to cluster longitudinal data, *Journal of Statistical Software*, Volume 65, Issue 4.

OECD/European Union (2018). *Health at a glance*.

Ηλεκτρονικές πηγές δεδομένων

a) https://ec.europa.eu/eurostat/databrowser/view/HLTH_SHA11_HCHF__custom_1246476/default/table?lang=en
b) https://ec.europa.eu/eurostat/databrowser/view/demo_pjan/default/table?lang=en
c) https://ec.europa.eu/eurostat/databrowser/view/nama_10_gdp/default/table?lang=en

Ελληνικό Στατιστικό Ινστιτούτο

Πρακτικά 34ου Πανελληνίου Συνεδρίου Στατιστικής (2022) (σελ. 113-121)

Σχέδια Δειγματοληψίας Χαμηλού Προϋπολογισμού σε Υποσύνολα του Επιπέδου R^2

***Χατζημιχαήλ Χριστίνα*[1], *Μπατσιάκα Μαρία*[1], *Φαρμάκης Νικόλαος*[1]**
ΠΜΣ Τμήμα Μαθηματικών Α.Π.Θ.
xristina.k.xatzimixail@gmail.com, marmpat@hotmail.com, farmakis@math.auth.gr

ΠΕΡΙΛΗΨΗ

Η Δειγματοληψία συμβάλλει στη μελέτη διαφόρων τυχαίων μεταβλητών με πολύ καλά αποτελέσματα από άποψη ακρίβειας και κυρίως ταχύτητας και χαμηλού κόστους. Υπάρχουν όμως και ειδικές περιπτώσεις όπου ο προϋπολογισμός της μελέτης μπορεί να έχει και υποπολλαπλάσιο μέγεθος από τα συνήθη μεγέθη. Τέτοια σχέδια δειγματοληψίας είναι περιπτώσεις όπου ο πληθυσμός είναι υποσύνολο (χωρίο) D του διδιάστατου επιπέδου R^2 και υποκαθίσταται από τμήμα μονοδιάστατου υποχώρου του (τόξο μονοδιάστατης καμπύλης C). Η δειγματοληψία γίνεται μόνο στο μονοδιάστατο τμήμα αυτό της C. Έτσι υποβιβάζεται το κόστος θεωρητικά άπειρες φορές και πρακτικά υποβιβάζεται 80 έως και 100 φορές. Στόχος είναι πάντα η εκτίμηση των διαφόρων παραμέτρων της τυχαίας μεταβλητής (τ.μ.) $Z=f(x,y)$ που μελετούμε και ιδίως η εκτίμηση της μέσης τιμής και της διασποράς. Εδώ (x,y) είναι σημείο του χωρίου D. Σε καταστάσεις σαν την ανωτέρω έχουμε συνήθως κάποιες πληροφορίες σχετικά με τη μορφή της συνάρτησης $f(x,y)$ με προσέγγιση ίσως κάποιας (-ων) παραμέτρου (-τρων). Σε μερικές περιπτώσεις μπορεί να επιλέξουμε μεταξύ 2 ή 3 πιθανών μορφών της $f(x,y)$ και φυσικά υιοθετούμε την πιο εύχρηστη, π.χ. πολυωνυμική, τριγωνομετρική, εκθετική, κλπ. Βρίσκουμε στη συνέχεια τις δύο μέσες τιμές της τ.μ. Z: Μία με βάση όλον τον πληθυσμό-χωρίο D και μία με βάση το τόξο της καμπύλης C και τις εξισώνουμε. Η λύση της εξίσωσης αυτής ως προς τις παραμέτρους της εξίσωσης της C προσδιορίζει την καμπύλη C και το τόξο-τομή της C με το χωρίο D. Το τόξο αυτό είναι ο πληθυσμός στόχος απ' όπου παίρνεται το δείγμα με κάποια μέθοδο δειγματοληψίας, π.χ. Συστηματική Δειγματοληψία.

Λέξεις κλειδιά: Δειγματοληψία, τυχαία μεταβλητή, μέση τιμή, διασπορά.

MSC: 62D05; 62E17

1.Εισαγωγή

Θεωρούμε το επίπεδο χωρίο $D \subseteq R^2$, π.χ. ένα ορθογώνιο παραλληλόγραμμο ΑΒΓΔ με διαστάσεις (ΑΒ)=α και (ΒΓ)=β. Κάθε σημείο Α του χωρίου D είναι ένα άτομο του πληθυσμού Π και σε αυτό αντιστοιχεί μία τιμή Z_A της τυχαίας μεταβλητής (τ.μ.) Z που μελετάμε. Επειδή το Α προσδιορίζεται από τις συντεταγμένες του (x,y) για την τ.μ. Z ισχύει $Z=f(x,y)=f(A)$, $\forall A \in D$, όπου η συνάρτηση f έχει σχετικά απλή μορφή. Από προηγούμενες επαφές μας με το πρόβλημα, θεωρούμε ότι έχουμε επαρκή πληροφορία για την μορφή της συνάρτησης της τ.μ. $Z=f(x,y)$. Με τον όρο επαρκής πληροφορία εννοούμε συνήθως ότι είναι γνωστή η μορφή της συνάρτησης Z με προσέγγιση κάποιων παραμέτρων και μία προδειγματοληψία θα βοηθήσει να διαπιστώσουμε ότι η υπόθεση μας ισχύει, ενώ με την κανονική φάση της δειγματοληψίας σε μεταγενέστερο στάδιο θα έχουμε πολύ πιο καθαρή την εικόνα της κατάστασης. Στο σημείο αυτό υιοθετούνται οι παρακάτω συμβολισμοί, οι οποίοι θα μας διευκολύνουν στην ανάπτυξη του θέματος:

(l_1): επίπεδο (O_y, O_z)
(l_2): επίπεδο (O_x, O_z)
(l_3): επίπεδο (O_x, O_y)

$|D|$: εμβαδό του χωρίου D
C: τόξο καμπύλης (C) στο επίπεδο του D
$|C_0|$: μήκος προβολής του τόξου C στον O_x
$|C|$: μήκος του τόξου C

Κύριος στόχος της παρούσας εργασίας είναι να εκτιμηθεί η μέση τιμή της τ.μ. Z, $\bar{Z}$, μόνο μέσω παρατηρήσεων που ανήκουν στο τόξο καμπύλης C και συνήθως υπάρχουν και άλλοι δευτερεύοντες στόχοι, όπως είναι ο υπολογισμός του $\iint_D f(x,y)\,dx\,dy$.

2.Βασική ιδέα της μεθόδου

Η μέθοδος που προτείνεται στην παρούσα εργασία για την εκτίμηση της μέσης τιμής της τ.μ. Z έχει ως εξής: θεωρούμε την συνάρτηση $Z=f(x,y)$ / $D \subseteq R^2$ ολοκληρώσιμη

στο D, υπολογίζουμε το ολοκλήρωμα $\iint_D f(\mathrm{x},\mathrm{y})\,\mathrm{dx}\,\mathrm{dy}$ και με βάση το Θεώρημα Μέσης Τιμής Ολοκληρωτικού Λογισμού (ΘΜΤΟΛ) η μέση τιμή της τ.μ. Z δίνεται από:

$$\bar{Z} = \frac{1}{|\mathrm{D}|}\iint_D f(\mathrm{x},\mathrm{y})\,\mathrm{dx}\,\mathrm{dy} \qquad (1)$$

Έστω τώρα επιφάνεια (l) κάθετη στο (l_3) και με γενέτειρα γραμμή την καμπύλη (C) στο επίπεδο (l_3). Συμβολίζουμε με τ την τομή του l με το γράφημα της Z. Η καμπύλη (C) είναι η προβολή της τομής τ επί του (l_3). Η τομή των (C) και D είναι το τόξο D ∩ (C) = C, με μήκος |C| και με εξίσωση ορισμού $\mathrm{y} = \varphi(\mathrm{x})$.

Στην συνέχεια, βρίσκουμε το ολοκλήρωμα $\int_{\mathrm{c}} f(\mathrm{x},\mathrm{y})\,\mathrm{dx} = \int_{\mathrm{c}} f\big(\mathrm{x},\varphi(\mathrm{x})\big)\,\mathrm{dx}$ και στηριζόμενοι στο ΘΜΤΟΛ συμπεραίνουμε ότι ο εκτιμητής της μέσης τιμής της τ.μ. Z ισούται με:

$$\bar{Z}' = \frac{1}{|\mathrm{C_0}|}\int_{\mathrm{c}} f\big(\mathrm{x},\varphi(\mathrm{x})\big)\,\mathrm{dx} \qquad (2)$$

Σε επόμενο στάδιο θεωρούμε τη διαφορά $\Delta\bar{Z} = |\bar{Z}' - \bar{Z}|$ και είναι προφανές ότι η ποιότητα του εκτιμητή $\bar{Z}'$ εξαρτάται από το μέγεθος του $\Delta\bar{Z}$. Η ιδανική περίπτωση είναι φυσικά να επιτύχουμε τέτοια τομή-τόξο C ώστε να ισχύει $\Delta\bar{Z} = 0$. Αυτό είναι αρκετά δύσκολο και έτσι στην πράξη αρκούμαστε να έχουμε αρκετά μικρή τιμή του $\Delta\bar{Z}$ ή αντίστοιχα του $\Delta\bar{Z} \,/\, \bar{Z}$, ώστε να έχουμε ικανοποιητικές προσεγγίσεις, με βάση πάντα τις εκάστοτε απαιτήσεις μας. Ο λόγος $\Delta\bar{Z} \,/\, \bar{Z}$ είναι η σχετική απόκλιση ως προς $\bar{Z}$ και αποτελεί μέτρο ποιότητας του εκτιμητή.

3.Εκτίμηση της μέσης τιμής της τ.μ. $\mathbf{Z = f(x,y) = ax^v + \beta x^{v-k} y^k}$

Έστω τα μέλη του πληθυσμού είναι στοιχεία του $D = [0,1] \times [0,1]$ με |D|=1 και σύμφωνα με την διαδικασία που περιγράφεται στην προηγούμενη παράγραφο λαμβάνεται υποπληθυσμός, Φαρμάκης (2015,2016) C= $\{\big(\mathrm{x},\mathrm{g}(\mathrm{x})\big), \mathrm{x} \in [0,1], \mathrm{g}(\mathrm{x}) \in [0,1]\}$ από τον οποίο προκύπτει η εκτίμηση για την μέση τιμή της τ.μ. Z. Είναι φανερό ότι ο αρχικός πληθυσμός ανήκει σε χώρο διάστασης 2, στο επίπεδο $(\mathrm{O_x},\mathrm{O_y})$, ενώ τα μέλη του υποπληθυσμού είναι στοιχεία μονοδιάστατου υποχώρου.

Υποθέτουμε τώρα ότι $\mathrm{Z}=f(\mathrm{x},\mathrm{y})= \mathrm{ax}^{\mathrm{v}}+\beta\mathrm{x}^{\mathrm{v}-\mathrm{k}}\mathrm{y}^{\mathrm{k}}$, $\mathrm{k},\mathrm{v} > 0$ και θεωρούμε αρχικά τόξο δειγματοληψίας:

C = $\{(\mathrm{x},\mathrm{y}), \mathrm{y} = \mathrm{g}(\mathrm{x}) = \lambda\mathrm{x}, \mathrm{x} \in [0,1], \mathrm{y} \in [0,1], \lambda \in [0,1]\}$ με $|\mathrm{C_0}|$=1, Φαρμάκης (2015,2016).

Η θεωρητική μέση τιμή της τ.μ. Z υπολογίζεται ως εξής:

$$\bar{Z} = \frac{1}{|D|}\iint_D f(x,y)\,dx\,dy = \int_0^1 \int_0^1 (ax^v + \beta x^{v-k}y^k)\,dxdy = \frac{a}{v+1} + \frac{\beta}{(k+1)(v-k+1)}$$

Συνεπώς,

$$\bar{Z} = \frac{a}{v+1} + \frac{\beta}{(k+1)(v-k+1)} \quad (3)$$

Ο εκτιμητής $\bar{Z}'$, ο οποίος προκύπτει από δείγμα ατόμων-σημείων μόνο από το τόξο C δίνεται από:

$\bar{Z}' = \frac{1}{|C_0|}\int_c f\big(x, g(x)\big)\,dx = \int_0^1 (ax^v + \beta x^{v-k}(\lambda x)^k)\,dx = \frac{a}{v+1} + \frac{\beta\lambda^k}{v+1}$. Επομένως,

$$\bar{Z}' = \frac{a}{v+1} + \frac{\beta\lambda^k}{v+1} \quad (4)$$

Ακολούθως, προσδιορίζεται η κλίση λ έτσι ώστε $\Delta\bar{Z} = 0$ ή τουλάχιστον $\Delta\bar{Z} \approx 0$. Ισχύει ότι,

$$\Delta\bar{Z} = 0 \Leftrightarrow \bar{Z} = \bar{Z}' \Leftrightarrow \frac{a}{v+1} + \frac{\beta}{(k+1)(v-k+1)} = \frac{a}{v+1} + \frac{\beta\lambda^k}{v+1} \Leftrightarrow \lambda^k = \frac{v+1}{(k+1)(v-k+1)}$$

Έχοντας υποθέσει ότι $\lambda \in [0,1]$ στο σημείο αυτό προκύπτει ο περιορισμός $k - v < 1$ για τους εκθέτες k, v. Συμπεραίνεται λοιπόν, ότι η τιμή της παραμέτρους λ για την οποία ισχύει $\Delta\bar{Z} = 0$ ισούται με:

$$\lambda = \left(\frac{v+1}{(k+1)(v-k+1)}\right)^{1/k} \quad (5)$$

Γενικεύοντας, θεωρούμε τόξο δειγματοληψίας, Φαρμάκης (2015, 2016)

$C = \{(x,y),\ y = g(x) = \lambda x^\rho,\ x \in [0,1], y \in [0,1], \lambda \in [0,1], \rho \in [0,1]\}$. Τότε,

$$\bar{Z}' = \frac{a}{v+1} + \frac{\beta\lambda^k}{v+1+(\rho-1)k} \quad (6)$$

και $\Delta\bar{Z} = 0$ όταν

$$\lambda^k = \frac{v+1+(\rho-1)k}{(k+1)(v-k+1)} \quad (7)$$

Εφόσον, $\rho \in [0,1]$ αποδεικνύεται ότι:

$$\frac{1}{(k+1)} \leq \lambda^k \leq \frac{v+1}{(k+1)(v-k+1)} \text{ ή } \left(\frac{1}{k+1}\right)^{1/k} \leq \lambda \leq \left(\frac{v+1}{(k+1)(v-k+1)}\right)^{1/k} \quad (8)$$

Σύμφωνα με τις σχέσεις (3) και (6) ισχύει ότι:

$$\Delta\bar{Z} = |\bar{Z}' - \bar{Z}| = \left|\frac{\beta\lambda^k}{v+1+(\rho-1)k} - \frac{\beta}{(k+1)(v-k+1)}\right| = |\beta| \left|\frac{\lambda^k}{v+1+(\rho-1)k} - \frac{1}{(k+1)(v-k+1)}\right|$$

Εφόσον έχουμε υποθέσει ότι $\lambda \in [0,1]$, $\rho \in [0,1]$, $v, k \in N^*$ και ειδικότερα για τους εκθέτες ν και k ισχύει ο περιορισμός $v - k + 1 > 0$, διαπιστώνεται ότι η ποσότητα

$$\frac{\lambda^k}{v+1+(\rho-1)k} - \frac{1}{(k+1)(v-k+1)}$$

ελαχιστοποιείται για $\rho = 1$. Συνεπώς, για $\rho = 1$ προκύπτει ότι:

$$\Delta\bar{Z} = |\beta| \left|\frac{\lambda^k}{v+1} - \frac{1}{(k+1)(v-k+1)}\right|.$$

Άρα για συγκεκριμένες τιμές των ν, κ και λ επιλέγουμε τη μεγαλύτερη τιμή του ρ (δηλ. την ρ=1) και εργαζόμαστε με την καμπύλη (C): y=λx, Συγκεκριμένα η (C) είναι ο (υπο)πληθυσμός στόχος από τον οποίο θα πάρουμε το δείγμα για την καλύτερη προσέγγιση της μέσης τιμής της τυχαίας μεταβλητής Z(x,y), Φαρμάκης (2001, 2015, 2016), (βλέπε παράδειγμα παρακάτω).

Έχοντας προσδιορίσει την βέλτιστη τιμή του ρ για την οποία η διαφορά $\Delta\bar{Z}$ ελαχιστοποιείται, στόχος είναι να προσδιοριστεί η τιμή της παραμέτρου λ για την οποία ισχύει $\Delta\bar{Z}$=0 ή τουλάχιστον $\Delta\bar{Z} \approx 0$, δηλαδή:

$$\left|\frac{\lambda^k}{v+1} - \frac{1}{(k+1)(v-k+1)}\right| = 0 \text{ ή τουλάχιστον } \left|\frac{\lambda^k}{v+1} - \frac{1}{(k+1)(v-k+1)}\right| \approx 0.$$

Συμπεραίνεται λοιπόν ότι για $\rho = 1$ η διαφορά $\Delta\bar{Z}$ ελαχιστοποιείται ως προς λ όταν

$\lambda^k = \frac{v+1}{(k+1)(v-k+1)}$, δηλαδή όταν $\lambda = \left(\frac{v+1}{(k+1)(v-k+1)}\right)^{1/k}$,

όπου σύμφωνα με την σχέση (8) αποδεικνύεται ότι είναι η μέγιστη τιμή που δύναται να λάβει η παράμετρος λ.

Κατόπιν των ανωτέρω σχετικά με τις παραγώγους ενδιαφέρον παρουσιάζει ο παρακάτω Πίνακας 1 όπου εμφανίζονται τα όρια των τιμών του λ για τις διάφορες τιμές των εκθετών ν και k. Το γεγονός ότι έχουμε τη μορφή της $g(x)=y=\lambda x^{\rho}$ σημαίνει ότι η καμπύλη (C) περνάει από την αρχή των αξόνων (0,0). Επίσης ονομάζουμε:

$A=\left(\frac{1}{k+1}\right)^{1/k}$ και $B=\left(\frac{\nu+1}{(k+1)(\nu-k+1)}\right)^{1/k}$.

Ακολουθεί ο Πίνακας 1 για τιμές ν=1,2,3,4 και τιμές του k=1,2,…,ν.

Να σημειωθεί ότι η καμπύλη (C) περνάει οριακά προς τα πάνω από το σημείο (1,B) γενικά. Οριακά προς τα πάνω πάλι από το σημείο (1,1), όταν είναι k=ν, όπως φαίνεται από την έκφραση του B ανωτέρω. Η οριακή τιμή προς τα κάτω είναι η ½ όταν έχουμε k=1 για κάθε τιμή του ν διότι το ν δεν εμφανίζεται στην έκφραση του A, του κάτω ορίου του λ.

Πίνακας 1

ν	***k***	**A**	$\mathbf{A \le \lambda \le B}$	**B**
1	1	$\frac{1}{2}$	**λ**	1
2	1	$\frac{1}{2}$	**λ**	$\frac{3}{4}$
2	2	$\frac{\sqrt{3}}{3}$	**λ**	1
3	1	$\frac{1}{2}$	**λ**	$\frac{2}{3}$
3	2	$\frac{\sqrt{3}}{3}$	**λ**	$\frac{\sqrt{6}}{3}$
3	3	$\frac{\sqrt[3]{2}}{2}$	**λ**	1
4	1	$\frac{1}{2}$	**λ**	$\frac{5}{8}$
4	2	$\frac{\sqrt{3}}{3}$	**λ**	$\frac{\sqrt{5}}{3}$
4	3	$\frac{\sqrt[3]{2}}{2}$	**λ**	$\frac{\sqrt[3]{5}}{2}$
4	4	$\frac{\sqrt[4]{125}}{5}$	**λ**	1

4.Εφαρμογή για ν=2 και k=1

Στον ακόλουθο πίνακα δίνονται οι εκτιμητές για την μέση τιμή της τ.μ. $Z = f(x,y) = ax^2 + \beta xy$ για τις περιπτώσεις όπου

$C = \{(x,y), y = g(x) = \lambda x, x \in [0,1], y \in [0,1], \lambda \in [0,1]\}$ και
$C = \{(x,y), y = g(x) = \lambda\sqrt{x}, x \in [0,1], y \in [0,1], \lambda \in [0,1]\}$ με $|C_0|=1$.

Σε κάθε περίπτωση υπολογίζεται η διαφορά $\Delta\bar{Z}$ και σύμφωνα με την μεθοδολογία που αναπτύχθηκε παραπάνω προσδιορίζεται η κλίση λ, έτσι ώστε $\Delta\bar{Z} = 0$ ή τουλάχιστον $\Delta\bar{Z} \approx 0$. Για την τ.μ. $Z = f(x,y) = ax^2 + \beta xy$ ισχύει ότι $\bar{Z} = \frac{4a+3\beta}{12}$

Πίνακας 2

C	$\bar{Z}'$	λ	$\Delta\bar{Z}$
$y = \lambda x$	$\frac{a+\beta\lambda}{3}$	$3/4$	$\frac{\beta(4\lambda-3)}{12}$
$y = \lambda\sqrt{x}$	$\frac{5a+6\beta\lambda}{15}$	$5/8$	$\frac{\beta(8\lambda-5)}{20}$

Συμβολίζουμε $\bar{Z}_1' = \frac{a+\beta\lambda}{3}$, $\bar{Z}_2' = \frac{5a+6\beta\lambda}{15}$ και κατ' επέκταση $\Delta\bar{Z}_1 = \frac{\beta(4\lambda-3)}{12}$ και $\Delta\bar{Z}_2 = \frac{\beta(8\lambda-5)}{20}$.

Επειδή, $\Delta\bar{Z}_1 - \Delta\bar{Z}_2 = \frac{-\beta\lambda}{15}$ προκύπτει ότι για $\beta > 0, \Delta\bar{Z}_1 < \Delta\bar{Z}_2$. Συμπεραίνεται δηλαδή, ότι μεταξύ των περιπτώσεων όπου $C = \{(x,y), y = g(x) = \lambda x, x \in [0,1], y \in [0,1], \lambda \in [0,1]\}$ και
$C = \{(x,y), y = g(x) = \lambda\sqrt{x}, x \in [0,1], y \in [0,1], \lambda \in [0,1]\}$
η μέση τιμή της τ.μ. $Z = f(x,y) = ax^2 + \beta xy$ είναι προτιμότερο να εκτιμηθεί από άτομα-μέλη που ανήκουν στο τόξο $C = \{(x,y), y = g(x) = \lambda x, x \in [0,1], y \in [0,1], \lambda \in [0,1]\}$.

<u>Παράδειγμα:</u> Στο χωρίο $D=\{(x,y): x \in [0,1], y \in [0,1]\}$ ορίζεται η τ.μ. $Z(x,y)= ax^2 + \beta xy$. Να εκτιμηθεί η μέση τιμή της Z στο χωρίο *D*. Δίνεται ότι είναι $\beta>0$.

<u>Απάντηση</u>: Σύμφωνα με τα παραπάνω (βλέπε σελίδα 5), η μέση τιμή της Z είναι προτιμότερο να εκτιμηθεί από την $y=\lambda x$, δηλαδή είναι $\rho=1$. Συμπεραίνουμε ότι πρέπει να πάρουμε δείγμα από το τμήμα της ευθείας με εξίσωση $y=\lambda x$ με το μέγιστο επιτρεπόμενο λ που προβλέπεται από την (8) για $\nu=2$ και $k=1$. Από τη δεύτερη γραμμή του πίνακα 1 (δηλαδή $\nu=2$ και $k=1$) προκύπτει $\lambda=0.75$, ήτοι το τόξο που θα είναι ο πληθυσμός στόχος θα είναι τμήμα της ευθείας

(C): y=0.75x.

Η μέση τιμή της Z = $f(x,y) = ax^2 + \beta xy$ στο χωρίο *D* είναι $\bar{Z} = \frac{a}{3} + \frac{\beta}{4} = \frac{4a+3\beta}{12}$. Θα πάρουμε ένα συστηματικό δείγμα από n+1 σημεία της y=0.75x, Φαρμάκης (2015, 2016). Τα σημεία αυτά θα είναι της μορφής:

$(x_i, y_i) = \left(\frac{i}{n}, \frac{3i}{4n}\right)$,i=0,1,2,3…,n, οπότε η εκτίμηση της μέσης τιμής της τ.μ. *Z*, η $\bar{Z}'_{(n)}$, προκύπτει

$z_i = ax_i^2 + \beta x_i y_i = \frac{ai^2}{n^2} + \frac{3\beta i^2}{4n^2} = \frac{(4a+3\beta)i^2}{4n^2}$ και $\bar{Z}'_{(n)} = \frac{4a+3\beta}{4n^2(n+1)} \sum_{i=0}^{n} i^2 = (4a+3\beta)\frac{2n+1}{24n}$

Η εκτίμηση της μέσης τιμής είναι συνάρτηση του μεγέθους n+1 του δείγματος. Μάλιστα όταν η δειγματοληψία είναι συστηματική λέγεται ότι η εκτίμηση της μέσης τιμής της τ.μ. *Z*, η $\bar{Z}'_{(n)}$, είναι συνάρτηση του πλήθους n των τμημάτων ίσου μήκους στα οποία χωρίζεται η πλευρά [0, 1] του χωρίου *D* πάνω στον άξονα των x. Το όριο της εκτίμησης αυτής καθώς το n τείνει στο άπειρο είναι:

$\bar{Z}' = \lim_{n\to\infty} \bar{Z}'_{(n)} = \frac{4a+3\beta}{12}$

Για μέγεθος δείγματος n+1 η διαφορά $\Delta\bar{Z}$ συναρτήσει του n είναι:

$\Delta\bar{Z} = \bar{Z}'_{(n)} - \bar{Z} = (4a+3\beta)\frac{2n+1}{24n} - \frac{4a+3\beta}{12} = \frac{4a+3\beta}{12}\left(\frac{2n+1}{2n} - 1\right) = \frac{4a+3\beta}{24n}$.

Προφανώς το όριο καθώς το n τείνει στο άπειρο είναι $\Delta\bar{Z} = \bar{Z}'_{(n)} - \bar{Z} = 0$.

Η σχετική απόκλιση ως προς $\bar{Z}$ είναι $\frac{\Delta\bar{Z}}{\bar{Z}} = \frac{1}{2n}$.

ABSTRACT

Sampling contributes to the study of various random variables with very good results in terms of accuracy and mainly speed and low budget. However, there are special cases when the study budget may be smaller than the usual size. Such sampling designs where the population is a subset D of the two-dimensional level R^2 and is substituted by a part of its one-dimensional subspace (arc of the one-dimensional curve C). Sampling applies only to the one-dimensional part of the curve C. Thus, the budget is reduced theoretically infinite times and practically it is reduced 80 to 100 times. The aim is the estimation of the various parameters of the random variable *Z=f(x,y)* that we study and especially the estimation of the mean and the variance. Here (*x,y*) is a point of the D. In situation like the one above we usually have some information above the form of the function *f*(*x,y*) by approaching perhaps some parameter(s). In some cases, we can choose between 2 or 3 possible forms of f (x, y) and of course we adopt the easiest to

use, e.g., polynomial, trigonometric, exponential, etc. We find the two means of Z: One based on the whole population D and one based on the arc of the curve C and equate them. The solution of this equation determines the curve C and the arc-cut of C with the D. This arc is the target population from which the sample is taken by some sampling method, e.g., Systematic Sampling.

Key words: sampling, random variable, mean, variance

ΑΝΑΦΟΡΕΣ

Φαρμάκης Ν. (2001) «*ΣΤΑΤΙΣΤΙΚΗ, Περιληπτική Θεωρία-Ασκήσεις*», Α & Π Χριστοδουλίδη, Θεσσαλονίκη

Φαρμάκης Ν. (2015) «*Δειγματοληψία και Εφαρμογές*» e-book, https://repository.kallipos.gr/ & www.kallipos.gr

Φαρμάκης Ν. (2016) «*Εισαγωγή στη Δειγματοληψία*», Εκδόσεις Αφοί Κυριακίδη Α.Ε, Θεσσαλονίκη.


Greek Statistical Institute
Proceedings of the 34rd Panhellenic Statistics Conference (2022), pp. 122–136

STOCHASTIC EPIDEMIC MODELLING OF COVID-19

A. Apsemidis[1], **N. Demiris**[1]
[1]Athens University of Economics and Business
{apsemidis, nikos}@aueb.gr

ABSTRACT

Covid-19 is undoubtedly a pandemic that the international society will remember and one of the most important events of the 21st century that has costed many lives and has had a large social impact and economic consequences. The disease responsible for a large number of deaths worldwide raised much scientific interest from a statistician's perspective and different models are proposed in the literature. The novelty of our methods lie on the fact that they are based on the hidden information of the total – rather than the observed – cases, since the disease is really expanding via cases that are not recorded. Using the compartmental type of epidemic models, we propose various ways of tackling the problem taking into account characteristics of Covid-19 that are known in the literature. The fitting procedure has been performed on data from many countries other than Greece using the Bayesian way of thinking via Hamiltonian Monte Carlo (HMC).

Keywords: Covid-19, epidemic, Bayesian, HMC

1. INTRODUCTION

Since December of 2019, Covid-19 has been spread on the whole world with millions of cases that suffer from simple symptoms till serious, chronic ones and even death. Due to the high transmissibility of Covid-19 (its R_0 was estimated to be approximately 3.8 in Flaxman et al., 2020), it is imperative that we monitor its progress for performing proper decision-making regarding intervention measures against it. The statistical models focus on estimating useful quantities about the epidemic in many countries and guide this procedure. There exists

a spectacularly large amount of literature regarding Covid-19, but the majority of epidemic modelling in general works on the level of observed cases (see for example Andersson and Britton, 2012 and Bjornstad, 2018). In this article we present such kind of models, but we extend their foundational ideas to models that work on the total number of cases, which include the unobserved ones. The most basic model, which is always used as a starting point, is the SIR model on the continuous-time domain. In this article, we are concerned with discrete-time models, but it is useful to present the original SIR to gain insight about the interpretation and intuition of the equations that follow next.

The simple SIR model is given by the following system of ordinary differential equations

$$\begin{aligned} \dot{S} &= -\lambda \cdot S \cdot I/N \\ \dot{I} &= \lambda \cdot S \cdot I/N - I/\tau \\ \dot{R} &= I/\tau \end{aligned} \tag{1}$$

where λ is the infection rate, τ is the infectious period, S and I are the number of susceptible and infectious individuals respectively and N is the size of the population. The dot symbol on top of a variable indicates differentiation with respect to t. This model sets the base for our proposed ones in the present article.

All of the described models posit a Poisson or Negative Binomial distribution on the data (either daily cases or daily deaths) and we refer to the mean of the assumed distribution at time t as θ_t. When both daily cases c_t and daily deaths d_t are utilized, we make the distinction $\theta_t^{(c)}$ and $\theta_t^{(d)}$ to denote the mean cases and deaths respectively. From now on, we use the Negative Binomial distribution when describing a likelihood, but Poisson models have also been trained for comparison. The Negative Binomial is expressed in terms of mean and dispersion parameter. The goodness of fit and predictions were assessed by information criteria, criteria regarding predictions accuracy and visual inspection.

At the beginning of the epidemic, Google and Apple started publishing mobility information for every country, which could potentially be used in the inference or prediction procedure. However, these mobility data (of eight variables) needed to be merged and processed in a sensible way in order to be utilized. We briefly discuss about this in the next Section, but actually the mobility variables took a lot of our time and effort in order to gain insights about them and relate

them to the epidemic. We do not include such an analysis to this article in order to focus on the epidemic modelling. Other data sources we use for training our models are the number of daily vaccinations, the age distribution of cases and the number of daily tests performed as discussed in the next Sections.

The remainder of the article is organised as follows. In Section 2, many models capable of inference or prediction are presented based on the observed Covid-19 data, while we also provide more advanced models that work on the latent level of total (both registered and unregistered) cases. In Section 3, some results of the best models are shown. Section 4 concludes the article. For all the analyses we use the R programming language and the model fitting procedures were performed using the Stan probabilistic programming language through R. Last but not least, we use the Bayesian methodology for all our models, which we train using HMC and the algorithm NUTS (see Hoffman and Gelman, 2014).

2. DISCRETE-TIME STOCHASTIC MODELS

As far as the mobility data are concerned, it seemed sensible that they could be used as a covariate on the infection rate λ_t, since larger mobility would correspond to more contacts. After experimenting on different methods, such as Principal Components Analysis (PCA), Independent Components Analysis and Factor Analysis we investigated whether such a procedure performs better than just keeping the variable with highest correlations with the others. We concluded that PCA is a straightforward and sensible way to proceed, so the next step was to decide on the number of Principal Components (PC) to keep. After trying models with different number of PC's, as well as an analysis regarding on the information discarding when dimensions are reduced, we decided to work with the first PC. Thus, when we refer to the mobility variable m_t, we mean the first PC of the eight variables, which are "Retail and recreation", "Grocery and pharmacy", "Parks", "Transit stations", "Workplaces", "Residential", "Driving" and "Walking". The first six of the original variables refer to percent change of mobility in places like restaurants (for the first variable), groceries (for the second variable) etc. gathered via cell phones by Google, while the last two refer to percent change in driving and walking as captured by requests on the maps application gathered by Apple.

