2023

ΠΡΑΚΤΙΚΑ
35ου Πανελληνίου & 1ου Διεθνούς Συνεδρίου Στατιστικής

ΣΤΑΤΙΣΤΙΚΗ ΣΤΙΣ ΕΠΙΣΤΗΜΕΣ ΤΗΣ ΥΓΕΙΑΣ

Ελληνικό Στατιστικό Ινστιτούτο
Greek Statistical Institute

ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ

Ελληνικό Στατιστικό Ινστιτούτο
Greek Statistical Institute

ΠΡΑΚΤΙΚΑ

35ου Πανελληνίου & 1ου Διεθνούς Συνεδρίου Στατιστικής

Αθήνα, 25-28 Μαΐου 2023

ΣΤΑΤΙΣΤΙΚΗ ΣΤΙΣ ΕΠΙΣΤΗΜΕΣ ΤΗΣ ΥΓΕΙΑΣ

Οργάνωση

Ελληνικό Στατιστικό Ινστιτούτο

Πανεπιστήμιο Δυτικής Αττικής

ISSN 1792-2461

Περιεχόμενα– Contents

Πρόλογος (Preface In Greek)…….....…**5**
Επιτροπές (Committees In Greek)……....……..**9**
Πρόγραμμα Συνεδρίου / Conference Program…….…...**14**

Εργασίες στα Ελληνικά– Papers in Greek

Κούτρας Μ. Β., Δαφνής Σπ..: Μελέτη των ιδιοτήτων μιας ευρείας οικογένειας συνέχων κατανομών…… …..**25**
Μάρκου Μ., Καρυώτη Β.: Πρόβλεψη Χρονοσειρών και μελλοντικών ακροτάτων σε αγχώδεις διαταραχές στον πληθυσμό της Ελλάδας……**35**
Μπομποτάς Π., Κουρούκλης Σ..: Βελτιωμένη Εκτίμηση παραμέτρων σε Inverse Gaussian Κατανομή……**46**
Παπαγεωργίου Β., Τσακλίδης Γ. : Ένα Επεκτεταμένο Επιδημιολογικό Φίλτρο Σωματιδίων Για την περιγραφή της μετάδοσης μολυσματικών ασθενειών. Εφαρμογή στα δεδομένα Covid-19 στην Ιταλία…….....**58**
Παπαγεωργίου Γ., Χαλικιάς Μ..: Σύγκριση εκτιμητών για την πρόβλεψη εκλογικών αποτελεσμάτων……...**74**
Ρακιτζής Αθ., Μαμζερίδου Ε.: Άνω μονόπλευρα διαγράμματα ελέγχου με εκτιμημένες παραμέτρους για μηδενοδιογκωμένες διεργασίες.....**92**

Εργασίες Στα Αγγλικά – Papers In English

Anastasiou A., Tsimikas J.: Optimal threshold selection on a two-cutpoint optimal r curv.....**107**
Chasiotis V., Karlis D.: Subdata selection for big data regression based on leverage scores……....**122**
Felipe A., Jaenada M., Miranda P., Pardo L.: Applying renyi's pseudodistance for robust model selection for independent not identically distributed observations **136**
Fountoukidis K. G., Antzoulakos D. L., and Rakitzis A. C. : A new adaptive run sum max chart.....**151**
Georgiou K., Charmanas K., Angelis L.: Thematic analysis and research trends on medical diagnosis patents using state-of-the-art topic modeling.....**165**
Kyrana Z, Pratsinakis E, Papafilippou N, Dordas C, Markos A, Menexes G.:.Comparison of dimensionality reduction and clustering methods on The Multidimensional Dataset Forest Cover Type" with mixed-type data **181**
Tasias K.,, Mpistintzanos P., Pekridis F.: A Mission Reliability Redundancy Modelfor An Aircraft Fleet With Cold Standby Spare Parts……**198**

Πρόλογος

Το *Πανελλήνιο Συνέδριο Στατιστικής* αποτελεί την κύρια επιστημονική εκδήλωση στο χώρο της Στατιστικής και των Πιθανοτήτων, στην Ελλάδα. Συνδιοργανώνεται σε ετήσια βάση, από το 1985, από το Ελληνικό Στατιστικό Ινστιτούτο, το οποίο είναι και μέλος της Ομοσπονδίας Ευρωπαϊκών Εθνικών Στατιστικών Εταιρειών (FENStatS), και ένα Ελληνικό ή Κυπριακό Ακαδημαϊκό Ίδρυμα. Το 35ο Εθνικό και το 1ο Διεθνές Συνέδριο Στατιστικής που πραγματοποιήθηκε τον Μάιο του 2023, διοργανώθηκε από το Ελληνικό Στατιστικό Ινστιτούτο και το Πανεπιστήμιο Δυτικής Αττικής, στην Αθήνα. Η εν λόγω εκδήλωση έφερε κοντά Στατιστικούς, Αναλυτές Δεδομένων, Μαθηματικούς, Πληροφορικούς, καθώς και επιστήμονες από όλο το φάσμα των θετικών επιστημών, με σκοπό την ανταλλαγή ιδεών και τη συζήτηση των πρόσφατων εξελίξεων στην ευρύτερη περιοχή της Στατιστικής Επιστήμης και ειδικότερα στην εφαρμογή της στατιστικής στις επιστήμες Υγείας. Η ιστορία του Συνεδρίου είναι μεγάλη και σπουδαία. Καταγράφει την ιστορία της Στατιστικής επιστήμης στην Ελλάδα, αναδεικνύοντας κάθε χρόνο νέες τάσεις τόσο στη θεωρία, αλλά και στα εφαρμοσμένα προβλήματα που εμφανίζονται συνεχώς σε άλλες επιστήμες και αναζητούν λύσεις στις Πιθανότητες και τη Στατιστική. Το Συνέδριο 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.

Η καινοτομία της φετινής εκδήλωσης ήταν η **διεθνοποίηση του Συνεδρίου**. Για πρώτη φορά το ΔΣ του ΕΣΙ, σε συνεργασία με το ανάδοχο Πανεπιστήμιο Δυτικής Αττικής, διοργάνωσε το 1ο Διεθνές Συνέδριο Στατιστικής, με κυρίαρχο θέμα του Συνεδρίου την «*Στατιστική στις Επιστήμες της Υγείας*». Η εν λόγω εκδήλωση έσμιξε Στατιστικούς, Βιοστατιστικούς, Επιδημιολόγους, Μαθηματικούς, Αναλυτές Δεδομένων, Πληροφορικούς, καθώς και επιστήμονες από το ευρύτερο πεδίο των

Εφαρμοσμένων Επιστημών, **από διάφορα μέρη του κόσμου**, με σκοπό την ανταλλαγή ιδεών και τη συζήτηση των πρόσφατων εξελίξεων στην ευρύτερη περιοχή της Στατιστικής Επιστήμης και των εφαρμογών της στην υγεία. Στο Συνέδριο συμμετείχαν 6 Έλληνες και ξένοι Κεντρικοί Ομιλητές, καθώς επίσης και 37 Έλληνες και ξένοι ομιλητές στο πλαίσιο *Ειδικών Συνεδριών*, με θεματολογία από όλο το φάσμα της Στατιστικής. Συνολικά συμμετείχαν περισσότεροι από 150 σύνεδροι, και παρουσιάστηκαν 106 εργασίες. **Για το Ελένειο Βραβείο Καλύτερης Διδακτορικής Διατριβής** στη Στατιστική που απονέμεται στη μνήμη της μικρής Ελένης με πρόταση και χορηγία των γονέων της , του συναδέλφου Τρύφωνα Δάρα και της συζύγου του Πολυξένης οι φετινοί υποψήφιοι ήταν οι : Θωμαδάκης Χρήστος με τη διατριβή του «Statistical modeling of CD4 cell count evolution in HIV-positive individuals before and after antiretroviral treatment initiation under missing data due to various dropout mechanisms», Καλλιγέρης Εμμανουήλ-Νεκτάριος με τη διατριβή του «New Developments on Modelling Techniques for Dynamical Systems with Applications», Λύκου Ρόδη με τη διατριβή της «Φίλτρα Γενικών κρυφών μαρκοβιανών μοντέλων, βελτιώσεις στο φίλτρο σωματιδίων με ελλιπή δεδομένα και εφαρμογές», Μεσελίδης Χρήστος με τη διατριβή του «Mathematical Modelling of Categorical Data with Actuarial and Financial Applications», Μπουραζάς Κωνσταντίνος με τη διατριβή του «Self-Starting methods in Bayesian statistical process control & monitoring», και Χατζηλένα Αναστασία με τη διατριβή της «Essays on Epidemic Models and their Statistical Analysis». Η Επιτροπή αξιολόγησης που αποτελούνταν από τους Καθηγητές Μαρία Κατέρη, Παναγιώτη Παπασταμούλη, Σάμη Τρέβεζα, Αθανάσιο Ρακιτζή και Δημήτριο Φουσκάκη, απένειμε το βραβείο στον κ. Χ. Θωμαδάκη ο οποίος και παρουσίασε σε ολομέλεια την εργασία του. Το Βραβείο **Καλύτερης Εργασίας Νέου Έλληνα Στατιστικού,** το οποίο έχει δωροθετήσει ο Καθηγητής Balakrishnan (McMaster University, Canada), απονεμήθηκε, μετά από απόφαση της Επιτροπής Βραβείου, στους κ. Βασίλειο Παπαγεωργίου και Maria Jaenada. Οι υποψήφιοι ήταν οι Maria Jaenada με την εργασία «Robust estimation for step-stress experiments with non-destructive one-shot devices under lognormal lifetime distribution», Εμμανουήλ-Νεκτάριος Καλλιγέρης με την εργασία «On Stochastic Dynamic Modeling of Incidence Data», Τζουμέρκας Γιώργος με την εργασία «A Comparison of Power-expected-posterior Priors in Shrinkage Regression», και Παπαγεωργίου Βασίλειος με την εργασία του «Ένα επεκτεταμένο επιδημιολογικό φίλτρο σωματιδίων για την περιγραφή της μετάδοσης

μολυσματικών ασθενειών. Εφαρμογή στα δεδομένα COVID-19 στην Ιταλία». Η Επιτροπή του Βραβείου αποτελείτο από τους Καθηγητές Μπατσίδη Α., Μπουρνέτα Α., και Ψαρράκο Γ. Στα χέρια σας κρατάτε τα **Πρακτικά του Συνεδρίου** στα οποία υποβληθήκαν 15 εργασίες, και έγιναν δεκτές 13 εργασίες. Η Επιτροπή Έκδοσης Πρακτικών του ΕΣΙ εκφράζει τις ευχαριστίες της προς τους κριτές: κ.κ. Αφένδρας Γ., Βόντα Ι., Γεωργακόπουλος Σ., Δημόπουλος Α., Θωμαδάκης Χ., Καραγιάννης Β., Κούτρας Β., Μακρίδης Α., Μαραβελάκης Π., Μαχαιράς Ν., Μαυρίδης Δ., Μεσελίδης Χ., Μπατσίδης Α., Μπομποτάς Π., Μωυσιάδης Π, Οικονόμου Π., Παπαγεωργίου Γ., Παπασταμούλης Π., Πετρόπουλος Κ., Σαχλάς Α., Σύψα Β, Σωτηράκογλου Κ., Τζαβελάς Γ., Τριανταφυλλόπουλος Κ., Χελιώτης Δ., Toma Α. 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

Επιτροπές

Οργανωτική Επιτροπή

Αγγελής Ε., Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης.
Βαμβακάρη Μ., Χαροκόπειο Πανεπιστήμιο.
Καλλιβωκάς Δ., Πανεπιστήμιο Δυτικής Αττικής.
Καραγρηγορίου Α., Πανεπιστήμιο Αιγαίου.
Κοσσιέρη Ε., Πανεπιστήμιο Δυτικής Αττικής.
Μαλεφάκη Σ., Πανεπιστήμιο Πατρών.
Μπερσίμης Σ., Πανεπιστήμιο Πειραιώς.
Παναγιωτάκος Δ., Χαροκόπειο Πανεπιστήμιο.
Παπαγεωργίου Ε., Πανεπιστήμιο Δυτικής Αττικής.
Παπαγεωργίου Γ., Πανεπιστήμιο Δυτικής Αττικής.
Χαλικιάς Μ., Πανεπιστήμιο Δυτικής Αττικής.
Ψαρράκος Γ., Πανεπιστήμιο Πειραιώς.

Επιστημονική Επιτροπή

Alin A., Dokuz Eylul University, Turkey.
Balakrishnan N., McMaster University Canada.
Barbu V., University of Rouen - Normandy, France.
Bayramoglu I., Izmir University of Economics.
Caballero Diaz F. F., Autonomus University of Madrid, Spain.
Dellaportas P., Athens University of Economics & Business / University College London.
Markatou M., University at Buffalo, USA.
Oliveira A., University of Aberta, Portugal.
Oliveira T., University of Aberta, Portugal.
Pierri F., University of Perugia, Italia.
Ram M., Graphic Era, India & Peter the Great St. Petersburg Polytechnic University, Saint Petersburg, Russia.
Refinetti P., Ecole Polytechnique Fédérale de Lausanne, Switzerland.
Trandafir C., Universidad Pública de Navarra, ES.
Viana M., University of Illinois at Chicago.

Yiannoutsos C. T., Indiana University, USA.
Αγγελής Ε., Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης.
Βαμβακάρη Μ., Χαροκόπειο Πανεπιστήμιο.
Βασδέκης Β., Οικονομικό Πανεπιστήμιο Αθηνών.
Δάφνη Ο., Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών.
Δάρας Τ., Πολυτεχνείου Κρήτης.
Δαμιανού Χ., Πολυτεχνείου Αθηνών.
Ευαγγελάρας Χ., Πανεπιστήμιο Πειραιώς.
Ζωγράφος Κ., Πανεπιστήμιο Ιωαννίνων.
Ηλιόπουλος Γ., Πανεπιστήμιο Πειραιώς.
Καλαματιανού Α., Πάντειο Πανεπιστήμιο Κοινωνικών και Πολιτικών Επιστημών.
Καραγιάννη Β., Πανεπιστήμιο Δυτικής Αττικής.
Καραγρηγορίου Α., Πανεπιστήμιο Αιγαίου.
Καρλής Δ., Οικονομικό Πανεπιστήμιο Αθηνών.
Καρώνη Χ., Εθνικό Μετσόβειο Πολυτεχνείο.
Κατέρη Μ., RWTH Aachen University Germany.
Κίτσος Χ., Πανεπιστήμιο Δυτικής Αττικής.
Κουνιάς Σ., Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών, Επίτιμος Πρόεδρος του ΕΣΙ.
Κούτρας Μ., Πανεπιστήμιο Πειραιώς.
Κωνσταντινίδης Δ., Πανεπιστήμιο Αιγαίου.
Μακρή Φ., Πανεπιστήμιο Πατρών.
Μαλεφάκη Σ., Πανεπιστήμιο Πατρών.
Μηλιένος Φ., Πάντειο Πανεπιστήμιο Κοινωνικών και Πολιτικών Επιστημών.
Μοσχονά Θ., Πανεπιστήμιο Δυτικής Αττικής.
Μπασιάκος Ι., Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών.
Μπατσίδης Α., Πανεπιστήμιο Ιωαννίνων.
Μπερσίμης Σ., Πανεπιστήμιο Πειραιώς.
Μπουρνέτας Α., Πανεπιστήμιο Αθηνών.
Μπραγουδάκης Ζ., Τράπεζα της Ελλάδος και Επισκέπτης Καθηγητής, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών.
Μωυσιάδης Χ., Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης.
Ντζούφρας Ι., Οικονομικό Πανεπιστήμιο Αθηνών.
Οικονόμου Π., Πανεπιστήμιο Πατρών.
Παναγιωτάκος Δ., Χαροκόπειο Πανεπιστήμιο.
Παπαγεωργίου Ε., Πανεπιστήμιο Δυτικής Αττικής.
Παπαδάτος Ν., Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών.
Παπαδόπουλος Γ., Γεωπονικό Πανεπιστήμιο Αθηνών.

Παπαϊωάννου Τ., Πανεπιστήμιο Πειραιώς και Πανεπιστήμιο Ιωαννίνων, Επίτιμος Πρόεδρος του ΕΣΙ.
Πετρόπουλος Κ., Πανεπιστήμιο Πατρών.
Ρακιτζής Α., Πανεπιστήμιο Πειραιώς.
Συρακούλης Κ., Πανεπιστήμιο Θεσσαλίας.
Συψα Β., Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών.
Σωτηράκογλου Κ., Γεωπονικό Πανεπιστήμιο Αθηνών.
Τζαβελλάς Γ., Πανεπιστήμιο Πειραιώς.
Τουλούμη Γ., Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών.
Τριανταφύλλου Ι., Πανεπιστήμιο Πειραιώς.
Τσακλίδης Γ., Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης.
Φατούρος Σ., Πανεπιστήμιο Δυτικής Αττικής.
Φουσκάκης Δ., Εθνικό Μετσόβειο Πολυτεχνείο.
Χαλικιας Μ., Πανεπιστήμιο Δυτικής Αττικής.
Χαραλαμπίδης Χ., Πανεπιστήμιο Αθηνών.
Χατζηπαντελής Θ., Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης.
Χριστοφίδης Τ., Πανεπιστήμιο Κύπρου.
Ψαρράκος Γ., Πανεπιστήμιο Πειραιώς.

Διαλέξεις

Κεντρικοί Ομιλητές

Ch. Charalambides, Emeritus Professor, National and Kapodistrian University of Athens.
N. Balakrishnan, Distinguished Professor, McMaster University.
N. Papadatos, Professor, National and Kapodistrian University of Athens.
S. Bar-Lev, Professor, Holon Institute of Technology.
S. Datta, Professor, Department of Biostatistics, University of Florida.
S. Denaxas, Professor, University College London.