According to the SIR formulation (1), the mean Covid-19 cases θ_t can be estimated by the quantity $\lambda_{t-1}S_{t-1}I_{t-1}/N$. Using backward differences, we trans-

form the cumulative number of cases, deaths and recoveries provided by the Johns Hopkins University (`https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series`) into daily data and construct the number of susceptible individuals by $S_t = N - \sum_{i=1}^{t} c_i$ and the number of infectious individuals by $I_t = c_t^c - d_t^c - r_t^c$, where c_t^c, d_t^c and r_t^c are the cumulative number of cases, deaths and recoveries at time t respectively. The parameter λ_t is written in various ways, some of which utilize mobility information.

One class of models we have trained is the following:

$$\begin{aligned} c_t &\sim NB(\theta_t, \psi) \\ \theta_t &= \eta_t / N \\ \eta_t &= \lambda_t S_{t-1} I_{t-1} + \phi_1 \eta_{t-1} \\ \log(\lambda_t) &= \beta_0 + \beta_1 m_t + \phi_2 log(\lambda_{t-1}) \end{aligned}$$

where we investigated removing terms such as any of the two autoregressive ones, or the mobility effect. The last equation is the one that incorporates the mobility data and we also tried to regress on many past days using $log(\lambda_t) = \beta_{0,t} + \sum_{i=1}^{7} \beta_i m_{t-i} + \phi log(\lambda_{t-1})$ (i.e. we put coefficients on the whole past week before each time t) or write it in a mixed effects fashion as $log(\lambda_t) = \beta_{0,t} + \beta_{m,t} + \phi log(\lambda_{t-1})$. Moreover, we fitted models with alternative combinations of fixed and random intercepts and fixed and random mobility effects. The mixed effects formulation allowed to use the mobility data in a special way. Instead of placing a non-informative Gaussian prior on the coefficient of m_t, we performed a moving-blocks Bootstrap scheme to estimate the error of the mobility PC and used a Gaussian prior centred at the observed PC with variance the one obtained by Bootstrap.

Another class of models was one that makes a distinction between cases that are led to death or recovery. Thus, we trained a mixture of two Negative Binomials – one for each group of either death ($j = 1$) or recovery ($j = 2$) – with mean θ_j / N where

$$\begin{aligned} \theta_{jt} &= \lambda_{jt} \cdot S_{t-1} I_{t-1} + \phi_{1j} \theta_{jt-1} \\ log(\lambda_{jt}) &= \beta_{0j} + \beta_{1j} m_t + \beta_{2j} x_{jt} + \phi_{2j} log(\lambda_{jt-1}) \end{aligned}$$

with x_{jt} being the number of deaths or recoveries at time t for $j = 1$ or $j = 2$ respectively. Again, removing terms from this general form was under study in order to build a type of SIRD model.

The construction of the reproduction number is straightforward by $R_t = \lambda_t \cdot \tau$. For the infectious period τ we tested either random draws from Half-Normal or Gamma distributions with a mean of approximately 7 days (which was a plausible choice), but we also tried to estimate it from the available data. To this end, we designed a "first-infected, first-died" or "first-infected, first-recovered" method to approximate the time-to-death, time-to-recovery and time-to-removal by using only information of the daily cases, deaths and recoveries. Then, we fitted a Gamma distribution by numerically maximizing the likelihood using the BFGS method and used random draws from this Gamma to scale the infection rate. Other versions of tested models just use a fixed number of six days.

It is known for Covid-19 that when one is infected they do not become infectious immediately, but they experience an initial exposed period. The exposed period is a time interval during which the infected individual is not yet capable to transmit the disease. In our initial models we incorporated such a period of four days by shifting the number of susceptible individuals that are available at each time t, i.e. $S_t^{new} = S_{t+4}$. Thus, the SIR and SIRD models become SEIR and SEIRD respectively.

Another important part of our research was the number and position of change-points to the model parameters. Covid-19 is an epidemic that lasts for more than two years and affects the population differently during different time intervals. Due to this change in the disease characteristics change-points on the model parameters are essential to accurately describe it. For instance, one way to break λ_t was to train different intercepts $c_{1,t}$ and $c_{2,t}$ in the log-equation, i.e.

$$\begin{aligned} log(\lambda_t) &= \beta_{0,t} + \beta_1 m_t + \phi log(\lambda_{t-1}) \\ \beta_{0,t} &= p \cdot c_{1,t} + (1-p) \cdot c_{2,t} \\ p &= (1 + \exp(-K))^{-1} \\ K &= 2 \cdot (t - T) \end{aligned}$$

where T is the time of change.

Two cases when we used different kind of likelihoods were first, when we experimented on data from other countries, where we assumed exchangeability

of their parameters and second, when we used a bivariate Negative Binomial that incorporates information from both recorded cases and deaths. Regarding the first case, we investigated the cases of Greece, Italy, Germany, Sweden, Cyprus, Finland, Netherlands and United Kingdom, by placing the likelihood

$$L = \prod_{i=1}^{W} \prod_{t=1}^{n_i} NB(c_{t,i}; \theta_{t,i}, \psi_i)$$

where W is the number of countries considered and i indicates a specific country with sample size n_i. Regarding the bivariate likelihood, we tried versions of the following model:

$$\begin{aligned} c_t &\sim NB(\theta_t^{(c)}, \psi_1) \\ d_t &\sim NB(\theta_t^{(d)}, \psi_2) \\ \theta_t^{(c)} &= \lambda_t S_{t-1} I_{t-1}/N \\ log(\lambda_t) &= \beta_0 + \beta_1 m_t + \phi log(\lambda_{t-1}) \\ \theta_t^{(d)} &= \sum_{j=1}^{t-1} p_t \pi_{t-j} \theta_j^{(c)} \end{aligned}$$

where p_t is the infection fatality ratio (IFR) and π_s is the percentage of deaths due to cases from s days ago. This percentage is taken by the discretized density of the infection-to-death distribution $\pi(t)$ (taken from Flaxman et al., 2020) as

$$\pi_s = \int_{s-0.5}^{s+0.5} \pi(t) dt$$

Finally, we performed one-week-ahead predictions for the number of cases and deaths using the predictive distribution of the parameters. Thus, we trained the model until time n_1 and predicted the period $n_1 + 1, ..., n_1 + 7$, then we repeated this procedure training until $n_2 > n_1$ and so on, therefore predicting the next week after several time periods.

While working with the recorded cases can be helpful and direct for decision making, there exists a large number of unregistered cases that drive the epidemic. Thus, we followed many paths to estimate the total number of cases C_t, like

connecting linearly the total cases with the mean registered cases as in $C_t = \theta_t^{(c)} + x_t$ or $\theta_t^{(c)} = \omega_t \cdot C_t$. The former assumes that x_t is the number of unregistered cases and the latter assumes that ω_t is the proportion of the cases we observe. Last but not least, the mean daily deaths were changed to $\theta_t^{(d)} = p_t \cdot \sum_{j=1}^{t-1} \pi_{t-j} \cdot C_j$.

We placed different prior distributions on the parameters and tested the performance of the models when adding change-points on the infection rate λ_t, the infection fatality ratio p_t and the proportion of observed cases ω_t. As far as the mobility variables is concerned, we tried to pre-process the data by fitting a generalized additive model (GAM) on them and then use the fitted mobility as the covariate m_t so that we reduce some noise.

All of the above formulations have two drawbacks: first, they use the recorded number of susceptible and infected individuals and, second they use the recovery data provided by Johns Hopkins University, which they turned out to be highly untrustworthy. Thus, we present a whole new class of models, that work on the latent total cases utilizing only information by the observed ones.

2.1 Estimating the total number of cases

The discrete-time stochastic epidemic models we presented so far set the foundations for even more complicated and accurate ones, which span a larger time series and take into consideration vaccination, demography and variants of the initial virus. The first novel approach to an epidemic model suitable for Covid-19 is given by the following equations

$$\begin{aligned} d_t &\sim NB(\theta_t, \psi) \\ \theta_t &= p_{(b)} \cdot I\big(t \in (l_b, l_{b+1} - 1)\big) \cdot \sum_{k=1}^{t-1} \pi_{t-k} \cdot C_k \\ C_t &= \lambda_{t-1} S_{t-1} I_{t-1} / N \\ S_t &= S_{t-1} - C_t \\ I_t &= \sum_{i=t-\tau}^{t} C_i \end{aligned} \tag{2}$$

where $\lambda_t = \lambda_{(j)} \cdot I\big(t \in (u_j, u_{j+1} - 1)\big)$. The indices b and j determine the number of change-points and the l_b's and u_j's determine their positions for IFR and the

infection rate respectively. The above formulation of an SIR model acts on the total number of cases assuming that individuals stay in the Infectious state for τ days. Then, it estimates the mean number of deaths as a proportion of the cases determined by the IFR, while past cases also contribute via suitable weights. Once again, many models have been tested based on this formulation, which test its performance placing covariates and autoregressive terms on λ_t, or simply testing different change-points.

The best change-point strategy for the infection rate turned out to be the following: we keep it constant for the first four weeks, then we change it once every two weeks and for the final interval we keep it constant for four weeks. The first and last periods are larger because of the larger uncertainty around the estimates. Every interval received a Log-Normal prior. The prior distribution for the IFR parameter $p_{(b)}$ needed some extra work. We placed an informative Gaussian prior distribution with standard deviation 10^{-4} and mean computed as

$$\mathbb{E}[p_{(b)}] = \frac{1}{4} \sum_{t=l_b}^{l_{b+1}-1} \sum_{k=1}^{4} p_k^{(0)} \frac{c_{t,k}}{\sum_{i=1}^{4} c_{t,i}}$$

where $p_k^{(0)}$ is the age group-specific IFR given in Ward et al., 2020, which we consider to be close to real and $c_{t,k}$ are the cases at time t for age group k in a given country. Thus, we scale the estimated IFR by the age distribution of the cases in a specific country and then average the result over the four groups 0-17, 18-39, 40-64 and 65+. The times l_b are determined by inspection of the country-specific IFR series $\sum_{k=1}^{4} p_k^{(0)}(c_{t,k})/(\sum_{i=1}^{4} c_{t,i})$.

The idea of model 2 works surprisingly fine on the real Covid-19 data, since it could re-create published estimates For instance, R_t took plausible values according to non-pharmaceutical measures applied at every interval, lock-downs or the effect of summer. However, we further generalize it using more realistic assumptions, such as taking into consideration the vaccination effect, the exposed period before infectiousness and the demography. To this end, we enhance the previous model with a few extra steps before the estimation of the daily deaths

d_t using the following equations

$$\begin{aligned} d_t &\sim NB(\theta_t^{(d)}, \psi_1) \\ \theta_t &= p_{(b)} \cdot I\big(t \in (l_b, l_{b+1})\big) \cdot \sum_{k=1}^{t-1} \pi_{t-k} \cdot C_k \\ C_t &= \lambda_{t-h-1} S_{t-h-1} I_{t-h-1}/N \\ S_t &= S_{t-1} - C_t - V_t + A \cdot N - A \cdot S_{t-1} \\ I_t &= \sum_{k=0}^{\tau-1} C_{t-k} - A \cdot I_{t-1} \end{aligned} \tag{3}$$

where h is the length of the exposed period, A is the birth rate (which we assume equals the death rate) and

$$V_t = \begin{cases} 0 & \text{, if} \quad t = 1, ..., 14 \\ 0.4 \cdot v_{t-14} & \text{, if} \quad t = 15, ..., 35 \\ (0.4 \cdot v_{t-14} + 0.1 \cdot v_{t-35}) & \text{, if} \quad t = 36, ..., n \end{cases} \tag{4}$$

where v_t is the number of vaccinations at time t. Thus, we assume that 40% of the vaccinated individuals move to the Removed state two weeks after vaccination and 10% more follow after three weeks to make a total of 50% protection. The 0.4 and 0.1 percentages were initially 0.68 and 0.95 respectively, but these had to be changed due to the effect of the delta variation.

One final model based on the new formulation is one that utilizes the observed cases data into the likelihood and, thus adding the steps

$$\begin{aligned} c_t &\sim NB(\theta_t^{(c)}, \psi_2) \\ \theta_t^{(c)} &= \omega_t \cdot C_{t-6} \end{aligned} \tag{5}$$

to the aforementioned model. Using this formulation we expect to have feedback on the total number of cases from the observed ones, instead of indirectly estimate them via the daily deaths. Moreover, we tried writing ω_t in a log-linear fashion to relate it with the number of tests performed, since we expect that there is the correlation: the more tests, the higher ω_t. This model is very hard to train, so we do not have any trustful results yet.

Putting covariate information on the infection rate for this class of models was very difficult, since proper training needed very sensitive algorithm parameters that led to slow completion of the desired iterations. Many times the fitting procedure had to be interrupted because it took over two weeks for less than one forth of the iterations. For the models we managed to fit, the results were not satisfactory (in contrast with the results when we worked with the observed cases), so we decided to perform a post-processing procedure in order to examine prediction accuracy of the infection rate given the mobility variable m_t. Thus, we assumed that the mean estimated λ_t has no error and considered it as a response variable with mobility being the independent one in a regression-type scenario. We tested a few methods, such as linear regression, regression trees, GAM and gradient boosted regression trees to discover the one with the most accurate predictions (based on a 5-times repeated 10-fold cross validation). The best among them was the gradient boosted trees, but the correlation between the two variables is so low that it is not advisable to make such predictions for any decision-making process.

3. MAIN RESULTS

In this Section, we provide some of our results based on models (2), (3) and (5). The most complete version of the proposed epidemic models is the SEIR with vaccination, demography and the bivariate likelihood, which produced very plausible results. The proportion of observed cases c_t/C_t can be seen in Figure 1 calculated by the median estimated C_t. Thus, at the beginning of the epidemic in Greece, we observed approximately one third of the Covid-19 cases, while this percentage increased due to more testing. Figure 2 shows the estimated reproduction number R_t as a piecewise constant function of time, produced by model (3) with fitted deaths given in Figure 3. Finally, the simple SIR model (2) has been proven to be capable of very accurate results, since its estimates come in agreement with the REACT study (Ward et al., 2020) in United Kingdom, which was thorough enough to be considered very close to reality. In Figure 4, we plot the cumulative total cases for United Kingdom and we also indicate the estimated cumulative total cases of Ward et al., 2020 at 17/7/2020. It seems that our model is very close to reality, although much calibration for the model was needed at the time this plot had been created. Therefore, we believe that our more complex models updated with the latest modifications will be even

more accurate by the time we end this research. All of our plots use a black line for median values, while the dark and light blue ribbons correspond to 50% and 95% credible intervals respectively.

4. CONCLUSION

In the present article some new approaches for epidemic modelling are provided to deal with the Covid-19 pandemic. We propose models that act on the observed level of registered cases, but also models that estimate useful quantities through the latent level of the total cases. Using HMC, we obtain plausible estimates that can be used for monitoring purposes and train models capable of either accurate inference or prediction. The models that estimate the number of total cases are hard to train and they typically need three to eight days, in contrast with the models acting on the observed cases, which need a few hours. Thus, we suggest that the former be used for inference (since they are way closer to reality), but one can use the latter formulation for fast predictions. This research is still ongoing (as part of a PhD thesis) and was initiated at the beginning of the epidemic. Thus, we have obtained numerous results and comparisons that cannot be fitted in the present article. However, we have presented the core of our work, as well as the main intuitions that lead the way for our approach. We are currently working on better calibration and tuning of our models on data from Greece and a few more countries, while we also train models with deeper hierarchies that make the inference procedure more robust. Moreover, using our proposed methods, the behaviour of the virus can easily be studied as a dynamics system, so that further intuition and help in decision-making processes be provided. Since the dynamics analysis is not yet completed, we did not include results in the present article.

ΠΕΡΙΛΗΨΗ

Η πανδημία Covid-19 είναι αδιαμφισβήτητα μία που θα μείνει στη μνήμη της παγκόσμιας κοινότητας ως ένα από τα σημαντικότερα γεγονότα του 21ου αιώνα, το οποίο κόστισε πολλές ζωές και είχε μεγάλο κοινωνικό αντίκτυπο και οικονομικές επιπτώσεις. Η θανατηφόρα ασθένεια κίνησε το επιστημονικό ενδιαφέρον από την οπτική γωνία της Στατιστικής και διαφορετικά μοντέλα προτείνονται στη βιβλιογραφία. Η καινοτομία των μεθόδων μας βρίσκεται στο γεγονός ότι αυτές βασίζονται

στην πληροφορία που αποσπάται από τα συνολικά – και όχι στα παρατηρούμενα – κρούσματα. Χρησιμοποιώντας τον τμηματικό τύπο επιδημικών μοντέλων, προτείνουμε ποικίλους τρόπους να αντιμετωπίσει κανείς το στατιστικό πρόβλημα λαμβάνοντας υπόψη τα ιδιαίτερα χαρακτηριστικά του Covid-19, γνωστά από τη βιβλιογραφία. Η διαδικασία της προσαρμογής των μοντέλων έγινε σε δεδομένα πολλών χωρών εκτός της Ελλάδας χρησιμοποιώντας τη Μπεϋζιανή μεθοδολογία μέσω του Hamiltonian Monte Carlo (HMC) αλγορίθμου.

REFERENCES

Andersson, H. and Britton, T. (2012). Stochastic epidemic models and their statistical analysis. *Springer Science & Business Media* **151**.

Bjørnstad, O. N. (2002). Epidemics: models and data using R. *Springer*.

Flaxman, S. and Mishra, S. and Gandy, A. and Unwin, H. J. T. and Mellan, T. A. and Coupland, H. and Whittaker, C. and Zhu, H. and Berah, T. and Eaton, J. W. et al. (2020). Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature Publishing Group* **584**(7820) 257-261.

Flaxman, S. and Mishra, S. and Gandy, A. and Unwin, H. and Coupland, H. and Mellan, T. and Zhu, H. and Berah, T. and Eaton, J. and Perez Guzman, P. et al. (2020). Report 13: Estimating the number of infections and the impact of non-pharmaceutical interventions on COVID-19 in 11 European countries.

Hoffman, M. D. and Gelman, A. (2014). The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.* **15**(1) 1593-1623.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Stan Development Team (2020). RStan: the R interface to Stan. R package version 2.21.2. http://mc-stan.org/.

Ward, H. and Atchison, C. J. and Whitaker, M. and Ainslie, K. E.C. and Elliott, J. and Okell, L. C. and Redd, R. and Ashby, D. and Donnelly, C. A. and Barclay, W. et al. (2020). Antibody prevalence for SARS-CoV-2 in

England following first peak of the pandemic: REACT2 study in 100,000 adults.

Figure 1: *Proportion of observed cases using the median values of C_t from model (5). We also smooth the output using a local regression (seen in blue color) for a better visual inspection.*

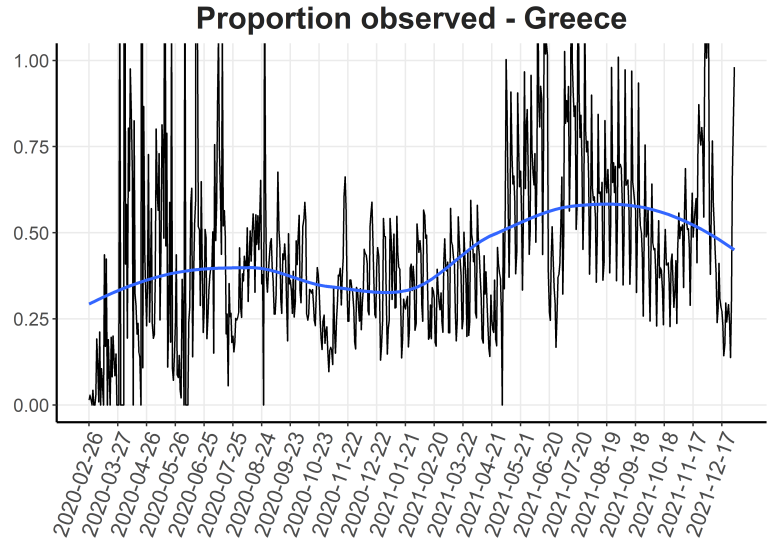


Figure 2: *Reproduction number estimates obtained from model (3).*

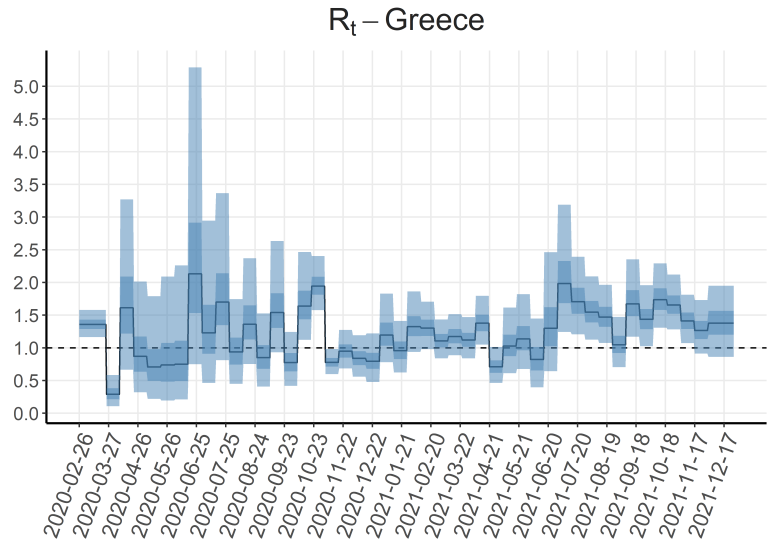


Figure 3: *Fitted deaths of model (3) for Greece. It seems that a fine fit has been achieved. The connected dots correspond to the daily deaths data.*

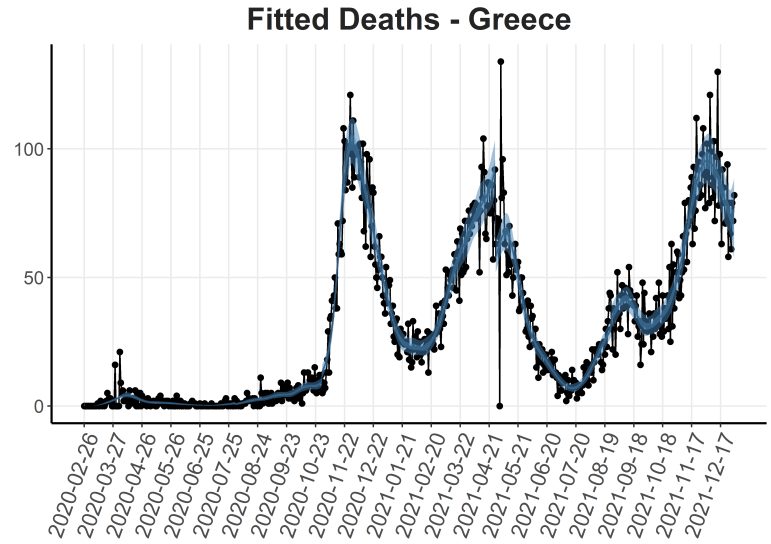
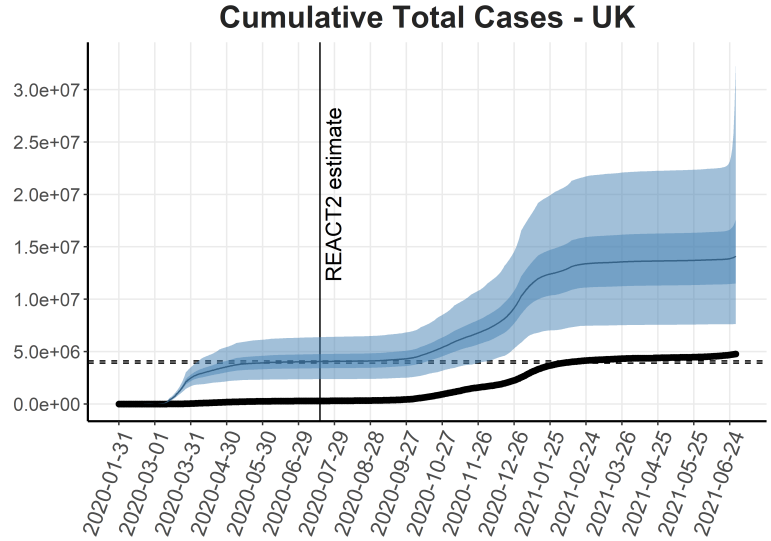


Figure 4: *Cumulative total cases obtained from model (2). Although this model was trained some time ago and it corresponds to the simplest SIR version, it seems to be close to reality. The two horizontal dashed lines indicate the 95% confidence interval of Ward et al., 2020 at 17/7/2020 (indicated by a vertical line). The dots correspond to the cumulative daily observed cases data.*



Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 34ου Πανελληνίου Συνεδρίου Στατιστικής (2022) (σελ.137-153)

Greek GDP forecasting using Bayesian multivariate models

***Z. Bragoudakis*[1], *I. Krompas*[2]**
[1]Bank of Greece & Department of Economics,
National and Kapodistrian University of Athens
zbragoudakis@bankofgreece.gr
[2]NBG Economic Research
krompas.ioannis@nbg.gr

ABSTRACT

Building on a proper selection of macroeconomic variables for constructing a Gross Domestic Product (GDP) forecasting multivariate model (Kazanas, 2017), this paper evaluates whether alternative Bayesian model specifications can provide greater forecasting accuracy compared to a standard Vector Error Correction model (VECM). To that end, two Bayesian Vector Autoreggression models (BVARs) are estimated, a BVAR using Litterman's prior (1979) and a BVAR with time-varying parameters (TVP-BVAR) (Primiceri, 2005). The out of sample forecasting performance of all three models is then evaluated over a period of 21 quarters (2016:Q1 to 2021:Q1), where the Litterman BVAR is found to greatly outperform both the VECM and the TVP-VAR.

***Keywords**:* Bayesian VARs, Forecasting,GDP,VECM

1.Introduction

Macro-econometrics according to Stock and Watson (2001), serve a quadruple purpose: Data description, forecasting, structural inference, and policy analysis. To that end, several types of models, from single-equation to large models with hundreds of equations have been used, like Klein's LINK model in 1980 (Klein, 1976) and more recently, Dynamic Stochastic General Equilibrium (DSGE) models (Christiano et al., 2018).

Lucas and other new classical economists were especially critical of the use of large-scale macro-econometric models to evaluate policy impacts when they were purportedly sensitive to policy changes (Lucas, 1976). Given that the optimal decision rules vary systematically with changes in the structure of series relevant to the decision maker, it follows that any policy change will systematically alter the structure of econometric models.

Based on Sim's (1980) framework of Vector Autoregressive models (VARs), the motivation for this paper is to build a set of small-scale flexible models that can be easily estimated, maintained, and change specifications when economic conditions change, in order to capture quickly and efficiently the new dynamics of the economy. VARs are n-equation, n-variable linear models in which each variable is in turn explained by its own lagged values, plus current and past values of the remaining n - 1 variables. This simple framework provides a systematic way to capture rich dynamics in multiple time series, and the statistical toolkit that came with VARs was easy to use and interpret. As Sims (1980) and others argued in a series of influential early papers, VARs held out the promise of providing a coherent and credible approach to data description, forecasting, structural inference, and policy analysis.
The remainder is organized as follows. Sections 2 and 3 provide a theroritical description of the Vector Error Correction Models (VECM) and Bayesian VAR models (BVAR and TVP-BVAR). Section 4 deals with the data and unit roots tests and present the main empirical findings. In Section 5, a forecast evaluation across the selected model is provided and Section 6 summarizes the conclusions.

2. Vector Error Correction Models (VECM)

Based on the works of Granger (1981) and Engle and Granger (1987), Vector error correction models are essentially restricted VARs, which contain a set of variables both in differences and in levels. The differences of the variables included in the model represent the short-run interrelations of the variables, whereas the linear combination of the levels of the variables, commonly referred to as the cointegrating vector (or vectors, as more than one linear combination of a set of variables can be included), represents the long-run dynamics of the variables. Mathematically, a representative VECM model can be written in matrix notation as follows:

$$\Delta y_t = m + \sum_{i=1}^{p-1} B_i \Delta y_{t-i} + A y_{t-1} + \varepsilon_t \quad (1)$$

where m is the vector containing the constants of the equations system, B_i is the matrix that contains the coefficients that describe the short-run impact of the variables' lag i and A is the matrix that contains the coefficients that describe the long-run relationship between the variables. The model can also be expanded to include exogenous variables. VECMs are very useful in modeling non-stationary time-series without having to exclude their long-run behavior, but, like their unrestricted counterparts (VARs), they suffer from the "curse of dimensionality", as the addition of a variable significantly increases the number of coefficients to be estimated.

3. Bayesian VAR models (BVAR and TVP-BVAR)

Bayesian VARs (BVAR) are an alternative to ordinary least squares (OLS) VARs initially proposed by Sims (1980), Doan, Sims and Litterman (1984), and others as an attempt to improve the forecasting performance of the econometric models available at the time. Under the Bayesian approach to econometrics, the estimated coefficients of a model are not an attempt to estimate their true value, but instead, they are perceived as a summary of the posterior distribution, which in its turn is proportional to the likelihood function times the prior. Priors represent any knowledge the researcher has beforehand about the coefficients. Following this technique results in the coefficients being essentially a matrix weighted average between the imposed priors and a regular OLS estimation (Ouliaris et al, 2016):

$$\hat{b} = [V^{-1} + \Sigma_e^{-1} \otimes (X'X)]^{-1}[V^{-1}\bar{b} + (\Sigma_e^{-1} \otimes X')Y] \quad (2)$$

where $\hat{b}$ is the matrix of the estimated VAR coefficients, V is the variance matrix of the prior distribution of model's coefficients, Σ_e is the variance-covariance matrix of the residuals and $\bar{b}$ is a diagonal matrix containing the prior means of each variable's own first lag coefficients. X and Y are the variables included in the model.

The error variance-covariance matrix necessary for the coefficient estimation is either estimated by fitting an AR(1) model on every variable and getting the error variances, by estimating an AR(1) and a VAR to obtain the diagonal elements of the variance-covariance matrix, or by estimating all variances-covariances as implied by a full VAR (an option not commonly used, as it can lead to a singular matrix). Imposing priors limits the parameter space that OLS would have to "search" for coefficient estimation and results in a more parsimonious and superior model in terms of forecasting. It must be noted that the priors imposed on the coefficient estimation do not need to reflect a specific economic theory, but only need to be consistent with the time-series properties of the variables included in the BVAR.

It is evident that imposing such restrictions on data that have many observations with unusual behavior, such as the Greek GDP after the economic crisis of 2008 may result in more accurate forecasts than those obtained by using simple OLS. In this paper we focus on the Minnesota prior, formulated by Litterman (Litterman, 1979) and his peers at the Minnesota University, which is a BVAR prior that formulates coefficients that force variables in the model look as if they were random walks (Del Negro and Schorfheide, 2010). Under this prior the researcher is required to specify a set of hyper parameters to estimate the model's coefficients: μ_1, λ_1, λ_2, and λ_3.

μ_1 is used as the prior mean of the coefficients in matrix $\bar{b}$ and it usually takes the value of 0 (if the variables of the model are stationary) or 1 (if the variables of the model have a unit root). λ_1, λ_2, and λ_3 are used to formulate diagonal elements of the V matrix (with non-diagonal elements being set to 0). More specifically, each diagonal element of V matrix for the j-th variable in the i-th equation at lag k is formulated as follows:

$$\left(\frac{\lambda_1}{k^{\lambda_3}}\right)^2 \text{ for } i = j,$$

$$\left(\frac{\lambda_1 \lambda_2 \sigma_i}{k^{\lambda_3} \sigma_j}\right)^2 \text{ for } i \neq j$$

where σ_i, σ_j are the square roots of the corresponding elements of the Σ_e matrix. This way:

λ_1 determines how binding the restrictions are. The closer to zero the value of λ_1 is, the more binding the restrictions are in the estimation of the coefficients. A value over 10 implies an uninformative prior.

λ_2 determines the cross-variable effects in the equations and is set between 0 and 1. The closer the value is to 1 the more lags of variable j impact variable *i* (for j≠i) in the BVAR.

Finally, λ_3 determines the decay rate of the own lags of a variable, excluding the first lag. As this hyper-parameter approaches zero, higher order lags decay at a slower rate.