Προσκεκλημένοι Ομιλητές

A. Burnetas, Professor, National and Kapodistrian University of Athens.
A. Kumar, Symbiosis Institute of Technology.
C. Heuchenne, Professor, University of Liège.
C. Ley, Professor, University of Luxembourg.
C. Weiss, Professor, Helmut Schmidt University.
Ch. Charalambides, Emeritus Professor, National and Kapodistrian University of Athens.
Ch. Thomadakis, National and Kapodistrian University of Athens.
Ch. T. Nakas, Professor, University of Thessaly.
D. Jeske, Professor, University of California.
E. Polyzos, Vrije Universiteit Brussel.
E.-N. Kalligeris, Postdoctoral Researcher, University of Rouen.
F. F. Caballero, Assistant Professor, Autonomous University of Madrid.
G. Bakoyannis, Assistant Professor, Athens University of Economics and Business.
G. D' Amico, Professor, University "Gabriele d'Annunzio".
G. Touloumi, Professor, National and Kapodistrian University of Athens.
J. M. Poggi, Professor, Paris-Saclay University.
K. P. Tran, Senior Associate Professor, University of Lille.
M.-P Bourguignon, Servier Research Institut.
M. Jaenada, Predoctoral Researcher, Complutense University of Madrid.
M. Limnios, Postdoctoral Researcher, University of Copenhagen.

M. Markatou, Distinguished Professor, University at Buffalo.
M. Stehlik, Professor, Valparaiso University.
N. Eleftheroglou, Assistant Professor, Delft University of Technology.
N. Pantazis, Assistant Professor, National and Kapodistrian University of Athens.
N. Papadatos, Professor, National and Kapodistrian University of Athens.
N. Pavlidis, Senior Lecturer, Lancaster University.
O. Jones, Professor, Cardiff University.
P. C. Trandafir, Associate Professor, Public University of Navarre.
P. Miranda, Associate Professor, Complutense University of Madrid.
P. Otto, Academic Staff, Karlsruhe Institute of Technology.
S. B. Singh, Professor, G. B. Pant University of Agriculture and Technology.
S. Georgiadis, Statistician, Copenhagen Center for Arthritis Research.
S. Knoth, Professor, Helmut Schmidt University.
T. Perdikis, Postdoctoral Researcher, Athens University of Economics and Business.
V. Sypsa, Associate Professor, National and Kapodistrian University of Athens.
V.-S. Barbu, Associate Professor, University of Rouen.
V. Bisht, G. B. Pant University of Agriculture and Technology.

Πρόγραμμα του Συνεδρίου

ΠΕΜΠΤΗ 25 ΜΑΪΟΥ 2023/ THURSDAY MAY 25 2023

Αίθουσα Α/ Hall A: Κεντρικό αμφιθέατρο/ Central Amphitheater

13:30 - 14:30	Εγγραφή Συνέδρων/ Registration of the Participants
Αίθουσα Α / Hall A 14:30 – 15:00	**ΕΝΑΡΞΗ ΤΟΥ ΣΥΝΕΔΡΙΟΥ – ΧΑΙΡΕΤΙΣΜΟΙ CONFERENCE OPENING – WELCOME ADDRESSES Προεδρεύων/Chair: Μ. Χαλικιάς, Ε. Παπαγεωργίου/ M. Chalikias, E. Papageorgiou**
	- Χαιρετισμοί – Welcome Addresses
Αίθουσα Α/ Hall A 15:00 - 17:30	**ΕΝΑΡΚΤΗΡΙΕΣ ΚΕΝΤΡΙΚΕΣ ΟΜΙΛΙΕΣ / PLENARY TALKS Προεδρεύουσα/Chair : Μ. Βαμβακάρη/M. Vamvakari**
15:00-15:30	**N. Balakrishnan** Linear Prediction
15:30-16:00	**S. Bar-Lev** An infinite class of exponential dispersion models for count data - a survey of current research
16:00-16:30	**S. Datta** Regression Analysis of a Future State Entry Time Distribution Conditional on a Past State Occupation
16:30-17:00	**N. Papadatos** The characteristic function of the discrete Cauchy distribution and the Cauchy-Cacoullos family of discrete distributions
17:00-17:30	**Ch. Charalambides** A class of power series q-distributions
17:30-18:00	**ΔΙΑΛΕΙΜΜΑ – ΚΑΦΕΣ – COFFEE BREAK**

Αίθουσα Α/ Hall A 18:00-19:20	**ΕΙΔΙΚΗ ΣΥΝΕΔΡΙΑ / SPECIAL SESSION Βραβείο Καλύτερου Νέου Στατιστικού/Best Young Statistician Award Προεδρεύων/Chair: Γ. Ψαρράκος / G. Psarrakos**
18:00 – 18:20	**M. Jaenada, N. Balakrishnan and L. Pardo** Step-stress test experiments under interval censored data with lognormal lifetime distribution
18:20 – 18:40	**G. Tzoumerkas and D. Fouskakis** Objective shrinkage priors via imaginary data
18:40 – 19:00	**V. E. Papageorgiou and G. Tsaklidis** An extended epidemiological particle filter for assessing infectious disease transmission. Application to covid-19 data in Italy
19:00 – 19:20	**E. N. Kalligeris, A. Karagrigoriou and Ch. Parpoula** On Stochastic Dynamic Modeling of Incidence Data

ΠΑΡΑΣΚΕΥΗ 26 ΜΑΪΟΥ 2023/ FRIDAY MAY 26 2023

Αίθουσα Α/ Hall A: Κεντρικό αμφιθέατρο/ Central Amphitheater
Αίθουσα Β: / Hall B: Σ109Α

Αίθουσα Α/ Hall A	**«Ελένειο» Βραβείο Καλύτερης Διδακτορικής Διατριβής 2021-2022** **«Eleneio» Award for Best Doctoral Thesis 2021-2022** **Προεδρεύων/Chair: Δ. Φουσκάκης, Μ. Κατέρη / D. Fouskakis, M. Kateri**
08:45 – 09:15	**Ch. Thomadakis** Joint modeling of longitudinal and competing-risks data using cumulative incidence functions for the failure submodels and accounting for potential misclassified cause of failure through double sampling
Αίθουσα Α/ Hall A **09:20 - 11:00**	**ΠΡΟΣΚΕΚΛΗΜΕΝΕΣ ΟΜΙΛΙΕΣ / INVITED TALKS** **Stochastic Modeling and Statistical Methods: Advances and Applications Part I** **Προεδρεύων/Chair: Ε. Παπαγεωργίου / E. Papageorgiou**
09:20-09:45	**E. N. Kalligeris, V. S. Barbu, G. Hacques, L. Seifert and N. Vergne** Assessing the Dynamic of Visual-Motor Skill in Climbing via Drifting Markov Modeling
09:45-10:10	**P. M. Menendez, A. Felipe, M. Jaenada, L. Pardo** Applying Rényi's pseudodistance for robust Model Selection for independent not identically distributed observations
10:10-10:35	**D. Bagkavos, M. Guillen and J.P. Nielsen** Nonparametric conditional survival function estimation with plug-in bandwidth and robust model selection
10:35-11:00	**Ch. T. Nakas, L. E. Bantis, B. Brewer and B. Reiser** Statistical inference for ROC curves after the Box-Cox transformation with the use of the package `rocbc' in R
Αίθουσα Β/ Hall B **09:20 - 11:00**	**ΕΙΔΙΚΗ ΣΥΝΕΔΡΙΑ / SPECIAL SESSION** **Techniques & Applications of Data Analytics & Machine Learning I** **Προεδρεύων/Chair : Π. Οικονόμου / P. Economou**
09:20-09:40	**K. Skarlatos, P. E. Maravelakis, S. Bersimis** Techniques for detecting change points in multivariate data streams with an application in shipping
09:40-10:00	**E. Skamnia, P. Economou, S. Bersimis** Detecting outliers, break points and level shifts with an application in shipping industry
10:00-10:20	**G. Papageorgiou, P. Economou, S. Bersimis** Detecting hidden trends and patterns in shipping industry using text visualization
10:20-10:40	**G. Arvanitopoulos, P. Economou, S. Bersimis** Applications of text mining in health
10:40-11:00	**Z. Bartsioka, A. Charalabidis, P. E. Maravelakis and S. Bersimis** Monitoring learning outcomes using runs and scans
11:00-11:30	***ΔΙΑΛΕΙΜΜΑ – ΚΑΦΕΣ – COFFEE BREAK***

Αίθουσα Α/ Hall A **11:30 - 13:30**	**ΠΡΟΣΚΕΚΛΗΜΕΝΕΣ ΟΜΙΛΙΕΣ / INVITED TALKS** **Statistical Methods and Applications in Systems Assurance & Quality Part I** **Προεδρεύων/Chair : Π. Οικονόμου / P. Economou**
11:30-12:00	**S. Knoth** Some Stylized Facts of the Conditional Expected Delay (CED)
12:00-12:30	**C. Weiss** Stein EWMA Control Charts for Count Processes
12:30-13:00	**M. Stehlik** Data transformations and transfer functions
13:00-13:30	**N. Pavlidis** Dimensionality reduction and clustering
Αίθουσα Β/ Hall B **11:30 – 13:30**	**ΣΥΝΕΔΡΙΑ / CONTRIBUTED SESSION** **Actuarial and Financial Mathematics** **Προεδρεύων / Chair: Δ. Κωνσταντινίδης / D. Konstantinides**
11:30-11:50	**Ch. Skiadas, Y. Dimotikalis**

	Expanding the Life Tables to Include the Healthy Life Expectancy. The Case of Norway
11:50-12:10	**A. Papachristos, A. Bozikas** Actuarial Aspects of Subjective Survival Probabilities
12:10-12:30	**M.V. Boutsikas, D. J. Economides. and E. Vaggelatou** On the time and aggregate claim amount until the surplus drops below zero or reaches a safety level in a jump diffusion risk model
12:30-12:50	~~**D. Ioannides**~~ **Cancelled** ~~Estimation the Probability of Default Cascades in Financial Markets~~
12:50-13:10	**S. Tzaninis** Martingales, change of measures and ruin probabilities for an inhomogeneous renewal risk model
13:10-13:30	**Ming Cheng, Dimitrios G. Konstantinides, Dingcheng Wang** Multivariate regularly varying insurance and financial risks in multi-dimensional risk models

13:30- 14:30	***ΜΕΣΗΜΒΡΙΝΗ ΔΙΑΚΟΠΗ – MIDDAY BREAK***

Αίθουσα Α/ Hall A **14:30 - 15:30**	**ΠΡΟΣΚΕΚΛΗΜΕΝΕΣ ΟΜΙΛΙΕΣ / INVITED TALKS** **Εφαρμοσμένη Βιοστατιστική / Applied Biostatistics** **Προεδρεύων/Chair : Δ. Παναγιωτάκος/D. Panagiotakos**
14:30 – 15:00	**V. Sypsa** Modelling transmission of infectious diseases and assessing the impact of interventions: From HIV and antibiotic resistant bacteria to SARS-CoV-2
15:00 – 15:30	**F. F. Caballero** Elastic Net Regression models and their use to obtain a continuous measure of multimorbidity

Αίθουσα Α/ Hall A **15:40 – 17:00**	**ΕΙΔΙΚΗ ΣΥΝΕΔΡΙΑ / SPECIAL SESSION** **Asymptotic Statistics** **Προεδρεύων / Chair : Σ. Τρέβεζας / S. Trevezas**
15:40 – 16:00	**P. Karkalakis, P.-H. Cournède, G. Hermange, S. Trevezas** A Moment-type Estimator for the Generalized Gamma Model with Complete and Interval Censored Data –Application to First Division Time of HSC
16:00 – 16:20	**Ι. Οικονομίδης, Σ. Τρέβεζας** The Method of Moments vs Maximum Likelihood -An overestimated difference?
16:20 – 16:40	**Δ. Κορδαλής, Σ. Τρέβεζας** The Convolution Algebra of Multi-time Markov Renewal Chains and elements of Statistical Estimation
16:40 – 17:00	**S. Trevezas, G. Gavrilopoulos, I. Votsi** Nonparametric Maximum Likelihood Estimation for discrete time Denumerable Markov Chains

Αίθουσα Β/ Hall B **15:40 – 17:00**	**ΣΥΝΕΔΡΙΑ / CONTRIBUTED SESSION** **Χρονοσειρές – Προβλέψεις** **Προεδρεύων / Chair : Χ. Κίτσος / C. Kitsos**
15:40 – 16:00	**Γ. Κοσμά, Β. Καρυώτη** Ανάλυση και Πρόβλεψη Χρονοσειρών με δεδομένα εποχικής γρίπης
16:00 – 16:20	**Μ. Μάρκου, Β. Καρυώτη** Πρόβλεψη χρονοσειρών και μελλοντικών ακροτάτων σε αγχώδεις διαταραχές στον πληθυσμό της Ελλάδας.
16:20 – 16:40	**Σ. Μαλεφάκη, Π. Οικονόμου** Παρακολούθηση μακροπρόθεσμης σχέσης μεταξύ συνολοκληρωμένων χρονοσειρών
16:40 – 17:00	**Μ. Χαλικιάς** Βελτιστοποίηση Σχεδιασμών Επαναλαμβανόμενων μετρήσεων 2 αγωγών εκτίμηση άμεσων επιδράσεων (direct effect)

18:00 - 19:30	***ΙΣΤΟΡΙΚΟΣ ΠΕΡΙΠΑΤΟΣ ΣΤΗΝ ΑΘΗΝΑ – HISTORIC WALKING TOUR IN ATHENS***
20:30	***ΕΠΙΣΗΜΟ ΔΕΙΠΝΟ – CONFERENCE DINNER***

ΣΑΒΒΑΤΟ 27 ΜΑΪΟΥ 2023/ SATURDAY MAY 27 2023

Αίθουσα Α/ Hall A: Σ109Α
Αίθουσα Β/ Hall B: Σ109Β

Αίθουσα Α/ Hall A **09:00 - 11:00**	**ΣΥΝΕΔΡΙΑ / CONTRIBUTED SESSION** **Applied Probability – Applied Statistics** **Προεδρεύων/Chair : Α. Μπατσίδης / A. Batsidis**
09:00 – 9:20	**K. Bourazas, K. Fokianos, Ch. Panayiotou, M. Polycarpou** An Adaptive Kernel-based Multivariate CUSUM for Location Shifts
09:20-09:40	**K. Fountoukidis, D. L. Antzoulakos, A. C. Rakitzis** The Variable Sample Size and Sampling Interval Run Sum Max Chart
09:40-10:00	**T. Gkelsinis, V. S. Barbu** Recent advances in reliability indexes for semi-Markov repairable systems with applications to financial credit scoring
10:00-10:20	**K. Tasias, P. Mpistintzanos, F. Pekridis** A mission reliability redundancy model for an aircraft fleet with cold standby spare parts
10:20-10:40	**I. Triantafyllou** Combined m-consecutive-k-out-of-n: F & consecutive kc-out-of-n: F structures with cold standby redundancy
10:40-11:00	**D. Lyberopoulos, N. Macheras** Martingale characterizations of mixed compound Poisson processes with applications in risk theory
Αίθουσα Β/ Hall B **09:00-11:00**	**ΣΥΝΕΔΡΙΑ / CONTRIBUTED SESSION** **Εφαρμοσμένη Στατιστική – Εφαρμογές Στατιστικής** **Προεδρεύουσα : Θ. Μοσχονά/ Th. Moschona**
09:00 - 09:20	**Τ. Δάρας, Β. Παντούλα, Β. Μάνδικας** Τρόποι εφαρμογής της Γραμμικής Διαχωριστικής Ανάλυσης σε δεδομένα από το χώρο της Υγείας
09:20 – 09:40	**Σ. Δαφνής, Χ. Ε. Ζώτος, Α. Ραυτοπούλου, Γ. Κ. Παπαδόπουλος** Η Επίδραση Των Περιόδων Ανεργίας Στην Ατομική Υγεία
09:40 – 10:00	**Σ. Φόρτης, Σ. Ντελίκου, Α. Γιαννάκη, Α. Ξυδάκη, Κ. Μαγγανάς, Χ. Φούντζουλα, Ε. Παπαγεωργίου, Α. Κριεμπάρδης** Επικοινωνία κυττάρων του αίματος μέσω μικροκυστιδίων: Ο ρόλος τους στη δρεπανοκυτταρική αναιμία
10:00 – 10:20	**Π. Δρόσος, Ε. Παύλου, Σ. Φόρτης, Ε. Παπαγεωργίου, Κ. Σταμούλης, Σ. Βαλσάμη, Μ. Πολίτου, Α. Κριεμπάρδης** Αιμοστατικό δυναμικό αιμοπεταλίων αποθηκευμένων στο ψύχος
10:20 – 10:40	**Ε. Παύλου, Σ. Φόρτης, Μ. Σιουμάλα, Σ. Δασκαλάκη, Α. Γιαννιώτη, Β. Μπίρτσας, Ε. Παπαγεωργίου, Δ. Πετράς, Ε. Νομικού, Α. Κριεμπάρδης** Η επίδραση της αιμοκάθαρσης στην αιμόσταση
10:40 - 11:00	**Π. Κ. Ρεβέλου, Χ. Παππάς, Γ. Παπαδόπουλος, Π. Ταραντίλης** Αναγνώριση της βοτανικής προέλευσης του ελληνικού ελαιολάδου με υπέρυθρη φασματοσκοπία και εφαρμογή μοντέλων επιβλεπόμενης μάθησης
11:00-11:30	**ΔΙΑΛΕΙΜΜΑ – COFFEE BREAK**

Αίθουσα Α/ Hall A **11:30 – 13:30**	**ΠΡΟΣΚΕΚΛΗΜΕΝΗ ΣΥΝΕΔΡΙΑ / INVITED SESSION** **Statistical Methods and Applications in Systems Assurance & Quality Part II** **Προεδρεύων/Chair: A. Ρακιτζής/A. Rakitzis**
11:30-12:00	**C. Heuchenne** Anomaly detection for compositional data using support vector data description
12:00-12:30	**K. P. Tran** Secure and Robust Federated Learning with Explainable Artificial Intelligence for Healthcare Systems
12:30-13:00	**P. Otto** A Dynamic Spatiotemporal Stochastic Volatility Model with an Application to Environmental Risks
13:00-13:30	**T. Perdikis** Distribution-free Control Charts for Monitoring Dispersion in Finite Horizon Productions

Αίθουσα Β/ Hall B 11:30 – 13:30	ΣΥΝΕΔΡΙΑ / CONTRIBUTED SESSION Στατιστική - Statistics Προεδρεύων/Chair : Σ. Δαφνής / S. Dafnis
11:30-11:50	**V. Chasiotis , D. Karlis** Subdata selection for big data regression based on leverage scores
11.50-12:10	**G. Seitidis, S. Nikolopoulos, I. Ntzoufras, D. Mavridis** Inconsistency identification in network meta-analysis via stochastic search variable selection
12:10-12:30	**K. Kontouli, O. Koutsiouroumpa, Ch. Christogiannis, S. Nikolakopoulos, D. Mavridis** Single-Arm Trials In Evidence Synthesis
12:30-12:50	**F. Milienos, N. Balakrishnan, S. Pal** A multiple stage destructive cure rate model
12:50-13:10	**A. Anastasiou, J. Tsimikas** Optimal threshold selection on a two-cut point optimal ROC curve
13:10-13:30	**P. Papastamoulis** Estimating mixtures of multinomial logistic regressions

13:30 - 15:45	ΜΕΣΗΜΒΡΙΝΗ ΔΙΑΚΟΠΗ – MIDDAY BREAK

Αίθουσα Α/ Hall A 15:45 - 17:45	ΠΡΟΣΚΕΚΛΗΜΕΝΗ ΣΥΝΕΔΡΙΑ / INVITED SESSION Stochastic Modeling and Statistical Methods: Advances and Applications Part II Προεδρεύων / Chair: Α. Καραγρηγορίου/A. Karagrigoriou
15:45 – 16:15	**V. Bisht and S. B. Singh** A Study on Dynamic Reliability Measures of Multi-state k-out-of-n systems
16:15 – 16:45	**V. P. Koutras, S. Malefaki, P. Psomas and A. N. Platis** Modelling wind intensity effects on wind turbine maintenance policies
16:45 – 17:15	**V. S. Barbu** Nonparametric estimation of semi-Markov processes and applications to reliability
17:15 – 17:45	**N. Eleftheroglou** Adaptive Prognostics of Engineering Assets Utilizing Markov Models
15:45-17:45	POSTER SESSION
	T. Tsiampalis, D. Panagiotakos The AFFINITY method: A methodological Framework for Imputing missing spatial data at an aggregate level and guaranteeing personal data privacy; implementation in the context of the official spatial Greek census data
	N. Nikolaou, M. Valizadeh, S. Behzadi, J. Staab, M. Dallavalle, A. Peters, A. Schneider, H. Taubenböck, K. Wolf A machine learning framework for cardiovascular health prediction modeling the interplay between various environmental, neighborhood and socio-economic features: a German-wide application
	T. Moysiadis, D. Koparanis, K. Liapis, M. Ganopoulou, G. Vrachiolias, I. Katakis, Ch. Moyssiadis, I. Vizirianakis, L. Angelis, K. Fokianos, I. Kotsianidis A novel personalized stepwise dynamic predictive algorithm in Chronic Lymphocytic Leukemia
	M. Ganopoulou, D. Koparanis, K. Liapis, E. Lamprianidou, S. Papadakis, K. Fokianos, I. Kotsianidis, L. Angelis, T. Moysiadis Causal Structure assessment in Health-Related Quality of Life questionnaires
	K. Gourgoura, P. Rivadeneyra, E. Stanghellini, Ch. Caroni, F. Bartolucci, G. Pucci, R. Curcio, M. Cavallo, L. Sanesi, G. Morgana, S. Bartoli, R. Ferranti, M. B. Pasticci, G. Vaudo Modeling the Health Impact of COVID-19 using Mixed Interaction Models and Chain Graph Models
	Z. Kyrana, E. Pratsinakis, N. Papafilippou, A. Markos, G. Menexes Comparison of dimensionality reduction and clustering methods on the multidimensional dataset "Forest Cover Type" with mixed-type data
	Μ. Τριανταφύλλου Συγκριτική ανάλυση κόστους νοσηλείας χειρουργικού τομέα σε ένα δημόσιο νοσοκομείο την πενταετία 2016-2020: Η περίπτωση του Γενικού Νοσοκομείου Αθηνών «Ιπποκράτειο»
	Δ.-Δ. Βάρσου, Α. Τσουμάνης, Μ. Αρτεμίου, Ν. Χειμαριός, Α. Παπαδιαμάντης, Α. Αφαντίτης Isalos Analytics Platform: Ένα λογισμικό μηχανικής μάθησης για μη-προγραμματιστές

17:45-18:15	*ΔΙΑΛΕΙΜΜΑ – COFFEE BREAK*

Αίθουσα Α/ Hall A 18:15-20:15	ΠΡΟΣΚΕΚΛΗΜΕΝΗ ΣΥΝΕΔΡΙΑ / INVITED SESSION Stochastic Modeling and Statistical Methods: Advances and Applications Part III Προεδρεύων/Chair: Ι. Τριανταφύλλου / I. Triantafyllou
18:15 – 18:45	**E. Polyzos, L. Pyl** Stochastic modeling of the elastic properties of carbon-fiber-reinforced 3D printed filaments using polynomial chaos expansion
18:45 – 19:15	**A. Kumar, P. Kumar** Understanding a system's performance in the Presence of *k*-out-of-*n: F* and Standby redundancy: A Reliability approach through Markov process
19:15 – 19:45	**G. D'Amico, F. Gismondi, F. Petroni** On some generalization of the ROCOF for general multi-state systems with applications
19:45 – 20:15	**P. C. Trandafir, Ch. P. Kitsos** Applying LSI to Probability and Statistics
20:15 - 20:45	**M. Markatou** Likelihood Methods in Pharmacovigilance

Αίθουσα Β/ Hall B 18:15-20:35	ΣΥΝΕΔΡΙΑ / CONTRIBUTED SESSION Εφαρμοσμένη Στατιστική/ Applied Statistics Chair/Προεδρεύων: Δ. Μαυρίδης / D. Mavridis
18:15 – 18:35	**K. Georgiou, K. Charmanas, N. Mittas, L. Angelis** Thematic analysis and research trends on medical diagnosis patents using state-of-the-art topic modelling
18:35 – 18:55	**V. Georgakis, P. Xenos** Healthcare Risk Management: Machine learning in Cardiovascular intensive care unit (CICU)
18:55-19:15	**A. Nalpantidi, D. Kallis** Multinomial mixture model for spatial data with an application to demographics
19:15-19:35	**E. Panagiotopoulos, I. Oikonomides, S. Trevezas** A Comparison of Generalized Linear Mixed-Effects Models and Random Forest Models on Crop Progress Monitoring
19:35-19:55	**L. Champezou, D. Kallis** Spatiotemporal clustering with an application on COVID-19 deaths in the provinces of the Netherlands
19:55-20:15	**D. Kouloumpou, G. Anastassiou** Brownian Motion Approximation by Neural Networks
20:15-20:35	**M. Ravani, K. Georgiou, G. Liantas, I. Chatzigeorgiou, L. Angelis, G. K. Ntinas** Carbon footprint and human toxicity study of greenhouse products from aquaponic cultivation
20:35-20:55	**M. Ravani, K. Georgiou, G. Liantas, L. Angelis, G. K. Ntinas** A comparative study of carbon footprint of sheep milk production using archetypal analysis

ΚΥΡΙΑΚΗ 28 ΜΑΪΟΥ 2023/ SUNDAY MAY 28 2023

Αίθουσα Α/ Hall A: Σ109Α
Αίθουσα Β/ Hall B: Σ109Β

Αίθουσα Α/ Hall A **09:00 - 11:00**	**ΠΡΟΣΚΕΚΛΗΜΕΝΗ ΣΥΝΕΔΡΙΑ / INVITED SESSION** **Stochastic Modeling and Statistical Methods: Advances and Applications Part IV** **Προεδρεύουσα / Chair: Σ. Μαλεφάκη/ S. Malefaki**
09:00 – 09:30	**O. Jones, L. Hayes, J. Cable** Fitting a detailed stochastic population model using Approximate Bayesian Computation
09:30 – 10:00	**M. Limnios, N. R. Hansen** Nonparametric Modeling of Event Processes with Applications to Conditional Local Independence Testing
10:00 – 10:30	**S. Georgiadis, D. Di Giuseppe, A. Scherer, M. Lund Hetland, K. Pavelka, J. Vencovsky, Z. Rotar, K. Perdan Pirkmajer, G. T. Jones, B. Glintborg, A. G. Loft, B. Gudbjornsson, O. Palsson, B. Michelsen, E. Klami Kristianslund, H. Relas, J. Huhtakangas, A. Ciurea, M. J. Nissen, J. K. Wallman, A. Yazici, M. Birlik, L. Ørnbjerg** A registry-based simulation study evaluating interchangeability of patient reported outcomes in axial spondyloarthritis
10:30 – 11:00	**A. Burnetas** Recursive Computation of Equilibrium and Optimal Strategies in Unobservable Feed-Forward Queueing Networks with Delay-Sensitive Customers.
Αίθουσα Β/ Hall B **09:00 - 11:00**	**ΣΥΝΕΔΡΙΑ / CONTRIBUTED SESSION** **Εφαρμοσμένες Πιθανότητες – Εφαρμοσμένη Στατιστική** **Προεδρεύων / Chair: Φ. Μηλένιος/ F. Milenios**
09:00 – 09:20	**Α. Ρακιτζής, Ε. Μαμζερίδου** Άνω μονόπλευρα Διαγράμματα Ελέγχου με Εκτιμημένες Παραμέτρους για Μηδενοδιογκωμένες Διεργασίες
09:20 – 09:40	**Γ. Παπαγεωργίου, Μ. Χαλικιάς** Σύγκριση εκτιμητών για την πρόβλεψη εκλογικών αποτελεσμάτων
09:40 -10:00	**Χ. Ευαγγελάρας, Β. Τραπουζανλής** Σχεδιασμοί για definitive screening με ελάχιστη συσχέτιση μεταξύ των τετραγωνικών επιδράσεων
10:00 – 10:20	**Σ. Τζανίνης** Η πιθανότητα χρεοκοπίας σε ένα γενικευμένο ανανεωτικό μοντέλο κινδύνου
10:20 – 10:40	**Π. Βλιώρα, Γ. Ψαρράκος, A. Toomaj** Μια οικογένεια μέτρων μεταβλητότητας που βασίζεται στην αθροιστική υπολειπόμενη εντροπία και σε στρεβλές συναρτήσεις
10:40 - 11:00	**Σ. Δαφνής, Μ. Β. Κούτρας** Μελέτη των ιδιοτήτων μιας ευρείας οικογένειας συνεχών κατανομών

11:00 - 11:30	***ΔΙΑΛΕΙΜΜΑ – COFFEE BREAK***

Αίθουσα Α/ Hall A **11:30 - 12:00**	**ΚΕΝΤΡΙΚΗ ΟΜΙΛΙΑ / PLENARY TALK** **e-Health** **Προεδρεύουσα/Chair : Γ. Τουλούμη / G. Touloumi**
11:30 – 12:00	**S. Denaxas** Methods for using electronic health records to study human diseases

Αίθουσα Α/ Hall A **12:10-13:30**	**ΕΙΔΙΚΗ ΣΥΝΕΔΡΙΑ / SPECIAL SESSION** **Το Βήμα των Υποψηφίων Διδακτόρων** **Προεδρεύων /Chair: Α. Μπουρνέτας, Γ. Τουλούμη**
12:10 – 12:30	**Σ. Ζαφειράτου, M. Stafoggia, Ε. Σαμόλη, Α. Αναλυτής, Χ. Γιαννακόπουλος, Κ. Β. Βαρώτσος, A. Schneider, K. Κατσουγιάννη** Επιδράσεις υψηλών θερμοκρασιών στη θνησιμότητα στο νομό Αττικής, εφαρμόζοντας μη γραμμικό μοντέλο κατανεμημένης χρονικής υστέρησης στο πλαίσιο του προγράμματος EXHAUSTION

12:30 – 12:50	**Σ. Ρούσσος, Β. Σύψα** Μεθοδολογία εκτίμησης του μεγέθους δύσκολα προσεγγίσιμων πληθυσμών με εφαρμογή στον πληθυσμό των χρηστών ενέσιμων ναρκωτικών στην Αθήνα
12:50 – 13:10	**Β. Μπαραλού, Χ. Θωμαδάκης, Ν. Δεμίρης, Γ. Τουλούμη** Στατιστικές μέθοδοι ανίχνευσης σε πραγματικό χρόνο της έναρξης και της εξέλιξης HIV επιδημικών εκρήξεων σε άτομα που κάνουν χρήση ενδοφλέβιων ναρκωτικών
13:10 – 13:30	**Ν. Καλπουρτζή, Γ. Τουλούμη** Προβλέψεις αριθμού θανάτων από καρδιαγγειακά νοσήματα μέχρι το 2035 στην Ελλάδα, χρησιμοποιώντας μοντέλο ατομικών προσομοιώσεων σε περιπτώσεις με έλλειψη εθνικών δεδομένων κοορτής

Αίθουσα Β/ Hall B **12:10-13:50**	**ΕΙΔΙΚΗ ΣΥΝΕΔΡΙΑ / SPECIAL SESSION** **Techniques & Applications of Data Analytics & Machine Learning II** **Προεδρεύων / Chair : Π. Μπομποτάς / P. Bobotas**
12:10 – 12:30	**A. Karaminas, P. Economou, S. Bersimis** Improving healthcare services using machine learning techniques
12:30 – 12:50	**F. Bersimis, P. Bagos** Polygenic risk scores and applications in clinical practice
12:50 – 13:10	**E. Nika, T. Tsiampalis, D. Georgakellos** Health budgets using data analytical techniques
13:10 – 13:30	**Ch. Boudoulis, O. Antonopoulou, A. Fousteris, P. Economou and S. Bersimis** The use of data analytics in Insurance: an overview
13:30 - 13:50	**Ch. Spyropoulos, P. Economou, S. Bersimis** The value of data exploitation in sports industry: an overview

13:50 - 16:00	***ΜΕΣΗΜΒΡΙΝΗ ΔΙΑΚΟΠΗ – MIDDAY BREAK***

Αίθουσα Α/ Hall A **16:00 - 18:00**	**ΠΡΟΣΚΕΚΛΗΜΕΝΗ ΣΥΝΕΔΡΙΑ / INVITED SESSION** **Missing data in longitudinal health studies: Methods to analyze cohort studies and surveillance data.** **Προεδρεύουσα/Chair : Γ. Τουλούμη / G. Touloumi**
16:00-16:30	**N. Pantazis** Missing data and reporting delay in surveillance data
16:30-17:00	**G. Touloumi, Ch. Thomadakis, L. Meligkotsidou, N. Pantazis** Longitudinal and Time-to-Drop-out Joint Models under at Random Drop-out Mechanism
17:00-17:30	**Ch. Thomadakis, N. Pantazis and G. Touloumi** Issues with the expected information matrix of linear mixed models provided by popular statistical packages under MAR dropout
17:30-18:00	**G. Bakoyannis** Semiparametric regression for competing risks data with missing not at random cause of failure

Αίθουσα Β/ Hall B **16:00-18:00**	**ΣΥΝΕΔΡΙΑ / CONTRIBUTED SESSION** **Στοχαστικές και Στατιστική** **Προεδρεύων / Chair: Π. Παπασταμούλης / P. Papastamoulis**
16:00 – 16:20	**Γ. Ψαρράκος** Στοχαστική διάταξη και ταυτότητες συνδιακύμανσης
16:20 – 16:40	**Ρ. Λύκου, Γ. Βασιλειάδης, Γ. Τσακλίδης** Φίλτρο Κλειστού Κρυφού Ομογενούς Μαρκοβιανου Συστήματος Με Πεπερασμένες Χωρητικότητες Στις Καταστάσεις
16:40 – 17:00	**Α. Μπατσίδης, M. D. Jiménez Gamero και B. Milošević** Έλεγχος καλής προσαρμογής για τη γενικευμένη Poisson κατανομή
17:00 – 17:20	**Π. Οικονόμου** Ένας νέος αλγόριθμος συσταδοποίησης παρουσία επικαλυπτόμενων Γκαουσιανών υποπληθυσμών
17:20 - 17:40	**Π. Μπομποτάς, Σ. Κουρούκλης** Βελτιωμένη εκτίμηση παραμέτρων σε inverse Gaussian κατανομή
17:40 - 18:00	**Δ. Πανάρετος, Α. Σαχλάς, Γ. Τζαβελλάς, Μ. Βαμβακάρη, Δ. Παναγιωτάκος** Αξιολόγηση της ακρίβειας των παραγόντων στην διερευνητική παραγοντική ανάλυση: ορισμός προβλήματος

18:00 - 18:30	***ΔΙΑΛΕΙΜΜΑ - COFFEE BREAK***
Αίθουσα Α/ Hall A: Κεντρικό αμφιθέατρο/ Central Amphitheater	
Αίθουσα Α/ Hall A **18:30-20:30**	**ΠΡΟΣΚΕΚΛΗΜΕΝΗ ΣΥΝΕΔΡΙΑ / INVITED SESSION** **Statistical Methods and Applications in Systems Assurance & Quality Part III** **Προεδρεύων/Chair: Σ. Μπερσίμης / S. Bersimis**
18:30 – 19:00	**J. M. Poggi** Statistics and Machine Learning in industry: combining heterogeneous or multi-scale model outputs
19:00 – 19:30	**C. Ley** Sport Statistics – When Figures are more than Numbers
19:30 – 20:00	**M. Bourguignon** A parametric quantile beta regression for modeling case fatality rates of COVID-19
20:00 - 20:30	**D. Jeske** Design and Inference for a RCT when Treatment Observations Follow a Two-Component Mixture Model
Αίθουσα Α/ Hall A **20:30-21:00**	**ΛΗΞΗ ΣΥΝΕΔΡΙΟΥ – CONFERENCE CLOSING** **Προεδρεύων/Chair: Ε. Αγγελής/ E. Angelis**

Εργασίες στα Ελληνικά

Papers in Greek

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 35ου Πανελληνίου Συνεδρίου Στατιστικής 2023) σελ.25-34

ΜΕΛΕΤΗ ΤΩΝ ΙΔΙΟΤΗΤΩΝ ΜΙΑΣ ΕΥΡΕΙΑΣ ΟΙΚΟΓΕΝΕΙΑΣ ΣΥΝΕΧΩΝ ΚΑΤΑΝΟΜΩΝ

Μ. Β. Κούτρας, Σ. Δ. Δαφνής
Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης, Πανεπιστήμιο Πειραιώς, Πειραιάς, Ελλάδα
mkoutras@unipi.gr, dafnisspyros@gmail.com

ΠΕΡΙΛΗΨΗ

Στην παρούσα εργασία μελετάμε τις ιδιότητες μίας ευρείας οικογένειας συνεχών μονομεταβλητών κατανομών με στήριγμα το διάστημα $(0, \infty)$, η οποία προσφέρει σημαντική ευελιξία για την προσαρμογή δεδομένων. Η μελέτη περιλαμβάνει και μέτρα κινδύνου, τα οποία έχουν ιδιαίτερη σημασία στον τομέα του Αναλογισμού.