3.1 The TVP-BVAR

Introduced by Primiceri (2005), the time-varying parameter VAR is a tool that allows model coefficients to change over time. This is particularly useful in capturing nonlinear relationships in the data as any model with time-varying parameters can successfully represent any nonlinear functional form (Swamy, 1975 and Granger, 2008). Macroeconomic variables are known to impact differently each other across the business cycle or after structural changes, hence the TVP-VAR is an interesting approach to econometric modeling. It must be noted that apart from time-varying parameters, TVP-VARs include stochastic volatility. This approach makes the model heavily parametrized but is necessary to avoid bias in the coefficients across potential volatility clusters, that is when a change in error variance is falsely attributed to coefficient variation (Sims, 2002).

To estimate TVP-VAR priors both Kalman filter and Gibbs Sampler are incorporated. Priors to be decided by the researcher include scaling of the variance matrices of the coefficients and the innovations but it is common practice to input the priors initially reported by Primiceri. Further common practice is to use a ten-year period as sample to formulate coefficient and error variance priors and to include two lags per variable. For more information on how TVP-VARs are formulated and estimated, one can look up Primiceri (2005) or Lubik and Matthes (2015).

4. Empirical Methodology and estimations

4.1 Data and unit roots test

The variable selection for the Greek GDP forecasting follows Kazanas (2017), where a VECM is utilized including data for real GDP (Y), unemployment rate (U), GDP deflator (P), 10-year government bond yield (GB), and exports as a percentage of GDP (XY). The data sample ranges from 2000Q1 to 2021Q1. All data are adjusted for seasonality and sourced from Eurostat's national accounts (Eurostat database code: na10), labour market (Eurostat database code: labour) and interest rate (Eurostat database code: irt) databases.

Table 1. *Descriptive Statistics*

	Mean	Median	Max.	Min.	Std. Dev.
Real GDP	5.07E+10	4.88E+10	6.34E+10	3.96E+10	6.89E+09
Unemployment rate	0.157	0.126	0.279	0.075	0.066690
GDP deflator	94.13	99.25	105.28	74.32	9.380585
Gov. Bond yield	0.068	0.052	0.254	0.008	0.048461
Exports (% GDP)	0.28	0.24	0.46	0.18	0.066456

In Table 1 the descriptive statistics of the selected variables is presented, while in Figures 1 and 2 we can see the long run evolution of the series over time. It is evident that the variables have increased volatility from 2009 to 2015, which reflects the impact of the economic crisis. Furthermore, real GDP, unemployment rate, GDP deflator, and exports (% GDP) display increased volatility during the second quarter of 2020, which reflects the impact of the pandemic and the corresponding lockdown.

Figure 1. *Variables in levels*

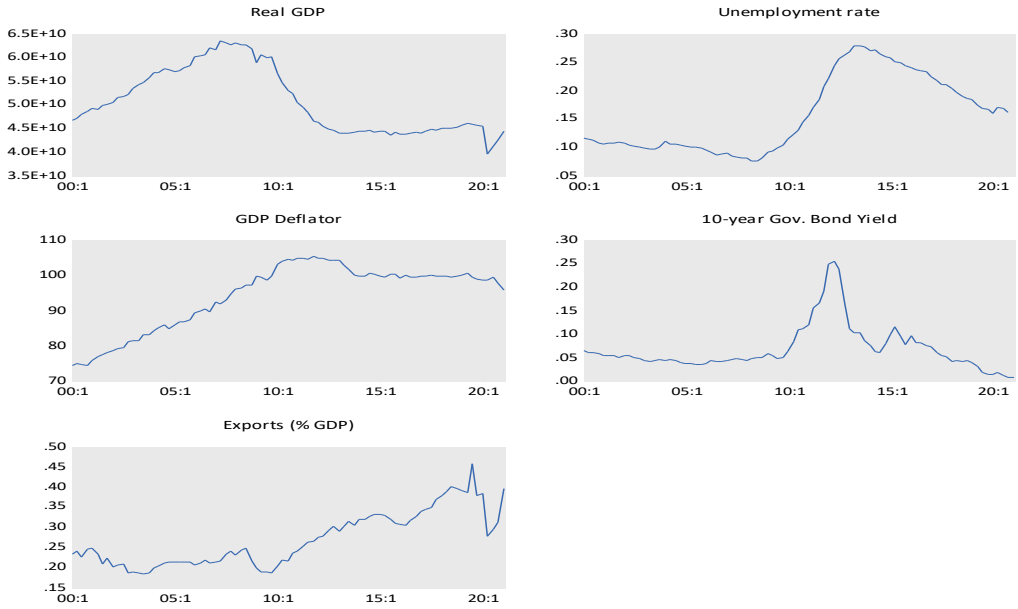
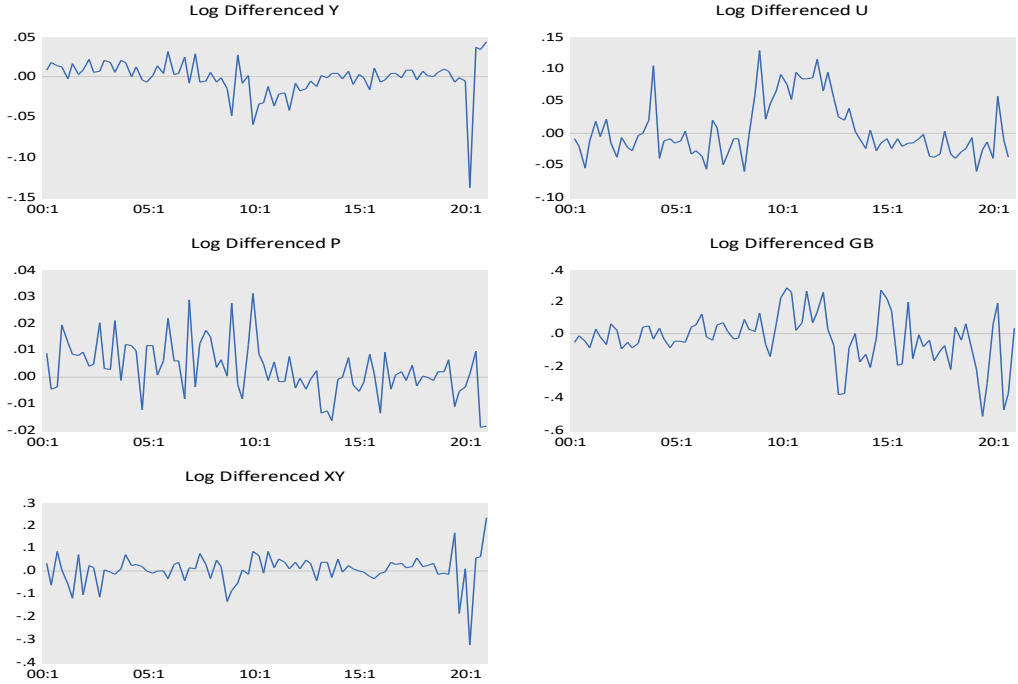


Figure 2. *Variables in Log Differences*



For each variable, the ADF unit root test (Dickey and Fuller, 1981) was conducted. All variables have a unit root in levels but are stationary if they are transformed into log differences.

Table 2. *ADF test p-values*

	Real GDP	Unemployment rate	GDP deflator	Gov. Bond yield	Exports (% GDP)
Levels	0.5357	0.1718	0.9976	0.1124	0.1308
Log differences	0.0000	0.0273	0.0000	0.0000	0.0000

4.2 Models Estimation

Three models are estimated using the abovementioned variables. A standard VECM model that incorporates long-run dynamics of the variables, which will serve as a forecasting evaluation benchmark, a Bayesian VAR with Minnesota (Litterman) priors, and a TVP-VAR. The models are estimated over a sixteen-year period (2000Q1 to 2015Q4), whereas the remaining sample (2016Q1 to 2021Q1) is to be used for evaluating the forecasting performance of the models.

4.2.1 The VECM Model

The estimation of this model follows the Johansen procedure (Johansen, 1995). A VAR is estimated in levels (including a constant and a trend) and by incorporating the lag length criteria it is found that two lags are optimal. The max eigenvalue cointegration test indicates the existence of two cointegrating vectors at the 5% level. Hence a VECM is estimated, with 1 lag per variable and 2 cointegrating vectors.

Table 3. *Maximum Eigenvalue Cointegration Test*

Hypothesized Number of Cointegrating equations	Eigenvalue	Max-Eigenvalue statistic	5% Critical value	P-value
None *	0.567460	51.96106	38.33101	0.0008
At most 1 *	0.450144	37.08209	32.11832	0.0114
At most 2	0.296067	21.76650	25.82321	0.1571
At most 3	0.255644	18.30464	19.38704	0.0713
At most 4	0.123745	8.190059	12.51798	0.2366

* Denotes rejection of the hypothesis at the 0.05 level

**MacKinnon-Haug-Michelis (1999) p-values

Hence, the model is formulated as follows:

$$\Delta y_t = m_1 + a_{11}(c_1 + c_2T + c_3y_{t-1} + c_4p_{t-1} + c_5gb_{t-1} + c_6u_{t-1} + c_7xy_{t-1}) + \\ +a_{12}(d_1 + d_2T + d_3y_{t-1} + d_4p_{t-1} + d_5gb_{t-1} + d_6u_{t-1} + d_7xy_{t-1}) + \\ +b_{11}\Delta y_{t-1} + b_{12}\Delta p_{t-1} + b_{13}\Delta gb_{t-1} + b_{14}\Delta u_{t-1} + b_{15}\Delta xy_{t-1} + \varepsilon_{1t}$$

$$\Delta p_t = m_2 + a_{21}(c_1 + c_2T + c_3y_{t-1} + c_4p_{t-1} + c_5gb_{t-1} + c_6u_{t-1} + c_7xy_{t-1}) + \\ +a_{22}(d_1 + d_2T + d_3y_{t-1} + d_4p_{t-1} + d_5gb_{t-1} + d_6u_{t-1} + d_7xy_{t-1}) + \\ +b_{21}\Delta y_{t-1} + b_{22}\Delta p_{t-1} + b_{23}\Delta gb_{t-1} + b_{24}\Delta u_{t-1} + b_{25}\Delta xy_{t-1} + \varepsilon_{2t}$$

$$\Delta gb_t = m_3 + a_{31}(c_1 + c_2T + c_3y_{t-1} + c_4p_{t-1} + c_5gb_{t-1} + c_6u_{t-1} + c_7xy_{t-1}) + \\ +a_{32}(d_1 + d_2T + d_3y_{t-1} + d_4p_{t-1} + d_5gb_{t-1} + d_6u_{t-1} + \\ + d_7xy_{t-1}) + b_{31}\Delta y_{t-1} + b_{32}\Delta p_{t-1} + b_{33}\Delta gb_{t-1} + b_{34}\Delta u_{t-1} + b_{35}\Delta xy_{t-1} + \\ + \varepsilon_{3t}$$

$$\Delta u_t = m_4 + a_{41}(c_1 + c_2T + c_3y_{t-1} + c_4p_{t-1} + c_5gb_{t-1} + c_6u_{t-1} + c_7xy_{t-1}) + \\ +a_{42}(d_1 + d_2T + d_3y_{t-1} + d_4p_{t-1} + d_5gb_{t-1} + d_6u_{t-1} + d_7xy_{t-1}) + \\ +b_{41}\Delta y_{t-1} + +b_{42}\Delta p_{t-1} + b_{43}\Delta gb_{t-1} + b_{44}\Delta u_{t-1} + b_{45}\Delta xy_{t-1} + \varepsilon_{4t}$$

$$\Delta xy_t = m_5 + a_{51}(c_1 + c_2T + c_3y_{t-1} + c_4p_{t-1} + c_5gb_{t-1} + c_6u_{t-1} + c_7xy_{t-1}) + \\ +a_{52}(d_1 + d_2T + d_3y_{t-1} + d_4p_{t-1} + d_5gb_{t-1} + d_6u_{t-1} + d_7xy_{t-1}) + \\ +b_{51}\Delta y_{t-1} + b_{52}\Delta p_{t-1} + b_{53}\Delta gb_{t-1} + b_{54}\Delta u_{t-1} + b_{55}\Delta xy_{t-1} + \varepsilon_{5t}$$

where *T* denotes a deterministic trend, lowercase names of the variables denote natural logarithms of said variables and $\varepsilon_{it} \sim N(0, \sigma^2)$.

Table 4.*VECM estimation output*

Cointegrating Equation	1	2			
y_{t-1}	1.000000	0.000000			
p_{t-1}	0.000000	1.000000			
gb_{t-1}	0.438604	0.098740			
	[6.15673]	[3.67792]			
u_{t-1}	0.387905	0.125085			
	[3.75521]	[3.21326]			
xy_{t-1}	-1.078572	-0.315739			
	[-4.42947]	[-3.44083]			
T	0.000578	-0.004727			
	[0.32659]	[-7.09289]			
c	-24.27234	-4.300468			
Variables:	Δy_t	Δp_t	Δgb_t	Δu_t	Δxy_t
Coint.Equation 1	0.045493	0.102953	-1.060690	0.047901	0.015096
	[0.90996]	[3.72949]	[-3.30049]	[0.43005]	[0.10368]
Coint.Equation 2	-0.141837	-0.318281	1.839943	0.061796	0.395072
	[-1.15367]	[-4.68851]	[2.32813]	[0.22561]	[1.10337]
Δy_{t-1}	-0.095724	-0.143579	1.821461	-0.253899	0.897779
	[-0.50739]	[-1.37829]	[1.50193]	[-0.60406]	[1.63396]
Δp_{t-1}	0.302848	-0.111501	1.635784	-0.541680	0.410560
	[1.32557]	[-0.88387]	[1.11383]	[-1.06419]	[0.61703]
Δgb_{t-1}	-0.025246	-0.007381	0.646362	0.035366	0.075213
	[-1.45907]	[-0.77252]	[5.81137]	[0.91744]	[1.49258]
Δu_{t-1}	-0.147127	0.089740	1.568756	0.333778	-0.333700
	[-2.00657]	[2.21657]	[3.32835]	[2.04323]	[-1.56268]
Δxy_{t-1}	-0.030706	0.016100	0.006938	-0.045525	-0.058512
	[-0.72875]	[0.69199]	[0.02561]	[-0.48496]	[-0.47681]
m	-0.000495	0.003948	-0.025286	0.010383	0.006950
	[-0.21058]	[3.03999]	[-1.67230]	[1.98135]	[1.01453]
t-statistics in [ ]					
R^2	0.366886	0.370907	0.520396	0.521789	0.263906
Adjusted R^2	0.284815	0.289357	0.458225	0.459798	0.168487

4.2.2 THE BVAR

Apart from μ_1, which is set to zero as the model is estimated in log differences which are stationary, the rest hyper-parameters necessary for obtaining coefficient estimates based on the Minnesota-Litterman prior are chosen to reflect key macroeconomic stylized facts, namely strong cross variable effect and the AR(1) nature of macroeconomic data. Exact values are chosen through forecasting sensitivity checks as suggested by Canova (2007) while ensuring that the selected model has white noise residuals. The resulting set of hyper-parameters is the following: $\lambda_1 = 5, \lambda_2 = 1, \lambda_3 = 0.01$

To obtain an estimation of the variance-covariance matrix an AR(1) model is fitted through each variable to estimate the variances of the residuals, while the covariances

(the diagonal elements of the matrix) are obtained from the equivalent matrix of the corresponding OLS VAR. Furthermore, most lag-length criteria indicate that 2 is the optimal number of lags. Hence, the model is formulated as follows:

$$\Delta y_t = m_1 + b_{11}\Delta y_{t-1} + b_{12}\Delta y_{t-2} + b_{13}\Delta p_{t-1} + b_{14}\Delta p_{t-2} + b_{15}\Delta gb_{t-1} + \\ + b_{16}\Delta gb_{t-2} + b_{17}\Delta u_{t-1} + b_{18}\Delta u_{t-2} + b_{19}\Delta xy_{t-1} + b_{110}\Delta xy_{t-2} + \varepsilon_{1t}$$

$$\Delta p_t = m_2 + b_{21}\Delta y_{t-1} + b_{22}\Delta y_{t-2} + b_{23}\Delta p_{t-1} + b_{24}\Delta p_{t-2} + b_{25}\Delta gb_{t-1} + \\ + b_{26}\Delta gb_{t-2} + b_{27}\Delta u_{t-1} + b_{28}\Delta u_{t-2} + b_{29}\Delta xy_{t-1} + b_{210}\Delta xy_{t-2} + \varepsilon_{2t}$$

$$\Delta gb_t = m_3 + b_{31}\Delta y_{t-1} + b_{32}\Delta y_{t-2} + b_{33}\Delta p_{t-1} + b_{34}\Delta p_{t-2} + b_{35}\Delta gb_{t-1} + \\ + b_{36}\Delta gb_{t-2} + b_{37}\Delta u_{t-1} + b_{38}\Delta u_{t-2} + b_{39}\Delta xy_{t-1} + b_{310}\Delta xy_{t-2} + \varepsilon_{3t}$$

$$\Delta u_t = m_4 + b_{41}\Delta y_{t-1} + b_{42}\Delta y_{t-2} + b_{43}\Delta p_{t-1} + b_{44}\Delta p_{t-2} + b_{45}\Delta gb_{t-1} + \\ + b_{46}\Delta gb_{t-2} + b_{47}\Delta u_{t-1} + b_{48}\Delta u_{t-2} + b_{49}\Delta xy_{t-1} + b_{410}\Delta xy_{t-2} + \varepsilon_{4t}$$

$$\Delta xy_t = m_5 + b_{51}\Delta y_{t-1} + b_{52}\Delta y_{t-2} + b_{53}\Delta p_{t-1} + b_{54}\Delta p_{t-2} + b_{55}\Delta gb_{t-1} + \\ + b_{56}\Delta gb_{t-2} + b_{57}\Delta u_{t-1} + b_{58}\Delta u_{t-2} + b_{59}\Delta xy_{t-1} + b_{510}\Delta xy_{t-2} + \varepsilon_{5t}$$

where lowercase letters denote the natural logarithms of the corresponding variables and $\varepsilon_{it} \sim N(0, \sigma^2)$.

Table 5. *BVAR estimation output*

Variables:	Δy_t	Δp_t	Δgb_t	Δu_t	Δxy_t
Δy_{t-1}	-0.034524	0.126019	0.389792	-0.331051	0.724426
	[-0.21784]	[1.22333]	[0.35692]	[-0.87277]	[1.44722]
Δy_{t-2}	0.298321	0.245310	0.392504	-0.496588	0.267113
	[1.86716]	[2.36212]	[0.35650]	[-1.29861]	[0.52931]
Δp_{t-1}	0.242653	0.018258	1.496880	-0.625650	-0.238986
	[1.12448]	[0.13017]	[1.00662]	[-1.21138]	[-0.35064]
Δp_{t-2}	0.044532	0.122010	1.866501	0.053188	0.006473
	[0.19997]	[0.84292]	[1.21631]	[0.09979]	[0.00920]
Δgb_{t-1}	-0.011045	0.003975	0.676470	0.013206	0.093930
	[-0.58420]	[0.32346]	[5.19246]	[0.29184]	[1.57302]
Δgb_{t-2}	-0.005766	0.010653	-0.363924	0.026184	-0.060995
	[-0.31098]	[0.88397]	[-2.84851]	[0.59008]	[-1.04162]
Δu_{t-1}	-0.170970	0.079585	1.311978	0.449508	-0.036018
	[-2.66942]	[1.91170]	[2.97261]	[2.93236]	[-0.17805]
Δu_{t-2}	0.063818	-0.037706	-0.787786	0.010723	0.547315
	[0.98767]	[-0.89777]	[-1.76924]	[0.06934]	[2.68177]
Δxy_{t-1}	-0.055087	-0.006095	0.217406	-0.021990	0.006686
	[-1.32978]	[-0.22637]	[0.76159]	[-0.22179]	[0.05110]
Δxy_{t-2}	-0.068490	0.027163	0.539662	0.010031	0.122699
	[-1.66260]	[1.01445]	[1.90106]	[0.10174]	[0.94302]
m	-0.000407	0.003774	-0.023079	0.008447	7.96E-05

	[-0.15910]	[2.27045]	[-1.30977]	[1.38025]	[0.00986]
t-statistics in [ ]					
R^2	0.439684	0.232014	0.521314	0.516892	0.223831
Adjusted R^2	0.327620	0.078416	0.425577	0.420271	0.068597

4.2.3 THE TVP-VAR

Unlike the two previous models, the TVP-VAR has no constant coefficients to be reported, as there are no point estimations. It is interesting however to report coefficient variation over time. The TVP-VAR is formulated as follows:

$$\Delta y_t = m_t^1 + b_t^{11}\Delta y_{t-1} + b_t^{12}\Delta y_{t-2} + b_t^{13}\Delta p_{t-1} + b_t^{14}\Delta p_{t-2} + b_t^{15}\Delta gb_{t-1} + \\ + b_t^{16}\Delta gb_{t-2} + b_t^{17}\Delta u_{t-1} + b_t^{18}\Delta u_{t-2} + b_t^{19}\Delta xy_{t-1} + b_t^{110}\Delta xy_{t-2} + \varepsilon_{1t}$$

$$\Delta p_t = m_t^2 + b_t^{21}\Delta y_{t-1} + b_t^{22}\Delta y_{t-2} + b_t^{23}\Delta p_{t-1} + b_t^{24}\Delta p_{t-2} + b_t^{25}\Delta gb_{t-1} + \\ + b_t^{26}\Delta gb_{t-2} + b_t^{27}\Delta u_{t-1} + b_t^{28}\Delta u_{t-2} + b_t^{29}\Delta xy_{t-1} + b_t^{210}\Delta xy_{t-2} + \varepsilon_{2t}$$

$$\Delta gb_t = m_t^3 + b_t^{31}\Delta y_{t-1} + b_t^{32}\Delta y_{t-2} + b_t^{33}\Delta p_{t-1} + b_t^{34}\Delta p_{t-2} + b_t^{35}\Delta gb_{t-1} + \\ + b_t^{36}\Delta gb_{t-2} + b_t^{37}\Delta u_{t-1} + b_t^{38}\Delta u_{t-2} + b_t^{39}\Delta xy_{t-1} + b_t^{310}\Delta xy_{t-2} + \\ + \varepsilon_{3t}$$

$$\Delta u_t = m_t^4 + b_t^{41}\Delta y_{t-1} + b_t^{42}\Delta y_{t-2} + b_t^{43}\Delta p_{t-1} + b_t^{44}\Delta p_{t-2} + b_t^{45}\Delta gb_{t-1} + \\ + b_t^{46}\Delta gb_{t-2} + b_t^{47}\Delta u_{t-1} + b_t^{48}\Delta u_{t-2} + b_t^{49}\Delta xy_{t-1} + b_t^{410}\Delta xy_{t-2} + \varepsilon_{4t}$$

$$\Delta xy_t = m_t^5 + b_t^{51}\Delta y_{t-1} + b_t^{52}\Delta y_{t-2} + b_t^{53}\Delta p_{t-1} + b_t^{54}\Delta p_{t-2} + b_t^{55}\Delta gb_{t-1} + \\ + b_t^{56}\Delta gb_{t-2} + b_t^{57}\Delta u_{t-1} + b_t^{58}\Delta u_{t-2} + b_t^{59}\Delta xy_{t-1} + b_t^{510}\Delta xy_{t-2} + \varepsilon_{5t}$$

Lowercase variable names denote natural logarithms of the variables. Time-varying coefficients are formulated as a random walk process:

$$b_t^{ij} = b_{t-1}^{ij} + \eta_t^{ij}$$

With $\varepsilon_{it} \sim N(0, \sigma_{i,t}^2)$ and $\log \sigma_{i,t}^2 = \log \sigma_{i,t-1}^2 + u_{i,t}$. Furthermore, η_t^{ij} and $u_{i,t}$ are normally distributed.

The above equations imply that there is no mechanism in the model to produce future values of the coefficients of the model, as in the absence of new shocks, coefficients remain the same. It is an interesting approach, however, to attempt a forecast based on the most recent interrelations between the variables and neglect coefficient values of the past that may not adequately represent the dynamics of the system anymore.

Figure 3. *Δy_t time varying coefficients*

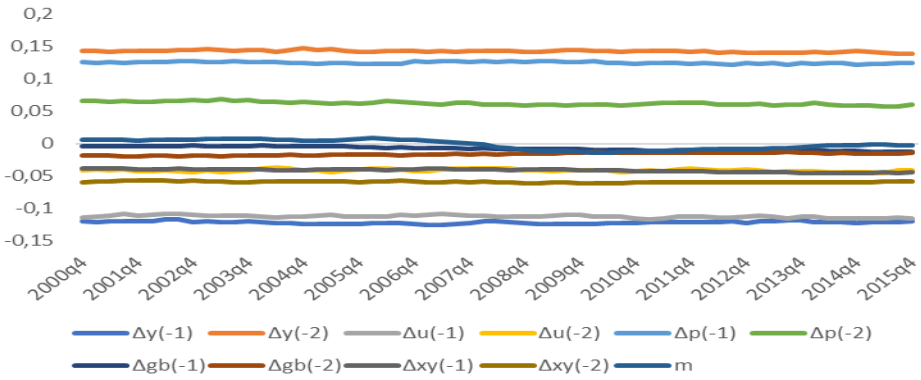


Figure 4. *Δp_t time varying coefficients*

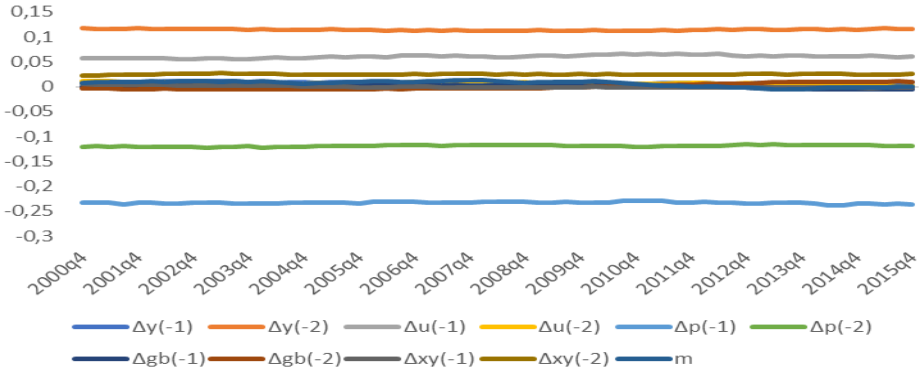


Figure 5. *Δgb_t time-varying coefficients*

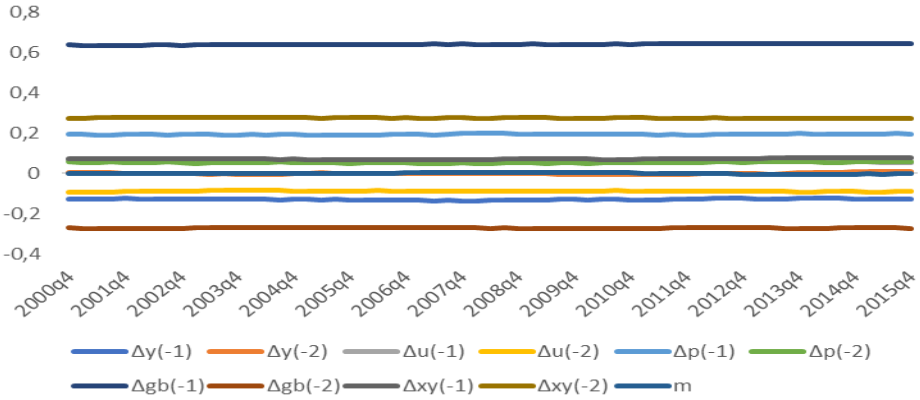


Figure 6. *Δu_t time-varying coefficients*

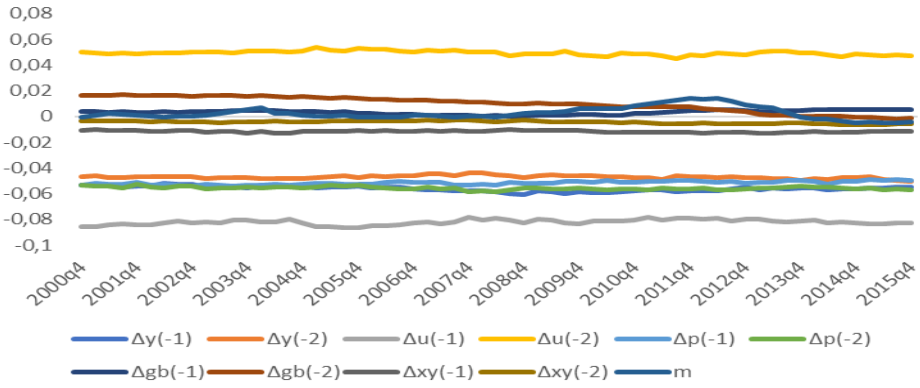
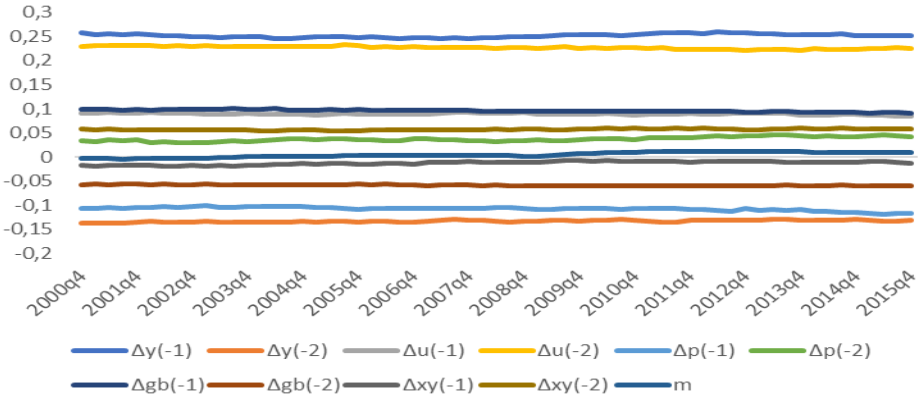


Figure 7. *Δxy_t time-varying coefficients*



As it can be seen in Figures 3-7, most time-varying coefficients are relatively stable in the model. The greatest interest is drawn to the time-varying constants of the GDP equation, which turns negative to accommodate for the economic crisis period and the time-varying constant in the unemployment equation. In general, the unemployment equation experiences the most significant time variations, as GDP and bond yield's impact on unemployment change over time according to the model.

5. Forecast evaluation

To evaluate the forecasting performance of the models the mean absolute percentage error (MAPE) is computed, over a forecasting period of 21 quarters (2016q1-2021q1). The MAPE is calculated as follows and the evaluation of the forecasts can be found in table 6:

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{\left|Y_t - \widehat{Y_t}\right|}{Y_t} = \frac{1}{n}\sum_{t=1}^{n}\frac{|e_t|}{Y_t}$$

Where n is the periods of the forecasting horizon $\widehat{Y_t}$ is the forecasted value of a variable and Y_t is the actual value of the same variable.

Table 6. *Mean Absolute Percentage Error of forecasts (%)*

	VECM	BVAR	TVP-VAR
Y	4.248611	**2.434411**	3.054219
P	7.697283	6.272289	**1.186609**
GB	56.90365	**48.39754**	190.3814
U	**12.84760**	16.76608	20.49910
XY	22.78990	19.49644	**11.06880**

It turns out that the Bayesian VAR estimated with Minnesota prior outperforms the other models in terms of GDP forecasting. The reasons for this superior performance may rely on the fact that instead of trying to capture the long-run dynamics in a sample of macroeconomic variables which contains a significant percentage of irregular behavior (due to the economic crisis), it may be more beneficial to restrict the parameter space of the coefficients to be estimated, by imposing a prior and hyperparameters that are consistent with the theorized behavior macroeconomic variables have (as the Minnesota prior does). In addition, the TVP-VAR although is usually used for ex-post analysis, also outperforms the VECM, but does not outperform the Bayesian VAR in terms of GDP forecasting.

6. Conclusions

Three VARs were estimated using a given set of variables aiming to examine whether Bayesian estimations could provide real GDP forecasting gains. Using two different Bayesian VAR estimation methods, namely Bayesian estimation using a Minnesota-Litterman prior and a TVP-VAR it is found that Bayesian estimation methods outperform the corresponding VECM. Specifically, the BVAR that was estimated using Minnesota Litterman prior outperformed the VECM model by 42% and the TVP-VAR by 20% in terms of forecasting error, over 21 quarters. This forecasting exercise demonstrated that there the most basic of Bayesian priors provided significant gains, but it is only one of the available priors a researcher has available to choose from. One could extend this research to include more advanced Bayesian priors such as the Sims-Zha prior (Sims and Zha, 1998) that incorporates the existence of unit roots and cointegrating relationships in the priors (as it is found in Table 3 that cointegration relationships exist between the variables of the given set); or the GLP prior (Giannone et al., 2015) that treats hyperparameters not as arbitrary inputs of the user but as parameters to be determined from an optimization procedure. Another way this research could be extended is by using the TVP-VAR estimated above (possibly using a larger sample if available, to account for the model's intensive parameterization), to compute the variation in the relations of the macroeconomic variables, as expressed by the time-varying coefficients, and thus examine structural changes of the Greek Economy over time. This model can also be used to perform impulse response analysis on specific dates, which allows examining how differently exogenous shocks would affect the Greek economy, at different points in time.