Λέξεις κλειδιά: Μετασχηματισμοί κατανομών, βαθμίδα αποτυχίας, προσαρμογή δεδομένων, μέτρα κινδύνου.

1. ΕΙΣΑΓΩΓΗ

Στη στατιστική βιβλιογραφία υπάρχει μεγάλο πλήθος εργασιών με κύριο περιεχόμενο την αναζήτηση νέων κατανομών που μπορούν να χρησιμοποιηθούν ως κατάλληλα παραμετρικά μοντέλα για εμπειρικά δεδομένα (βλ. π.χ. Lee et al., 2013).

Τις τελευταίες δεκαετίες το ενδιαφέρον προς την προαναφερόμενη κατεύθυνση είναι ακόμα πιο έντονο και αφορά κυρίως στην ανάπτυξη πιο ευέλικτων κατανομών, με ιδιαίτερη έμφαση να δίνεται στις κατανομές που παρουσιάζουν ασυμμετρία και βαριές ουρές, οι οποίες εφαρμόζονται σε περιοχές όπως τα οικονομικά, τα χρηματοοικονομικά και η αναλογιστική επιστήμη (βλ. π.χ. Ahmad et al., 2022).

Οι περισσότερες μεθοδολογίες που χρησιμοποιούνται μπορούν να χαρακτηρισθούν ως 'μέθοδοι συνδυασμού', καθώς συνδυάζουν υπάρχουσες κατανομές για να δημιουργήσουν νέες ή προσθέτουν παραμέτρους σε υπάρχουσες κατανομές. Σύμφωνα με τους Lee et al. (2013) οι μεθοδολογίες αυτές μπορούν να κατηγοριοποιηθούν στις εξής περιπτώσεις:

1. Τη δημιουργία ασύμμετρων κατανομών.
2. Τη δημιουργία νέων κατανομών με προσθήκη παραμέτρων στις υπάρχουσες κατανομές.

3. Την οικογένεια των Βήτα-παραγόμενων κατανομών ή γενικότερα, τη μέθοδο ολοκληρωτικού μετασχηματισμού πιθανότητας μίας τυχαίας μεταβλητής με στήριγμα το διάστημα $(0,1)$. Σύμφωνα με αυτή, ξεκινάμε με μία τυχαία μεταβλητή ορισμένη στο $(0,1)$ (π.χ. τη Βήτα) και θεωρούμε το ολοκλήρωμα της συνάρτησης πυκνότητάς της από το 0 ως το $F(x)$, όπου $F(x)$ είναι η αθροιστική συνάρτηση κατανομής (ασκ) μίας άλλης τυχαίας μεταβλητής.
4. Τη σύνθετη μέθοδο, η οποία συνδυάζει δύο περικομμένες κατανομές με επικαλυπτόμενα στηρίγματα για τη δημιουργία μιας νέας κατανομής.

Στην παρούσα εργασία προτείνουμε μία τεχνική, με την οποία έχοντας μία κατανομή (από τις κλασσικές ή και άλλη) και μία συνάρτηση g η οποία πληροί κάποιες απλές ιδιότητες, 'παράγουμε' μία νέα οικογένεια συνεχών μονομεταβλητών κατανομών με στήριγμα το διάστημα $(0,\infty)$. Στην Ενότητα 2, ορίζουμε τη νέα οικογένεια κατανομών. Στην Ενότητα 3, προσφέρουμε επιπλέον ευελιξία στην προαναφερόμενη οικογένεια κατανομών για την προσαρμογή δεδομένων, ορίζοντας μία μετασχηματισμένη οικογένεια. Στην Ενότητα 4, μελετάμε μέτρα κινδύνου τα οποία έχουν ιδιαίτερη σημασία στον τομέα του Αναλογισμού.

2. ΟΡΙΣΜΟΙ

Στις περισσότερες κλασσικές κατανομές, οι οποίες έχουν στήριγμα το θετικό ημιάξονα $(0,\infty)$, μπορεί εύκολα να διαπιστώσει κανείς ότι υπάρχει η δυνατότητα η ασκ $F(x)$ να μετασχηματισθεί κατάλληλα έτσι ώστε να καταλήξουμε σε μία γραμμική συνάρτηση του x. Αυτό σημαίνει ότι μπορεί να βρεθεί μία πραγματική συνάρτηση $g(x)$ τέτοια ώστε

$$g(F(x)) = cx + d,\ c > 0. \tag{1}$$

Τέτοιοι μετασχηματισμοί δίνονται στον Πίνακα 1.

Πίνακας 1: Μετασχηματισμοί της ασκ κλασσικών κατανομών στη μορφή $g(F(x)) = cx + d$

Κατανομή	$F(x)$	Μετασχηματισμός	$g(x)$	c	d
Εκθετική	$1 - e^{-\lambda x}$	$-\ln(1-F) = \lambda x$	$-\ln(1-x)$	λ	0
Weibull	$1 - e^{-\lambda x^{\alpha}}$	$(-\ln(1-F))^{\frac{1}{\alpha}} = \lambda^{\frac{1}{\alpha}} x$	$(-\ln(1-x))^{\frac{1}{\alpha}}$	$\lambda^{\frac{1}{\alpha}}$	0
Pareto	$1 - (\frac{\lambda}{x+\lambda})^{\alpha}$	$(\frac{1}{1-F})^{\frac{1}{\alpha}} = \frac{x}{\lambda} + 1$	$(\frac{1}{1-x})^{\frac{1}{\alpha}}$	$\frac{1}{\lambda}$	1
Gompertz	$1 - e^{-\alpha(e^{\lambda x}-1)}$	$\ln(\frac{-\ln(1-F)}{\alpha} + 1) = \lambda x$	$\ln(\frac{-\ln(1-x)}{\alpha} + 1)$	λ	0
Dagum	$\left(1 + (\frac{x}{\lambda})^{-\alpha}\right)^{-p}$	$(F^{-1/p} - 1)^{-\frac{1}{\alpha}} = \frac{x}{\lambda}$	$(x^{-1/p} - 1)^{-\frac{1}{\alpha}}$	$\frac{1}{\lambda}$	0
Εκθετική-λογαριθμική	$1 - \frac{\ln(1-(1-p)e^{-\lambda x})}{\ln p}$	$-\ln(\frac{1-p^{1-F}}{1-p}) = \lambda x$	$-\ln(\frac{1-p^{1-x}}{1-p})$	λ	0
Log-logistic	$\frac{x^{\alpha}}{\lambda^{\alpha}+x^{\alpha}}$	$(\frac{F}{1-F})^{\frac{1}{\alpha}} = \frac{x}{\lambda}$	$(\frac{x}{1-x})^{\frac{1}{\alpha}}$	$\frac{1}{\lambda}$	0

Στη συνέχεια, με τη βοήθεια της σχέσης (1), θα ορίσουμε μια νέα οικογένεια συνεχών μονομεταβλητών κατανομών με στήριγμα το διάστημα $(0,\infty)$ την οποία θα συμβολίζουμε με $D_g^+(c,d)$.

Ορισμός 1 Μία κατανομή ανήκει στην οικογένεια $D_g^+(c,d)$, $c > 0$, αν η ασκ $F(x)$ έχει στήριγμα το διάστημα $(0,\infty)$ και γράφεται στη μορφή

$$F(x) = g^{-1}(cx + d), \tag{2}$$

όπου για τη συνάρτηση $g : (0,1) \to \mathbb{R}$ ισχύουν οι εξής συνθήκες:

- Η g είναι γνησίως αύξουσα και παραγωγίσιμη.
- $lim_{x\to 1^-} g(x) = \infty$.
- $lim_{x\to 0^+} g(x) = d$ ή $lim_{x\to 0^+} g(x) = -\infty$.

Στη συνέχεια θα χρησιμοποιούμε τους συμβολισμούς $X \sim D_g^+(c,d)$, $F \in D_g^+(c,d)$ για να δηλώνουμε ότι μια τυχαία μεταβλητή ακολουθεί την κατανομή ${D_g}^+(c,d)$ ή ισοδύναμα η ασκ F ανήκει στην οικογένεια ${D_g}^+(c,d)$.

Βάσει των ιδιοτήτων της συνάρτησης g του Ορισμού 1, είναι άμεσο ότι η F είναι πράγματι ασκ.

Αφού έχουμε ορίσει την οικογένεια ${D_g}^+(c,d)$, μπορούμε στην συνέχεια να παρατηρήσουμε ότι η αντίστοιχη συνάρτηση πυκνότητας $f(x)$ μπορεί να εκφραστεί ως

$$f(x) = \frac{c}{g'(F(x))}. \tag{3}$$

Ο παραπάνω τύπος μπορεί να προκύψει παραγωγίζοντας την (2), οπότε θα πάρουμε

$$f(x) = F'(x) = (g^{-1})'(cx+d)(cx+d)' = c(g^{-1})'(cx+d)$$

και χρησιμοποιώντας στη συνέχεια τη γνωστή σχέση

$$(g^{-1})'(x) = \frac{1}{g'(g^{-1}(x))}.$$

Η βαθμίδα αποτυχίας $r(x)$ μιας τυχαίας μεταβλητής $X \sim D_g^+(c,d)$ μπορεί να εκφραστεί ως (βλ. π.χ. **Barlow and Proschan** (1965))

$$r(x) = \frac{f(x)}{1-F(x)},\ x > 0$$

και με χρήση της (3) προκύπτει

$$r(x) = \frac{c}{g'(F(x))(1-F(x))},$$

ή ισοδύναμα, θέτοντας $Q(x) = g'(x)(1-x)$, $0 < x < 1$,

$$r(x) = \frac{c}{Q(F(x))},\ x > 0.$$

3. Η ΜΕΤΑΣΧΗΜΑΤΙΣΜΕΝΗ D_g-ΟΙΚΟΓΕΝΕΙΑ ΚΑΤΑΝΟΜΩΝ

Έστω $F \in D_g^+(c,d)$ και μία γνησίως αύξουσα και παραγωγίσιμη συνάρτηση $g_0 : (0,1) \to \mathbb{R}$ τέτοια ώστε

$$lim_{x\to 0^+} g_0(x) = 0 \ \text{ και } \ lim_{x\to 1^-} g_0(x) = 1. \tag{4}$$

Αν ορίσουμε τη συνάρτηση $g_1 = g \circ g_0$ παρατηρούμε ότι

- η $g_1 : (0,1) \to \mathbb{R}$ είναι γνησίως αύξουσα και παραγωγίσιμη.
- $lim_{x\to 0^+} g_1(x) = lim_{x\to 0^+} g(g_0(x)) = lim_{y\to 0^+} g(y) = d.$
- $lim_{x\to 1^-} g_1(x) = lim_{x\to 1^-} g(g_0(x)) = lim_{y\to 1^-} g(y) = \infty.$

Συνεπώς, η συνάρτηση $g_1 : (0,1) \to \mathbb{R}$ μπορεί να χρησιμοποιηθεί ως γεννήτορας μίας νέας D_g-οικογένειας κατανομών. Η οικογένεια αυτή θα ονομάζεται μετασχηματισμένη D_g-οικογένεια και θα συμβολίζεται με $D^+_{g,g_0}(c,d)$.

Κάνοντας χρήση του τύπου (2) συμπεραίνουμε ότι η ασκ $F^*(x)$ της $D^+_{g,g_0}(c,d)$ θα έχει τη μορφή

$$\begin{aligned} F^*(x) &= {g_1}^{-1}(cx+d) = (g\circ g_0)^{-1}(cx+d) = ({g_0}^{-1}\circ g^{-1})(cx+d) \\ &= {g_0}^{-1}(g^{-1}(cx+d)) \end{aligned}$$

δηλ.

$$F^*(x) = {g_0}^{-1}(F(x)). \tag{5}$$

Σημειώνουμε ότι συνήθως, η συνάρτηση g_0 θα περιέχει και παραμέτρους, οπότε η νέα κατανομή F^* θα έχει πρόσθετες παραμέτρους (σε σχέση με την $F(x)$), κάτι το οποίο θα της προσδίδει αυξημένη ευελιξία για προσαρμογή σε πραγματικά δεδομένα.

Θα εξετάσουμε στη συνέχεια συγκεκριμένες επιλογές της συνάρτησης $g_0(x)$ οι οποίες οδηγούν σε οικογένειες κατανομών που ήδη έχουν εμφανισθεί στη διεθνή βιβλιογραφία.

Παράδειγμα 1 Έστω $g_0 : (0,1) \to \mathbb{R}$ με

$$g_0(x) = \frac{1}{1+\theta(\frac{1}{x}-1)},\ \theta > 0.$$

Μπορεί εύκολα να ελεγχθεί ότι η g_0 είναι γνησίως αύξουσα με

$$lim_{x\to 0^+} g_0(x) = 0 \ \text{ και } \ lim_{x\to 1^-} g_0(x) = 1.$$

Αν $F \in D^+_g(c,d)$, η μετασχηματισμένη κατανομή $D^+_{g,g_0}(c,d)$ μπορεί να προσδιορισθεί από τον τύπο (5,) αφού εύκολα διαπιστώνεται ότι

$${g_0}^{-1}(x) = \frac{1}{1+\frac{1}{\theta}(\frac{1}{x}-1)}.$$

Επομένως η ασκ της $D^+_{g,g_0}(c,d)$ θα έχει τη μορφή

$$F^*(x) = \frac{1}{1+\frac{1}{\theta}(\frac{1}{F(x)}-1)},$$

ή ισοδύναμα, χρησιμοποιώντας τη νέα παράμετρο $\sigma = \frac{1}{1-\theta} < 1$,

$$F^*(x) = 1 - \frac{\sigma(1-F(x))}{\sigma - F(x)}. \tag{6}$$

Η κατανομή αυτή έχει προταθεί πρόσφατα από τους *Ahmad et al.* (2022).

Παράδειγμα 2 Έστω $g_0 : (0,1) \to \mathbb{R}$ με

$$g_0(x) = \frac{ln((\theta-1)x+1)}{ln\theta},\ 0 < \theta \neq 1.$$

Μπορεί εύκολα να ελεγχθεί ότι η g_0 είναι γνησίως αύξουσα με

$$lim_{x\to 0^+} g_0(x) = 0 \ \text{ και } \ lim_{x\to 1^-} g_0(x) = 1.$$

Αν $F \in D_g^+(c,d)$, η μετασχηματισμένη κατανομή $D_{g,g_0}^+(c,d)$ μπορεί εύκολα να προσδιορισθεί από τον τύπο (5,) αφού η αντίστροφη συνάρτηση ${g_0}^{-1}$ δίνεται ως εξής

$${g_0}^{-1}(x) = \frac{\theta^x - 1}{\theta - 1}.$$

Επομένως η ασκ της $D_{g,g_0}^+(c,d)$ θα έχει τη μορφή

$$F^*(x) = \frac{\theta^{F(x)} - 1}{\theta - 1}.$$

Η κατανομή αυτή έχει προταθεί από τους ***Mahdavi and Kundu*** (2016) (βλ., επίσης, ***Khalil et al.*** (2021)).

Παράδειγμα 3 Έστω $g_0 : (0,1) \to \mathbb{R}$ με

$$g_0(x) = \Big(1 - \frac{\theta}{ln(1 - x^{\frac{1}{\alpha}})}\Big)^{-1},\ \theta > 0,\ \alpha > 0.$$

Μπορεί εύκολα να ελεγθεί ότι η g_0 είναι γνησίως αύξουσα με

$$lim_{x\to 0^+} g_0(x) = 0 \ \text{ και } \ lim_{x\to 1^-} g_0(x) = 1.$$

Αν $F \in D_g^+(c,d)$, η μετασχηματισμένη κατανομή $D_{g,g_0}^+(c,d)$ μπορεί εύκολα να προσδιορισθεί από τον τύπο (5) χρησιμοποιώντας ότι

$${g_0}^{-1}(x) = (1 - e^{-\theta\frac{x}{1-x}})^{\alpha}.$$

Επομένως η ασκ της $D_{g,g_0}^+(c,d)$ θα έχει τη μορφή

$$F^*(x) = \Big(1 - e^{-\theta\frac{F(x)}{1-F(x)}}\Big)^{\alpha}.$$

Η κατανομή αυτή έχει προταθεί πρόσφατα από τους ***Barati and Rashidi*** (2022).

Σημειώνουμε ότι οι ιδιότητες που απαιτούνται για τη συνάρτηση μετασχηματισμού g_0 (μονοτονία και συνθήκες (4)) ικανοποιούνται στην περίπτωση που ως g_0 επιλέξουμε

1. την ασκ μιας τυχαίας μεταβλητής με στήριγμα το διάστημα $(0,1)$,
2. την αντίστροφη συνάρτηση της ασκ μιας τυχαίας μεταβλητής με στήριγμα το διάστημα $(0,1)$.

Για παράδειγμα, θα μπορούσε ως $g_0(x)$ να επιλεγεί η ασκ (ή η αντίστροφη της ασκ) μιας κατανομής Βήτα με παραμέρους α, β, δηλαδή

$$g_0(x) = \frac{1}{B(\alpha,\beta)} \int_0^x x^{\alpha-1}(1-x)^{\beta-1}dx.$$

Ωστόσο αυτή η επιλογή εμπεριέχει δυσκολίες για τη μελέτη της οικογένειας $D^+_{g,g_0}(c,d)$, αφού δεν υπάρχει κλειστός τύπος για το ολοκλήρωμα που εμφανίζεται παραπάνω και η έκφραση (5) καθίσταται δύσχρηστη. Η τελευταία παρατήρηση δεν ισχύει αν επιλεγούν συγκεκριμένες ειδικές περιπτώσεις για τα α, β, ή άλλες κατανομές με στήριγμα το $(0,1)$, οι οποίες έχουν απλές εκφράσεις για την ασκ τους.

Σημειώνουμε επίσης ότι οι ιδιότητες της συνάρτησης μετασχηματισμού διατηρούνται αν χρησιμοποιήσουμε δύο τέτοιες συναρτήσεις και στη συνέχεια δημιουργήσουμε τη σύνθεσή τους.

Το επόμενο παράδειγμα βοηθά στην κατανόηση της χρησιμότητας των παραπάνω παρατηρήσεων.

Παράδειγμα 4 Έστω $g_{0,1} : (0,1) \to \mathbb{R}$ με

$$g_{0,1}(x) = 1 - (1 - x^{\frac{1}{\beta}})^{\frac{1}{2}}$$

και $g_{0,2} : (0,1) \to \mathbb{R}$ η συνάρτηση που χρησιμοποιήθηκε στο Παράδειγμα 2, δηλαδή

$$g_{0,2}(x) = \frac{ln((\theta-1)x+1)}{ln\theta}.$$

Τότε, ορίζοντας τη σύνθεση, $g_0 = g_{0,1} \circ g_{0,2}$, η μετασχηματισμένη κατανομή $D^+_{g,g_0}(c,d)$ θα έχει ασκ της μορφής

$$F^*(x) = \Big(1 - (1 - \frac{\theta^{F(x)} - 1}{\theta - 1})^2\Big)^{\beta},\ \beta > 0,\ 0 < \theta \neq 1.$$

Η κατανομή αυτή έχει προταθεί από τους ***Hozaien et al.*** (2020).

Στον Πίνακα 2 δίνεται ένας ενδεικτικός κατάλογος συναρτήσεων $g_0(x)$ οι οποίες μπορούν να χρησιμοποιηθούν για τη δημιουργία μετασχηματισμένων κατανομών $D^+_{g,g_0}(c,d)$, ξεκινώντας από κάποιο γεννήτορα g (ή ισοδύναμα από κάποια κατανομή $F \in D^+_g(c,d)$). Σημειώνεται ότι το ρόλο της g_0 εδώ μπορεί να τον διαδραματίσει είτε η συνάρτηση που δίνεται στην πρώτη στήλη, είτε η συνάρτηση που δίνεται στη δεύτερη στήλη, είτε η σύνθεση δύο ή περισσοτέρων επιλογών από την πρώτη ή τη δεύτερη στήλη του Πίνακα 2. Είναι φανερό ότι, κάνοντας χρήση αυτής της δυνατότητας, καθίσταται εφικτό να εισαχθούν πολλές νέες κατανομές της οικογένειας που προτείνουμε στην παρούσα εργασία.

Πίνακας 2: Ενδεικτικός κατάλογος συναρτήσεων $g_0(x)$ για τη δημιουργία μετασχηματισμένων κατανομών ${D^+}_{g,g_0}(c,d)$

$g_0(x)$	${g_0}^{-1}(x)$	Παράμετροι	Σχόλια
x^α	$x^{\frac{1}{\alpha}}$	$\alpha > 0$	Ειδική περίπτωση της κατανομής $B(\alpha,\beta)$ με $\beta = 1$
$1-(1-x)^\beta$	$1-(1-x)^{\frac{1}{\beta}}$	$\beta > 0$	Ειδική περίπτωση της κατανομής $B(\alpha,\beta)$ με $\alpha = 1$
$1-(1-x^\alpha)^\beta$	$(1-(1-x)^{\frac{1}{\beta}})^{\frac{1}{\alpha}}$	$\alpha,\beta > 0$	Κατανομή Kumaraswamy, βλ. π.χ. τον Jones (2009)
$(1+\theta(\frac{1}{x}-1))^{-1}$	$(1+\frac{1}{\theta}(\frac{1}{x}-1))^{-1}$	$\theta > 0$	Παράδειγμα 1
$\frac{ln((\theta-1)x+1)}{ln\theta}$	$\frac{\theta^x-1}{\theta-1}$	$0 < \theta \neq 1$	Παράδειγμα 2
$\frac{1}{1-\frac{\theta}{ln(1-x^{\frac{1}{\alpha}})}}$	$(1-e^{-\theta\frac{x}{1-x}})^\alpha$	$\theta > 0, \alpha > 0$	Παράδειγμα 3
$\frac{1}{1-lnx}$	$e^{1-\frac{1}{x}}$	-	-

4. ΜΕΤΡΑ ΚΙΝΔΥΝΟΥ

Στην ενότητα αυτή θα μελετήσουμε μέτρα κινδύνου τα οποία έχουν ιδιαίτερη σημασία στον τομέα του Αναλογισμού.

4.1 Αξία σε κίνδυνο

Στην αναλογιστική επιστήμη, το μέτρο της αξίας σε κίνδυνο (value at risk, quantile risk measure, quantile premium principle) είναι ένα ευρέως χρησιμοποιούμενο μέτρο χρηματοοικονομικού κινδύνου. Εκτιμά τη μέγιστη δυνατή απώλεια ενός χαρτοφυλακίου (με μία δοσμένη πιθανότητα, έστω q) σε μία συγκεκριμένη χρονική περίοδο, θεωρώντας ότι επικρατούν κανονικές συνθήκες στην αγορά. Για παράδειγμα, έστω ότι ένα χαρτοφυλάκιο έχει αξία σε κίνδυνο μίας ημέρας ίση με €2,000,000 σε επίπεδο εμπιστοσύνης 95% ($q = 95\%$). Αυτό σημαίνει πως (προϋποθέτοντας πως θα επικρατούν κανονικές συνθήκες για μια ημέρα) η τράπεζα αναμένει, με πιθανότητα 95%, η αξία του χαρτοφυλακίου να μην μειωθεί περισσότερο από €2,000,000 κατά τη διάρκεια μίας ημέρας, ή ισοδύναμα υπάρχει πιθανότητα 5% να μειωθεί η αξία του χαρτοφυλακίου περισσότερο από 2,000,000 κατά τη διάρκεια μιας ημέρας.

Αν η τυχαία μεταβλητή X περιγράφει τη μείωση στην αξία του χαρτοφυλακίου σε συγκεριμένη χρονική περίοδο και συμβολίσουμε με x_q την αξία σε κίνδυνο σε

επίπεδο εμπιστοσύνης q, θα ισχύει:

$$P(X > x_q) = 1 - q \Rightarrow F(x_q) = q,$$

δηλαδή η αξία σε κίνδυνο μιας τυχαίας μεταβλητής X είναι το q-οστό ποσοστιαίο σημείο της κατανομής της X.

Αν $X \sim D^+_{g,g_0}(c,d)$ και συμβολίσουμε με VaR^*_q την αξία σε κίνδυνο για την X σε επίπεδο εμπιστοσύνης q, θα έχουμε

$$P(X > VaR^*_q) = 1 - q,$$

ή ισοδύναμα

$$F^*(VaR^*_q) = q$$

απ΄ όπου, με βάση την (5), προκύπτει ότι θα πρέπει να ισχύει

$$g_0(q) = F(VaR^*_q).$$

Επομένως η αξία σε κίνδυνο για την X μπορεί να βρεθεί από τον τύπο

$$VaR^*_q = F^{-1}(g_0(q)), \tag{7}$$

ή ισοδύναμα, είναι η αξία σε κίνδυνο για την αρχική κατανομή $F \in D^+_g(c,d)$ που αντιστοιχεί σε επίπεδο εμπιστοσύνης $g_0(q)$.

Παράδειγμα 1 (συνέχεια) Αν η X ακολουθεί την κατανομή που προτάθηκε από τους Ahmad et al. (2022), με ασκ την (6), η αξία σε κίνδυνο VaR^*_q θα δίνεται από τον τύπο

$$VaR^*_q = F^{-1}(t),$$

όπου $t = \frac{q\sigma}{q+\sigma-1}$. Το αποτέλεσμα αυτό συμπίπτει με τον τύπο (20) των Ahmad et al. (2022).

4.2 Αναμενόμενη ζημία αξίας σε κίνδυνο

Εκτός από την αξία σε κίνδυνο, σημαντικό επίσης μέτρο χρηματοοικονομικού κινδύνου είναι η αναμενόμενη ζημία αξίας σε κίνδυνο (Tail Value at Risk, $TVaR$). Παρά το γεγονός ότι η αξία σε κίνδυνο αποτελεί το δημοφιλέστερο μέτρο κινδύνου, έχει το σημαντικό μειονέκτημα ότι δεν δίνει κάποια πληροφορία για το πόσο μεγάλη μπορεί να γίνει η ζημία στο χειρότερο σενάριο. Για το λόγο αυτό χρησιμοποιείται εκτεταμένα και η αναμενόμενη ζημία αξίας σε κίνδυνο ή αναμενόμενο έλλειμμα (expected shortfall).

Έστω $F \in D^+_g(c,d)$ και $F^* \in D^+_{g,g_0}(c,d)$ η αντίστοιχη μετασχηματισμένη κατανομή. Θα δώσουμε στη συνέχεια ορισμένες εκφράσεις για το μέτρο $TVaR_q$ της F και $TVaR^*_q$ της F^*, μέσω του γεννήτορα g και της συνάρτησης μετασχηματισμού g_0.

Σύμφωνα με τον ορισμό του $TVaR_q$ (βλ. π.χ. Dowd (2005)) θα έχουμε

$$TVaR_q = \frac{1}{1-q}\int_{VaR_q}^{\infty} xf(x)dx$$

και κάνοντας χρήση της (3) καταλήγουμε στην έκφραση

$$TVaR_q = \frac{c}{1-q}\int_{VaR_q}^{\infty} \frac{x}{g'(F(x))}dx.$$

Εκτελώντας το μετασχηματισμό $Y = F(x)$ στο ολοκλήρωμα, προκύπτει

$$TVaR_q = \frac{1}{1-q}\int_{F(VaR_q)}^{1} F^{-1}(y)dy$$

ή ισοδύναμα, λόγω της ισότητας $F(x) = g^{-1}(cx+d)$ η οποία οδηγεί στον τύπο $F^{-1}(x) = (g(x)-d)/c$,

$$TVaR_q = \frac{1}{c(1-q)}\int_{g^{-1}(cVaR_q+d)}^{1} (g(x)-d)dx. \tag{8}$$

Σημειώνουμε ότι στις περισσότερες εφαρμογές έχουμε $d = 0$ οπότε ο τύπος (8) παίρνει την απλούστερη μορφή

$$TVaR_q = \frac{1}{c(1-q)}\int_{g^{-1}(cVaR_q)}^{1} g(x)dx.$$

Για την ποσότητα $TVaR_q^*$ θα έχουμε, κατ΄ αναλογία, μία έκφραση όμοια με την (8), όπου στη θέση της συνάρτησης g θα πρέπει να χρησιμοποιηθεί ο γεννήτορας $g_1 = g \circ g_0$ της $F^* \in D^+_{g,g_0}(c,d)$, πιο συγκεκριμένα

$$TVaR_q^* = \frac{1}{c(1-q)}\int_{g_0^{-1}(g^{-1}(cVaR_q^*+d))}^{1} (g(g_0(x))-d)dx.$$

Κάνοντας τέλος χρήση του τύπου (7), ο οποίος δίνει $F(VaR_q^*) = g_0(q)$ και λαμβάνοντας υπόψη ότι $F(x) = g^{-1}(cx+d)$ καταλήγουμε στην έκφραση

$$TVaR_q^* = \frac{1}{c(1-q)}\int_{q}^{1} (g(g_0(x))-d)dx.$$

ΕΥΧΑΡΙΣΤΗΡΙΟ: Η παρούσα εργασία έχει χρηματοδοτηθεί από το Πρόγραμμα Μεταπτυχιακών Σπουδών στην «Αναλογιστική Επιστήμη και Διαχείριση Κινδύνων» του Πανεπιστημίου Πειραιώς μέσω του έργου C908: «Ανάπτυξη Νέων Μεθοδολογιών στις Ασφαλίσεις και στην Διαχείριση Κινδύνων».

ABSTRACT

In the present work we study the properties of a wide family of continuous univariate distributions with support $(0, \infty)$, which offers flexibility for fitting to real data. Our study also includes risk measures, which are important in the field of Actuarial Science.

ΑΝΑΦΟΡΕΣ

Ahmad, Z., Mahmoudi E. and Dey S. (2022). A new family of heavy tailed distributions with an application to the heavy tailed insurance loss data, *Commun. Stat. Simul.*, **51:8**, 4372-4395.

Barati, R. and Rashidi, S. F. (2022). On the Coronavirus disease death rate modeling utilizing generalized exponential Kumaraswamy, *New Mathematics and Natural Computation*, p. 20 (to appear).

Barlow R. E. and Proschan F. (1965). *Mathematical Theory of Reliability*, Wiley, New York.

Dowd, K. (2005). *Measuring Market Risk*, Wiley, New York.

Hozaien, H. E., Al-Dayian, G. R., Yehia, E. G. and El-Helbawy, A. A. (2020). Topp-Leone alpha power family of distributions, *Egyptian Journal of Commerce Studies*, 10.21608/ALAT.2022.260618.

Jones, M. C. (2009). Kumaraswamy's distribution: A beta-type distribution with some tractability advantages, *Stat. Methodol.*, **6:1**, 70-81.

Khalil, A., Ahmadini, A. A. H., Ali, M., Mashwani, W. K., Alshqaq, S. S. and Salleh, Z. (2021). A novel method for developing efficient probability distributions with applications to engineering and life science data, *Journal of Mathematics*, 1-13, https://doi.org/10.1155/2021/447927.

Lee, C., Famoye, F. and Alzaatreh, A. Y (2013). Methods for generating families of univariate continuous distributions in the recent decades, *Comput. Stat.*, **5**, 219-238, https://doi.org/10.1002/wics.1255.

Mahdavi, A. and Kundu D. (2016). A new method for generating distributions with an application to exponential distribution, *Commun. Stat. - Theory Methods*, **46:13**, 6543-6557.

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 35ου Πανελληνίου Συνεδρίου Στατιστικής (2023), σελ.35-45

ΠΡΟΒΛΕΨΗ ΧΡΟΝΟΣΕΙΡΩΝ ΚΑΙ ΜΕΛΛΟΝΤΙΚΩΝ ΑΚΡΟΤΑΤΩΝ ΣΕ ΑΓΧΩΔΕΙΣ ΔΙΑΤΑΡΑΧΕΣ ΣΤΟΝ ΠΛΗΘΥΣΜΟ ΤΗΣ ΕΛΛΑΔΑΣ

Μ.Μάρκου[1], Β.Καρυώτη[2]
[1,2]Τμήμα Διοίκησης Τουρισμού, Πανεπιστήμιο Πατρών
m_markou@upatras.gr, vaskar@upatras.gr

ΠΕΡΙΛΗΨΗ

Οι κρίσεις άγχους και η αύξηση περιστατικών κατάθλιψης και άλλων διαταραχών της ψυχικής υγείας, αποτελεί πλέον σοβαρό ζήτημα παγκοσμίως, το οποίο καθιστά αναγκαία τη μελέτη και την ανάλυσή του. Ως πιθανές αιτίες διερευνώνται, τόσο η ανεξέλεγκτα επιταχυνόμενη ανάπτυξη της τεχνολογίας, συναρτήσει των απαιτήσεων που επιφέρει στην καθημερινή μας ζωή, όσο και η νέα τάξη πραγμάτων ως αποτέλεσμα της πανδημίας Covid-19. Στην παρούσα εργασία, μελετήθηκαν δύο διαφορετικές χρονοσειρές: τα ποσοστά των ατόμων που πάσχουν από άγχος & κατάθλιψη στην χώρα μας, από το 1990-2017.Η μελέτη των χρονοσειρών αυτών, επικεντρώθηκε στην πρόβλεψη, αλλά διαφοροποιήθηκε ως προς το αντικείμενο της πρόβλεψης. Στόχος των παραγόμενων προβλέψεων, ήταν ο χρονικός εντοπισμός του μελλοντικού σημείου, στο οποίο η χρονοσειρά αναμένεται να βελτιστοποιηθεί τοπικά.Για τη μοντελοποίηση των δεδομένων χρησιμοποιήθηκαν η γνωστή Box-Jenkins μεθοδολογία για την ανάλυση χρονοσειρών και ο αλγόριθμος LipschitzINdicator για την πρόβλεψη του μελλοντικού τοπικού βέλτιστου. Συγκεκριμένα,εντοπίστηκαν οι χρονικές στιγμές που θα εμφανιστούν τα τοπικά βέλτιστα των ποσοστών των ατόμων που πάσχουν από άγχος και κατάθλιψη στην Ελλάδα.

Λέξεις Κλειδιά: Χρονοσειρές, Πρόβλεψη, Ακρότατα, Βελτιστοποίηση, Μοντέλα ARIMA, LipschitzINdicator

1. ΕΙΣΑΓΩΓΗ

Η κατάθλιψη είναι μια διαταραχή ψυχικής υγείας που επηρεάζει περίπου το 5% των ενηλίκων παγκοσμίως, ποσοστό που αντιστοιχεί σε 350 εκατομμύρια ανθρώπους παγκοσμίως. Σύμφωνα με τον Παγκόσμιο Οργανισμό Υγείας (ΠΟΥ), η κατάθλιψη αποτελεί την κύρια αιτία αναπηρίας στον κόσμο, με περισσότερους από 264 εκατομμύρια ανθρώπους να πάσχουν με μέτρια ως και σοβαρή κατάθλιψη (WHO, 2023).