Περίληψη

Χρησιμοποιώντας ένα αξιολογημένο σύνολο μακροοικονομικών μεταβλητών για την κατασκευή ενός πολυμεταβλητού υποδείγματος πρόβλεψης του Ελληνικού ΑΕΠ (Καζανάς, 2017), η παρούσα εργασία εξετάζει το κατά πόσο εναλλακτικά μπεϋζιανά μοντέλα μπορούν να παρέχουν μεγαλύτερη προβλεπτική ικανότητα σε σύγκριση με ένα πολυμεταβλητό υπόδειγμα διόρθωσης σφαλμάτων. Για τον σκοπό αυτό, εκτιμώνται δύο μπεϋζιανά υποδείγματα, ένα χρησιμοποιώντας τις παραμετροποιήσεις του Litterman (1979) και ένα μπεϋζιανό υπόδειγμα χρονικά μεταβαλλόμενων συντελεστών (Primiceri,2005). Η εκτός δείγματος επίδοση των τριών υποδειγμάτων αξιολογήθηκε χρησιμοποιώντας μία περίοδο 21 τριμήνων (από το 2016:Q1 έως το 2021:Q1), με το υπόδειγμα με τις παραμετροποιήσεις του Litterman να υπερτερεί σε προβλεπτική ακρίβεια του Ελληνικού ΑΕΠ σε σχέση με τα υπόλοιπα.

ΑΝΑΦΟΡΕΣ

Canova, F. (2007). "10. Bayesian VARs", *Methods for Applied Macroeconomic Research*, Princeton University Press, Princeton, 373-417.

Christiano, J. L., Eichenbaum, S.M. and Trabandt, M. (2018). On DSGE models, *NBER Working Paper,* **24811**.

Del Negro, M. and Schorfheide, F. (2010). "Bayesian Macroeconometrics", *Handbook of Bayesian Econometrics*.

Dickey, D. and Fuller, W. (1981). Likelihood Ratio Statistics for Autoregressive Time Series with a Unit Root, *Econometrica*, **49**, 1057-1072.

Doan, T., Litterman, R. and Sims, C. (1983). Forecasting and Conditional Projection Using Realistic Prior Distributions, *NBER Working Paper*, **1202**.

Engle, R. and Granger, C. (1987). Co-Integration and Error Correction: Representation, Estimation, and Testing, *Econometrica*, **55**, 251-276.

Giannone, D., Lenza, M. and Primiceri, G. (2015). Prior Selection for Vector Autoregressions, *Review of Economics and Statistics*, **97**, 436-451.

Granger, C. (1981). Some properties of time series data and their use in econometric model specification, *Journal of Econometrics*, **16**, 121-130.

Granger, C. (2008). Non-linear Models: Where Do We Go Next? – Time-Varying Parameter Models, *Studies in Nonlinear Dynamics and Econometrics*, **12**,1-10.

Johansen, S. (1995). *Likelihood-based Inference in Cointegrated Vector Autoregressive Models*, Oxford: Oxford University Press.

Kazanas, T. (2017). A Vector Error Correction Forecasting Model of the Greek Economy, *Hellenic Fiscal Council Working Paper*, **2**.

Klein, L. R. (1976). Project LINK: Linking National Economic Models. Challenge, **19**, 25–29.

Litterman, R., (1979). Techniques of forecasting with Bayesian vector autoregressions, *Federal Reserve Bank of Minneapolis Working Papers*, 115.

Lubik, T. and Matthes, C. (2015). Time-Varying Parameter Vector Autoregressions: Specification, Estimation and an Application, *Federal Reserve Bank of Richmond Economic Quarterly*, **101**, 323-352.

Lucas, R. (1976). Econometric policy evaluation: A critique, *Carnegie-Rochester Conference Series on Public Policy*, **1**, 19-46.

Mackinnon, J., Haug, A., and Michelis, L. (1999). Numerical Distribution Functions of Likelihood Ratio Tests for Cointegration, *Journal of Applied Econometrics*, **14**, 563-577.

Ouliaris, S., Pagan, A. R., Restrepo J. (2016). "Bayesian Vars", *Quantitative Macroeconomic Modelling with Structural Vector Autoregressions - An Eviews Implementation*, 49-62.

Primiceri, G. (2005). Time Varying Structural Vector Autoregressions and Monetary Policy, *The Review of Economic Studies*, **72**, 821-852.

Sims, C. (1980). Macroeconomics and Reality, *Econometrica*, **48**, 1-48.

Sims, C. (2002). Comments on Cogley and Sargent's 'Evolving Post-World War II U.S. Inflation Dynamics., *NBER Macroeconomics Annual 2001*, **16**, 373-379.

Sims, C. and Zha, T. (1998). Bayesian Methods for Dynamic Multivariate Models, *International Economic Review*, **39**, 949-968.

Stock, J. and Watson, W. (2001). Vector Autoregressions, *The Journal of Economic Perspectives*, **15**, 101-115.

Swamy, P. (1975). Bayesian and Non-Bayesian Analysis of Switching Regressions and a Random Coefficient Regression Model, *Journal of the American Statistical Association*, **70**, 593-602.


Greek Statistical Institute
Proceedings of the 34th Panhellenic Statistics Conference (2022), pp. 154−166

SYNTHESIS OF TIME-SERIES WITH MISSING OBSERVATIONS USING GENERATIVE ADVERSARIAL NETWORKS

***Owen D. Jones*[1], *Thomas Poudevigne-Durance*[1], *Yipeng Qin*[2]**
[1] School of Mathematics, Cardiff University, U.K.
{JonesO18, Poudevigne-DuranceT}@cardiff.ac.uk
[2] School of Computer Science and Informatics, Cardiff University, U.K.
QinY16@cardiff.ac.uk

ABSTRACT
We introduce a new method for the data synthesis of time-series using Generative Adversarial Networks (GANs) that can be applied directly to data with missing observations. Unlike previous GAN-based time-series models which use Recurrent Neural Networks (RNNs), our approach directly models the conditional distribution of the current observation given past observations, which it does using an auxiliary GAN trained on the joint distribution of the current and past observations. A benefit of this approach is that it allows us to use a Masked Wasserstein GAN (MaWGAN) to train the model, which can directly accommodate missing values, unlike existing time-series GANs. The veracity of the approach is demonstrated with a simulation experiment, for which we get good results even with high levels of missing data.

Keywords: time-series; data synthesis; missing data; generative adversarial network.

1. INTRODUCTION

Data synthesis refers to the simulation of data while preserving privacy. Many national statistics organisations and state-owned enterprises are interested in data synthesis as a way of distributing the analytic value of data without revealing confidential personal details (Kaloskampis et al. 2019, Sallier 2020). A promising development for data synthesis has been the advent of Generative Adversarial Networks (GANs: Goodfellow et al. 2014); see e.g. Hitawala (2018) for a review. GANs use two neural nets, one to generate synthetic data, and the other to build a critic which is used to train the generator (also called a discriminator). The generator and critic are trained iteratively, so that as the generator improves the critic becomes more discerning, allowing further refinement of the generator. GANs are capable of reproducing complex dependencies in data, and are amenable to privacy protection approaches such as *differential privacy* (Campbell 2019, Jordon et al. 2022). In what follows we will focus on using GANs to model temporal dependencies and will not explicitly consider privacy issues, however we will

assess the performance of our generator using distribution based measures, implicitly judging its ability to reproduce some underlying distribution rather than a specific realisation from it (the data).

Traditional time-series models focus on predictive power rather than data synthesis. Formally they model the signal but not the noise (or texture) around the signal, and for data synthesis both are required. An advantage of a non-parametric model such as a GAN in this setting is that it models both noise and signal simultaneously. Mogren (2016), Esteban et al. (2017) and Yoon et al. (2019) have all previously used GANs for time-series prediction and synthesis. Their approaches all use recurrent neural network architectures (RNNs). RNN-GANs effectively model the conditional distribution of the current observation conditioned on past observations. Information from past observations is encoded as features in one or more hidden layers that are fed back as inputs into the generator for the current observation. Unfortunately there is no clear way to translate missing values into features in these hidden layers, so we take a different approach.

We introduce a two-stage approach to build a model for the conditional distribution of the current observation given past observations. Firstly we build a GAN model for the joint distribution of present and past observations, then secondly we leverage the generator and critic from the joint distribution to build a forecaster that directly models the target conditional distribution.

Missing data is ubiquitous and is as much of an issue for data synthesis as elsewhere. Until recently missing data has been a problem for GANs as existing training/fitting algorithms require complete observations, so users have had to either first impute the missing data or just discard incomplete observations. However in a recent paper the authors introduced a novel GAN algorithm that can directly train a synthetic data generator from datasets with missing values, which we have called MaWGAN for *Masked Wasserstein GAN* (Poudevigne-Durance et al. 2022). MaWGAN is based on a modification of the Wasserstein distance and is easily implemented by incorporating into the critic masks generated from the pattern of missing data in the original dataset. Moreover we will see that this approach also works in our current setting, so that our approach to modelling the conditional distribution of the present given the past can deal directly with missing data.

The Wasserstein distance or Kantorovich–Rubinstein metric (also called the Earth-Mover distance) is a distance function defined between probability distributions. In the context of a GAN the critic is trained to estimate the Wasserstein distance between the distribution that the data were sampled from and the distribution represented by the generator. Kantorovich & Rubinstein (1958) famously showed that the Wasserstein distance can be written as a Lipschitz metric, which is the form we use (see Section 2.1).

The original GAN effectively used the Kullback-Leibler divergence to measure the distance between data and generator (Arjovsky et al. 2017), however—unlike the Wasserstein distance—this approach is susceptible to the so-called vanishing gradients problem, whereby the critic rejects all samples from the generator and does not allow it to

learn. Other approaches to measuring the distance between data and generator have been proposed, for example Variational Divergence (Nowozin et al. 2016) and Maximum Mean Discrepancy (Li et al. 2017), the latter being a special case of scoring rule minimisation (Pacchiardi et al. 2021).

In what follows we give some background on MaWGAN before describing our two-stage approach for applying GANs to time series. We then provide pseudo-code showing how to incorporate MaWGAN so that the method is applicable to time-series data with missing values. We also give some preliminary test results in which we test the method using data simulated from an auto-regressive AR(3) model. The results are promising and show that the method can cope with missing data, though there is scope for further tuning of the parameters and architecture of the GAN nets (generator, critic and forecaster).

2. METHODOLOGY

2.1 MaWGAN

A basic description of the *Masked Wasserstein GAN* is needed for what follows. See Poudevigne-Durance et al. (2022) for more details. In what follows our vectors are all row vectors by default.

WGAN-GP MaWGAN builds on the WGAN-GP algorithm (Arjovsky et al. 2017, Gulrajani et al. 2017). Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$ be an i.i.d. sample from some (unknown) distribution $\mathcal{P}$, and let $G : (0,1)^d \to \mathbb{R}^d$ be our generator. G takes a vector of i.i.d. $U(0,1)$ random variates and returns a vector with distribution $\mathcal{Q}$ say. The WGAN-GP critic calculates an estimate of the Wasserstein distance, so that the generator is trained to minimise the distance between $\mathcal{P}$ and $\mathcal{Q}$ as measured by the Wasserstein distance.

The Wasserstein distance can be written as

$$W(\mathcal{P}, \mathcal{Q}) = \sup_{\|f\|_L \leq 1} \left(\mathbb{E}_{X \sim \mathcal{P}} f(X) - \mathbb{E}_{Y \sim \mathcal{Q}} f(Y) \right)$$

where $\|f\|_L$ is the Lipschitz constant of f. Let $C : \mathbb{R}^d \to \mathbb{R}_+$ be our critic, let $\mathbf{y}_1, \ldots, \mathbf{y}_n$ be a sample from the generator G, and for $\epsilon_i \sim U(0,1)$ put $\mathbf{z}_i = \epsilon_i \mathbf{x}_i + (1 - \epsilon_i)\mathbf{y}_i$, then we train the critic to maximise

$$\frac{1}{n} \sum_i C(\mathbf{x}_i) - \frac{1}{n} \sum_i C(\mathbf{y}_i) - \lambda \frac{1}{n} \sum_i \left(\|\nabla C(\mathbf{z}_i)\|_2 - 1 \right)^2 .$$

The key idea here is that the regularisation term will restrict the critic C to be close to a Lipschitz function with Lipschitz constant 1. $\lambda > 0$ controls the degree of regularisation and can be tuned to improve the convergence of the critic.

MaWGAN MaWGAN is based on a variation of the Wasserstein distance that incorporates a random mask to capture the effect of missing data. For our purposes a mask $\mathbf{m} = (m_1, \ldots, m_d)$ is an element of $\{0,1\}^d$ and a random mask is just a measure $\mathcal{M}$ on $\{0,1\}^d$. Given a data point $\mathbf{x} = (x_1, \ldots, x_d)$ and a mask $\mathbf{m}$, x_j is treated as missing if and only if $m_j = 0$. We define the $\mathcal{M}$-Wasserstein distance as

$$W_{\mathcal{M}}(\mathcal{P}, \mathcal{Q}) = \sup_{\|f\|_L \leq 1} \mathbb{E}_{M \sim \mathcal{M}} \left(\mathbb{E}_{X \sim \mathcal{P}} f(X \odot M) - \mathbb{E}_{Y \sim \mathcal{Q}} f(Y \odot M) \right)$$

where $\odot$ represents pointwise multiplication. It can be shown that if the data is *Missing Completely At Random* (MCAR, see Rubin (1976) for classifications of missing data) then $W_{\mathcal{M}}$ and W generate the same topology on the space of measures on $\mathbb{R}^d$, so that a sequence of measures $\mathcal{Q}_i$ (representing a sequence of improving generators) will converge to $\mathcal{P}$ w.r.t. the Wasserstein distance if and only if they converge to $\mathcal{P}$ w.r.t. the $\mathcal{M}$-Wasserstein distance.

We approximate the $\mathcal{M}$-Wasserstein distance analogously to the WGAN-GP approach. Let $\mathbf{m}_i$ be the mask corresponding to data point $\mathbf{x}_i$, then using our previous notation, we train the critic to maximise

$$\frac{1}{n} \sum_i C(\mathbf{x}_i \odot \mathbf{m}_i) - \frac{1}{n} \sum_i C(\mathbf{y}_i \odot \mathbf{m}_i) - \lambda \frac{1}{n} \sum_i \left(\|\nabla C(\mathbf{z}_i \odot \mathbf{m}_i)\|_2 - 1 \right)^2 .$$

Here we interpret $\mathbf{x}_i \odot \mathbf{m}_i$ as replacing the missing values in $\mathbf{x}_i$ with zeros, and $\mathbf{y}_i \odot \mathbf{m}_i$ replaces the corresponding values of $\mathbf{y}_i$ with zeros. It is this modified critic that defines the MaWGAN methodology.

2.2 Two-stage GAN model for time-series data

Let $\ldots, X_{-1}, X_0, X_1, \ldots$ be a real-valued stationary time-series with dependency of lag k. That is, X_i is conditionally independent of X_{i-k-j} for $j \geq 1$, conditioned on $(X_{i-1}, \ldots, X_{i-k})$. Let $M_i = 0$ if X_i is missing and 1 if not. We will assume that the M_i are independent of each other and of the X_i (so the X_i are MCAR). Our target is the conditional distribution of $X_i | (X_{i-1}, \ldots, X_{i-k})$. This is completely determined by the joint distribution of $(X_i, X_{i-1}, \ldots, X_{i-k})$, and so our approach is to use MaWGAN to fit a generator G and critic C to this joint distribution—which we can do in the presence of missing values—then use them to train a model for the conditional distribution.

Let $x_1, \ldots, x_n$ and $m_1, \ldots, m_n$ be observations of $X_1, \ldots, X_n$ and $M_1, \ldots, M_n$ respectfully, and put $\mathbf{u}_i = (x_i, \ldots, x_{i+k})$ and $\mathbf{v}_i = (m_i, \ldots, m_{i+k})$ for $i = 1, \ldots, n - k$. Given the $\mathbf{u}_i$ and $\mathbf{v}_i$ we use MaWGAN to train a generator $G : (0,1)^{k+1} \to \mathbb{R}^{k+1}$ and critic $C : \mathbb{R}^{k+1} \to \mathbb{R}_+$. If $\mathbf{z}$ is a vector of independent $U(0,1)$ random variables, then the distribution of $G(\mathbf{z})$ will approximate that of $(X_i, \ldots, X_{i+k})$. Our goal is to train a forecaster:

$$F : \mathbb{R}^k \times (0,1) \to \mathbb{R}.$$

For $Z \sim U(0,1)$ we want $F(x_1, \ldots, x_k, Z) \sim X_{k+1} | (X_1 = x_1, \ldots, X_k = x_k)$. To train F we generate $\mathbf{w} = (w_1, \ldots, w_{k+1})$ from G and z from a $U(0,1)$ then put

$$\mathbf{y} = (w_1, \ldots, w_k, F(w_1, \ldots, w_k, z))$$

Repeat this m times to get a synthetic sample $\mathbf{y}_i$, $i = 1, \ldots, m$, where m is the batch size. The $\mathbf{y}_i$ should look like realisations of $(X_1, \ldots, X_{1+k})$, so we can measure the performance of F by comparing the sample $\mathbf{y}_1, \ldots, \mathbf{y}_m$ to a batch of m of the $\mathbf{u}_i$, and we can do this comparison using the critic C.

As is usual for GANs, we iteratively train F and C, but we don't continue training G. Continuing to train C should encourage the critic to concentrate on the distribution of the last element of the sample given the other k elements, because the training of F will have no effect on the distribution of the first k elements being passed to the critic. Once F is trained we can use it to forecast using the most recent k observations, or generate synthetic series (starting with a single sample from G).

2.3 Pseudo-code

We suppose that for $i = 1, \ldots, n-k$ we have data $\mathbf{u}_i$ and masks $\mathbf{v}_i$, which are contiguous subsequences of length $k+1$ taken from a time-series with missing values. We have a generator $G : (0,1)^{k+1} \to \mathbb{R}^{k+1}$, critic $C : \mathbb{R}^{k+1} \to \mathbb{R}_+$ parameterised by weights $\boldsymbol{\theta}_C$, and a forecaster $F : \mathbb{R}^k \times (0,1) \to \mathbb{R}$ parameterised by weights $\boldsymbol{\theta}_F$. We assume that G and C have already been trained with data $\mathbf{u}_i$ and masks $\mathbf{v}_i$ using MaWGAN.

For any $\mathbf{w} = (\mathbf{w}(*), w(k+1)) \in \mathbb{R}^{k+1}$ we will write $\mathbf{w}(*)$ for the first k elements and $w(k+1)$ for the last element.

Require: forecaster weights $\boldsymbol{\theta}_F$ and critic weights $\boldsymbol{\theta}_C$, learning rates α_F and α_C.
Require: num. epochs t_F, forecaster batch size m_F, critic iterations t_C, critic batch size m_C, critic regularisation λ.

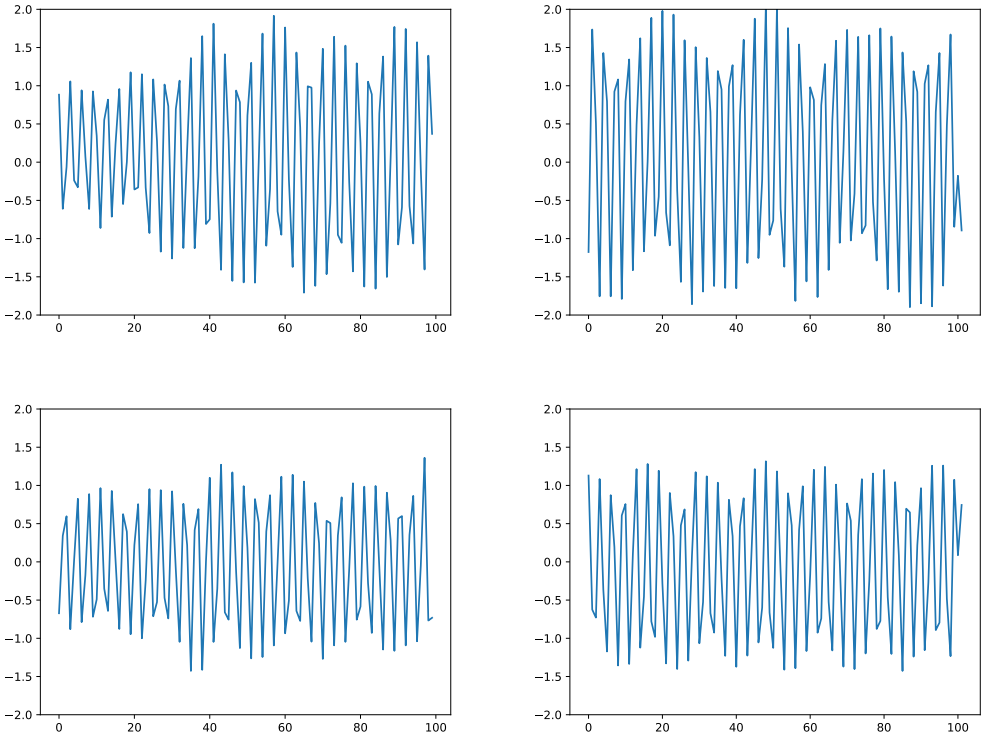
for $s = 1, \ldots, t_F$ **do** ▷ update the forecaster
  for $t = 1, \ldots, t_C$ **do** ▷ update the critic
    choose a batch σ of size m_C from $\{1, \ldots, n-k\}$
    for $i = 1, \ldots, m_C$ **do** ▷ calculate critic loss
      $\bar{\mathbf{u}}_i \leftarrow \mathbf{u}_{\sigma(i)} \odot \mathbf{v}_{\sigma(i)}$
      $\mathbf{w}_i = (\mathbf{w}_i(*), w_i(k+1)) \leftarrow G(\mathbf{a})$ for $\mathbf{a} \sim U(0,1)^{k+1}$
      $w_i(k+1) \leftarrow F(\mathbf{w}_i(*), b)$ for $b \sim U(0,1)$
      $\bar{\mathbf{w}}_i \leftarrow \mathbf{w}_i \odot \mathbf{v}_{\sigma(i)}$
      $\mathbf{z}_i \leftarrow \epsilon \bar{\mathbf{u}}_i + (1-\epsilon)\bar{\mathbf{w}}_i$ for $\epsilon \sim U(0,1)$
      $L_C^i \leftarrow C(\bar{\mathbf{u}}_i) - C(\bar{\mathbf{w}}_i) - \lambda(\|\nabla C(\mathbf{z}_i)\|_2 - 1)^2$
    end for
    $L_C \leftarrow \frac{1}{m_C} \sum_{i=1}^{m_C} L_C^i$
    update $\boldsymbol{\theta}_C$ using gradient of L_C (increasing L_C) and learning rate α_C
  end for
  for $i = 1, \ldots, m_F$ **do** ▷ calculate forecaster loss

16: $\mathbf{w}_i = (\mathbf{w}_i(*), w_i(k+1)) \leftarrow G(\mathbf{a})$ for $\mathbf{a} \sim U(0,1)^{k+1}$
17: $w_i(k+1) \leftarrow F(\mathbf{w}_i(*), b)$ for $b \sim U(0,1)$
18: $L_F^i \leftarrow C(\mathbf{w}_i)$
19: **end for**
20: $L_F \leftarrow \frac{1}{m_F} \sum_{i=1}^{m_F} L_F^i$
21: update $\boldsymbol{\theta}_F$ using negative gradient of L_F (decreasing L_F) and learning rate α_F
22: **end for**

3. TEST CASE

To test the method we used sequences of length 100 generated from an AR(3) model with parameters $(0.1, -0.3, 0.9)$ and error variance 0.1^2. That is $X_n = 0.1X_{n-1} - 0.3X_{n-2} + 0.9X_{n-3} + \epsilon_n$ where the ϵ_n are i.i.d. $N(0, 0.1^2)$. Some typical training sequences are plotted in Figure 1.

Figure 1: *LEFT: Samples of length 100 from an $AR(3)$ with parameters $(0.1, -0.3, 0.9)$ and error variance 0.1^2. RIGHT: Samples of length 100 from a GAN forecaster trained using the sample to the left (with no missing data).*



In Figure 1 we also plot some synthetic data from the models trained using each sample (a separate model was trained for each sample). Qualitatively the GAN forecaster

performs well, capturing both the high and low frequency oscillations in the data. We also observe that a sequence of length 100 is not long enough to capture the full variety of behaviour that the AR(3) process can exhibit. For example in Figure 1 the scale of the output top left is clearly different to the scale of the output bottom left, even though both are realisations of the same process. In both cases the GAN forecaster has got the scale correct, however we only expect the GAN forecaster to synthesise output at the scale to which it has been trained. That is, while the same AR(3) model produced both the upper and lower output on the left, we don't expect the GAN forecaster that produced the top right output to be able to produce the bottom right output, and vice versa. To get a single GAN forecaster that could reproduce both we would need to train it on a sequence long enough that it contained both types of behaviour.

3.1 Performance measures

Quantifying the performance of the GAN forecaster is not straight-forward, as we are not interested in its performance as a forecaster, but rather how well it captures the (temporal) dependence structure of the data. We used two performance measures. The first is the *Likeness Score* L introduced by Guan & Loew (2020). Suppose we have observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ from some distribution and observations $\mathbf{y}_1, \ldots, \mathbf{y}_m$ from a second distribution, then to calculate L we first generate three auxiliary sets of information

$$\begin{aligned} S_{\mathbf{x}} &= \{\|\mathbf{x}_i - \mathbf{x}_j\|_2\}_{i \neq j} \\ S_{\mathbf{y}} &= \{\|\mathbf{y}_i - \mathbf{y}_j\|_2\}_{i \neq j} \\ S_{\mathbf{x},\mathbf{y}} &= \{\|\mathbf{x}_i - \mathbf{y}_j\|_2\}_{i,j} \end{aligned}$$

For $A, B \subset \mathbb{R}$ let $\kappa(A, B) \in [0, 1]$ be the Kolmogorov-Smirnov distance between A and B, namely the maximum absolute difference between the empirical cumulative distribution functions of A and B. The Likeness Score for our two sets of observations is then

$$L = 1 - \kappa(S_{\mathbf{x}}, S_{\mathbf{x},\mathbf{y}}) \vee \kappa(S_{\mathbf{y}}, S_{\mathbf{x},\mathbf{y}}),$$

where $\vee$ indicates the maximum. Note that $L \in [0, 1]$ and the two sets of observations have likeness one if and only if they are identical, with lower scores indicating greater dissimilarity. To have a likeness of zero the two datasets would need to have disjoint ranges.

In our application the $\mathbf{x}_i$ will always be the $\mathbf{u}_i$ defined in Section 2.2, and the $\mathbf{y}_i$ will be the analogous subsequences taken from a sample of synthetic data generated by a GAN forecaster (the same length as the original sample).

Our second performance measure is more ad-hoc. Let $\boldsymbol{\theta} = (0.1, -0.3, 0.9)$ be the coefficients of our AR(3) process and let $\hat{\boldsymbol{\theta}}$ be the usual maximum likelihood estimate of $\boldsymbol{\theta}$, calculated using a synthetic sample from the GAN forecaster, then our performance measure is just the mean-squared error $M := \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|_2^2/3$, which we denote the *AR Coefficient Score*. Clearly smaller values are better.

To reduce the variation due to sampling from the generator we calculate L and M 100 times using different samples of synthetic data, then take the average. In what follows we take L and M to be these averaged values. To allow for the variation in performance due to the original sample used and the stochastic nature of the GAN fitting, we generated 30 different samples from the AR(3) process and fitted a GAN to each one.

When estimating the performance of the GAN forecaster there are three sources of variation:

- From the sampling of the data. As this is a simulation experiment we can gauge this by using multiple independent data samples (30 in our case).
- From the sampling of the generator. We mitigate this by comparing each data sample with 100 independent synthetic samples and averaging the results.
- From the fitting of the generator. The process of fitting a GAN is stochastic due to the random selection of observation batches, and for each separate data sample we re-fit the GAN. We mitigate this by using a very large number of training iterations, to be reasonably confident that the GAN has converged.

The variation represented by the confidence intervals in Figures 2 and 3 below is mainly due to the sampling of the data, though is necessarily confounded with the other two sources of variation. The length n of the data samples will also clearly effect the performance of the GAN model, however we do not explore this here and just fix $n = 100$.

3.2 Results

To assess the effect of missing data, for each original AR(3) sequence we generated six auxiliary sequences with increasing levels of missing data: 10%, 20%, ..., 60%. Observations were removed uniformly and independently, so the data is Missing Completely At Random (MCAR). The auxiliary sequences were nested so that all the points missing in one are also missing in those with higher levels of missing data.

For our experiments we took $k = 3$, and the critic and generator both had 1 hidden layer with 100 nodes. For the initial training of the generator and critic we used learning rate $\alpha = 0.0001$; batch size 30; iterations $t_G = 5000$ and $t_C = 10$; and critic regularisation $\lambda = 10$ (using the notation of Poudevigne-Durance et al. 2022). For the training of the forecaster we found that the same parameters worked, namely the learning rates were $\alpha_F = 0.0001$ and $\alpha_C = 0.0001$; batch sizes were $m_F = 30$ and $m_C = 30$; iterations were $t_F = 5000$ and $t_C = 10$; and the critic regularisation was $\lambda = 10$. Some optimisation of these parameters is required, as for any WGAN-GP based algorithm, but this was not done particularly systematically.

The performance of the GAN forecaster is summarised in Figures 2 (Likeness Score) and 3 (AR Coefficient Score). For both figures we give the average performance of the GAN forecaster for different levels of missing data (with 95% confidence intervals). For both measures the performance of the GAN forecaster is not much effected by levels of missingness up to 40%.

Figure 2: *Likeness Scores. Each point is the mean of 30 calculations of the Likeness Score L, comparing a sequence of length 100 from an AR(3) process with synthetic data generated using a GAN forecaster, trained using the AR(3) sample. The error bars give 95% confidence intervals. The level of missing data is the proportion of observations removed at random from the original sample before training the GAN forecaster, though note that the Likeness Score is always calculated using the original data with no missing observations.*

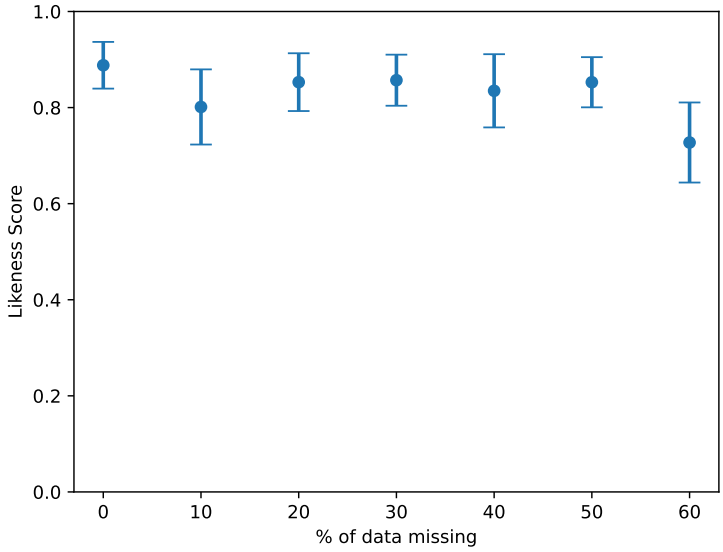
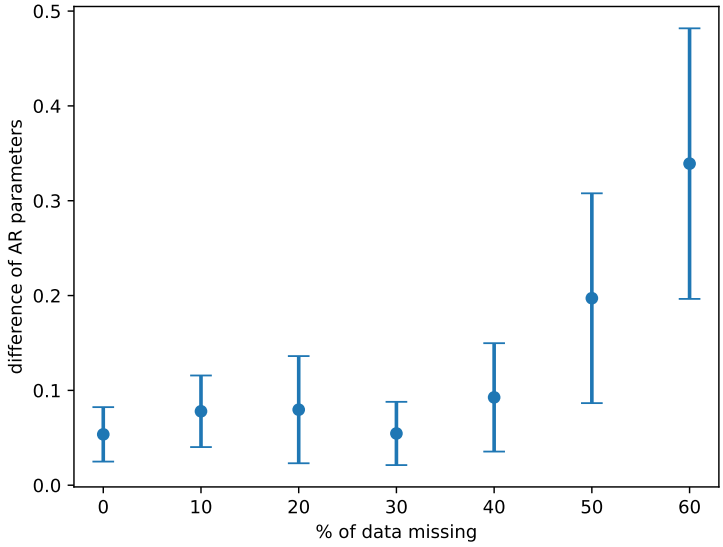


Figure 3: *AR Coefficient Scores. Each point is the mean of 30 calculations of M, the average squared distance between the true parameters of an AR(3) process and the estimated parameters from a synthetic sample generated using a GAN forecaster, trained using a sample of size 100 from the AR(3) model. The error bars give 95% confidence intervals. The level of missing data is the proportion of observations removed at random from the original sample before training the GAN forecaster.*



To have a basis for comparison we fitted an AR(3) process to each of the original samples we generated, and then calculated L and M as above. For the AR Coefficient Score we got a mean of 0.0027 with 95% CI $(0.0012, 0.0042)$, and for the Likeness Score we got a mean of 0.809 and 95% CI $(0.751, 0.867)$. That is, according to the AR Coefficient Score the fitted AR(3) processes gives a better fit, which is to be expected as we are using a correctly specified model, however according to the Likeness Score the GAN forecaster gives a better fit. Possibly this is because the GAN forecaster fits to the sample it is given and the Likeness Score is measuring how well you match that sample, rather than the parameters of the original process. Formally we can interpret this as saying the GAN forecaster is doing a good job "locally" but could do better "globally". As noted at the start of Section 3, a sample of length 100 is probably too short for the GAN forecaster to learn all the possible behaviours that this AR(3) process can exhibit, so better global performance would require a larger data sample. In future work we will compare the performance of our GAN forecaster with the methods of Esteban et al. (2017) and Yoon et al. (2019), in the case of no missing data.