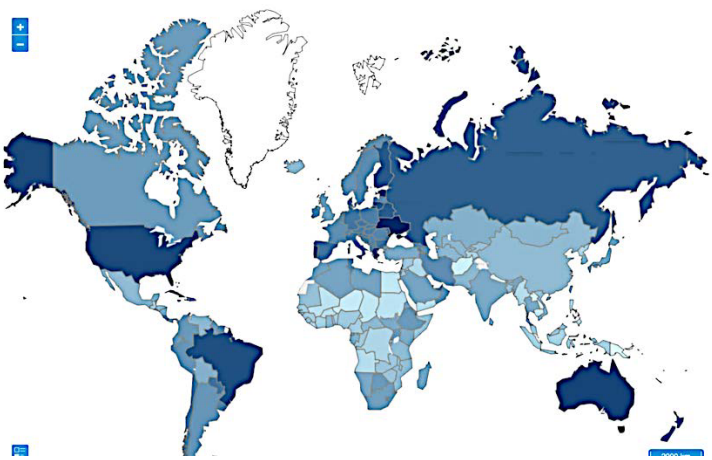
Συχνά, η κατάθλιψη συνδέεται με αγχώδεις διαταραχές. Το άγχος αποτελεί ένα φυσιολογικό μέρος της ζωής, ως αποτέλεσμα της ανησυχίας για την υγεία, την

οικονομική ευμάρεια, τα οικογενειακά προβλήματα, την προσωπική εξέλιξη αλλά και την επιβίωση .Όταν όμως το άγχος, η ανησυχία και το αίσθημα του φόβου είναι έντονα στην καθημερινότητα και δεν υποχωρούν, με αποτέλεσμα να επιδεινώνονται, τότε γίνεται λόγος για αγχώδεις διαταραχές (National Institute of Mental Health).

Οι καταθλιπτικές διαταραχές επηρεάζουν τον τρόπο με τον οποίο ενεργούν, αισθάνονται και σκέφτονται οι άνθρωποι, οι οποίοι πάσχουν από αυτές. Συμπτώματα τα οποία εκδηλώνουν είναι, η θλίψη, η κόπωση, η απελπισία, η αϋπνία ή ο υπερβολικός ύπνος, η ταραχή, το άγχος, η αδυναμία συγκέντρωσης και κοινωνικοποίησης, προβλήματα στην εργασιακή απόδοση και στην οικογενειακή συνεισφορά. Σε σοβαρές περιπτώσεις εκδηλώνονται σκέψεις θανάτου ή αυτοκτονίας (Beaglehole, Mulder, Frampton, Boden, Newton-Howes, &Bell, 2018).

Έρευνες έχουν δείξει ότι χώρες χαμηλού εισοδήματος τείνουν να εμφανίζουν υψηλότερα ποσοστά κατάθλιψης σε σχέση με εκείνες υψηλού εισοδήματος. Επίσης, είναι ανησυχητική η αύξηση των καταθλιπτικών διαταραχών στους εφήβους την τελευταία δεκαετία. Η Ελλάδα ανήκει στις χώρες με τα υψηλότερα ποσοστά κατάθλιψης παγκοσμίως. Η πτωτική πορεία της οικονομίας, η αύξηση των επιπέδων φτώχειας και των ποσοστών ανεργίας σε συνάρτηση με το αυξημένο κόστος ζωής και τη στασιμότητα των μισθών αποτελούν βασικό παράγοντα για την εκδήλωση συμπτωμάτων άγχους και κατάθλιψης σε υψηλότερα επίπεδα από άλλες χώρες τις Ευρώπης. Όσον αφορά τους νέους, η ανασφάλεια επαγγελματικής εξέλιξης και προσωπικής ανεξαρτητοποίησης από τον οικογενειακό ιστό λόγω της οικονομικής αρχικά κατάστασης στην χώρα και της εμφάνισης του Covid-19, που επιβράδυνε και άλλαξε τα δεδομένα ζωής, τα ποσοστά αυξήθηκαν κυρίως από την ηλικία των 17 ετών. Στον παρακάτω χάρτη, σκιαγραφημένες με έντονο χρώμα,παρουσιάζονται οι χώρες με τα υψηλότερα ποσοστά των ατόμων που πάσχουν από κατάθλιψη. (Wisevoter & Institute for Health Metrics and Evaluation (IHME), 2023).

Εικόνα 1. *Παγκόσμιος Χάρτης με τα ποσοστά κατάθλιψης ανά χώρα*



Για την σύνταξη μιας λειτουργικής κοινωνίας και για τη βελτίωση της ψυχικής υγείας είναι απαραίτητο να μελετηθούν και να επιβεβαιωθούν οι παράγοντες που είναι κύριοι υπεύθυνοι για την εκδήλωση ψυχικών και αγχωδών διαταραχών, όπως επίσης και η αξιολόγηση της νέας τάξης πραγμάτων παγκοσμίως και στη χώρα μας μετά την πανδημία.

Για την κατανόηση και την αντιμετώπιση των αγχωδών διαταραχών ένα ισχυρό εργαλείο αποτελεί η προβλεπτική ικανότητα των χρονοσειρών. Μέσω της πρόβλεψης των μελλοντικών ακρότατων των χρονοσειρών διαφαίνεται μια δυναμική προσέγγιση για τον προσδιορισμό προληπτικών μέτρων και την προσαρμογή των υπηρεσιών ψυχοκοινωνικής υγείας. Αυτή η προσέγγιση, επιτρέπει την εστίαση σε ευάλωτες ομάδες του πληθυσμού και τη βελτίωση της κοινωνικής ευαισθητοποίησης για τα θέματα της ψυχικής υγείας. Η πρόβλεψη, συνδυασμένη με μια ολοκληρωμένη κατανόηση της εξέλιξης των αγχωδών διαταραχών, μπορεί να αποτελέσει το θεμέλιο λίθο για προγράμματα παρέμβασης, που ανταποκρίνονται αποτελεσματικά στις ανάγκες της κοινότητας προωθώντας την ψυχοκοινωνική ευημερία.

Στόχος της μελέτης μας, αποτελεί ο εντοπισμός του χρόνου παρουσίασης του τοπικού βέλτιστου της εξεταζόμενης χρονοσειράς. Δηλαδή, μας ενδιαφέρει η γνώση του χρονικού σημείου, κατά το οποίο οι ελάχιστες και μέγιστες τιμές της χρονοσειράς θα προκύψουν. Το μαθηματικό ισοδύναμο πρόβλημα είναι η εύρεση της χρονικής στιγμής που θα εμφανιστεί ένα τοπικό ακρότατο (μέγιστο ή ελάχιστο). Από τη βιβλιογραφία προκύπτει, ότι η πρόβλεψη του μελλοντικού χρόνου κατά τον οποίο αναμένεται να βελτιστοποιηθεί τοπικά μια χρονοσειρά δεν έχει ερευνηθεί εκτενέστερα.

Στην παρούσα εργασία, επιλέχθηκαν κατάλληλα μοντέλα ARIMA και εφαρμόστηκε ο αλγόριθμοςLipschitz.

Η εργασία δομείται ως εξής: στην ενότητα 2, παρουσιάζεται η περιγραφή των δεδομένων και οι μέθοδοι που χρησιμοποιήθηκαν, ενώ στην ενότητα 3 προχωρούμε στη μελέτη των δεδομένων. Έχοντας επιλέξει, τα κατάλληλα μοντέλα, στην ενότητα 4 προσδιορίζεται το μελλοντικό βέλτιστο και παρατίθενται τα συμπεράσματα της μελέτης. Τέλος, η εργασία ολοκληρώνεται με τις μελλοντικές προτάσεις για περαιτέρω έρευνα.

2. ΔΕΔΟΜΕΝΑ ΚΑΙ ΜΕΘΟΔΟΙ

2.1 Δεδομένα Ανάλυσης

Για την ανάλυση που ακολουθεί, χρησιμοποιήθηκαν δεδομένα από την ηλεκτρονική βάση Kaggle. Το Kaggle είναι μια κοινότητα δεδομένων που απευθύνεται σε επιστήμονες με σκοπό την ανάπτυξη μοντέλων και μεθοδολογιών για την επίλυση προβλημάτων που αφορούν μεγάλα δεδομένα και την ανάλυση τους. Για τους σκοπούς της παρούσας εργασίας, επιλέχθηκε δείγμα από τη βάση δεδομένων που περιελάβανε τα ποσοστά νοσηρότητας σε ψυχικές διαταραχές παγκοσμίως, ανά χώρα

και περίπτωση διαταραχής. Ως δείγμα επιλέχθηκαν τα ποσοστά περιστατικών κατάθλιψης και άγχους στην Ελλάδα, από το 1990-2017.Δεδομένα μετά το 2017 δεν είναι δημοσιευμένα.

2.1 Τεχνικές Εντοπισμού του χρονικού σημείου βελτιστοποίησης μιας χρονοσειράς

Σύμφωνα με τη Λισγαρά (2011), το φαινόμενο που αναπαρίσταται από μια χρονοσειρά δύναται να αντιμετωπιστεί ως μια μαθηματική συνάρτηση με ρ περιορισμούς που επηρεάζουν τη συμπεριφορά της.

Έστω, $F: \mathbb{R}^{\rho} \to \mathbb{R}$ είναι μια άγνωστη, αλλά υπαρκτή συνάρτηση, που περιγράφει το εξεταζόμενο φαινόμενο ως $Z = F(x_1, x_2, \dots, x_{\rho})$, όπου $x_1, x_2, \dots, x_{\rho}$ οι τιμές που επηρεάζουν το φαινόμενο. Επιπλέον, παρατηρείται ότι η τιμή κάθε περιορισμού εξαρτάται από το χρόνο.

Επομένως, η $Y = Y\{t: t \in T\}$ είναι ισότιμη της παραπάνω σχέσης, καθώς και οι τιμές τους. Επίσης, $F(x_{1,t}, x_{2,t}, \dots, x_{\rho,t}) = Y_t, \forall t \in T$, όπου $x_{i,t}$ υποδηλώνει τις τιμές των x_i στον χρόνο T. Έτσι, η F μπορεί να υπολογιστεί για κάθε πιθανή τιμή της μεταβλητής x_i. Καθώς το φαινόμενο εξελίσσεται οι ρ παράμετροι ορίζονται αριθμητικά, άρα η χρονοσειράY_t ταυτίζεται με τη γραφική αναπαράσταση των διακριτών σημείων που απαρτίζουν τη συνάρτηση των $x_{1,t}, x_{2,t}, \dots, x_{\rho,t}$. Άρα κάθε χρονοσειρά μπορεί να θεωρηθεί ως συνάρτηση υποκείμενη σε πολυάριθμους περιορισμούς, ανάλογα με τη φύση της εξεταζόμενης χρονοσειράς, και να προσεγγιστεί μέσω της θεωρίας της ανάλυσης.

Η Μέθοδος Οπισθοδρόμησης αποτελεί το δεύτερο σκέλος της μελέτης μας. Βασικός σκοπός της μεθόδου, είναι να προβλεφθεί ο μελλοντικός χρόνος που θα προκύψει η τοπική βέλτιστη τιμή μιας χρονοσειράς, μέσω της χρήσης παρελθοντικών τιμών.

Ανάλογα με τη φύση του κάθε προβλήματος, τα στάδια της διαδικασίας πρόβλεψης του μελλοντικού χρόνου όπου θα προκύψει η τοπική βέλτιστη τιμή της χρονοσειράς διαφοροποιούνται.

Τα βασικότερα βήματα της διαδικασίας είναι τα ακόλουθα (Λισγαρά, 2011):

1. *Διαδικασία Εύρεσης Παρελθοντικού Τοπικού Ελάχιστου*

   Σε αυτήν τη φάση ελέγχονται όλα τα παρελθοντικά σημεία της υπό εξέτασης χρονοσειράς, ώστε να βρεθούν εκείνα τα σημεία με τιμή μικρότερη από το τελευταίο γνωστό σημείο.

2. *Διαδικασία Προσέγγισης Μελλοντικού Βήματος*

Το μέγεθος του μελλοντικού βήματος που θα εφαρμοστεί στο τελευταίο γνωστό σημείο της χρονοσειράς προσεγγίζεται από την επεξεργασία των σημείων της ακολουθίας. Γι' αυτό, προσεγγίζεται η τιμή της σταθεράς Lipschitz.

Η σταθερά Lipschitz χρησιμοποιείται ως μέτρο για την παρακολούθηση της μέγιστης αλλαγής της συνάρτησης. Έτσι, η σταθερά Lipschitz μέσω της βελτιστοποίησης μπορεί να βοηθήσει στην πιο γρήγορη εύρεση του βέλτιστου.

Μέσω προσεγγίσεων, εκτιμάται το βέλτιστο βήμα k, το οποίο θα εφαρμοστεί στο αρχικό σημείο της εξεταζόμενης χρονοσειράς, με σκοπό την εύρεση του μελλοντικού χρονικού σημείου στο οποίο προβλέπεται το τοπικό βέλτιστο (Στρατογιάννη, 2017).

3. *Διαδικασία Εύρεσης Μελλοντικού Τοπικού Μέγιστου*

   Σε αυτό το βήμα, με τη βοήθεια της τεχνικής βελτιστοποίησης χρησιμοποιείται το μέγεθος μελλοντικού βήματος που υπολογίστηκε στο προηγούμενο βήμα, το οποίο πλέον έχει γνωστή κατεύθυνση.

   Η ακολουθία τιμών που προέκυψε από το πρώτο στάδιο ξεκινά από το παρελθοντικό ελάχιστο σημείο, βρίσκει και προσπερνά το τελευταίο γνωστό σημείο, και με τη χρήση του κατάλληλου βήματος θα καταλήξει στο μελλοντικό μέγιστο.

   Παρατηρούμε ότι βασικός στόχος της αναζήτησης της χρονικής στιγμής κατά την οποία βελτιστοποιείται μια χρονοσειρά με τη βοήθεια της τεχνικής της οπισθοδρόμησης είναι ο συνδυασμός των τεχνικών πρόβλεψης μελλοντικών τιμών με στόχο την καλύτερη πρόβλεψη.

2.2 Αλγόριθμος Lipschitz

Στον εν λόγω αλγόριθμο, η σταθερά Lipschitz, αποτελεί δείκτη επιτάχυνσης της χρονοσειράς. Ο αλγόριθμος χρησιμοποιεί την εικονική σειρά L_t, η οποία αναπαριστά τις προσεγγίσεις της σταθεράς Lipschitz ενός υποσυνόλου, για τις παραγόμενες προβλέψεις σε αντίθεση με τις τιμές της αρχικά εξεταζόμενης χρονοσειράς.

Αναλυτικότερα, στη δεδομένη τεχνική η σταθερά υπολογίζεται να τείνει σε μεγάλες τιμές, και δεν είναι εφικτό να χρησιμοποιηθεί για την πρόβλεψη χρονικών περιόδων κρίσης. Για τον λόγο αυτόν η μέθοδος εφαρμόζεται σε μία προκαθορισμένη περιοχή [n-m, n], όπου n τα τελευταία γνωστά σημεία της χρονοσειράς και m τα σημεία του παρελθόντος, στα οποία εφαρμόστηκε ο αλγόριθμος. Όσο πιο ευρεία είναι η περιοχή τόσο πιο ομαλές είναι οι μεταβολές που θα υποστεί η σταθερά L.

Κατά την εφαρμογή του, δημιουργείται αρχικά συνάρτηση για τη μελέτη των χρονοσειρών και για την εύρεση του μελλοντικού χρονικού σημείου. Στη συνέχεια,

πραγματοποιείται υπολογισμός των παραγώγων για κάθε σημείο της συνάρτησης και κατ΄ επέκταση η κλίση της ευθείας για κάθε νεοεισερχόμενο σημείο.

Ακολουθεί ο υπολογισμός της μέγιστης τιμής των παραγώγων και ο ορισμός της μέγιστης εξ αυτών ως σταθερά L. Τελευταίο βήμα του αλγορίθμου, είναι ο υπολογισμός της μεταβλητής α, η οποία αντιπροσωπεύει τη χρονική στιγμή, στην οποία εμφανίζεται το ελάχιστο ή το μέγιστο του της εξεταζόμενης χρονοσειράς.

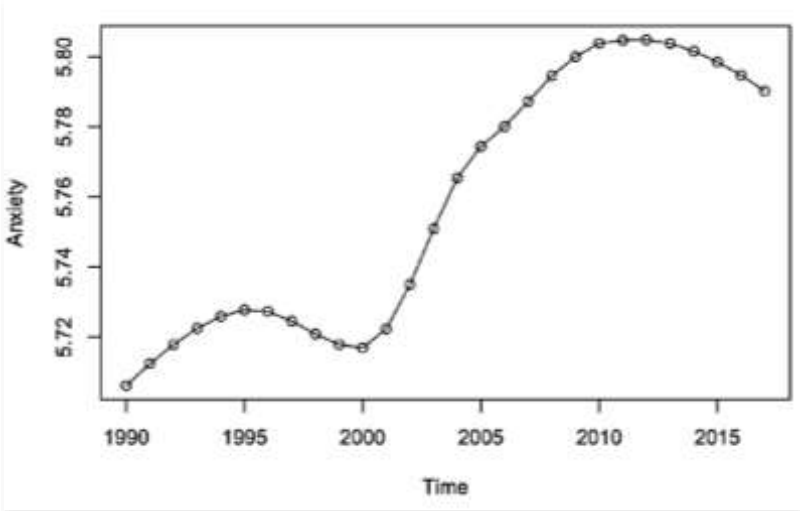
Ο συγκεκριμένος αλγόριθμος επιλέχθηκε λόγω της ευαισθησίας που παρουσιάζει στον εντοπισμό έναρξης μιας κρίσης, αλλά και στην ικανότητα του να προσπερνά εσφαλμένες προειδοποιήσεις (Λισγαρά, 2011).