4. DISCUSSION

Our GAN forecaster has produced some promising results, however more testing is required, and there is scope for generalising and tuning the method. At the time of writing our systematic experimentation has been restricted to the case presented in Section 3, however clearly of interest for further investigation is the effect on performance of the length of the data sequence n and the lag k. Moreover our method easily generalises to multivariate time-series, and indeed can be used to model any conditional distribution, so it would also be of interest to see how well it can capture the dependencies of time-series models such as the Periodic Autoregressive or Spatial Autoregressive (see e.g. Holan et al. 2010, LeSage & Pace 2009).

In practice securing the convergence of a GAN requires a balance between the speed at which the generator and critic converge. Our methodology requires the convergence of the generator, critic and forecaster, so we have three things balance. We simplified the interplay of the three nets by training the generator and critic first, then switching to the critic and forecaster, however allowing simultaneous training of all three nets could improve the overall speed at which they converge, and improve their performance, but at the expense of more tuning parameters. We also need to be mindful that the initial training of the critic may mean it is too specialised for the early training of the forecaster, causing it to focus on details rather than broad features. One way of alleviating this problem somewhat is to pause training of the critic while the forecaster "catches up".

It is clear that in practice the choice of lag is important and may not be as straightforward as for our case study. The obvious guideline is that the lag should be large enough to encompass any features in the data, such as seasonal effects or irregular cycles, however the larger the lag the more complicated the forecaster has to be, which translates into more and larger internal layers for all the GAN nets (generator, critic and forecaster), making the fitting slower and more temperamental.

Finally we note that the effect of the dependencies between the $\mathbf{u}_i$ on the convergence of the GAN nets is something that warrants investigation, as is the question of how we use the architecture of these nets to exploit the temporal structure of the $\mathbf{u}_i$. For example dimension-reducing feature layers have proved effective in recurrent neural networks, and may do so here as well (Yoon et al. 2019).

ΠΕΡΙΛΗΨΗ

Στην παρούσα εργασία προτείνεται μια νέα μέθοδος σύνθεσης χρονοσειρών με τη χρήση των Generative Adversarial Networks (GANs) η οποία μπορεί να αξιοποιηθεί και στην περίπτωση ελλειπουσών παρατηρήσεων. Η μέθοδος μοντελοποιεί την δεσμευμένη κατανομή της τρέχουσας παρατήρησης δοθέντων των παρελθόντων παρατηρήσεων, κάτι που επιτυγχάνει με ένα βοηθητικό GAN εκπαιδευμένο στην από κοινού κατανομή της τρέχουσας και των παρελθόντων παρατηρήσεων. Ένα πλεονέκτημα της μεθόδου είναι η χρήση ενός Masked Wasserstein GAN (MaWGAN) για

την εκπαίδευση του μοντέλου το οποίο μπορεί να συνεκτιμήσει και τις ελλείπουσες παρατηρήσεις. Η αποτελεσματικότητα της μεθόδου ακόμα και για μεγάλα ποσοστά ελλιπουσών παρατηρήσεων, επιβεβαιώνεται με ένα πείραμα προσομοίωσης.

REFERENCES

Arjovsky, M., Chintala, S. and Bottou, L. (2017). Wasserstein Generative Adversarial Networks. *Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017*, pp.214–223.

Campbell, M. (2019). Synthetic data: How AI is transitioning from data consumer to data producer and why that's important. *Computer*, **52**(10):89–91.

Esteban, C., Hyland, S.L. and Rätsch, G. (2017). Real-valued (medical) time series generation with recurrent conditional GANs, *arXiv*:1706.02633.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014). Generative Adversarial Nets. *Advances in Neural Information Processing Systems*, **27**.

Guan, S. and Loew, M.H. (2020). Measures to evaluate Generative Adversarial Networks based on direct analysis of generated images. *arXiv*: 2002.12345.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V. and Courville, A.C. (2017). Improved training of Wasserstein GANs. *Advances in Neural Information Processing Systems*, **30**.

Hitawala, S. (2018). Comparative Study on Generative Adversarial Networks, *arXiv*: 1801.04271.

Holan, S.H., Lund, R. and Davis, G. (2010). The ARMA alphabet soup: A tour of ARMA model variants. *Statistics Surveys*, **4**:232–274.

Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., Cohen, S.N. and Weller, A. (2022). Synthetic Data–what, why and how?. *arXiv*:2205.03257.

Kaloskampis, I., Pugh, D., Joshi, C. and Nolan, L. (2019). Synthetic data for public good. *ONS Data Science Campus blog.* `https://datasciencecampus.ons.gov.uk/projects/synthetic-data-for-public-good/`

Kantorovich, L.V. and Rubinstein, G.Sh. (1958). On a space of completely additive functions. *Ser. Mat. Mekh. i Astron.*, **13**(7):52–59. (In Russian)

LeSage, J. and Pace, R.K. (2009). *Introduction to spatial econometrics.* Chapman and Hall/CRC.

Li, C.L., Chang, W.C., Cheng, Y., Yang, Y. and Póczos, B. (2017). MMD GAN: Towards deeper understanding of moment matching network. *Advances in Neural Information Processing Systems*, **30**.

Mogren, O. (2016). C-RNN-GAN: Continuous recurrent neural networks with adversarial training, *arXiv*:1611.09904.

Nowozin, S., Cseke, B. and Tomioka, R. (2016). f-GAN: Training generative neural samplers using variational divergence minimization. In *Proceedings of the*

30th International Conference on Neural Information Processing Systems, 2016, pp.271–279.

Pacchiardi, L., Adewoyin, R., Dueben, P. and Dutta, R. (2021). Probabilistic forecasting with generative networks via scoring rule minimization. *arXiv*:2112.08217.

Poudevigne-Durance, T., Jones, O.D. and Qin, Y. (2022). MaWGAN: a Generative Adversarial Network to create synthetic data from datasets with missing data. *Electronics*, **11**:837.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, **63**(3):581–592.

Sallier, K. (2020). Toward more user-centric data access solutions: producing synthetic data of high analytical value by data synthesis. *Statistical Journal of the IAOS*, **36**(4):1059–1066.

Yoon, J., Jordon, J. and van der Schaar, M. (2018). GAIN: Missing Data Imputation using Generative Adversarial Nets, *arXiv*:1806.02920.

Yoon, J., Jarrett, D. and van der Schaar, M. (2019). Time-series Generative Adversarial Networks. *Advances in Neural Information Processing Systems*, **32**.

Greek Statistical Institute
Proceedings of the 34th Panhellenic Statistics Conference (2022) pp.167-182

THE LEPTO-VARIANCE OF STOCK RETURNS

Vassilis Polimenis
Aristotle University
polimenis@auth.gr

ABSTRACT

The Regression Tree (RT) sorts the samples using a specific feature and finds the split point that produces the maximum variance reduction from a node to its children. Our key observation is that the best factor to use (in terms of MSE drop) is always the target itself, as this most clearly separates the target. Thus, using the target as the splitting factor provides an upper bound on MSE drop (or lower bound on the residual children MSE). Based on this observation, we define the k-bit lepto-variance λk^2 of a target variable (or equivalently the lepto-variance at a specific depth k) as the variance that cannot be removed by any regression tree of a depth equal to k. As the upper bound performance for any feature, we believe λk^2 to be an interesting statistical concept related to the underlying structure of the sample as it quantifies the resolving power of the RT for the sample. The max variance that may be explained using RTs of depth up to k is called the sample k-bit macro-variance. At any depth, total sample variance is thus decomposed into lepto-variance λ^2 and macro-variance μ^2. We demonstrate the concept, by performing 1- and 2-bit RT based lepto-structure analysis for daily IBM stock returns.

Keywords: regression tree, stock returns, financial factors, idiosyncratic variance, lepto-variance, macro-variance

1. INTRODUCTION

In many machine learning applications, from biostatistics, to feature selection there is an interest in finding features that explain a large fraction of the total variance of a target outcome.

In regression analysis, the residual sum of squares (RSS) or the idiosyncratic variance depend on the factors used in the regression. In general, we can get lower residual variance by adding an extra factor. Thus, idiosyncratic (or unexplained) variance depends on the factors used to explain the dependent variable.

Here, by using the Regression Tree (RT) methodology, we define a new type of idiosyncratic variance, the **lepto-variance** that is uniquely defined for a sample regardless of the set of factors used to explain the variable. We define the **macro-variance** of a sample as the upper bound of the sample variability that may be explained by any attribute. Lepto- and macro-variance are features that depend solely on sample structure. Similar to the resolving power of an optical system (such as a microscope or

telescope) that places a limit on its ability to distinguish images or reproduce fine detail, the lepto-variance of a target places a limit on the ability of a k-bit RT in explaining fine structure in the sample.

In financial analysis, there is an interest in finding the financial factors that explain a large fraction of the total stock price variability. Stock return variance that cannot be explained by broad market factors is considered idiosyncratic for the specific stock. As the concept of lepto-variance offers a new statistical approach to measuring idiosyncratic volatility, it can have many significant machine learning and statistical applications. In financial analysis, it may help to better understand the relation of idiosyncratic volatility to prices. In this paper, the concept is demonstrated by performing 1- and 2-bit RT based lepto-structure analysis for daily IBM stock returns.

2. THE RESOLVING POWER OF A REGRESSION TREE

Usually, Regression Trees (RT) provide a binary splitting of the sample space with minimization criterion the Residual Sum of Squares $\Sigma j\ (yj - \bar{y}j)^2$, with $\bar{y}j$ the average y for the subspace that includes yj. Finding the best binary partition of the feature space for a regression tree is computationally infeasible. Hence, a recursive greedy algorithm is used. When splitting a categorical predictor that takes q possible unordered values, $2^{q-1} - 1$ possible partitions of the q values into two groups need to be evaluated. Such computation becomes prohibitive for large q. There are various "concavity" theorems that simplify the problem (see [Ripley (1996)] and [Hastie et al. (2009)]). Even before the publication of the first RT algorithms, [Fisher (1958)] showed that for a continuous-valued target Y the least squares partition of a set is contiguous. [Breiman et al. (1984)] extended for a decision tree with binary (2-class) target Y.

Instead of performing a search in an exponential number of possible binary partitions, a simplified linear search of only q-1 partitions where attribute values are sorted based on their strength of correlation with the target outcome suffices. Starting at the root node of the tree (#0) which includes all samples, the algorithm will consider all possible splitting factors (x) and split points (c) that define a pair of half-planes. For each factor x, the best split point c can be very quickly determined and hence by scanning through all factors, the best pair (x, c) can be feasibly determined, and the algorithm recursively repeats the splitting process for each of the two children nodes.

Effectively, at each node #j the RT sorts the sub-sample of this node Sj using the chosen split factor xj and finds the split point cj that produces the maximum MSE drop from the node to its children Lj and Rj. We may think of this MSE drop (variance[i] reduction) as an (informational or impurity) gain from splitting via this factor. With no loss of generality, assume that the left child L contains the small y values (i.e. assume $\bar{y}_L$ = mean(y | in L sub-sample) < $\bar{y}_R$ = mean(y | in R sub-sample)). Generally, sorting a sample based on a factor xj will not produce a sorted target y.

Definition 1. A binary split of S into L and R is *sorted* if all target values y in L are smaller than all target values in R.

Equivalently, a split is sorted if the maximum target value in L is smaller than the minimum target value in R.

We provide an extension of the [Fisher (1958)] theorem on grouping, with the following lemma for Regression Trees.

Lemma 1. In a Regression Tree, using the target itself as the predictor provides an upper bound on the MSE reduction.

Proof. We first show that any unsorted split is strictly dominated by a sorted split. Let us assume an unsorted split of a sample S into L and R (with no loss of generality, assume $\bar{y}_L < \bar{y}_R$). Then the maximum value of the left sub-tree u1 is larger than the minimum value of the right sub-tree u0

$$u1 = \max(L) > u0 = \min(R)$$

But then, we can get a better split by swapping u0 with u1 into L and R respectively. Because moving u0 into L and u1 into R, will move the center of the L sub-sample farther to the left and the center of the right sub-sample farther to the right, thus producing a larger separation between the left and right sub-samples without changing the relative sample sizes. By the law of total variance, a larger between-group variability means a smaller within-group variability and thus a better split.

Since for any unsorted split there exists a strictly better sorted split, the best binary split is always a sorted split. Using the target on itself would most clearly separate the target, as it can generate all sorted splits. Thus, no predictor can ever perform better (in terms of MSE drop) than the target itself.

2.1 The 1-bit Lepto-structure of a sample

To clarify the ideas an example is presented below. On Table 1 we see the values for two financial factors f1 and f2 and a hypothetical stock y for 8 days. There is an outlier return for the stock at t=3.

When returns are simply sorted by the date they occur, the best split is a group of the first 5 samples. Table 2a depicts MSE for the left and right children for all possible splits from position 1 to 8 and the total weighted children MSE in the last column.[ii] Figure 1a shows the optimal resulting depth 1 tree.

In this case the weighted MSE drops from 3.172 to 1.896 with an MSE drop of 40.23% = 100% - 1.896 / 3.172. In terms of outcome values, the split is {-0.5, -2.0, 0.0} and {1.5, -1.0, 4.0, 2.0, 1.0} - an unsorted split strictly inferior to the one we would get by swapping 0.0 with -1.0.

Table 1. *Hypothetical returns of two factors f1 and f2 and a stock y for 8 days (in percentages)*

t	f1	f2	y	
1	2.0	2.0	1.5	
2	1.8	6.2	-1.0	
3	5.0	1.8	4.0	* outlier
4	7.0	4.0	2.0	
5	6.0	6.0	1.0	
6	4.8	5.8	-0.5	
7	2.2	5.0	-2.0	
8	1.0	1.0	0.0	

When the 1st factor f1 is used (see Table 2b), the split point is f1<4.9 (i.e. the middle point between 4.8 and 5.0). The 5 samples t=1, 2, 6, 7, 8 belong to the left branch, while t=3, 4, 5 are put in the right branch. In this case the minimum weighted children MSE becomes 1.421 with a 55.21% MSE drop. Figure 1b shows the optimal RT. When the 2nd factor f2 is used, the split point is f2 < 4.5 (i.e. the middle point between 4.0 and 5.0). The 4 samples t=1, 3, 4, 8 belong to the left branch, while t=2, 5, 6, 7 are put in the right branch. In this case the weighted MSE becomes 1.609 with a 49.26% MSE drop. Figure 1c shows the optimal RT.

As proven in Lemma 1, regressing a series on itself achieves the upper bound of MSE drop and the lowest children MSE. In particular, here when the factor used is y, the optimal split point is $y \leq 0$ (i.e. equivalently y < .5, the middle point between 0 and 1.0). The 4 samples t=2, 6, 7, 8 belong to the left branch, while t=1, 3, 4, 5 are put in the right branch (see Table 2d). In this case, the lowest possible MSE becomes $\lambda 1^2$ = .922 with a 70.94% MSE drop. The symbol $\lambda 1^2$ is used for depth 1 **lepto-variance**. For obvious reasons, $\lambda 1^2$ is also called the ***1-bit lepto-variance*** of the sample.

Definition 2. The 1-bit sample lepto-variance $\lambda 1^2$ of a target variable is defined as the MSE of a depth 1 RT of the target on itself.

Effectively one may think of the remaining variance fraction 29.06% as structure beyond the resolving power of depth 1 trees; no factor may ever explain more than 70.94% of the y variability with a binary depth 1 RT. Thus, the ability of a factor to explain the target via a depth 1 RT should be compared against the benchmark of the ability of the target to explain itself (i.e. 70.94% in this case). The remaining 29.06% is variance due to the ***lepto-structure*** of the target (the terms lepto-variance and lepto-structure are used interchangeably). Figure 1d shows the optimal resulting tree.

***Figure 1a**. Split for the optimal depth=1 RT when simple time (t) is used as a factor at t<5.5 (5.5 is the mid-point between t=5 and 6)*

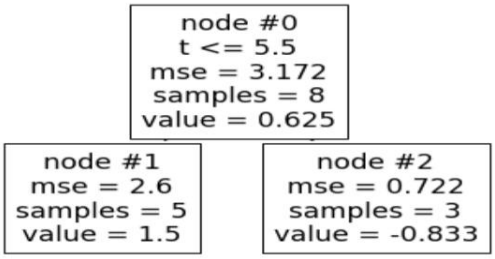


***Table 2a**. Optimal depth=1 RT when simple time (t) is used as the splitting factor; i.e. samples are sorted based on their t values.*

t	f1	f2	y	mse L	mse R	L+R
1	2.0	2.0	1.5	0.000	3.500	3.063
2	1.8	6.2	-1.0	1.563	3.646	3.125
3	5.0	1.8	4.0	4.167	1.840	2.713
4	7.0	4.0	2.0	3.172	1.172	2.172
5	6.0	6.0	1.0	2.600	0.722	1.896
6	4.8	5.8	-0.5	2.722	1.000	2.292
7	2.2	5.0	-2.0	3.561	0.000	3.116
8	1.0	1.0	0.0	3.172		
					MSE	1.896
					info gain	40.23%

***Figure 1b.** Split for the optimal depth=1 RT when the first factor is used at f1<4.9 (4.9 is used as the mid-value between 4.8 and 5.0).*

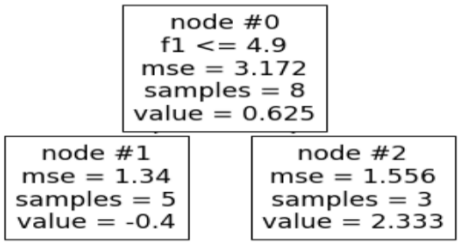


Table 2b. *Optimal depth=1 RT when the first factor f1 is used; i.e. samples are sorted based on their f1 values.*

t	f1	f2	y	mse L	mse R	L+R
8	1.0	1.0	0.0	0.000	3.561	3.116
2	1.8	6.2	-1.0	0.250	3.583	2.750
1	2.0	2.0	1.5	1.056	4.240	3.046
7	2.2	5.0	-2.0	1.672	2.672	2.172
6	4.8	5.8	-0.5	1.340	1.556	1.421
3	5.0	1.8	4.0	3.806	0.250	2.917
5	6.0	6.0	1.0	3.316	0.000	2.902
4	7.0	4.0	2.0	3.172		
					MSE	1.421
					info gain	55.21%

Figure 1c*. Split for the optimal depth=1 RT when the second factor is used at f2<4.5*

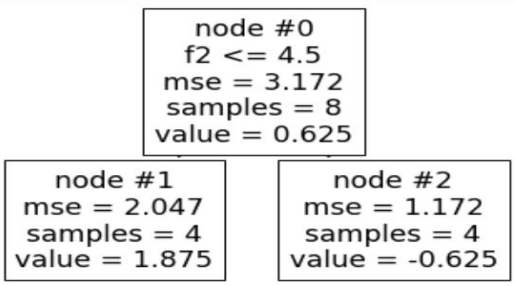


Table 2c. *Optimal depth=1 RT when the 2nd factor is used; i.e. samples are sorted based on their f2 values.*

t	f1	f2	y	mse L	mse R	L+R
8	1.0	1.0	0.0	0.000	3.561	3.116
3	5.0	1.8	4.0	4.000	2.056	2.542
1	2.0	2.0	1.5	2.722	2.040	2.296
4	7.0	4.0	2.0	2.047	1.172	1.609
7	2.2	5.0	-2.0	4.040	0.722	2.796
6	4.8	5.8	-0.5	3.722	1.000	3.042
5	6.0	6.0	1.0	3.194	0.000	2.795
2	1.8	6.2	-1.0	3.172		
					MSE	1.609
					info gain	49.26%

***Figure 1d**. Split for the optimal depth=1 RT when the target outcome is also used as the predictor (i.e. factor is y).*

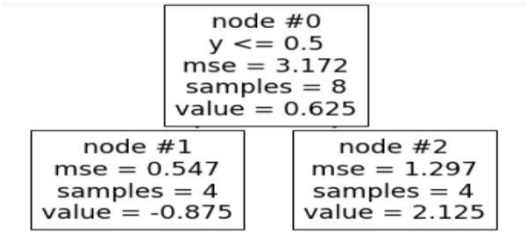


***Table 2d**. Optimal depth=1 RT when the stock return vector is regressed on itself (i.e. factor is y); i.e. samples are sorted based on their y values.*

t	f1	f2	y	mse L	mse R	L+R
7	2.2	5.0	-2.0	0.000	2.500	2.188
2	1.8	6.2	-1.0	0.250	2.139	1.667
6	4.8	5.8	-0.5	0.389	1.760	1.246
8	1.0	1.0	0.0	0.547	1.297	0.922
5	6.0	6.0	1.0	1.000	1.167	1.063
1	2.0	2.0	1.5	1.389	1.000	1.292
4	7.0	4.0	2.0	1.765	0.000	1.545
3	5.0	1.8	4.0	3.172		
					MSE	0.922
					info gain	70.94%

2.2 The 2-bit Lepto-structure of a sample

The concept of lepto-variance of a sample may also be defined for trees of a maximum depth larger than 1. As we move deeper down on a RT there will always be less residual variance. The argument of Lemma 1 will still be valid; at any node, the best split is always a sorted split. But the greediness of the RT may generate a situation where sorting in a split is sub-optimal (by not using the entire allowed max depth in some branches). Thus, although highly unlikely for realistic samples and relatively small depths, utilizing the target itself as the splitting variable may not always achieve the lowest residual error for an *allowed max depth*. For example, sorting the initial sample {0,-2,4,1} will isolate 4 prematurely. Thus, the RT will not be allowed to grow a full 2-bit tree structure that would explain the entire variability. But it will still achieve the lowest RSS at any *average* depth (1.75 bits for the example).[iii]

Definition 3. For any depth k, the k-bit sample lepto-variance λk^2 of a target variable is defined as the MSE of an (average) depth k RT of the target on itself.

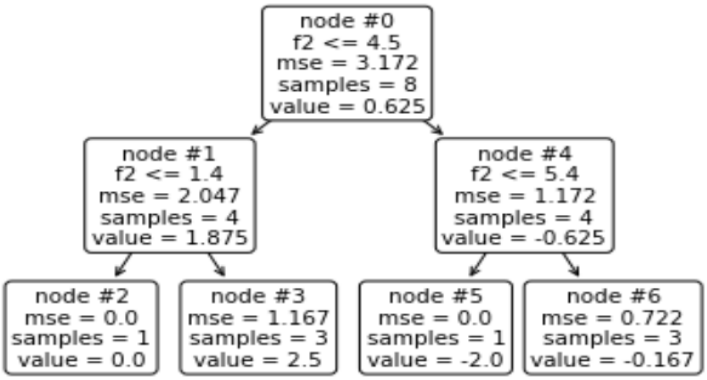
The notation $\mu 1^2$ is used to denote the max variance drop (**1-bit macro-variance**) on a root node; i.e. the MSE drop when the target vector is regressed on itself. Similarly, we

use $\mu 2^2$ to denote the max variance drop when using a 2-bit (depth 2) RT instead of a 1-bit RT etc. Total sample variance then equals

$$\sigma^2 = \lambda 0^2 = \mu 1^2 + \lambda 1^2 = \mu 2^2 + \lambda 2^2 = \ldots = \mu j^2 + \lambda j^2 \quad (1)$$

Thus, the k-bit lepto-variance is the minimum residual variance of any RT with depth k. The concept is made clear for 2-bit RTs with the hypothetical 8 sample stock returns of Table 1. On Figure 2a the optimal depth 2 RT split when the stock return vector is regressed on f2 twice is shown (i.e. the RT is restricted to use only f2 as a predictor at both levels). As before, the 1st split is based on f2 < 4.5, and then in internal nodes #1 and #4, split criteria f2 < 1.4 and f2 < 5.4 are respectively used. Residual MSE at depth 2 for the RT equals ⅜ · 1.167 + ⅜ · .722 = .70833. Similarly, on Figure 2b the optimal depth 2 RT split when the stock return vector is regressed on f1 twice is shown. In nodes #1 and #4, f1 < 2.1 and f1 < 5.5 are the splitting criteria. Residual MSE for the RT equals ⅜ · 1.056 + ¼ · .562 + ¼ · .25 = .60. On Figure 2c the optimal depth 2 RT split when the stock return vector is freely regressed on both factors (i.e. f1, and f2) is shown. Initially, factor f1 provides more information so the split on the 1st level is based as before on f1<=4.9. Then, in nodes #1 and #4, f2 is used.

Figure 2a*. Split for the optimal depth 2 RT when the target vector is regressed on f2 twice (only f2 at both levels).*



Residual MSE for the RT equals ¼ · .562 + ⅜ · .389 + ¼ · .25 = .348875 which is better than . 71 for f2 and .6 for f1, and even below the benchmark .92 1-bit RT with y regressed on itself (i.e., 1-bit lepto-variance). The 2-bit lepto-variance is calculated on Figure 2d via the optimal 2-bit benchmark RT when y is the predictor. Using y to sort, both extreme target values y = -2 and y = 4 are isolated. Residual MSE for the RT equals $\lambda 2^2$ = ⅜ · .167 + ⅜ · .167 = .125 which is lower than the .35 residual MSE for f1+f2 combined. The 2-bit macro-variance of y equals $\mu 2^2$=3.172 - .125 = 3.047.

***Figure 2b**. Split for the optimal depth=2 RT when the stock return vector is regressed on f1 twice (only f1 at both levels).*

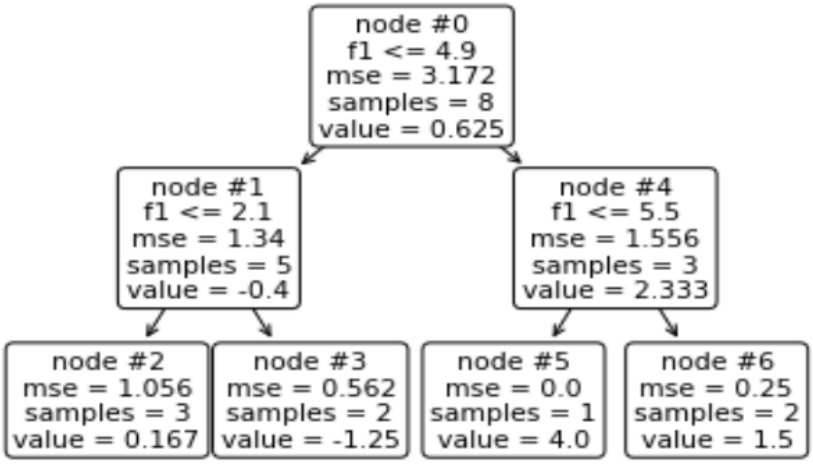


***Figure 2c**. Split for the optimal depth=2 RT when the stock return vector is freely regressed on both factors (i.e., f1, and f2).*

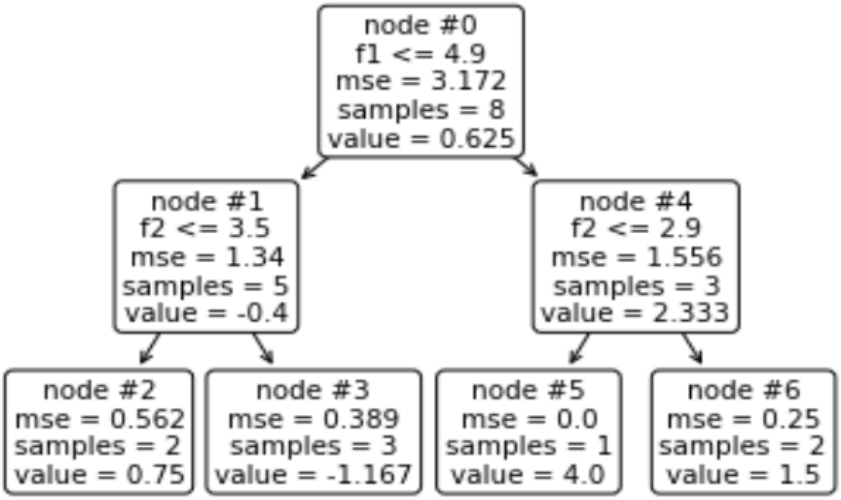
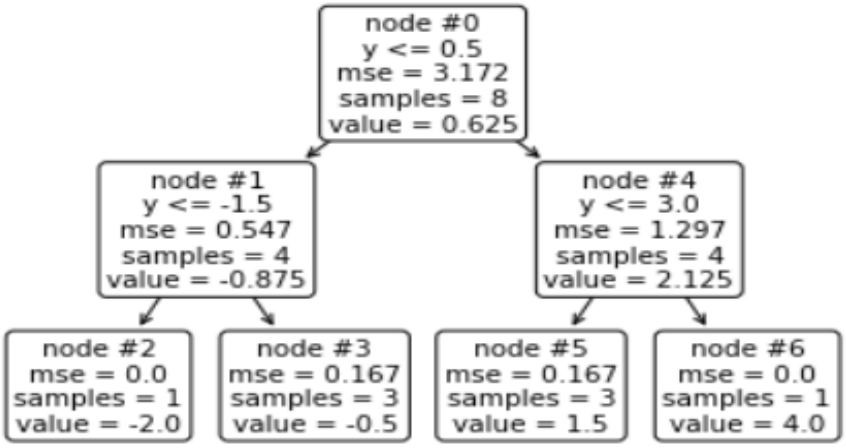


***Figure 2d**. Splits for the benchmark depth=2 RT when the stock return vector is twice regressed on itself. Residual MSE is the 2-bit lepto-variance of y.*



3. EXAMPLE LEPTO-VARIANCE ESTIMATION FOR A STOCK

The concepts of lepto- and macro-variance may be of interest for a host of machine learning applications from biostatistics to genomics. In finance, understanding the sources of stock volatility and how idiosyncratic volatility affects stock prices is a question of major significance for investors and portfolio managers and therefore it has been investigated empirically since virtually the inception of classical asset pricing theory. A well-known puzzle in finance is that stocks with greater idiosyncratic volatility are producing lower than expected returns. This is an anomaly because, according to economic theory, idiosyncratic volatility does not constitute a risk factor. See [Stambaugh et al. (2015)] for a discussion of the phenomenon and possible explanation.

Here, to further demonstrate the concept of lepto-variance an example from financial analysis, a 3-factor analysis with real stock return data, is performed. Specifically, 5 years of daily return data for a major US corporation, International Business Machines Corp. (IBM), for the period from 1/5/2015 to 30/4/2020 are analyzed. The sample comprises 1259 daily returns.[iv]

In financial asset pricing, a well-known set of features is the three-factor model of [Fama and French (1993)]. The three-factor model expands the simpler Capital Asset Pricing Model by [Sharpe (1964)] and [Lintner (1965)] based on a time-series linear regression of excess portfolio returns of the type

$$y(t) = a + b\cdot MEx(t) + s\cdot SMB(t) + h\cdot HML(t) + e(t) \quad (2)$$

y(t) is the excess return on a security for period t above the risk-free return. Factor MEx is the excess return on the value-weight (VW) market portfolio above the risk-free asset. Factor SMB is the return on a diversified portfolio of small stocks minus the

return on a diversified portfolio of big stocks for the period. Factor HML is the difference between the returns on diversified portfolios of high and low book-to-market (B/M) ratio stocks. [Fama and French (1996)] argue that the sensitivities b, s and h in (2) capture most variation in expected returns, so the true value of the intercept in (2) should be zero for all well -priced securities or portfolios.

Basic descriptive statistics for the daily returns of the 3 Fama-French factors and IBM stock for this 5-year period are shown in Table 1b. We utilize the lepto-variance framework developed here to demonstrate the relative power of the Fama and French three factors in explaining IBM return variability. Table 1c presents the correlation matrix of the three factors and IBM.

***Table 1b**. Descriptive statistics for the daily returns of the 3 Fama-French factors and IBM stock for the 5-year period from 1/5/2015 to 30/4/2020 (in percentages)*

	MEx	SMB	HML	IBM
count	1259			
mean	0.037	-0.018	-0.035	0.005
std	1.20	0.59	0.67	1.56
min	-12.00	-4.58	-4.71	-12.85
25%	-0.33	-0.34	-0.39	-0.62
50%	0.05	-0.04	-0.06	0.05
75%	0.52	0.31	0.30	0.67
max	9.34	5.73	3.19	11.30

***Table 1c**. Correlation matrix of the 3 factors and IBM returns.*

	MEx	SMB	HML	IBM
MEx	1			
SMB	0.145	1		
HML	0.137	0.219	1	
IBM	0.734	0.046	0.147	1

The 1-bit lepto-structure of IBM is recovered via a depth 1 RT of IBM returns on IBM itself. The optimal depth 1 RT when IBM is regressed on itself is shown in Figure 3a. At the root of the tree (node #0) the entire sample is included with mean return .005, and total MSE of 2.444. Optimal split is for IBM < -.365 which is valid for 32.6% of the sample. The remaining 67.4% of the samples belong to the right child. Remaining depth 1 lepto-variance equals $\lambda 1^2$ = .326 x 1.894 + .674 x 1.234 = 1.449. The 1-bit **macro-variance** of IBM equals the MSE drop from the original 2.444 total sample MSE, $\mu 1^2$ = 2.444 – 1.449 = .995.

Next we run a depth 1 RT where the IBM return vector is regressed on all 3 Fama-French factors (i.e., the market excess return MEx and the SMB and HML factors). The optimal depth 1 RT is shown on Figure 3b. The MEx factor dominates with a residual MSE = .137 x 2.984 + .863 x 1.771 = 1.937. Thus, MEx explains 2.444 - 1.937= .507 of IBM variance which represents .507/.995 = 51% of the total IBM 1-bit macro-variance.

Figure 3a*. The optimal depth 1 RT when the IBM return vector (IBM) is regressed on itself.*

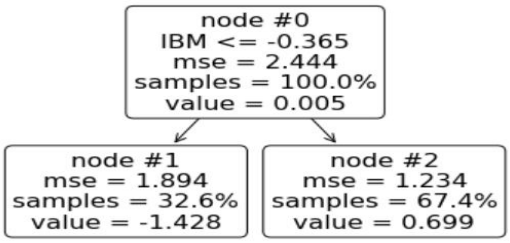


Figure 3b*. Optimal depth 1 RT when the IBM return vector is regressed on all 3 factors*

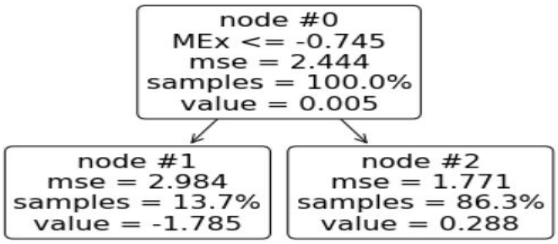
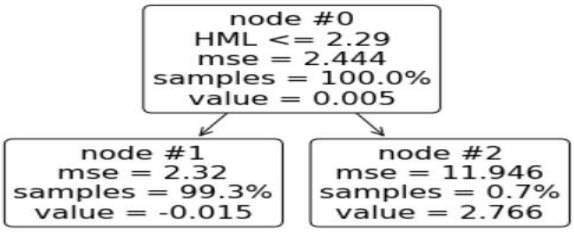


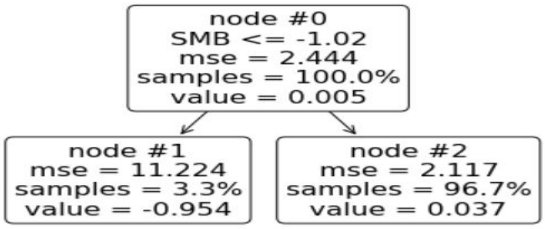
Figure 3c*. Optimal depth 1 RT for IBM regressed only on the Fama-French SMB and HML factors.*



Next, we run a depth 1 RT where the IBM return vector is regressed on the Fama-French SMB and HML factors (i.e., the MEx feature is excluded from the RT to force the use of SMB or HML). The optimal depth 1 RT is shown on Figure 3c. Between the 2 features, HML dominates SMB and has a residual MSE of .993 x 2.32 + .007 x 11.946 = 2.389. Thus, HML explains .055 of the total MSE which represents .055/.995 = 5.5% of IBMs 1-bit macro-structure. Finally, we run a depth 1 RT where the IBM return vector is regressed only on the SMB factor (i.e., both MEx and HML features are excluded from the RT to force the use of SMB). The optimal depth 1 RT is shown on

Figure 3d. Use of the SMB factor generates a residual MSE of .033 x 11.224 + .967 x 2.117 = 2.413. Thus, SMB only explains 3.12% of the 1-bit macro-variance for IBM.

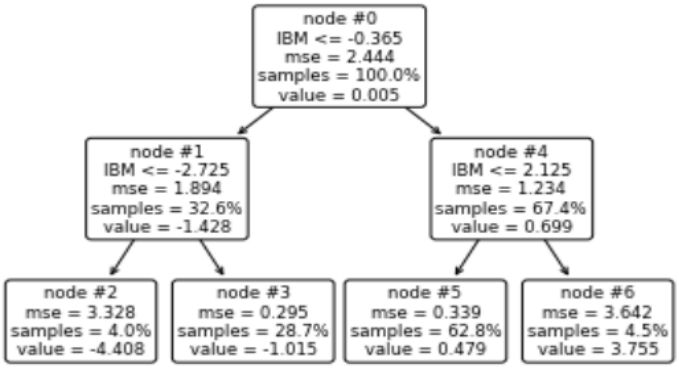
Figure 3d*. Optimal depth 1 RT when the IBM return vector is regressed only on the SMB factor*



3.1 The 2-bit Lepto-structure for IBM

We now focus on the 2-bit lepto-variance for IBM. On Figure 4a, the split for the optimal depth 2 RT when the IBM return vector is regressed on itself is shown. The 2-bit lepto-variance is the residual MSE $\lambda 2^2$ = .04 x 3.328 + .287 x .295 + .628 x .339 + .045 x 3.642 = .594. This means that a 2.444 - .594 = 1.85 has been explained using a 2-bit RT. The first bit explained $\mu 1^2$ = .995, and the 2nd level explained an extra $\mu 2^2$ - $\mu 1^2$ = 1.85 - .995 = .855. In total, 75.7% = 1.85/2.444 of total IBM variability is macro-structure at the 2-bits level, while the remainder 24.3% is the 2-bit lepto-structure.

Figure 4a*. Split for the optimal depth 2 RT when the IBM return vector is regressed on itself.*



On Figure 4b, the split for the optimal depth 2 RT when the IBM return vector is regressed on the 3 factors is shown. The Market return factor dominates with a depth 2 residual MSE = 1.466, implying a total explained variance .978/1.85 = 52.86% of the full 2-bit IBM macro-structure.

On Figure 4c, the split for the optimal depth 2 RT when the IBM return vector is regressed on the Fama-French SMB and HML factors is shown. Residual variance equals 2.325, implying a total explained variance at level 2 of only .119/1.85 = 6.43% of the 2-bit macro-variance. Finally, the optimal split (not shown) for the 2-bit RT when

the IBM return vector is regressed only on the SMB factor (the lowest correlation feature) has residual MSE = 2.380, implying a total explained variance at level 2 of only .064/1.85 = 3.46% of the 2-bit macro-variance.

***Figure 4b**. Split for the optimal depth 2 RT when the IBM return vector is regressed on the 3 factors.*

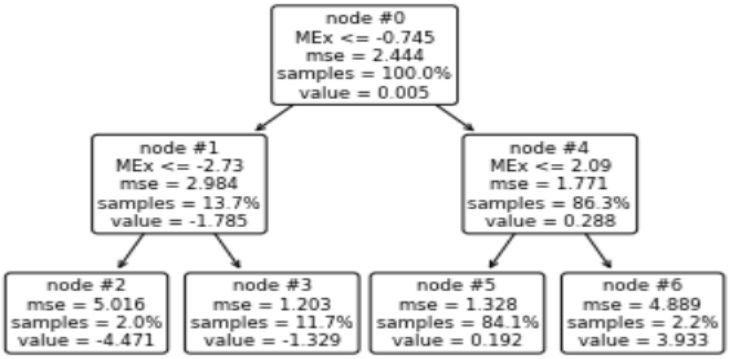
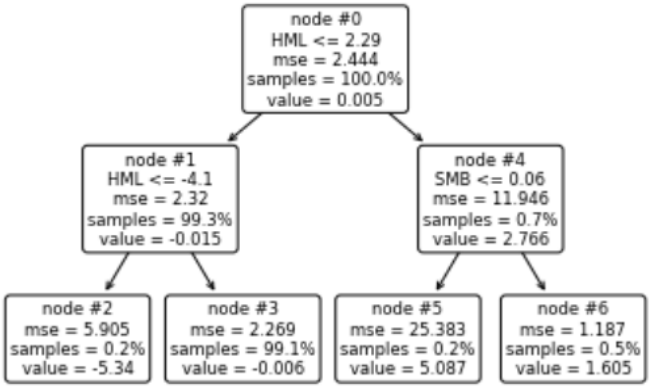


***Figure 4c**. Split for the optimal depth 2 RT when the IBM return vector is regressed on the Fama-French SMB and HML factors.*



4. CONCLUSION

It is shown here that since in a Regression Tree (RT) it is always beneficial to generate a sorted split of a sample S, using the target itself as predictor provides an upper bound in terms of the sample variability that can be explained. The k-bit lepto-variance of a sample is defined as the residual MSE after the sample has been regressed on itself up to k times. The k-bit macro-variance is the variance captured by the target itself, and thus represents the maximum variance that can be captured by any combination of features. Total sample variance is decomposed in macro- and lepto-variability. The

existence of lepto-variance in a sample places a limit on the resolving power of a k-bit RT in explaining its fine structure. As a demonstration, the lepto-structure analysis of 5-years of daily IBM stock returns is performed. It is shown that while market movements may explain slightly more than 50% of the return macro-structure, the SMB and HML factors can only explain negligible amounts of the 1-bit and 2-bit macro-structure of IBM returns. As the concept lepto-variance offers a new statistical approach to analyzing variance, in financial analysis, it may help to better understand the relation of idiosyncratic volatility to prices and further investigate the well-known idiosyncratic volatility puzzle in finance.

ΠΕΡΙΛΗΨΗ

Το δέντρο παλινδρόμησης (RT) ταξινομεί τα δείγματα χρησιμοποιώντας ένα συγκεκριμένο χαρακτηριστικό και βρίσκει το σημείο διαίρεσης που παράγει τη μέγιστη μείωση διακύμανσης από έναν κόμβο στα παιδιά του. Η βασική μας παρατήρηση είναι ότι ο καλύτερος παράγοντας (από την άποψη της μείωσης του MSE) είναι πάντα ο ίδιος ο στόχος, καθώς αυτός διαχωρίζει σαφώς τον εαυτό του. Έτσι, η χρήση του στόχου ως παράγοντα παρέχει ένα ανώτατο όριο πτώσεως του MSE (ή το κατώτατο όριο για το εναπομένον MSE). Με βάση αυτή την παρατήρηση, ορίζουμε τη k-bit λεπτοδιακύμανση λk^2 μιας μεταβλητής στόχου (λεπτοδιακύμανση σε συγκεκριμένο βάθος k) ως τη διακύμανση που δεν μπορεί να αφαιρεθεί από κανένα δέντρο παλινδρόμησης βάθους k. Ως το θεωρητικό όριο απόδοσης για οποιοδήποτε χαρακτηριστικό, πιστεύουμε ότι το λk^2 είναι μια ενδιαφέρουσα στατιστική έννοια που σχετίζεται με την υποκείμενη δομή του δείγματος, καθώς ποσοτικοποιεί την διακριτική ικανότητα του RT για το δείγμα. Η μέγιστη διακύμανση που μπορεί να εξηγηθεί χρησιμοποιώντας RT βάθους k ονομάζεται k-bit μακροδιακύμανση του δείγματος. Σε οποιοδήποτε βάθος, η συνολική διακύμανση του δείγματος αναλύεται σε λεπτοδιακύμανση λ^2 και μακροδιακύμανση μ^2. Επιδεικνύουμε την ιδέα εκτελώντας ανάλυση της 1- και 2-bit λεπτοδομής των ημερήσιων αποδόσεων της μετοχής της IBM. Καθώς η έννοια της λεπτοδιακύμανσης προσφέρει μια νέα στατιστική προσέγγιση για την ανάλυση της μεταβλητότητας, στη χρηματοοικονομική ανάλυση, μπορεί να βοηθήσει στην καλύτερη κατανόηση της σχέσης της ιδιοσυγκρατικής μεταβλητότητας με τις τιμές και στην διερεύνηση του ομώνυμου παζλ μεταβλητότητας στα χρηματοοικονομικά.

REFERENCES

Breiman L., Friedman J., Olshen R. and Stone C. (1984). *Classification and Regression Trees*. Wadsworth, Belmont, CA

Fama E. and French. K. R. (1993). Common Risk Factors in the Returns on Stocks and Bonds. *Journal of Financial Economics*, **33**, 3-56

Fama E. and French. K. R. (1996). Multifactor Explanations of Asset Pricing Anomalies. *The Journal of Finance*, **51**, 55-84

Fisher, W. D. (1958). On Grouping for Maximum Homogeneity. *Journal of the American Statistical Association*, **53**, 789-798.

Hastie T., Tibshirani R. and Friedman J. (2009). *Elements of Statistical Learning*, Springer
Lintner J. (1965). The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets. *Review of Economics and Statistics*, **47**, 13-37.
Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*, Cambridge University Press
Sharpe, W. F. (1964). Capital Asset Prices: A Theory of Market Equilibrium Under Conditions of Risk. *The Journal of Finance*, **19**, 425-442
Stambaugh R.F., Yu J. and Yuan Y. (2015). Arbitrage Asymmetry and the Idiosyncratic Volatility Puzzle. *The Journal of Finance*, **70**. 1903-1948.

[i] The term variance in this paper implies a population variance (simple sum of squared error division by sample size n) and not the variance estimate which corrects for degrees of freedom.
[ii] when the split is after position 8 all samples belong in the same child and the MSE is the total variance for the sample 3.172 (exactly 3.171875).
[iii] This conjecture seems intuitive but is not obvious and needs to be formally shown.
[iv] NYSE Price data were downloaded from https://finance.yahoo.com (Currency in USD)

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 34ουΠανελληνίου Συνεδρίου Στατιστικής(2022)(σελ.183-199)

GREEK STATISTICAL INSTITUTE ANALYTICS

Ioannis Stamatoulis
Undergraduate Student
Athens University of Economics and Business, Department of Statistics
johnystm96@gmail.com

ABSTRACT

The aim of this project is to support the Greek Statistical Institute (G.S.I.) with insights regarding all the conferences that have happened from 2005 through 2019. For the first time, the dataset has been digitalized and advanced analytics methods have been used in order to extract useful insights about the conferences. We were able to identify information about where the majority of its publications come from and who its strongest contributors are. Last but not least, by using machine learning methods we were able to generate meaningful clusters in order to identify a group of loyal participants who have been consistently engaging in the Institute's activities.

Keywords: Feature Engineering, Principal Components Analysis, Hierarchical Clustering, K-Means

1. INTRODUCTION

All of these years, a large amount of data pertaining to the Greek Statistical Institute conferences was only available in text format. For the first time, we have converted these data into a tabular format that can be used to harvest meaningful information about the activities of the Institute using machine learning techniques. This effort lays the groundwork for continuing to update this dataset with material from future conferences. The project was implemented in Python 3.7 using Jupyter notebook, pandas, numpy, scikit learn, matplotlib and seaborn.

The first phase of the project involved the creation of the dataset from scratch. All data were in an unprocessed form. Therefore, the data were transformed to a format suitable for statistical analysis. After creating the dataset we continued with the pre-processing

of the data. The pre-processing involved checking for any remaining mistakes, dealing with missing values, checking for duplicates and making the appropriate changes and improvements in order to have the dataset ready for analysis. Standard statistical methods were used to produce appropriate descriptive measures. The second phase of the project involved the analysis of the dataset. The analysis included two steps: feature engineering and modelling. In the feature engineering step, we created new features by combining the original available features. By creating new features, we could provide a richer input to the clustering algorithms used. For example the average number of coauthors per participant was calculated for our analysis. The final goal of this work is to identify clusters of participants with similar characteristics concerning their behaviour in G.S.I conferences and come up with a group of loyal participants who have been consistently engaging in and supporting the Institute's activities. In this final step, we used machine learning methods such as clustering techniques (K-means, Agglomerative hierarchical clustering etc.) by taking into account the number of publications in the last 15 years, the average number of publications per year, the topics of the publications and the average number of coauthors.

2. DATASET

Regarding the dataset, we collected information from raw data extracted from the digital documents of the Greek Statistical Institute proceedings. The raw data contained information about the conferences from 2005 to 2019 which were organized by G.S.I..

2.1 Detailed Description of the Dataset Features

We came up with a dataset which contained 1361 rows and 41 features. Each row represents one project/speech and provides the information about the following characteristics.

- **Event:** (int) Identification number of the conference.
- **Year:** (int) Year the conference took place.
- **Location:** (str) Location the conference took place.
- **Author:** (str) First author name.
- **Co-author1 ~ Co-author6:** (str) Names of the co-authors.
- **University:** (str) University/institution of the author.
- **Department:** (str) Department of the university/institution the author is affiliated to.
- **Universtity1 ~ University6:** (str) Universities/institutions of the co-authors.
- **Department1~Department6:** (str) Departments of the universities/institutions that the co-authors are affiliated to.
- **Language:** (str: gr or eng) Language that the project was written in.

- **Occupation:** (str) Occupation of the author/speaker. See page 5 for more details.
- **Occupation1- Occupation6:** (str) Occupation of the co-authors.
- **Published:** (str: yes or no) Binary indicates whenever the project was published in the Greek Statistical Institute proceedings.
- **Student:** (str: yes or no) Binary indicates whenever a student joined the project.
- **Awards:** (str: yes or no) Binary indicates whenever the project won an award.
- **Duration:** (int) Scheduled duration of the talk in minutes.
- **Topics:** (str) Topic of the project.
- **Abroad_cooperation:** (str: yes or no) Binary indicates whenever a co-author belonged to a university/institution abroad.
- **Non_university_cooperation:** (str: yes or no) Binary indicates whenever a co-author belonged to a nonacademic affiliation.
- **Chairman:** (str) Chairman of the presentation.
- **Invited:** (str: yes or no) Binary indicates whenever the author was invited to the conference.

2.2 Clarifications for some Features

Topics: Initially the presentations were classified to 51 different topics based on the section title of the conferences programs that the presentation belonged to. In order to facilitate the analysis we decided to merge them into 6 different new but brooder topics based on 2020 Mathematics Subject Classification [Dunne E. and Hulek K., 2020]. For more information about the topics, see Section 2.2.

Occupation variables: This feature has 21 different values and it refers only to projects that have been published in the proceedings of the Greek Statistical Institute. The encoding is the following:

- 0. Unknown
- 1. Professor
- 2. Associate Professor
- 3. Assistant Professor
- 4. Docent
- 5a. PhD Graduate
- 5b. PhD Student
- 6. Postgraduate Student
- 7. Undergraduate
- 8. Research Contributor
- 9. Colleague

- 10. Emeritus Professor
- 11. Economist
- 12. Statistician
- 13. Part-time Professor
- 14. Doctor
- 15. Researcher
- 16. Analyst
- 17. University employee
- 18. Contractor
- 19. Teacher

University/ Department variables: This information was available only for the presentations that have been published at the proceedings of the Greek Statistical Institute.

2.2 Dataset Insight

In this section we will refer to some statistics that provide useful information about the conferences, the talks and the participating authors.

One of the Institute's operations is the organising of conferences around Greece and Cyprus, as well as the establishment of competitions to award young statisticians for their success in Probability and Statistics and other scientific activities. The first figure depicts a map of the locations where the conferences were held throughout Greece and Cyprus. The size of the marks corresponds to the number of talks during the conference. In addition the lineplot in Figure 2 depicts the per year number of presentations given in the conferences. The black line represents the actual values of the presentations made each year and the red dashed line represents a forecast for each year. The forecast for each year in fact is the mean value of presentations for each previous year. The years are shown on the x-axis, while the number of talks is shown on the y-axis. furthermore a 95% confidence interval has been calculated for the forecast, which appears in the diagram as a shaded area around the red line.
We can observe that the conference in 2006, held in Kastoria, had the highest number of talks (137), while the conference in 2017, held in Larnaca, had the fewest talks (45).

Figure 1. *Map of the conferences locations*

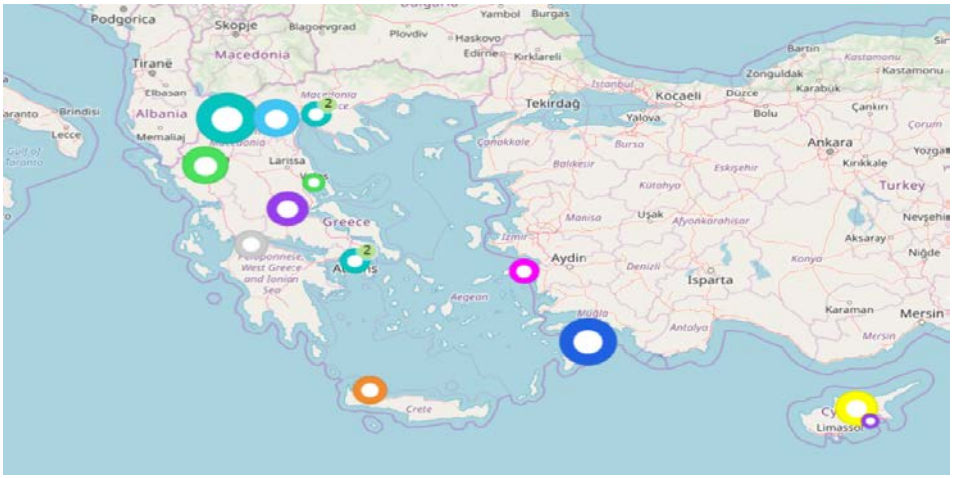


Figure 2. *Presentations per year*

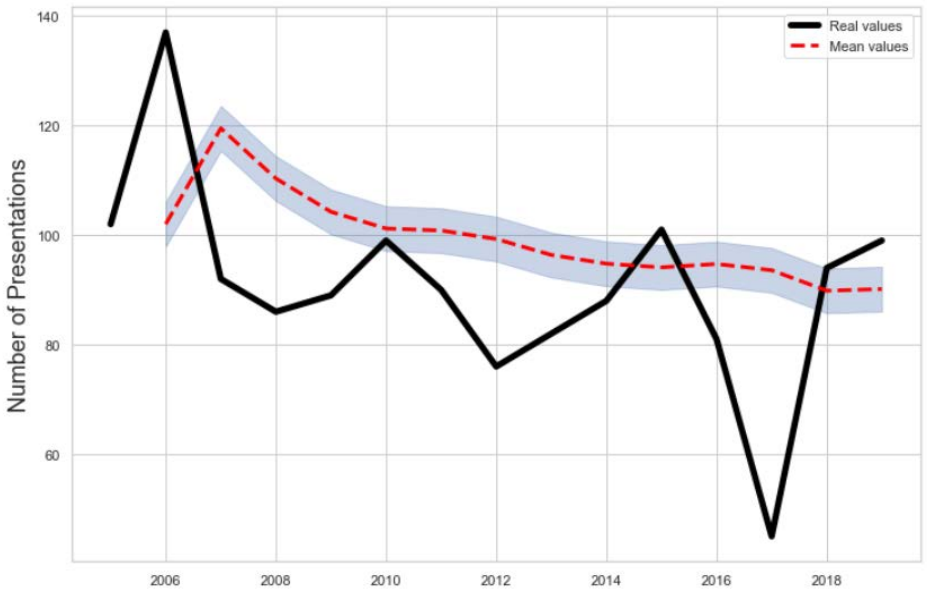
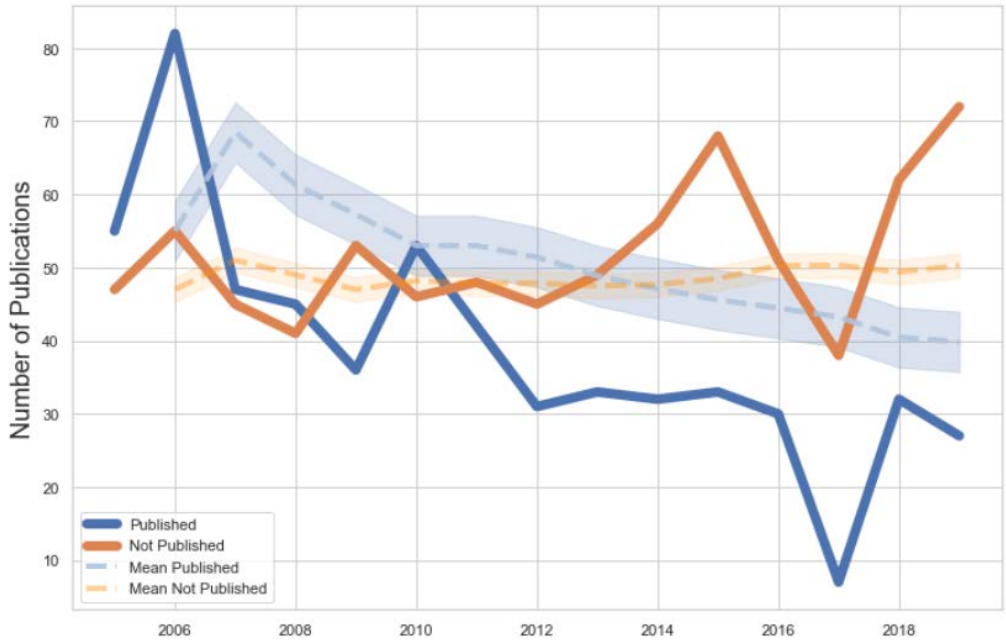


Figure 3 investigates the project's publications in the proceedings of the Greek Statistical Institute. The lineplot, in particular, illustrates the number of projects that were published (blue line) and the number of projects that were not published (orange line) from 2005 to 2019. As in the previous diagram there are two dashed lines that

represent an estimate of the actual values. More specifically the light blue dashed line despite the forecast of the projects that were published each year and the shadowed area is the 95% confidence interval. While the light orange dashed line despite the forecast of the projects that were not published and correspondingly the area around the line is the 95% confidence interval of this estimation. As we see, in 2006 the publications were the most (82) and the projects that weren't published in the proceedings were 55. On the other hand, the fewest publications were in 2017 with only 7 and that year the projects that weren't published were 38. We can also observe that each year after 2010 the number of the published projects in the proceedings is much smaller than the number of projects that were not published, so we can come up with the conclusion that the committee examining these projects became stricter after 2010.

Figure 3. *Publications per Year*



As indicated in the dataset description, initially the presentations were classified in 50 different topics which were combined into six final categories based on the 2020 Mathematics Subject Classification, as shown in Table 1. It is worth mentioning that the statistical topics were divided into two categories: "Statistics in Applications" and "Statistics" and that there is an additional (seventh) "topic" called "Other" that refers to general interest, invited speakers, awards, and posters.

Table 1 provides the details of this new classification. New categories and the coding of the 2020 Mathematics Subject Classification are shown in the left part and at the right part we can see the old topics.

Table 1. *Description of the Topics Based on 2020 Mathematics Subject Classification*

Probability Theory (AMS 60)	History of Probabilities And Statistics' 'Probability Theory' 'Probability Theory and Statistics' 'Applied Probabilities'
Stochastic Processes (AMS 60_GHJ)	Stochastic Process', 'Stochastic Reviews'
Statistics in Applications (AMS 62)	Statistics in Economics' 'Statistics and Education' 'Educational - Social Statistics' 'Environmental Statistics' 'Neural Network - Data Mining' 'Health Statistics' 'Statistics and Finance' 'Statistics in Business' 'Statistical Methods in Applied Geographical Analysis' 'Statistics in Seismology' 'Modeling, Mechanical Systems Design And Analysis' 'Statistics and Internet'
Statistics (AMS 62_b)	Statistics' 'Applied Statistics' 'Bayesian Statistics' 'Experimental Designs' 'Sampling' 'Biometrics' 'Biostatistics' 'Data Analysis' 'Multivariate Analysis' 'Linear Models, Regression, Computational Statistics' 'Computational Statistics' 'Statistical Models and Applications' 'Time Series'
Operations Research (AMS 90)	Operational Research', 'Reliability Theory'
Game Theory, Economics, Finance (AMS 91)	Actuarial' 'Demography And Actuarial' 'Actuarial-Insurance and Financial Mathematics' 'Finance and Economic Applications' 'Econometrics'
Other	General Interest' 'Invited Speakers' 'Awards' 'Young Statistician's Prize' 'Posters'

Figure 4 presents the total number of talks for each topic based on the merge of the Mathematics Subject Classification. As expected, the topic of Statistics is the most popular topic, with over 650 talks (48%), while the less popular topic is the Operations Research with less than 50 talks (2%).

Figure 4. *Presentations per topic*

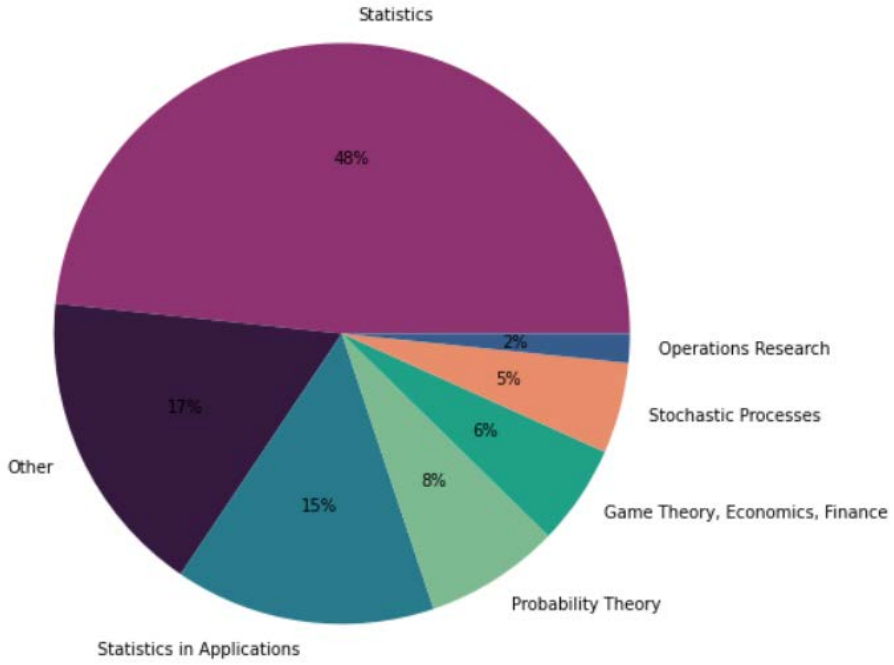
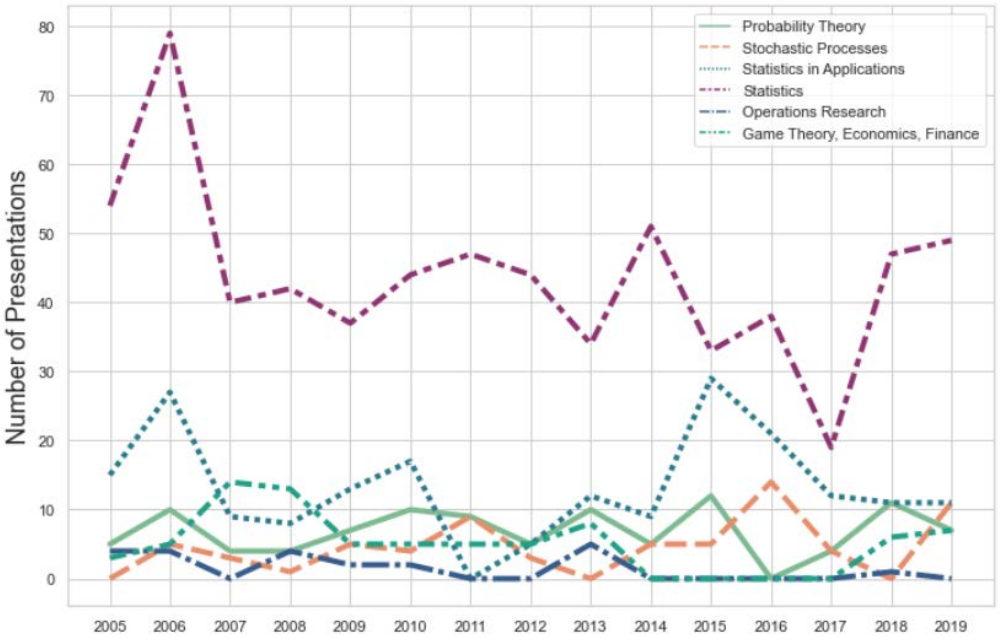


Figure 5, shows the topics that were featured in conferences from 2005 to 2019 and which ones were the most popular. The x-axis represents the years, while the y-axis represents the number of talks for each topic. Each line represents a different topic. Statistics has been the most dominant topic year after year, however it is interesting to note that after 2004, there was an increasing interest in the field of "Statistics in Applications". Several topics were also excluded for certain years, for example, in 2016 in Thessaloniki, the topics “Probability Theory”, “Operations Research,” and “Game Theory, Economics, Finance” were absent, at this point, the lines that counterbalance each of these topics overlap the zero. The years with the biggest variety topic-wise were 2006 in Kastoria, 2008 in Samos, 2009 in Chania and 2010 in Veria.