3. ΜΕΛΕΤΗ ΔΕΔΟΜΕΝΩΝ

Όπως, ήδη έχει αναφερθεί στόχος της συγκεκριμένης ανάλυσης δεν είναι ακριβώς η εύρεση της μελλοντικής τιμής της χρονοσειράς, αλλά μέσω της πρόβλεψης να εντοπιστεί το μελλοντικό χρονικό σημείο στο οποίο αναμένεται να βελτιστοποιηθεί τοπικά.

Μετά την καταχώρηση των δεδομένων, η γραφική απεικόνιση της χρονοσειράς που αφορά τα ποσοστά των ατόμων που πάσχουν από άγχος, υποδηλώνει μία μικρή αύξηση την πρώτη δεκαετία (1990-2000), ενώ ξαφνικά την πρώτη δεκαετία του 2000 σημειώνεται ραγδαία αύξηση, με ήπια πτωτική τάση από το 2012 και μετά. Επίσης, παρατηρεί κανείς ότι πριν την πανδημία το ποσοστό των ατόμων που έπασχαν από άγχος στην Ελλάδα προσέγγιζε το 5.75%.

Γράφημα 1:*Ποσοστά ατόμων που πάσχουν από άγχος στην Ελλάδα, για τα έτη 1990-2017*

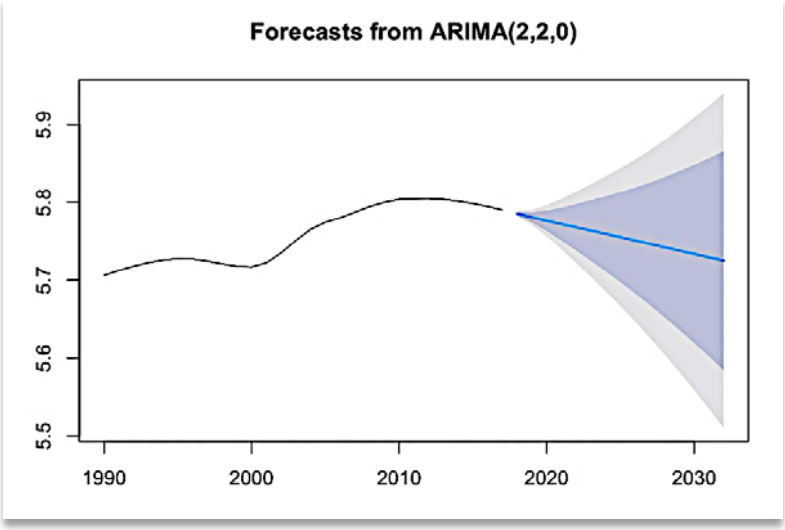


Για την προσαρμογή του κατάλληλου μοντέλου γίνεται χρήση της R, βασιζόμενοι στην γνωστή μεθοδολογία Box-Jenkins. Στα πλαίσια αυτής της εργασίας, κριτήριο επιλογής του κατάλληλου μοντέλου αποτέλεσαν τα μέτρα ακρίβειας, και συγκεκριμένα το Μέσο Απόλυτο Ποσοστιαίο Σφάλμα (MAPE) και η εκτίμηση του υποδείγματος με χρήση του πακέτου forecast του λογισμικού R.

Πραγματοποιώντας έλεγχο Dickey-Fuller για τη στασιμότητα του μοντέλου, προέκυψε με p-value 0.3063 ότι όντως η χρονοσειρά είναι στάσιμη και έπειτα από δοκιμές και με σύγκριση των μέτρων ακρίβειας, επιλέχθηκε ως καταλληλότερο μοντέλο το ARIMA(2,2,0).

Στη συνέχεια, όπως φαίνεται και στο Γράφημα 2, πραγματοποιήθηκε forecasting για 15 έτη και από τις εκτιμώμενες τιμές πρόβλεψης οι οποίες προέκυψαν για τα ποσοστά έως το 2032, παρατηρείται διαρκής πτωτική τάση της τάξεως 0.004% ετησίως.

Γράφημα 2:*Πρόβλεψη για τα ποσοστά των ατόμων που πάσχουν από άγχος*



Ακολούθησε, με τη βοήθεια του λογισμικού MATLAB η εφαρμογή του αλγορίθμου Lipschitz, και προέκυψε σταθερά L=0.0286 και ο υπολογισμός της μεταβλητής α. Ο Πίνακας 1 παρέχει τις τιμές της α, που δείχνει πότε θα προκύψει το μελλοντικό τοπικό βέλτιστο.

Πίνακας 1:*Αποτελέσματα της μεθόδου LIN για τα δεδομένα του άγχους*

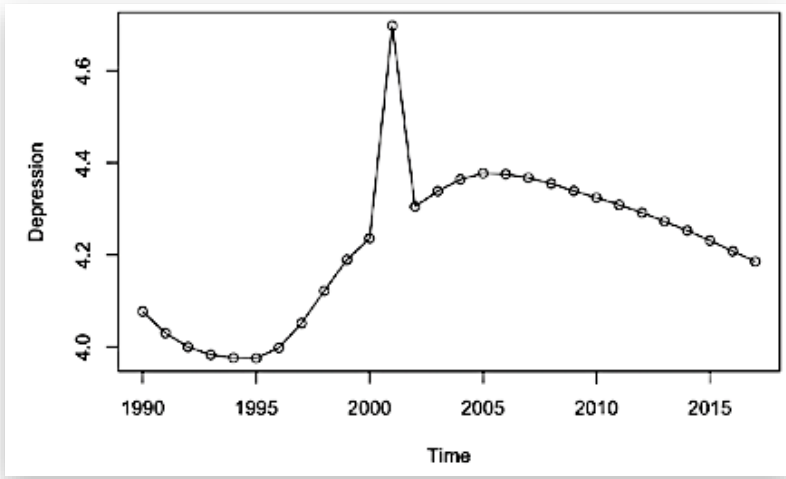
9.1075	30.3118	15.5472	2	7.9989	67.6544	18.7991
10.7972	122.1538	18.8176	3.9402	7.7780	429.8346	15.0403
12.3847	0	64.3784	6.3471	10.4198	56.5460	
16.7993	20.8642	10.6201	10.0968	15.0660	25.2844	

Σύμφωνα με τις τιμές αυτές το μελλοντικό βέλτιστο θα εμφανιστεί σε περίπου 429 μήνες, περίπου σε 35 έτη. Το ποσοστό των ατόμων που έπασχαν από άγχος θα αποκτήσει την ελάχιστη τιμή του σε αυτό το χρονικό διάστημα, γιατί η χρονοσειρά του άγχους εμφανίζει καθοδική πορεία, με πρόσκαιρη ανοδική τάση. Το ποσοστό των ατόμων που πάσχουν από άγχος ολοένα αυξάνεται με αποτέλεσμα την τόση μεγάλη αργοπορία εμφάνισης της ελάχιστης τιμής του.

Ακολουθώντας την ίδια διαδικασία, πραγματοποιήθηκε ανάλυση του ποσοστού των ατόμων που πάσχουν από κατάθλιψη, με δεδομένα είκοσι οκτώ ετών, 1990-2017.

Σε αντίθεση, με τα αντίστοιχα ποσοστά των ατόμων που πάσχουν από άγχος, τα ποσοστά των ατόμων που πάσχουν από κατάθλιψη, παρουσιάζει αρχικά μια πτωτική τάση στα πέντε πρώτα έτη, ενώ για την επόμενη δεκαετία (1995-2005) σημειώνεται αύξηση των ποσοστών με ένα peekto 2001, τιμή η οποία μπορεί να θεωρηθεί μεμονωμένο σημείο. Από το 2005 και μετά, φαίνεται να υπάρχει πτωτική τάση έως το 2017. Καταγράφηκε από τα περιγραφικά στατιστικά, ότι το ποσοστό περιστατικών κατάθλιψης πριν την πανδημία προσέγγιζε το 4.22%,δηλαδή 2.3% χαμηλότερο από τα σημερινά δεδομένα σύμφωνα με τον Παγκόσμιο Οργανισμό Υγείας.

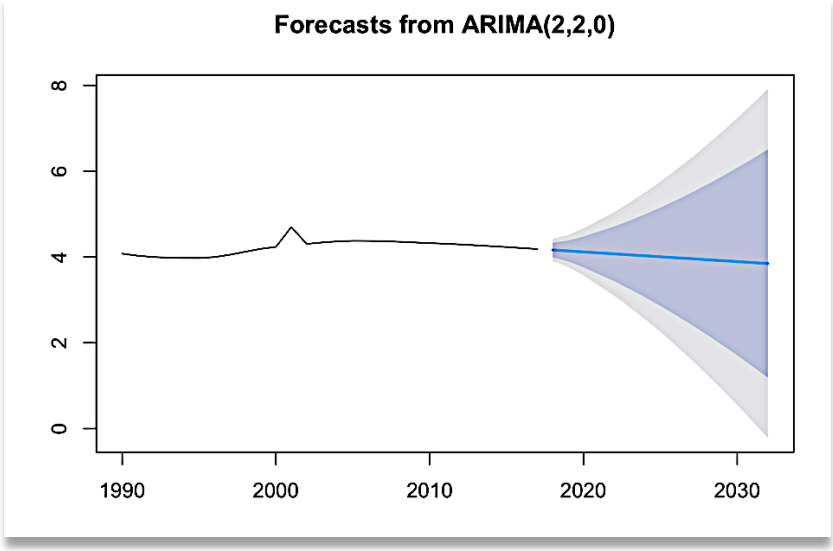
Γράφημα 3:*Ποσοστά ατόμων που πάσχουν από κατάθλιψη στην Ελλάδα, για τα έτη 1990-2017*



Η κατανομή αποδείχθηκε στάσιμη, μέσω του ελέγχου Dickey-Fuller, με p-valueίσο με 0.9203 και μετά από κατάλληλη διαδικασία προέκυψε κατάλληλο μοντέλο το ARIMA(2,2,0).

Πραγματοποιώντας πρόβλεψη για τα επόμενα 15 έτη, προέκυψε πως τα ποσοστά των ατόμων που πάσχουν από κατάθλιψη, θα έχει ελαφριά πτωτική τάση της τάξεως του 0.02% ανά έτος, όπως παρουσιάζεται στο Γράφημα 4.

Γράφημα 4:*Πρόβλεψη για τα ποσοστά των ατόμων που πάσχουν από κατάθλιψη*



Εν συνεχεία υπολογίστηκε η σταθερά L, να παίρνει τιμή ίση με 0.4620 και τιμές α του αλγορίθμου να είναι οι τιμές του Πίνακα 2:

Πίνακας 2: *Αποτελέσματα της μεθόδου LIN για τα δεδομένα της κατάθλιψης*

19.8	1133.8	13.7	27.6	123.7	60.2	44.1
30.1	41.2	20.0	36.4	73.3	54.3	39.3
55.3	17.3	2.0	70.0	57.6	48.4	
142.2	13.1	2.4	451.2	62.6	45.9	

Σύμφωνα με την ανάλυση της χρονοσειράς, η σταθερά L υπολογίστηκε να είναι 0.4620, με το τοπικό βέλτιστο να εμφανίζεται σε περίπου 38 μήνες. Στους τριάνταοκτώ μήνες, το ποσοστό των ατόμων που πάσχουν από κατάθλιψη θα αποκτήσει την ελάχιστη τιμή του γιατί η χρονοσειρά έχει πτωτική τάση.

4. ΑΠΟΤΕΛΕΣΜΑΤΑ-ΣΥΜΠΕΡΑΣΜΑΤΑ

Στην παρούσα εργασία, μελετήθηκαν τα ποσοστά των ατόμων που πάσχουν από άγχος και κατάθλιψη ως χρονοσειρές, με σκοπό την εύρεση του μελλοντικού βέλτιστου.Αξιοσημείωτο είναι το γεγονός ότι η συγκεκριμένη μελέτη επικεντρώνεται στην εύρεση του χρονικού σημείου κατά το οποίο θα εμφανιστεί το τοπικό βέλτιστο και όχι στην εύρεση της τιμής του.

Παρατηρούμε ότι το ποσοστό των ατόμων που πάσχουν από άγχος θα απαιτήσει μεγαλύτερο χρονικό διάστημα για να ελαχιστοποιηθεί σε σχέση με τα ποσοστά των ατόμων που πάσχουν από κατάθλιψη. Αυτό μπορεί να οφείλεται στο γεγονός ότι η διαφορετική φύση του άγχους σε σχέση με την κατάθλιψη μπορεί να οδηγήσει σε διαφορετικούς τρόπους αντιμετώπισης. Ενώ, η κατάθλιψη μπορεί να απαιτεί κυρίως φαρμακευτική αγωγή, το αγχωμένο άτομο μπορεί να επωφεληθεί περισσότερο από θεραπευτικές συζητήσεις και στρατηγικές αντιμετώπισης.

5. ΜΕΛΛΟΝΤΙΚΕΣ ΠΡΟΣΕΓΓΙΣΕΙΣ

Για την περαιτέρω ενίσχυση της προβλεπτικής ικανότητας των παραπάνω μοντέλων προτείνεται η διερεύνηση χρονικών σημείων που οδηγούν σε μελλοντικές βέλτιστες τιμές και η δημιουργία μεθοδολογιών σύμφωνα με τη φύση των δεδομένων. Συγκεκριμένα, επιδιώκουμε την ανάπτυξη νέων μεθοδολογιών εύρεσης βέλτιστου σημείου, έχοντας ως βάση της μέθοδο Newton-Rapshon.

HNewton-Raphson ανήκει στις Μεθόδους Ανίχνευσης κατά Γραμμή και βασίζεται σε δεδομένο αρχικό σημείο. Αρχικό βήμα της μεθόδου αποτελεί η κατασκευή τετραγωνικής προσέγγισης της αντικειμενικής, με σκοπό την ελαχιστοποίηση της. Ως σημείο ελαχιστοποίησης ορίζεται κάθε φορά το αρχικό σημείο της επόμενης επανάληψης. Η μέθοδος βασίζεται σε επαναληπτικό τύπο και επιλέχθηκε ως σημείο αναφορά για την ανάπτυξη νέων μεθοδολογιών, λόγω της ικανότητας της να συγκλίνει πιο γρήγορα από τη μέθοδο απότομης καθόδου.

ABSTRACT

Anxiety attacks, increasing incidence of depression and other mental health disorders, is now a serious issue worldwide, thus its study and analysis is necessary. Post Covid-19 era as well as the shocking technological advancement that demands our undivided attention on a daily bases are investigated as possible causes. In the present study, data on the rates of anxiety and depression in our country from 1990-2017 and data on sales of mental health drugs in our country from 2020 onwards are used. Time series analysis, ARIMA model and Lipschitz Indicator algorithm were used to model the data to predict the future optimum.

By applying the specific methods, the optimal percentages in Greece were identified, but this was not possible for the pharmaceutical sales data. Finally, the results and conclusions are recorded in order to study the change in the rate of future incidents, to investigate its dominant factors to be addressed and improve mental health in the coming years.

Keywords: Timeseries, Forecast, Optimization, ARIMA, LipschitzINdicator.

ΑΝΑΦΟΡΕΣ

Καρβέλης Χαράλαμπος (2008).*«Ανάλυση Χρονολογικών σειρών: Προβλέποντας το Μέλλον, Κατανοώντας το Παρελθόν»,* Τμήμα Μαθηματικών, Πανεπιστήμιο Πατρών.

Λισγαρά Γ. Ελένη (2011). *Τεχνικές Βελτιστοποίησης στην Πρόβλεψη Χρονοσειρών, Τμήμα Διοίκησης Επιχειρήσεων,* Πανεπιστήμιο Πατρών.

Στρατογιάννη Αικατερίνη (2017). *Πρόβλεψη Μελλοντικών Ακροτάτων Χρονοσειρών: «Η Περίπτωση της Ανεργίας»*, Τμήμα Μαθηματικών, Πανεπιστήμιο Πατρών.

Armijo L. (1996). *Minimization of function having Lipschitz continuous first partial derivatives,* Pacific Journal of Mathematics, 16:1-3.

Box, G.E.P., & G.M. Jenkins (1970).*Time Series Analysis: Forecasting and Control, San Francisco: Holden Day*(revised ed. 1976).

Beaglehole, B., Mulder, R. T., Frampton, C. M., Boden, J. M., Newton-Howes, G., & Bell, C. J. (2018). *Psychological distress and psychiatric disorder after natural disasters: Systematic review and meta-analysis.* The British Journal of Psychiatry*, 213*(6):716–722.

Kantz H., & Schreiber T. (2004), *Nonlinear Time Series Analysis,* Cambridge University.

National Institute of Mental Health, *https://www.nimh.nih.gov.*

Vrahatis N. M., Androulakis S. G., Lambrinos N. J., &Magoulas D.G. (2000), *"A class of gradient unconstrained minimization algorithms with adaptive stepsize",* Journal of Computational and Applied Mathematics.

Wisevoter, *https://wisevoter.com/country-rankings/most-depressed-countries/*.

Wood G. R. and Zhang B. P. (1996). *Estimation of the Lipschitz Constant of a Function,* *Journal of Global Optimization,* 8(1):91-103, January

World Health Statistics (2023). *Monitoring health for the SDGs, sustainable development goals,* 978-92-4-007433-0

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 35ου Πανελληνίου & 1ου Διεθνούς Συνεδρίου Στατιστικής (2023), σελ46-57

ΒΕΛΤΙΩΜΕΝΗ ΕΚΤΙΜΗΣΗ ΠΑΡΑΜΕΤΡΩΝ ΣΕ INVERSE GAUSSIAN ΚΑΤΑΝΟΜΗ

Π. Μπομποτάς[1], Σ. Κουρούκλης[2]
[1]Τμήμα Μαθηματικών, Πανεπιστήμιο Θεσσαλίας
pbobotas@uth.gr
[2]Τμήμα Μαθηματικών, Πανεπιστήμιο Πατρών
stavros@math.upatras.gr

ΠΕΡΙΛΗΨΗ

Η παρούσα εργασία ασχολείται με την εκτίμηση της μέσης τιμής μ και της παραμέτρου $1/\lambda$ μιας inverse Gaussian κατανομής $IG(\mu, \lambda)$ από την πλευρά της Στατιστικής Θεωρίας Αποφάσεων. Κατασκευάζονται νέες κλάσεις βελτιωμένων εκτιμητών και παρουσιάζονται συγκριτικά αποτελέσματα προσομοίωσης για επιλεγμένα μέλη αυτών των κλάσεων και εκτιμητές διαθέσιμους στη βιβλιογραφία.