Figure 5. *Presentations for each topic per year*



3. FEATURE ENGINEERING

The dataset we had at first did not provide us with a lot of information about authors and co-authors as it had no numerical data. Therefore we decided to use the existing features of the old dataset and combine them with each other in order to create a new dataset that would be more informative about the participants. The term "participants" refers to the authors/speakers and the co-authors of project contributors.

3.1 Description of the New Participants Dataset

The new dataset has 1145 cases and 13 new features. Each case refers to one participant and give us the following information:

- **Names:** Name of the participant.
- **Author_Counts**: Number of presentations/ projects the participant appears as first author.
- **Total_Counts**: Number of presentations/ projects the participant appears as first author or co-author in the last 15 years.

- **Average_Co-authors**: Mean of the number of co-authors. If someone is systematically working alone then the mean will be equal to zero.
- **Published**: Number of times a participant appeared as author or co-author in an article published at the proceedings.
- **AMS60**: Number of times a participant joined a project with the topic AMS60 ("Probability Theory").
- **AMS60_GH**: Number of times a participant joined a project with the topic AMS60_GH ("Stochastic Processes").
- **AMS62**: Number of times a participant joined a project with the topic AMS62 ("Statistics in Applications").
- **AMS62_b**: Number of times a participant joined a project with the topic AMS62_b ("Statistics").
- **AMS90**: Number of times a participant joined a project with the topic AMS90("Operation Research").
- **AMS91**: Number of times a participant joined a project with the topic AMS91("Game Theory").
- **General**: Number of times a participant joined a project with the "topic" Other.
- **Last5**: Number of times a participant appeared as an author or co-author in the last 5 years (from 2015 to 2019)

4. CLUSTERING

Clustering algorithms look for similarities or differences between data points in order to group together similar observations. The final goal is the clustering of the new dataset so that we could discover clusters that include participants with similar characteristics concerning their behaviour in G.S.I conferences and come up with a group of loyal participants.

4.1 Principal Component Analysis

We used Principal Component Analysis (PCA), in order to decide which features of the new dataset should be used and also to visualize the clustering. PCA is a dimensionality-reduction approach for reducing the dimensionality of big data sets by converting a large collection of variables into a smaller one that retains most of the information in the large set. Naturally, reducing the number of variables in a data set reduces accuracy. Nevertheless, the aim of dimensionality reduction is to simplify the data at the price of some accuracy loss. Smaller data sets are easier to examine and visualize, and machine learning algorithms can analyze data much more easily and quickly without having to deal with unnecessary variables. A simplified explanation is that at first we had to normalize the range of variables such that they all contribute

equally to the analysis. Second, we computed the covariance matrix, which is used to determine how the variables in the input dataset differ from the mean in relation to each other, or to check whether there is any link between them. Eigenvectors and eigenvalues must be computed from the covariance matrix in order to discover the data's major components. Principal components are new variables that are created by linearly combining variables. These combinations are made in such a way that the new variables (principal components) are uncorrelated, and the majority of the information contained in the starting variables is squeezed into the first components. Next, we had to decide whether to preserve all of these components or to eliminate those with less significance (low eigenvalues) and combine the remaining ones to produce a matrix of vectors known as the Feature vector. In the final stage, we utilized the feature vector constructed from the covariance matrix's eigenvectors to reorient the data from the original axis to the ones indicated by the principal components (hence the name Principal Components Analysis). This is accomplished by multiplying the original data set's transpose by the feature vector's transpose [Wold, et al., 1987].

In our case the starting features are Author_Counts, Total_Counts, Published, Average_Co-authors and Last5 as the other features do not provide as much information in total. It turns out that in PCA space, the variance is maximized along the first two components. PC1 (explaining 73% of the variance) and PC2 (explaining 22% of the variance). Together, they explain 95%.These components are derived from linear weighted combinations of strongly correlated features in the dataset. The importance of each feature is reflected by the magnitude of the corresponding values in the eigenvectors (higher magnitude means higher importance). Therefore, the most important features for the first component are Total_Counts, Published and last5. Similarly Author_Counts and Average_Co-authors are most important for the second feature.

As we mentioned before, our final goal was the clustering of the new dataset so that we could come up with clusters that include participants with similar characteristics. Clustering was performed using the components identified by PCA. So that takes the label of the cluster to which each participant belongs. The algorithms that we tried on our new dataset to find participants with similar characteristics were K-Means [Hartigan and Wong, 1979] and Hierarchical Clustering [Day and Edelsbrunner, 1984]. We determined the optimal number of clusters with the Elbow method for K-Means algorithm and the dendrograms for the different methods of Agglomerative hierarchical clustering. The elbow method performs K-means clustering on the dataset for a range of K values (for example, 1 to 10), then computes an average score for all clusters for each value of K. The frequently used score is the sum of the square distances from each

location to its assigned centre, then computes an average score for all clusters for each value of K. The frequently used score is the sum of the square distances from each location to its assigned centre. When the total metrics for each model are displayed, the optimal value for K may be visually determined and it is the point of inflection on the curve [Syakur, M.A., Khotimah, 2018]. On the other hand in a dendrogram if we locate the largest vertical difference between nodes, and in the middle pass an horizontal line. The number of vertical lines intersecting it is the optimal number of clusters (when affinity is calculated using the method set in linkage). We came up to create 3 or 4 clusters with the best results given by the 3 clusters as shown from the silhouette score of the clustering. We will go into further detail on the clustering in the next section.

4.2 Comparison of Different Methods

In order to compare the clustering that was implemented with the algorithms we used intra metrics, which evaluate how similar are the elements belonging to the same cluster and how dissimilar are elements belonging to different clusters. More specifically, we used the Silhouette score [K. R. Shahapure and C. Nicholas, 2020]. This measure compares the mean distance of a data point to all of the data points belonging to the same cluster and the average of minimum distances to all other clusters. Silhouette score for a datapoint **i** is given as:

$$s_i = \frac{b_i - a_i}{max(b_i, a_i)} \quad (1)$$

where,

b_i: is the inter cluster distance defined as the average distance to the closest cluster of datapoint **i** except for that it's a part of.

$$b_i = min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i,j) \quad (2)$$

a_i : is the intra cluster distance defined as the average distance to all other points in the cluster to which it's a part of.

$$a_i = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i,j) \quad (3)$$

Overall Silhouette Score for the complete dataset can be calculated as the mean of Silhouette Score for all data points in the dataset. As can be seen from the formula

Silhouette Score always lies between -1 to 1. The value 1 represents the best clustering while -1 means that the clustering is not successful.
The Silhouette Scores, for all the different techniques we used and for the 3 clusters, are:

- Agglomerative Hierarchical Clustering
  - Euclidean distance Methods [Day and Edelsbrunner, 1984]:
    - Ward (Silhouette Score = 0.7359)
    - Average (Silhouette Score = 0.8178)
    - Complete (Silhouette Score = 0.8332)
    - Single (Silhouette Score = 0.9023)
  - **Manhattan distance** Methods [Day and Edelsbrunner, 1984]:
    - **Average (**Silhouette Score = **0.8207)**
    - Complete (Silhouette Score = 0.8079)
    - Single (Silhouette Score = 0.9036)
- K-Means Clustering [Hartigan and Wong, 1979]
  - Euclidean distance (Silhouette Score = 0.6917)
  - Manhattan distance (Silhouette Score = 0.7589)

The method that provides the best clustering according to Silhouette Score is the Hierarchical Agglomerative Clustering after using it with the Average method and Manhattan distance. The Silhouette Score for this method is 0.8207. It is good to mention that some methods provide better Silhouette scores, however the two of three clusters for these methods contain only one or two participants, which is undesirable because we want as many evenly distributed clusters as possible.

4.3 PCA Visualization of the Clustering

We can see the PCA visualization of the clustering in the following plot and the names of participants with similar characteristics that belong to the same cluster. In Figure 6 the axes represent the two independent principal components (2 dimensions) of PCA. The principal components are linear weighted combinations of strongly correlated features in order to cram the majority of the information contained in the initial variables into the first two components and minimize the dimensions so that the clusters may be visualized. In Table 2, the 29 names displayed in the green box have been chosen randomly from the total of the participants that belong to cluster 2 and are placed in alphabetical order. The other two boxes represent the total number of participants in each respective cluster, again in alphabetical order.

Figure 6. *Clustering Visualization*

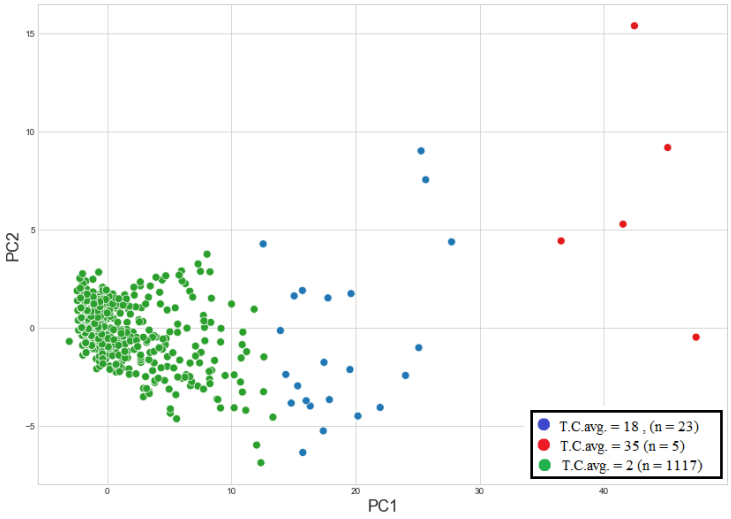


Table 2. *Participants for each Cluster*

Anastasiadou S.	Bersimis S.	Farmakis N.
Batsidis A.	Chalikias M.	Kougioumtzis D.
Bompotas P.	Charalambides Ch.	Kounias S.
Chasiotis V.	Chatzopoulos St.A.	Koutras M.V.
Chatzipantelis Th.	Dimitriou I.	Tsaklidis G.
Chorozoglou D.	Iliopoulos G.	
Dritsakis N.	Kakoullos Th.	
Evangelaras Ch.	Karagrigoriou A.	
Georgiou V.	Karlis D.	
Ioannides D.	Kolyva Machaira F.	
Kakavakis D.	Malefaki S.	
Ketzaki E.	Moysiadis Ch.	
Kitsos Ch.	Oikonomou P.	
Konstantinides D.G.	Panagiotakos D.	
Koutrouvelis I.A.	Papadimitriou E.	
Kyriakousis A.	Papaioannou T.	
Manatakis M.	Rakitzis A.	
Matalliotakis G.	Rigas A.	
Millionis A.E.	Skladas Ch.	
Papana A.	Triantafyllou I.S.	
Parpoula Ch.	Tzavelas G.	
Perikleous K.	Vamvakari M.	
Sachlas A.	Vasileladis G.	
Singiridou E.		
Sotiropoulos I.		
Spyridis Th.		
Theodosiadou O.		
Vasileladis V.		
Viana M.		

4.4 Clusters Characteristics

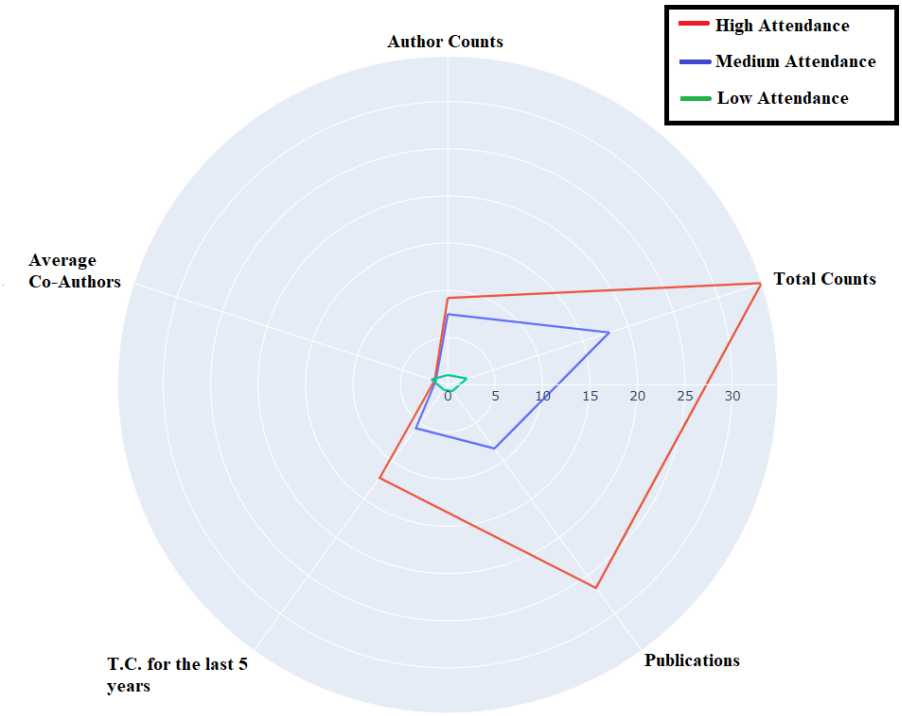
We came up with 3 clusters: low attendance cluster (Green), medium attendance cluster (Blue), high attendance cluster (Red) and the size of each cluster is 23, 5 and 1117 respectively.
As already mentioned, the clustering is implemented by taking into account the two components that came up from PCA. After taking the labels for each participant we integrate it with the starting information of the dataset. We can use these labels in order to calculate the average values of each feature for the clusters. In this way we can describe the characteristics of the clusters and observe the differences between them. Table 3 presents the average values of each variable for the clusters rounded.

Table 3. *Average Value of each Cluster Rounded for all of the Features*

	High attendance	Medium attendance	Low attendance
Total Counts	**35**	18	2
Publications	**27**	9	1
Average Co-authors	**1**	1	2
Total Counts for last 5 years	**13**	6	1
Author counts	**10**	8	1

In figure 7, we will be seeing the polar line plot, in which each feature of our dataset is represented as a vertex of a polygon mark in polar coordinates. In the polar line plot there are three polygons each one representing one of our clusters. So in this way we can visualize the differences of the clusters. It would be useful to mention that for the feature Average Co-Authors there is not a big difference because of the range of the feature.

Figure 7. *Polar Line Plot for the Visualization of the Characteristics of the Clusters*



4.5 Conclusion of the Clustering

To conclude we implemented Agglomerative Hierarchical Clustering by using the Average method and Manhattan distance. We took into account the first two components from PCA that we used on the new dataset. We came up with 3 clusters: low attendance cluster (Green), medium attendance cluster (Blue), high attendance cluster (Red), the size of each cluster is 23, 5 and 1117 respectively and the characteristics of each cluster are as we described above. The authors and co-authors of medium attendance cluster (Blue) and high attendance cluster (Red) reflect a small group of 28 academic members of the Greek Statistical Institute who have been consistently engaging in and supporting the Institute's activities. These professionals appear to have a significant effect on the Institute's character, work, and continuity but without disregarding of course the contribution of other participants. Despite the fact that it relates to more ephemeral and transitory partnerships, the huge number of participants in the low attendance cluster (Green) reflects the Institute's variety and the Greek and foreign scientific community's strong interest in the Institute's activities. The features that had the biggest impact on the clustering were the Total Counts and Publications.

ΠΕΡΙΛΗΨΗ

Στόχος του έργου αυτού είναι να υποστηρίξει το Ελληνικό Στατιστικό Ινστιτούτο (Ε.Σ.Ι.) με πληροφορίες σχετικά με όλα τα συνέδρια που έχουν πραγματοποιηθεί από το 2005 έως το 2019. Για πρώτη φορά, το σύνολο δεδομένων ψηφιοποιήθηκε και χρησιμοποιήθηκαν προηγμένες μέθοδοι ανάλυσης προκειμένου να εξαχθούν χρήσιμες πληροφορίες σχετικά με τα συνέδρια. Καταφέραμε να εντοπίσουμε πληροφορίες σχετικά με το από που προέρχεται η πλειονότητα των δημοσιεύσεών του και ποιοι είναι οι ισχυρότεροι συντελεστές του. Τέλος, με τη χρήση μεθόδων μηχανικής μάθησης καταφέραμε να δημιουργήσουμε ουσιαστικές συστάδες προκειμένου να εντοπίσουμε μια ομάδα πιστών συμμετεχόντων που συμμετέχουν σταθερά στις δραστηριότητες του Ινστιτούτου.

Acknowledgements: I would like to thank my academic supervisor, Xanthi Pedeli and the Institute's supervisors Malvina Vamvakari and Alexandros Karagrigoriou for all of their help and support.

REFERENCES

Hartigan, J.A. and Wong, M.A., 1979. Algorithm AS 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, *28*(1), pp.100-108.

Wold, S., Esbensen, K. and Geladi, P., 1987. Principal component analysis. *Chemometrics and intelligent laboratory systems*, *2*(1-3), pp.37-52.

Day, W.H. and Edelsbrunner, H., 1984. Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of classification*, *1*(1), pp.7-24.

Johnson, S.C., 1967. Hierarchical clustering schemes. *Psychometrika*, *32*(3), pp.241-254.

K. R. Shahapure and C. Nicholas, "Cluster Quality Analysis Using Silhouette Score," *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, 2020, pp. 747-748, doi: 10.1109/DSAA49011.2020.00096.

Syakur, M.A., Khotimah, B.K., Rochman, E.M.S. and Satoto, B.D., 2018, April. Integration k-means clustering method and elbow method for identification of the best customer profile cluster. In *IOP conference series: materials science and engineering* (Vol. 336, No. 1, p. 012017). IOP Publishing.

Dunne, E. and Hulek, K., 2020. Mathematics subject classification 2020. *European Mathematical Society Magazine*, (115), pp.5-6.


Greek Statistical Institute
Proceedings of the 34th Panhellenic Statistics Conference (2022), pp. 200−215

On the selection of optimal subdata for big data regression

V. Chasiotis, D. Karlis
Department of Statistics, Athens University of Economics and Business, Greece
{chasiotisv, karlis}@aueb.gr

ABSTRACT
In the big data era researchers face a series of problems. Even standard approaches, like linear regression, can be difficult or problematic with huge volumes of data. Traditional approaches for regression in big datasets may suffer due to the large sample size, since they involve inverting huge data matrices or even because the data cannot fit to the memory. Proposed approaches are based on selecting representative subdata to run the regression. Existing approaches select the subdata using information criteria and/or properties from orthogonal arrays. In the present paper we improve existing algorithms providing a new algorithm that is based on D-optimality approach. A simulation experiment as well as a real data application are also provided.

Keywords: Experimental designs; D-optimality; Information matrix; Linear regression; Subsampling.

1. INTRODUCTION

Recent research in various disciplines is characterized by the unprecedented demand of big data analysis. Typically, the scale of the datasets increases, so does the demand of computational resources for the statistical analysis and modeling process. Although the availability of computational power increases rapidly, it still falls far behind under the explosive increase in data volume.

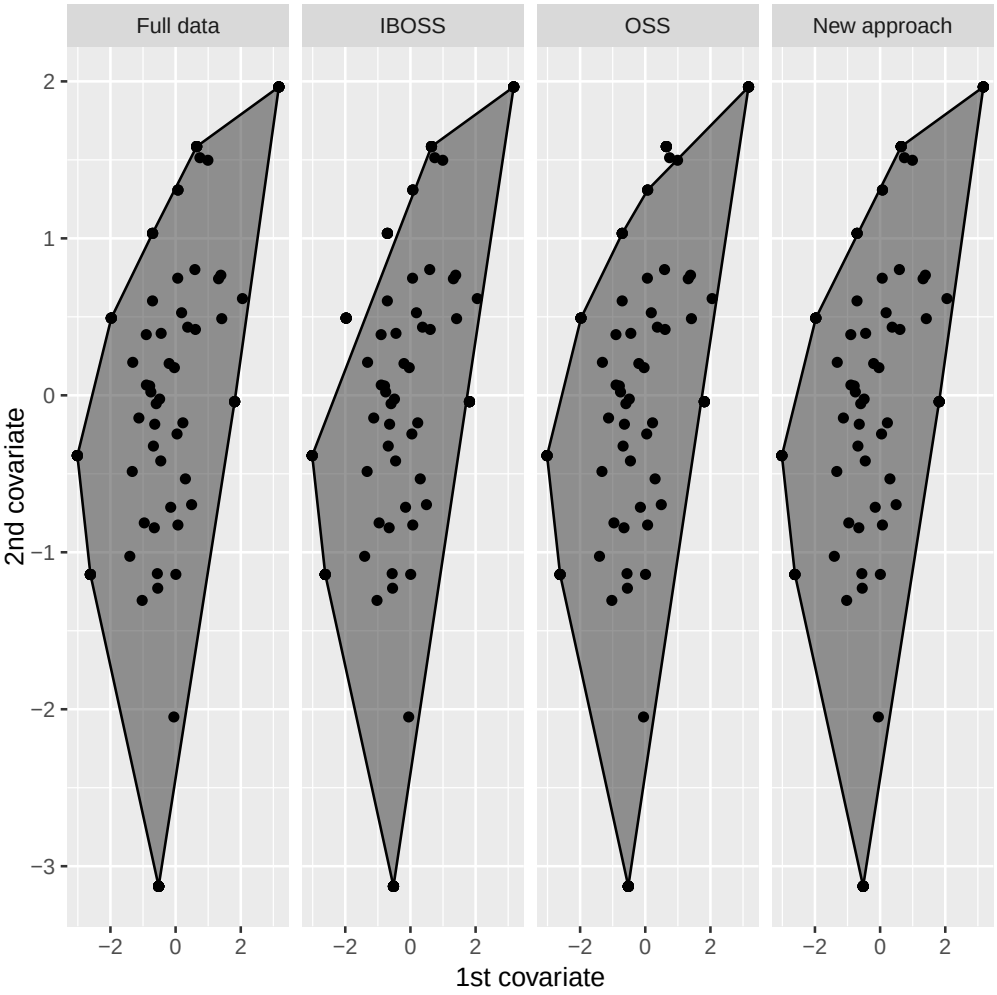
This creates new challenges to data storage and analysis. A standard approach is based on data reduction, or subsampling, where one selects a portion of the data to extract useful information. This is a crucial step in big data analysis. For massive data, subsampling techniques are popular to mitigate computational burden by reducing the data size considerably and bringing it back to a doable size. Consider the problem of regression where the sample size n is quite large and one needs to fit a model with standard least squares approach. In many circumstances, this can create a lot of computational issues, since the standard least square approach involves big matrices that perhaps do not fit in the memory. Working with less data is an option as far as the reduced dataset can keep as much information as possible. In most problems, while picking the necessary data with pure randomness is an option, improved approaches can be used to select subdata in an optimal way.

As a first attempt Drineas et al. (2011) proposed the idea of selecting part of the data with a random selection. In particular, they proposed to make a randomized Hadamard transform on data and then use uniform subsampling to take random subdata to approximate ordinary least squares estimators in linear regression models. Such an approach, based on the idea and theory of random matrices, has found a lot of applications. On the other hand, they suffer from the inherent randomness of the procedure. In recent years, an alternative approach attempts to use deterministic rather than random selection of data points based on certain criteria. Such approaches share significant relationship with optimal-design problems, traditionally known in statistics for many years. Methods of optimal design of experiments might also be applicable to the setting of big data by providing the methodology for selecting data points to create the optimal subdata.

The paper of Wang et al. (2019) is central to such approaches bringing ideas from optimal designs to the selection of data points proposing the information-based optimal subdata selection (IBOSS) approach. A recent extension is given by Wang et al. (2021) with the orthogonal subsampling (OSS) approach. Details about the aforementioned approaches will follow in Section 3.

To motivate the problem, consider the data in Figure 1. We have used two covariates and 50 data points. The shadowed area, in each subplot, is the convex hull generated by the selected subdata. Suppose that we want to select 8 observations. A plausible idea for selecting subdata, is to select data points in some sense with large convex hull as close as possible to the one generated by the full data (first panel). In such a case, the selected data points can have a large volume and hence the determinant of the information matrix will be large. Recall that the inverse of the information matrix appears in the variance of the estimated coefficient vector. The IBOSS approach (second panel) tries to select points at the extreme of the two covariates, whereas the OSS approach (third panel), using a different loss function, attempts to select data points at the corners of the two-dimensional space. A detailed description of the aforementioned approaches is given in Section 3. Our aim and contribution of the present paper, is to improve and create an approach (last panel) that balances the two ideas and with a few steps can improve a lot by selecting data points by the D-optimality criterion, namely to increase the determinant of the information matrix. In this selected motivating example, our approach succeeds in approximating the convex hull of the full data. Note that we present a main algorithm that starts from the one of the OSS approach and with a few additional considerations, it improves with respect to the generalized variance of the selected subdata. We also propose some extensions which at the cost of additional execution time can further improve the D-optimality criterion under the selected data points.

Figure 1: *An example for the different approaches. Data with two covariates and full data size of* 50 *data points were generated. The different approaches were used to select* 8 *data points. The full data can be seen in the first panel and then the IBOSS approach, the OSS approach and our new approach.*



The remaining of the paper proceeds as follows. Section 2 provides some theoretical arguments that will be the basis of the newly developed approach. Section 3 describes approaches already existing in the current literature. The algorithm and its variants are described in detail in Section 4. Simulation evidence to support the new approach is provided in Section 5. A real dataset is used for illustration in Section 6, while concluding remarks can be found in Section 7.

2. THEORETICAL CONSIDERATIONS

We assume that the full data are denoted by $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, n$. Let the following linear regression model:

$$y_i = \beta_0 + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_1 + \epsilon_i, \quad i = 1, 2, \ldots, n, \tag{1}$$

where y_i is a response, β_0 is the intercept parameter, $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^{\mathrm{T}}$ is a covariate vector, $\boldsymbol{\beta}_1 = (\beta_1, \beta_2, \ldots, \beta_p)^{\mathrm{T}}$ is a p-dimensional vector of unknown slope parameters, and ϵ_i's are the error terms that are uncorrelated satisfying $E(\epsilon_i) = 0$ and $V(\epsilon_i) = \sigma^2$.

We take into consideration the full data under model (1), and so the least-square estimator of $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1^{\mathrm{T}})^{\mathrm{T}}$, which is its best linear unbiased estimator as well, is

$$\hat{\boldsymbol{\beta}}_{\text{Full}} = \left(\sum_{i=1}^{n} \mathbf{z}_i \mathbf{z}_i^{\mathrm{T}}\right)^{-1} \sum_{i=1}^{n} \mathbf{z}_i y_i,$$

where $\mathbf{z}_i = (1, \mathbf{x}_i^{\mathrm{T}})^{\mathrm{T}}$.

The covariance matrix of $\hat{\boldsymbol{\beta}}_{\text{Full}}$ is equal to the inverse of

$$\mathbf{Q}_{\text{Full}} = \frac{1}{\sigma^2} \sum_{i=1}^{n} \mathbf{z}_i \mathbf{z}_i^{\mathrm{T}},$$

where $\mathbf{Q}_{\text{Full}}$ is the observed Fisher information matrix of $\boldsymbol{\beta}$ for the full data if ϵ_i are normally distributed. Even though we do not require the normality assumption, $\mathbf{Q}_{\text{Full}}$ will be still called the information matrix.

However, it is not always feasible to fully analyze the whole data, since the sample size n of the full data can be too large. Thus, an approach is to gain useful information from the full data given that computational resources are limited. An effective investigation can be focused on selecting a subset of the full data.

Let δ_i be a variable that indicates whether $(\mathbf{x}_i, y_i)$ is included in the subdata. Therefore, $\delta_i = 1$ if $(\mathbf{x}_i, y_i)$ is included in the subdata and $\delta_i = 0$ otherwise. In case of selecting subdata of size k, the least-square estimator of $\boldsymbol{\beta}$ remains the best linear unbiased estimator based on the subdata, that is,

$$\hat{\boldsymbol{\beta}}_{\text{Sub}} = \left(\sum_{i=1}^{n} \delta_i \mathbf{z}_i \mathbf{z}_i^{\mathrm{T}}\right)^{-1} \sum_{i=1}^{n} \delta_i \mathbf{z}_i y_i,$$

where $\sum_{i=1}^{n} \delta_i = k$.

The information matrix with subdata of size k can be written as

$$\mathbf{Q}_{\text{Sub}} = \frac{1}{\sigma^2} \sum_{i=1}^{n} \delta_i \mathbf{z}_i \mathbf{z}_i^{\mathrm{T}}. \tag{2}$$

The selected subdata should be optimal is some way. According to the theory of the optimal design of experiments, this is assessed with respect to a criterion, which is related to the covariance matrix of the estimated parameters. Such a popular criterion is the D-optimality criterion which seeks to maximize the determinant of the information matrix. Thus, as Wang et al. (2019) proposed, we are interested in maximizing the determinant of $\mathbf{Q}_{\text{Sub}}$ subject to $\sum_{i=1}^{n}\delta_i = k$.

However, the information matrix $\mathbf{Q}_{\text{Sub}}$ in (2) can be written in another way, and so the problem of maximizing its determinant can be finally redefined. The information matrix $\mathbf{Q}_{\text{Sub}}$ in (2) can be written as

$$\mathbf{Q}_{\text{Sub}} = \frac{k}{\sigma^2}\begin{bmatrix} 1 & \frac{\sum_{i=1}^{n}\delta_i x_{i1}}{k} & \cdots & \frac{\sum_{i=1}^{n}\delta_i x_{ip}}{k} \\ \frac{\sum_{i=1}^{n}\delta_i x_{i1}}{k} & \frac{\sum_{i=1}^{n}\delta_i x_{i1}^2}{k} & \cdots & \frac{\sum_{i=1}^{n}\delta_i x_{i1}x_{ip}}{k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\sum_{i=1}^{n}\delta_i x_{ip}}{k} & \frac{\sum_{i=1}^{n}\delta_i x_{i1}x_{ip}}{k} & \cdots & \frac{\sum_{i=1}^{n}\delta_i x_{ip}^2}{k} \end{bmatrix}.$$

In order to discriminate between the covariate vectors $\mathbf{x}_i$, $i = 1, 2, \ldots, n$ and the covariates, let $\mathbf{x}_j^*$, $j = 1, 2, \ldots, p$ be the jth covariate under the selected subdata. After some calculations, we get that:

$$\mathbf{Q}_{\text{Sub}} = \frac{k}{\sigma^2}\left(\mathbf{u}^{\text{T}}\mathbf{u} + \mathbf{U}\right),$$

where $\mathbf{u} = (1, \bar{\mathbf{x}}_1^*, \bar{\mathbf{x}}_2^*, \ldots, \bar{\mathbf{x}}_p^*)$, $\mathbf{U} = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & s^2_{\mathbf{x}_1^*} & s_{\mathbf{x}_1^*\mathbf{x}_2^*} & \cdots & s_{\mathbf{x}_1^*\mathbf{x}_p^*} \\ 0 & s_{\mathbf{x}_1^*\mathbf{x}_2^*} & s^2_{\mathbf{x}_2^*} & \cdots & s_{\mathbf{x}_2^*\mathbf{x}_p^*} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & s_{\mathbf{x}_1^*\mathbf{x}_p^*} & s_{\mathbf{x}_2^*\mathbf{x}_p^*} & \cdots & s^2_{\mathbf{x}_p^*} \end{bmatrix}$, $\bar{\mathbf{x}}_j^* = \frac{\sum_{i=1}^{n}\delta_i x_{ij}}{k}$, $s^2_{\mathbf{x}_j^*} = \frac{\sum_{i=1}^{n}\delta_i (x_{ij} - \bar{\mathbf{x}}_j^*)^2}{k}$, $j = 1, 2, \ldots, p$, and $s_{\mathbf{x}_o^*\mathbf{x}_j^*} = \frac{\sum_{i=1}^{n}\delta_i x_{ij}x_{it}}{k} - \bar{\mathbf{x}}_o^*\bar{\mathbf{x}}_j^*$, $o \neq j = 1, 2, \ldots, p$.

Based on the matrix determinant lemma (for more information, see Harville (1997)), since $\det(\mathbf{U}) = 0$, that is $\mathbf{U}$ is not invertible, we get that:

$$\det\left(\mathbf{Q}_{\text{Sub}}\right) = \frac{k^{p+1}}{\sigma^{2(p+1)}}\left(\det(\mathbf{U}) + \mathbf{u}\text{adj}(\mathbf{U})\mathbf{u}^{\text{T}}\right),$$

where adj($\mathbf{U}$) is the adjugate matrix of $\mathbf{U}$, or

$$\det\left(\mathbf{Q}_{\text{Sub}}\right) = \frac{k^{p+1}}{\sigma^{2(p+1)}}\det\left(\begin{bmatrix} s^2_{\mathbf{x}_1^*} & s_{\mathbf{x}_1^*\mathbf{x}_2^*} & \cdots & s_{\mathbf{x}_1^*\mathbf{x}_p^*} \\ s_{\mathbf{x}_1^*\mathbf{x}_2^*} & s^2_{\mathbf{x}_2^*} & \cdots & s_{\mathbf{x}_2^*\mathbf{x}_p^*} \\ \vdots & \vdots & \ddots & \vdots \\ s_{\mathbf{x}_1^*\mathbf{x}_p^*} & s_{\mathbf{x}_2^*\mathbf{x}_p^*} & \cdots & s^2_{\mathbf{x}_p^*} \end{bmatrix}\right). \tag{3}$$

The expression (3) is the generalized variance (Wilks, 1932) of covariates $\mathbf{x}_j^*$, $j = 1, 2, \ldots, p$. Therefore, the problem of maximizing the determinant of the information matrix in (2) can be addressed as a problem of maximizing the generalized variance of covariates under the selected subdata.

Let $\mathbf{A} = \begin{bmatrix} s^2_{\mathbf{x}_1^*} & s_{\mathbf{x}_1^*\mathbf{x}_2^*} & \cdots & s_{\mathbf{x}_1^*\mathbf{x}_p^*} \\ s_{\mathbf{x}_1^*\mathbf{x}_2^*} & s^2_{\mathbf{x}_2^*} & \cdots & s_{\mathbf{x}_2^*\mathbf{x}_p^*} \\ \vdots & \vdots & \ddots & \vdots \\ s_{\mathbf{x}_1^*\mathbf{x}_p^*} & s_{\mathbf{x}_2^*\mathbf{x}_p^*} & \cdots & s^2_{\mathbf{x}_p^*} \end{bmatrix}$. Applying Cholesky decomposition to $\mathbf{A} = (A_{jo})$, $j, o = 1, 2, \ldots, p$, we get that:

$$\mathbf{A} = \mathbf{L}\mathbf{L}^{\mathrm{T}},$$

where $\mathbf{L} = (L_{jo})$, $j, o = 1, 2, \ldots, p$ is a real lower triangular matrix such that:

$$L_{jj} = \sqrt{A_{jj} - \sum_{o=1}^{j-1} L_{jo}^2} \quad \text{and} \quad L_{jo} = \frac{A_{jo} - \sum_{h=1}^{o-1} L_{jh}L_{oh}}{L_{oo}}, j > o.$$

Therefore, we get that:

$$\det(\mathbf{A}) = \prod_{j=1}^{p} L_{jj}^2. \tag{4}$$

Theorem 2.1 The generalized variance of covariates under the subdata is maximized by the selection of data points for which $s^2_{\mathbf{x}_j^*}$ is maximized for any $j = 1, 2, \ldots, p$, and $s_{\mathbf{x}_o^*\mathbf{x}_j^*} = 0$ for any $j > o = 1, 2, \ldots, j-1$, simultaneously.

Proof. According to (4), the generalized variance of covariates under the subdata is maximized when L_{jj}^2, or $A_{jj} - \sum_{o=1}^{j-1} L_{jo}^2$, are maximized for any $j = 1, 2, \ldots, p$. Therefore, the maximization of A_{jj} for any $j = 1, 2, \ldots, p$ and $L_{jo}^2 = 0$, or $L_{jo} = 0$, for any $o = 1, 2, \ldots, j-1$ are required, simultaneously.

For $j = 1$ we get that $L_{11}^2 = A_{11}$, and so the maximization of $s^2_{\mathbf{x}_1^*}$ is derived. For $j = 2$ we get that $L_{21} = A_{21}/L_{11}$, and so $A_{21} = 0$ is required. Therefore, the maximization of $s^2_{\mathbf{x}_2^*}$ and $s_{\mathbf{x}_1^*\mathbf{x}_2^*} = 0$ are derived.

For any $j = 3, 4, \ldots, p$ we get that $L_{j1} = A_{j1}/L_{11}$ and so $A_{j1} = 0$ is required. Also, for any $j = 3, 4, \ldots, p$ and for any $o = 2, 3, \ldots, j-1$ we get that $\sum_{h=1}^{o-1} L_{jh}L_{oh} = 0$, since both $L_{jh} = 0$ and $L_{oh} = 0$ for any $h = 1, 2, \ldots, o-1$, and so $A_{jo} = 0$ is required. Therefore, the maximization of $s^2_{\mathbf{x}_j^*}$ and $s_{\mathbf{x}_o^*\mathbf{x}_j^*} = 0$ for any $j = 3, 4, \ldots, p$ and for any $o = 1, 2, \ldots, j-1$ are derived. This concludes the proof.

According to Theorem 2.1, the determinant of the information matrix in (2) is maximized collecting data points for which $s^2_{\mathbf{x}_j^*}$'s are maximized for any $j = 1, 2, \ldots, p$, and $s_{\mathbf{x}_o^*\mathbf{x}_j^*} = 0$ for any $j > o = 1, 2, \ldots, j-1$, simultaneously. Even though such a case may not be feasible, in order to maximize the determinant of the information matrix in (2), one should be interested in collecting data points for which $s^2_{\mathbf{x}_j^*}$, $j = 1, 2, \ldots, p$

are getting as maximum as possible and $s_{\mathbf{x}_o^*\mathbf{x}_j^*}$, $j > o = 1, 2, \ldots, j-1$ tends to zero, simultaneously.

3. EXISTING APPROACHES

We describe briefly existing approaches upon which we want to improve.

3.1 The IBOSS algorithm

Wang et al. (2019) proposed the IBOSS approach in order to select subdata. The scope of their approach is to obtain, given the size of the subdata, data points that provide most of the information contained in the full data. Their idea was motivated by the concept of optimal experimental designs, which aims at organizing, conducting, and interpreting results of experiments in an efficient manner in order to obtain as much useful information as possible, given a budget. More specifically, their idea was to apply the "maximization" of an information matrix, that takes place in optimal experimental designs, to recognize data points that provide most of the information contained in the full data. Therefore, they concluded that an optimal estimator of the unknown parameters of a linear regression model, based on the subdata, can be obtained by "maximizing" the inverse of the covariance matrix of the unknown parameters. They were interested in maximizing a matrix, and so the choice of an optimality criterion function of the matrix was needed. They preferred the popular D-optimality criterion that maximizes the determinant of the inverse of the covariance matrix of the unknown parameters given the subdata.

Furthermore, Wang et al. (2019) developed an algorithm that was based on their information-based optimal subdata selection framework. Their algorithm actually was motivated by their result for an upper bound of the determinant of the inverse of the covariance matrix of the unknown parameters given the subdata. The approach is based on collecting subdata with extreme covariate values, both small and large, occurring with the same frequency. Therefore, the algorithm of the IBOSS approach selects data points with the smallest as well as largest values of all covariates sequentially, given that previous selected data points are excluded. The time complexity of the algorithm of the IBOSS approach is $O(np + kp^2)$.

The IBOSS algorithm outperforms the uniform and the leverage-based subsampling approaches in selecting informative subdata from big data. Also, the IBOSS algorithm compares favorably to the leverage-based subsampling approach in terms of the execution time, both being more efficient than modeling on the full data. The basic approach in Wang et al. (2019) has found several extensions to other cases like Cheng et al. (2020) for logistic regression, Yao and Wang (2019) for multinomial logistic regression, Wang and Ma (2021) for quantile regression. Algorithmic extensions can be found in Wang (2019) and Lee et al. (2021). For further work on the topic see Yao and Wang (2021) and Deldossi and Tommasi (2022).

3.2 The OSS algorithm

Wang et al. (2021), motivated by Wang et al. (2019), proposed a method to select subdata, with a focus on linear regression models, based on an orthogonal array (OA). Their purpose was to obtain subdata that approach an OA. The approach is driven by the fact that a two-level OA represents an optimal design for linear regression, since it minimizes the average variance of the estimated parameters as well as provides the best predictions (Dey and Mukerjee, 1999). An OA represents an optimal design because of its combinatorial orthogonality, and so the OSS approach selects data points with maximum combinatorial orthogonality.

Based on the combinatorial orthogonality of an OA, Wang et al. (2021) developed a sequential addition algorithm that selects data points from the full data focusing on the maximization of the subdata's orthogonality. Also, some data points are deleted in each step of the algorithm in order to reduce the number of candidate data points, and so to speed up the algorithm. All covariates are scaled to $[-1, 1]$. The OSS approach, and so the developed algorithm, is based on a discrepancy function that was defined by Wang et al. (2021) in order to measure the distortion of data points on keeping two features simultaneously. The two features are connected with the optimality of orthogonal arrays. The first feature is that data points are located at the corners of the data domain $[-1, 1]^p$ and have large distances from the center. Thus, extreme data points provide more information about the model. The second feature is the combinatorial orthogonality, that is data points (their signs) are as dissimilar as possible. The computational complexity of the algorithm of the OSS approach is $O(np\log k)$.

A design is A-optimal when the trace of the inverse of its information matrix is minimized in the class of all designs. Wang et al. (2021) evaluated their approach by calculating, among others, the D- and A-efficiencies of the selected subdata, since a two-level OA represents a D-, A-optimal design for linear regression. The D- and A-efficiencies of the subdata can be calculated using:

$$\mathrm{D}_{eff} = \frac{\det(\mathbf{Q}_{\mathrm{Sub}})^{1/(p+1)}}{k}$$

and

$$\mathrm{A}_{eff} = \frac{p+1}{k\sum_{j=1}^{p+1}\lambda_j(\mathbf{Q}_{\mathrm{Sub}}^{-1})},$$

respectively, where $\lambda_j(\mathbf{Q}_{\mathrm{Sub}}^{-1})$ denotes the jth eigenvalue of $\mathbf{Q}_{\mathrm{Sub}}^{-1}$.

The OSS approach outperforms the uniform subsampling and the IBOSS approach in selecting informative subdata from big data. Also, the OSS approach is faster than the IBOSS one, and they are both faster than modeling on the full data. It is important to mention that the IBOSS approach selects data points with only the first aforementioned feature without taking into any consideration of the second feature. Therefore, the consideration of the second feature is an important improvement of the OSS approach compared with the IBOSS one.

Theorem 2.1 shows that the perfect case is achieved under orthogonality taking the extreme points. The OSS algorithm, since the covariates are continuous, cannot achieve this. Since it is too difficult in practice to achieve zero covariance between the covariates, Theorem 2.1 implies that a good strategy in practice towards a better subdata selection is to select observations that maximize the generalized variance.

4. THE NEW PROPOSED ALGORITHMS

In this section, we develop an algorithm with a primary focus on improving the algorithm of the OSS approach, such that a reselection of subdata leads to an increase of the generalized variance.

The input of our algorithm is the subdata obtained by the OSS algorithm, and so the implementation of our algorithm requires the implementation of the OSS algorithm. Our goal is to identify and interchange selected data points by the OSS algorithm with those that were not selected. The criterion that allows the aforementioned interchange is the increase of the generalized variance, denoted as V. We do not take into consideration all data points that were not selected by the OSS algorithm as candidate data points, but some of them. The candidate data points are obtained in the same way that the IBOSS algorithm includes data points for all covariates. The difference is that the candidate data points in our approach are simultaneously selected for all covariates, and so previously selected data points are not excluded. Such a method of selection can lead to at least duplicated data points, and so only one of them is kept. Therefore, we are not able to know the final size of the candidate data points, since it is not feasible to know in advance the existence of at least duplicated data points. However, the maximum final size of the candidate data points is equal Kp, where K is an even number of candidate data points selected for each covariate.

Our first algorithm (Alg1) aims at improving the OSS algorithm.

Remark 4.1 In Alg1, the data points $\mathbf{f}_w$, $w = 1, 2, \ldots, N_{\text{F}}$ that are interchanged with the data points $\mathbf{s}_i$, $i = 1, 2, \ldots, k$, are selected in the order in which $\mathbf{F} = (\mathbf{f}_w)$ is constructed in Step 2.

A consequence of Remark 4.1 is that the interchange of a data point $\mathbf{s}_i$ with a later data point of $\mathbf{F}$ could possibly lead to a greater generalized variance V_{new}. However, it is not feasible to determine if such a case can occur, since Alg1 interchanges a data point $\mathbf{s}_i$ with the first data point that locates in $\mathbf{F}$ given that $V_{\text{new}} > V$.

The aforementioned consequence of Remark 4.1 proposes a variation on Alg1, and so a new algorithm (VAlg1) seeks to improve the generalized variance of the final subdata. VAlg1 makes a change as to which data point of $\mathbf{F}$ is chosen to be interchanged with a data point of $\mathbf{S}$. Therefore, a data point $\mathbf{s}_i$ is interchanged with that data point of $\mathbf{F}$, among all remaining ones, for which V_{new} is the largest possible. The interchanging between data points in VAlg1 is applied to all data points of $\mathbf{S}$.

Algorithm 1 Alg1

Input: subdata $\mathbf{S} = (\mathbf{s}_i), i = 1, 2, \ldots, k$ of the OSS approach, initial full data $\mathbf{D}_{\text{Full}}$, subdata size k, candidate data points K from each covariate

Output: new obtained subdata $\mathbf{S}$

Step 1: Preperation

$\mathbf{S} = \text{convert}(\mathbf{S})$ ▷ convert subdata $\mathbf{S}$ to their initial values

$V = \det(\mathbf{Q}_{\text{Sub}})$ ▷ generalized variance of $\mathbf{S}$

$\mathbf{D} = \mathbf{D}_{\text{Full}} - \mathbf{S} = (d_{rj})$ ▷ remaining data points $\mathbf{d}_{r\cdot} = (d_{r1}, \ldots, d_{rp}) \notin \mathbf{S}$

$N_{\text{F}} = \text{nrow}(\mathbf{D})$ ▷ number of data points $\mathbf{d}_{r\cdot} \in \mathbf{D}$

$\mathbf{F} = \emptyset$ ▷ initialize the index set of candidate data points

Step 2: Find candidate data points

for j in $1, \ldots, p$ **do**

$\mathbf{d}_{\cdot j} = \text{sort}(\mathbf{d}_{\cdot j})$ ▷ sort $\mathbf{d}_{\cdot j} = (d_{1j}, \ldots, d_{N_{\text{F}}j})$

$\mathbf{D} = \text{sort}(\mathbf{D})$ ▷ sort $\mathbf{D}$ based on $\mathbf{d}_{\cdot j}$

$\mathbf{F} = \mathbf{F} \cup \mathbf{d}_{1\cdot} \cup \cdots \cup \mathbf{d}_{K/2\cdot}$

$\mathbf{F} = \mathbf{F} \cup \mathbf{d}_{N_{\text{F}}-K/2+1\cdot} \cup \cdots \cup \mathbf{d}_{N_{\text{F}}\cdot}$

end for

$\mathbf{F} = \text{unique}(\mathbf{F})$ ▷ keep unique data points of $\mathbf{F} = (\mathbf{f}_w)$

$N_{\text{F}} = \text{nrow}(\mathbf{F})$ ▷ number of data points $\mathbf{f}_w \in \mathbf{F}$

Step 3: Main algorithm

for i in $1, \ldots, k$ **do**

for w in $1, \ldots, N_{\text{F}}$ **do**

$\mathbf{s}_i \leftrightarrow \mathbf{f}_w$ ▷ interchange data points $\mathbf{s}_i$ and $\mathbf{f}_w$

$V_{\text{new}} = \det(\mathbf{Q}_{\text{Sub}})$ ▷ generalized variance of new $\mathbf{S}$

if $V_{\text{new}} > V$ **then**

$V = V_{\text{new}}$

break

else

$\mathbf{s}_i \leftrightarrow \mathbf{f}_w$

end if

end for

end for

return S

Algorithm 2 VAlg1

Steps 1 and 2: Same as in Alg1
Step 3: Main algorithm
for i in $1, \ldots, k$ **do**
  for w in $1, \ldots, N_{\mathrm{F}}$ **do**
    $\mathbf{s}_i \leftrightarrow \mathbf{f}_w$ ▷ interchange data points $\mathbf{s}_i$ and $\mathbf{f}_w$
    $V_{\text{new}} = \det(\mathbf{Q}_{\text{Sub}})$ ▷ generalized variance of new $\mathbf{S}$
    if $V_{\text{new}} > V$ **then**
      $V = V_{\text{new}}$
    else
      $\mathbf{s}_i \leftrightarrow \mathbf{f}_w$
    end if
  end for
end for
return S

The time complexity to construct $\mathbf{F}$ is $O((n-k)p)$. The procedure of interchanging data points between $\mathbf{S}$ and $\mathbf{F}$ has time complexity $O(kN_{\mathrm{F}})$. Thus, the time complexity of Alg1 is $O(np\log k + (n-k)p + kN_{\mathrm{F}})$. VAlg1 has the same time complexity as Alg1. In the remaining of this paper, in the cases that Alg1 is executed with more than one iteration, only Step 3 is repeated.

Note that both Alg1 and VAlg1 are, in fact, an attempt to implement what Theorem 2.1 indicates. Attempting to increase the variance of covariates under the subdata we collect as candidate data points, data points close to the extreme ones, i.e. data points that are considered as more probable to lead us close to orthogonality.

5. SIMULATION EXPERIMENT

In this section, we use simulated data in order to evaluate the performance of both Alg1 and VAlg1.

The observations $\mathbf{x}_i$'s follow a multivariate normal distribution, that is, $\mathbf{x}_i \sim N(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma} = (\Sigma_{ij})$, $i, j = 1, 2, \ldots, p$ is a covariance matrix. Also, $\Sigma_{ij} = 1$ for $i = j = 1, 2, \ldots, p$ and $\Sigma_{ij} = 0.5$ for $i \neq j = 1, 2, \ldots, p$.

The response data are generated from the linear model in (1) with the true value of $\boldsymbol{\beta}$ being a 11 dimensional vector with all elements equal to 1. The error terms ϵ_i's are normally distributed with $E(\epsilon_i) = 0$ and $V(\epsilon_i) = 3$. An intercept is included, so $p = 10$.

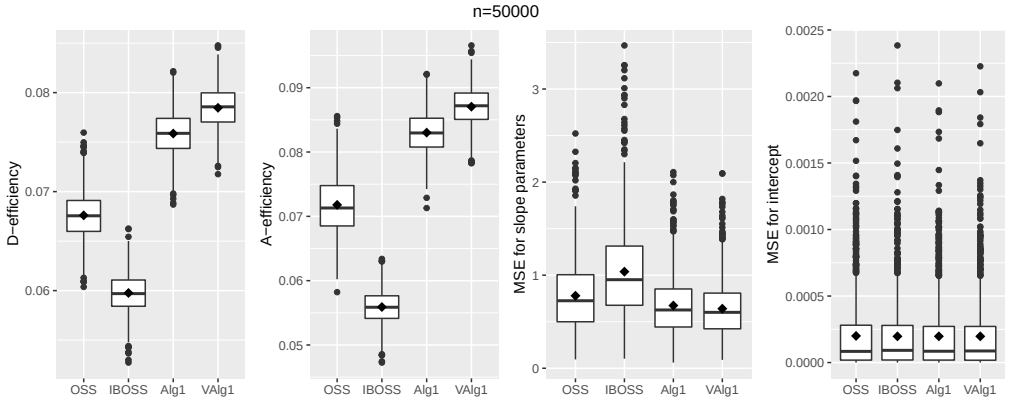
The simulation is repeated 1000 times. We calculate the D-, A-efficiencies and the mean squared error (MSE) of the subdata selected by our approach as well as the approaches of IBOSS and OSS. As shown in Wang et al. (2019) and Wang et al. (2021), we estimate the intercept with the adjusted estimator $\hat{\beta}_0 = \bar{y} - \bar{\mathbf{x}}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_1^{Sub}$, where $\bar{y}$ is the mean of the response full data, $\bar{\mathbf{x}}$ is the vector of means of all covariates in the full data,

and $\hat{\boldsymbol{\beta}}_1^{Sub}$ is the ordinary least squares estimate of $\boldsymbol{\beta}_1^{Sub}$ based on the subdata. Therefore, we consider $(\hat{\beta}_0^{(r)} - \beta_0)^2$ and $||\hat{\boldsymbol{\beta}}_1^{(r)} - \boldsymbol{\beta}_1||^2$ the MSE for intercept and slope estimators in the rth repetition, where $\hat{\beta}_0^{(r)}$ and $\hat{\boldsymbol{\beta}}_1^{(r)}$ are $\hat{\beta}_0$ and $\hat{\boldsymbol{\beta}}_1^{Sub}$ in the rth repetition.

We are interested in investigating the case when the full data size is $n = 5 \times 10^4$, and the subdata size is fixed at $k = 100$. Also, the maximum final size of the candidate data points is fixed at $Kp = 250$, that is $K = 25$. Alg1 is executed with 5 iterations, and VAlg1 is executed once.

Figure 2 shows the MSEs, D-efficiency, and A-efficiency for the subdata selected by different approaches. The mean values (♦) are also provided.

Figure 2: *The MSEs, D- and A-efficiencies for the subdata selected by different approaches, when the full data size is $n = 5 \times 10^4$, the subdata size is $k = 100$, and the number of candidate data points is $K = 25$. Alg1 is executed with 5 iterations, and VAlg1 is executed once.*



Both Alg1 and VAlg1 outperform the IBOSS and OSS algorithms in each of the D-, A-optimality criteria, and provide more accurate estimates for the model parameters as one can see in Figure 2. The important finding is the magnitude of progress in the D-, A-optimality criteria rather the absolute improvement which is sure since we start from the OSS algorithm and we proceed by improving the generalized variance. MSE for intercept is immutable among the three approaches. Also, VAlg1 can be considered to be more effective than Alg1. Such a result can find justification in the interchanging of data points between **S** and **F**. One could argue that our approach is more effective when each data point of **S** is interchanged with a data point of **F** for which V_{new} is the largest possible. However, Alg1 may be more effective than VAlg1 under other circumstances. Such a result could be achieved either executing Alg1 with more than 5 iterations or applying it under another set of n, p, k and K.

It is worth mentioning that our approach performs well under the A-optimality criterion, even though both Alg1 and VAlg1 are developed based on the increasing of the determinant of the information matrix of the subdata. Therefore, our approach could provide some insights on how to obtain subdata that approach an OA.

6. REAL DATA APPLICATION

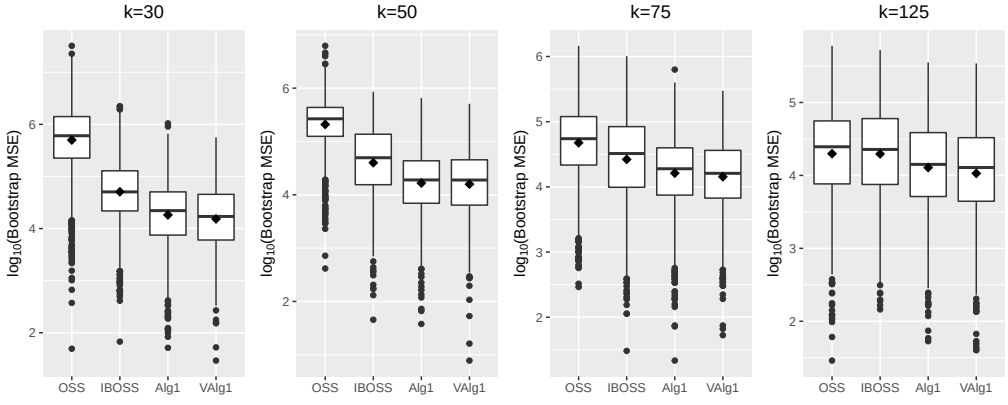
We evaluate the performance of our approach on a real data example, in which the accuracy of the ordinary least squares estimates of slope parameters in model (1) is examined, based on the subdata obtained by our approach.

The dataset of this example is related to power consumption of three different distribution networks of Tetouan city which is located in north Morocco. The full data contains $n = 52,416$ data points and contains five measurements: temperature, humidity, wind speed, diffuse flows and general diffuse flows, so the number of covariates in the model is $p = 5$. The response variable is the power consumption of the second zone of Tetouan city. Further information about the dataset can be found in Salam and Hibaoui (2018).

We are interested in comparing the performance of our algorithms with the algorithms of the approaches of IBOSS and OSS, and so we consider the MSE for the vector of slope parameters for each algorithm by using 1000 bootstrap samples, as in Wang et al. (2019) and Wang et al. (2021). Each bootstrap sample is a random sample of size n from the full data using sampling with replacement. For a bootstrap sample, we implement each algorithm in order to obtain the subdata and then from the selected subdata we estimate the parameters of the model.

We consider $k = 30, 50, 75, 125$, and $K = 10$. Alg1 is executed with 5 iterations, and VAlg1 is executed once. Figure 3 shows the bootstrap MSEs by different approaches. The mean values (⧫) are also provided. Our approach, that is both Alg1 and VAlg1, outperforms the IBOSS and OSS algorithms in minimizing the bootstrap MSEs. We notice that Alg1 and VAlg1 have a similar behavior, but VAlg1 being slightly better in this case. Also note that the improvement we get from our algorithms is much larger when the subdata size is smaller.

Figure 3: *The bootstrap MSEs for the subdata selected by different approaches, when the subdata size is $k = 30, 50, 75, 125$, and the number of candidate data points is $K = 10$. Alg1 is executed with 5 iterations, and VAlg1 is executed once.*



In Table 1, we present the mean execution time (in seconds) of Alg1 and VAlg1 in

the aforementioned real data application. All computations are carried out on a PC with 3.6 GHz Intel 8-Core I7 processor and 16GB memory.

Table 1: *The mean execution time (in seconds) of Alg1 and VAlg1 for the real data application, which is related to power consumption of three different distribution networks of Tetouan city. Alg1 is executed with* 5 *iterations, and VAlg1 is executed once. The subdata size is* $k = 30, 50, 75, 125$*, and the number of candidate data points is* $K = 10$*.*

Algorithm	k	Mean execution time (in seconds)
Alg1	30	0.115
	50	0.283
	75	0.433
	125	0.619
VAlg1	30	0.039
	50	0.077
	75	0.131
	125	0.183

VAlg1 is executed once, since it searches among all data points of $\mathbf{F}$. The implementation of VAlg1 with more than one execution may result to a time-consuming algorithm without any improved results. As in Alg1, the mean execution time of VAlg1 is getting slower as the number of k increases. Both Alg1 and VAlg1 start from the OSS approach. Even though the execution times of both Alg1 and VAlg1 are added to the execution time of the OSS approach in order to obtain an optimal subdata, the results in Figure 2 and Table 1 indicate that our approach is worth implementing.

7. CONCLUDING REMARKS

We have presented algorithms to select data points in an optimal way from a big dataset so as to be able to run regression and derive coefficients that share as much information as possible. The newly developed algorithms were compared with existing ones to show the kind of improvement that we can take back.

The theoretical results in Section 2 indicate the working direction of Wang et al. (2019) and Wang et al. (2021). However, as Wang et al. (2021) stated, finding subdata that exactly approach an OA may be impossible in many cases. Theorem 2.1 shows that an orthogonal array provides the best result but in practice this is not feasible with real data. The results in Sections 5 and 6 show that neither the selection of only extreme data points nor the approach of an OA can always lead to the most informative subdata. Therefore, our approach is generally based on the maximization of the generalized variance, without being interested in directly obtaining specific data points according to Theorem 2.1. However Theorem 2.1 implies that searching towards the maximization is a good strategy for practical purposes.

Also, Wang et al. (2021) discussed that the OSS approach could be improved by a

greedy modification of their algorithm, that is to interchange data points between the selected subdata and the remaining data points based on the minimization of a discrepancy function. Our approach is an updated version of what Wang et al. (2021) discussed, since we obtain some candidate data points in advance, and so a very time-consuming algorithm is avoided. Moreover, two features of our approach that can be modified are the number of iterations of Alg1 and which data points are interchanged. While here we pursue D-optimality the presented approach satisfies good properties with respect to A-optimality. Also, extensions to other models are straightforward, as for example GLM type of models.

We would like to mention that our `R` code was not optimized in anyway and thus perhaps further time savings are possible.

ΠΕΡΙΛΗΨΗ

Στην εποχή των μεγάλων δεδομένων οι ερευνητές αντιμετωπίζουν μια σειρά από προβλήματα. Ακόμη και οι κλασικές προσεγγίσεις, όπως η γραμμική παλινδρόμηση, μπορεί να είναι δύσκολες ή προβληματικές, όταν ο όγκος των δεδομένων είναι τεράστιος. Οι παραδοσιακές προσεγγίσεις για παλινδρόμηση σε μεγάλα σύνολα δεδομένων ενδέχεται να μην είναι αποτελεσματικές λόγω του μεγάλου μεγέθους του δείγματος, καθώς περιλαμβάνουν την αντιστροφή τεράστιων πινάκων ή ακόμα και επειδή τα δεδομένα δεν μπορούν να χωρέσουν στη μνήμη. Οι προτεινόμενες προσεγγίσεις βασίζονται στην επιλογή αντιπροσωπευτικών υποδεδομένων για την εκτέλεση της παλινδρόμησης. Οι υπάρχουσες προσεγγίσεις επιλέγουν τα υποδεδομένα χρησιμοποιώντας κάποια κριτήρια βελτιστοποίησης ή/και ιδιότητες από τους ορθογώνιους σχηματισμούς. Στην παρούσα εργασία βελτιώνουμε τους υπάρχοντες αλγόριθμους παρέχοντας έναν νέο αλγόριθμο που βασίζεται στο κριτήριο της D-βελτιστοποίησης. Επίσης, παρέχονται ένα πείραμα προσομοίωσης καθώς και μια εφαρμογή σε πραγματικά δεδομένα.

REFERENCES

Cheng, Q., Wang, H., and Yang, M. (2020). Information-based optimal subdata selection for big data logistic regression. *Journal of Statistical Planning and Inference*, **209**, 112–122.

Deldossi, L. and Tommasi, C. (2022). Optimal design subsampling from big datasets. *Journal of Quality Technology*, **54**(1), 93–101.

Dey, A. and Mukerjee, R. (1999). *Fractional Factorial Plans*. John Wiley & Sons.

Drineas, P., Mahoney, M. W., Muthukrishnan, S., and Sarlós, T. (2011). Faster least squares approximation. *Numerische mathematik*, **117**(2), 219–249.

Harville, D. A. (1997). *Matrix Algebra From a Statistician's Perspective*. New York: Springer-Verlag.

Lee, J., Schifano, E. D., and Wang, H. (2021). Fast optimal subsampling probability approximation for generalized linear models. *Econometrics and Statistics*.

Salam, A. and Hibaoui, A. E. (2018). Comparison of machine learning algorithms for the power consumption prediction: - case study of Tetouan city -. In: *2018 6th International Renewable and Sustainable Energy Conference (IRSEC)*, pages 1–5.

Wang, H. (2019). Divide-and-conquer information-based optimal subdata selection algorithm. *Journal of Statistical Theory and Practice*, **13**(3), 1–19.

Wang, H. and Ma, Y. (2021). Optimal subsampling for quantile regression in big data. *Biometrika*, **108**(1), 99–112.

Wang, H., Yang, M., and Stufken, J. (2019). Information-based optimal subdata selection for big data linear regression. *Journal of the American Statistical Association*, **114**(525), 393–405.

Wang, L., Elmstedt, J., Wong, W. K., and Xu, H. (2021). Orthogonal subsampling for big data linear regression. *Annals of Applied Statistics*, **15**(3), 1273–1290.

Wilks, S. S. (1932). Certain generalizations in the analysis of variance. *Biometrika*, 24(3-4), 471–494.

Yao, Y. and Wang, H. (2019). Optimal subsampling for softmax regression. *Statistical Papers*, **60**(2), 585–599.

Yao, Y. and Wang, H. (2021). A review on optimal subsampling methods for massive datasets. *Journal of Data Science*, **19**(1), 151–172.

ΤΡΑΠΕΖΑ ΤΗΣ ΕΛΛΑΔΟΣ

ΕΥΡΩΣΥΣΤΗΜΑ