Λέξεις Κλειδιά: συνάρτηση ζημίας, συνάρτηση κινδύνου, μη αποδεκτοί εκτιμητές, εκτιμητές τύπου Stein, εκτιμητές τύπου Brewster and Zidek, εκτιμητές τύπου Strawderman, ιδιότητα μονότονου λόγου πιθανοφανειών.

1. ΕΙΣΑΓΩΓΗ

Η inverse Gaussian κατανομή εισήχθη από τον Schrödinger (1915) για την ανάλυση της κίνησης Brown (Wiener process), ενώ για τη μελέτη της, πρωτοπόρες εργασίες είναι αυτές των Tweedie (1945, 1957a, b) και Wald (1945). Η κατανομή αυτή έχει εφαρμογές σε πολλά πεδία όπως η Βιολογία, η Ιατρική, τα Οικονομικά, η Θεωρία Αξιοπιστίας, τα προβλήματα χρόνου ζωής. Μερικές κλασικές πηγές σχετικά με τις ιδιότητες και τις εφαρμογές της κατανομής αποτελούν τα βιβλία των Jorgensen (1982), Chhikara and Folks (1989), και Seshardi (1993).

Μία τυχαία μεταβλητή X ακολουθεί inverse Gaussian κατανομή με παραμέτρους μ και λ, συμβολικά $IG(\mu, \lambda)$, εάν η συνάρτηση πυκνότητάς της είναι

$$f(x;\mu,\lambda) = \left(\frac{\lambda}{2\pi x^3}\right)^{1/2} exp\left\{-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right\}, \quad x > 0, \qquad \mu > 0, \qquad \lambda > 0.$$

Η παράμετρος μ είναι η μέση τιμή της κατανομής. Όταν η inverse Gaussian κατανομή προέρχεται από μία Wiener διαδικασία, η παράμετρος $1/\lambda$ εκφράζει την παράμετρο διασποράς της διαδικασίας.

Αντικείμενο της εργασίας είναι η μελέτη του προβλήματος βελτιωμένης εκτίμησης των μ και $\sigma = 1/\lambda$ από την πλευρά της Στατιστικής Θεωρίας Αποφάσεων, και συγκεκριμένα ως προς τις συναρτήσεις ζημίας (σ.ζ.) $L_\mu(\hat{\mu}, \mu) = (\hat{\mu} - \mu)^2/\mu^3$ (rescaled squared error) και $L_\sigma(\hat{\sigma}, \sigma) = \hat{\sigma}/\sigma - \ln(\hat{\sigma}/\sigma) - 1$ (σ.ζ. εντροπίας), αντίστοιχα. Δύο εκτιμητές συγκρίνονται με κριτήριο τη μέση ζημία τους (συνάρτηση κινδύνου) και καλύτερος είναι αυτός που έχει μικρότερη μέση ζημία για κάθε τιμή των άγνωστων παραμέτρων. Εξ αυτών, ο εκτιμητής με τη μεγαλύτερη μέση ζημία για κάθε τιμή των άγνωστων παραμέτρων λέγεται μη αποδεκτός (inadmissible), ενώ αυτός με τη μικρότερη αναφέρεται ως βελτιωμένος (improved) σε σχέση με τον μη αποδεκτό. Ο εκτιμητής με τη μικρότερη μέση ζημία σε μία κλάση εκτιμητών αναφέρεται ως βέλτιστος εκτιμητής της κλάσης.

Έστω $X_1, \dots, X_n$ ένα τυχαίο δείγμα από $IG(\mu, \lambda)$. Ακολουθούν συμβολισμοί και μερικά γνωστά αποτελέσματα που είναι χρήσιμα για τη συνέχεια.

- $\bar{X} = (1/n)\sum_{i=1}^n X_i \sim IG(\mu, n\lambda)$,
- $S = (1/n)\sum_{i=1}^n (1/X_i - 1/\bar{X}) \sim (n\lambda)^{-1}\chi^2_{n-1}$ (χι-τετράγωνο),
- $(\bar{X}, S)$ είναι μία επαρκής στατιστική συνάρτηση για το (μ, λ),
- $(\bar{X}, S)$ είναι ο εκτιμητής μέγιστης πιθανοφάνειας για το $(\mu, 1/\lambda)$,
- $\lambda(X_i - \mu)^2/(\mu^2 X_i) \sim \chi^2_1$, $i = 1, \dots, n$ (Shuster, 1968),
- $\lambda n(\bar{X} - \mu_0)^2/({\mu_0}^2 \bar{X})$, για $\mu \neq \mu_0 > 0$ έχει πυκνότητα $h(x; \mu, \sigma)$ πολύπλοκης μορφής (δεν είναι μη κεντρική χι-τετράγωνο),
- $\lambda n(\bar{X} - \mu_0)^2/({\mu_0}^2 \bar{X}) \sim \chi^2_1$, για $\mu = \mu_0 > 0$, δηλαδή $h(x; \mu_0, \sigma) = h(x)$ η πυκνότητα της χ^2_1,
- $nS/\sigma = \lambda\sum_{i=1}^n (1/X_i - 1/\bar{X}) \sim g(u)$ (πυκνότητα της χ^2_{n-1}),
- $T/\sigma = \lambda n(\bar{X} - \mu_0)^2/({\mu_0}^2 \bar{X}) \sim h(x; \mu, \sigma)$.

Επίσης, ισχύουν οι ακόλουθες ιδιότητες (τύπου) μονότονου λόγου πιθανοφανειών (μλπ),

(A1) $h(x; \mu, \sigma)/h(x)$ είναι γνησίως αύξουσα ως προς $x > 0$ για $\mu \neq \mu_0$ (Kourouklis, 1997),

(A2) $g(c_1 x)/g(c_2 x)$ είναι αύξουσα ως προς $x > 0$ για $0 < c_1 < c_2$.

Για τη βελτίωση του $\bar{X}$, οι Ma et al. (2014), βασιζόμενοι στη μορφή μιας κλάσης βελτιωμένων εκτιμητών του μ, που παρήγαγαν οι Hsieh et al. (1990) υπό μία Μπεϋζιανή οπτική, πρότειναν τον εκτιμητή συρρίκνωσης (shrinkage estimator)

$$\hat{\mu}_{a,c} = \left(1 - \frac{c}{n/S + 4a} exp\left\{-\frac{a}{\bar{X}}\right\}\right)\bar{X}$$

και έδειξαν ότι για δοθέν $a > 0$, ικανή συνθήκη ώστε ο $\hat{\mu}_{a,c}$ να είναι καλύτερος από τον $\bar{X}$ είναι όταν $0 < c \leq 2an/(n+1)$. Σημειώνεται ότι $EL_\mu(\bar{X}, \mu) = 1/(n\lambda)$, δηλαδή ο $\bar{X}$ έχει σταθερή μέση ζημία ως προς μ. Ο εκτιμητής

$$\hat{\mu}_a = \left(1 - \frac{2aS}{(n+1)(1 + (4a/n)S)} exp\left\{-\frac{a}{\bar{X}}\right\}\right)\bar{X}$$

είναι ο εκτιμητής με τη μεγαλύτερη συρρίκνωση μεταξύ των εκτιμητών της κλάσης $\hat{\mu}_{a,c}$, και προκύπτει για τη μέγιστη δυνατή τιμή του $c = 2an/(n+1)$, ενώ με βάση τα αριθμητικά αποτελέσματα προσομοιώσεων, οι Ma et al. (2014) προτείνουν στην πράξη τη χρήση του εκτιμητή

$$\hat{\mu}_1 = \left(1 - \frac{2S}{(n+1)(1+(4/n)S)}\, exp\left\{-\frac{1}{\bar{X}}\right\}\right)\bar{X}.$$

Από την άλλη πλευρά, όσον αφορά την εκτίμηση του $\sigma = 1/\lambda$, ένας κλασικός εκτιμητής είναι ο

$$\delta_0 = \frac{n}{n-1}S,$$

ο οποίος είναι ο βέλτιστος εκτιμητής στην κλάση των εκτιμητών $\{cS : c > 0\}$ ως προς τη μέση ζημία εντροπίας $EL_\sigma(cS, \sigma)$. Για τη βελτίωση του δ_0, μία γενική μορφή βελτιωμένων εκτιμητών συρρίκνωσης του δ_0 στη βιβλιογραφία είναι

$$\delta_\varphi = \varphi(W)S, \qquad W = \frac{T}{nS},$$

όπου φ είναι θετική αύξουσα συνάρτηση με $\lim_{w\to\infty} \varphi(w) = n/(n-1)$. Ο Kourouklis (1997), αποδεικνύοντας την ιδιότητα του μλπ της inverse Gaussian κατανομής και κάνοντας χρήση της τεχνικής Kubokawa (1994), παρήγαγε τους εκτιμητές τύπου Stein (1964) και Brewster and Zidek (1974),

$$\delta_{\varphi_S} = min\left\{\frac{n}{n-1}, 1+W\right\}S \quad \text{και} \quad \delta_{\varphi_{BZ}} = \frac{\int_0^1 x^{-1/2}(1+Wx)^{-n/2}dx}{\int_0^1 x^{-1/2}(1+Wx)^{-n/2-1}dx}S,$$

αντίστοιχα, οι οποίοι ικανοποιούν τις παρακάτω ικανές συνθήκες ώστε ο δ_φ να είναι καλύτερος από τον δ_0,

$$\frac{\int_0^1 x^{-1/2}(1+wx)^{-n/2}dx}{\int_0^1 x^{-1/2}(1+wx)^{-n/2-1}dx} \le \varphi(w) \le \frac{n}{n-1}.$$

Επίσης, οι Bobotas and Kourouklis (2009), επεκτείνοντας την τεχνική του Strawderman (1974) για την εκτίμηση της διασποράς ενός κανονικού πληθυσμού, στην εκτίμηση μιας γενικής παραμέτρου κλίμακας υπό την παρουσία μιας ενοχλητικής παραμέτρου (nuisance parameter), παρήγαγαν εκτιμητές τύπου Strawderman (1974) για το σ, υπό την τετραγωνική συνάρτηση ζημίας $(\hat{\sigma}/\sigma - 1)^2$.

Η δομή της εργασίας είναι η ακόλουθη. Στην Ενότητα 2 δίνονται ικανές συνθήκες για εκτιμητές με μια γενικότερη μορφή από αυτή των Ma et al. (2014) ώστε να βελτιώνουν τον $\bar{X}$, ενώ παρουσίαζονται αριθμητικά αποτελέσματα που υποδεικνύουν νέους εκτιμητές με καλύτερη συμπεριφορά μέσης ζημίας από τον προτεινόμενο εκτιμητή των Ma et al. (2014), $\hat{\mu}_1$. Στην Ενότητα 3 δίνονται ικανές συνθήκες για εκτιμητές απλής μορφής τύπου Strawderman (1974) ώστε να βελτιώνουν τον δ_0, η οποία ολοκληρώνεται με την παρουσίαση αριθμητικών αποτελεσμάτων της συγκριτικής συμπεριφοράς της μέσης ζημίας των σχετικών εκτιμητών.

2. ΒΕΛΤΙΩΜΕΝΗ ΕΚΤΙΜΗΣΗ ΓΙΑ ΤΟ μ

Ορμώμενοι από τη μορφή του εκτιμητή συρρίκνωσης των Ma et al. (2014), θεωρούμε τη γενική μορφή βελτιωμένων εκτιμητών

$$\delta_h = \left(1 - h(S) exp\left\{-\frac{a}{\bar{X}}\right\}\right)\bar{X}, \qquad a > 0, \qquad 0 \le h(s) \le 1.$$

Τα ακόλουθα αποτελέσματα δίνουν ικανές συνθήκες ώστε ο δ_h να βελτιώνει τον $\bar{X}$. Σημειώνεται ότι το δεύτερο αποτέλεσμα αποτελεί μία πιο χαλαρή ικανή συνθήκη από το πρώτο.

Πρόταση 1. Για δοθέν $a > 0$, ο εκτιμητής δ_h είναι καλύτερος από τον $\bar{X}$ εάν

$$Eh^2(S) \le 2\frac{\sqrt{n\lambda + 2a} - \sqrt{n\lambda}}{\sqrt{n\lambda + 4a}} Eh(S).$$

Σκιαγράφηση απόδειξης. Ισχύει ότι

$$E\left[\bar{X}\exp\left(-\frac{a}{\bar{X}}\right)\right] = \mu\ \exp\left(\frac{n\lambda - \sqrt{n\lambda(n\lambda + 2a)}}{\mu}\right)$$

$$E\left[\bar{X}^2\exp\left(-\frac{a}{\bar{X}}\right)\right] = \mu^2\left(\frac{\mu}{n\lambda} + \frac{\sqrt{n\lambda + 2\alpha}}{\sqrt{n\lambda}}\right)\exp\left(\frac{n\lambda - \sqrt{n\lambda(n\lambda + 2a)}}{\mu}\right)$$

$$E\left[\bar{X}^2\exp\left(-\frac{2a}{\bar{X}}\right)\right] = \mu^2\left(\frac{\mu}{n\lambda} + \frac{\sqrt{n\lambda + 4\alpha}}{\sqrt{n\lambda}}\right)\exp\left(\frac{n\lambda - \sqrt{n\lambda(n\lambda + 4a)}}{\mu}\right)$$

$$\frac{\mu}{n\lambda} + \frac{\sqrt{n\lambda + 2\alpha}}{\sqrt{n\lambda}} - 1 \ge \left(\frac{\sqrt{n\lambda + 2\alpha} - \sqrt{n\lambda}}{\sqrt{n\lambda + 4\alpha}}\right)\left(\frac{\mu}{n\lambda} + \frac{\sqrt{n\lambda + 4\alpha}}{\sqrt{n\lambda}}\right)$$

ενώ για τη διαφορά των μέσων ζημιών των δ_h και $\bar{X}$ έχουμε

$$\begin{aligned} RD(\delta_h, \bar{X}) &= EL_\mu(\delta_h, \mu) - EL_\mu(\bar{X}, \mu) \\ &= \mu^{-3}\left\{Eh^2(S)E\left[\bar{X}^2\exp\left(-\frac{2a}{\bar{X}}\right)\right] - 2Eh(S)E\left[\bar{X}^2\exp\left(-\frac{a}{\bar{X}}\right)\right]\right. \\ &\qquad \left. + 2\mu Eh(S)E\left[\bar{X}\exp\left(-\frac{a}{\bar{X}}\right)\right]\right\} \\ &\le \mu^{-1}\exp\left(\frac{n\lambda - \sqrt{n\lambda(n\lambda + 2a)}}{\mu}\right)\left\{Eh^2(S)\left(\frac{\mu}{n\lambda} + \frac{\sqrt{n\lambda + 4\alpha}}{\sqrt{n\lambda}}\right)\right. \\ &\qquad \left. - 2Eh(S)\left(\frac{\mu}{n\lambda} + \frac{\sqrt{n\lambda + 2\alpha}}{\sqrt{n\lambda}} - 1\right)\right\} \end{aligned}$$

$$\leq \mu^{-1} \exp\left(\frac{n\lambda - \sqrt{n\lambda(n\lambda+2a)}}{\mu}\right)\left(\frac{\mu}{n\lambda} + \frac{\sqrt{n\lambda+4\alpha}}{\sqrt{n\lambda}}\right)\left\{Eh^2(S) - 2\left(\frac{\sqrt{n\lambda+2\alpha}-\sqrt{n\lambda}}{\sqrt{n\lambda+4\alpha}}\right)Eh(S)\right\}.$$

Πρόταση 2. Για δοθέν $a > 0$, ο εκτιμητής δ_h είναι καλύτερος από τον $\bar{X}$ εάν

$$Eh^2(S) \leq \frac{2a}{n\lambda+3a} Eh(S).$$

Σκιαγράφηση απόδειξης. Από Πρόταση 1, παρατηρώντας ότι ισχύει

$$\frac{\sqrt{n\lambda+2\alpha}-\sqrt{n\lambda}}{\sqrt{n\lambda+4\alpha}} = \frac{2\alpha}{\sqrt{n\lambda+4\alpha}}\frac{1}{\sqrt{n\lambda+2\alpha}+\sqrt{n\lambda}} > \frac{\alpha}{\sqrt{n\lambda+4\alpha}\sqrt{n\lambda+2\alpha}} > \frac{\alpha}{n\lambda+3\alpha}.$$

Στη συνέχεια, θεωρούμε την ακόλουθη ειδική μορφή των παραπάνω βελτιωμένων εκτιμητών δ_h

$$\delta_\psi = \left(1 - \frac{\psi(S)\,S}{1+\varepsilon\,S} exp\left\{-\frac{a}{\bar{X}}\right\}\right)\bar{X},$$

όπου $a > 0, \varepsilon > 0, \psi(s)$ φθίνουσα με $0 \leq \psi(s) \leq r$. Η ιδέα πίσω από αυτή τη θεώρηση της ειδικής μορφής για τη συνάρτηση $h(s)$, βρίσκεται στη γενίκευση της μορφής του εκτιμητή $\hat{\mu}_a$ ο οποίος μπορεί να γραφεί ως

$$\hat{\mu}_a = \left(1 - \frac{(c/n)S}{1+(4a/n)S)} exp\left\{-\frac{a}{\bar{X}}\right\}\right)\bar{X}.$$

Το ακόλουθο αποτέλεσμα δίνει ικανές συνθήκες ώστε ο δ_ψ να βελτιώνει τον $\bar{X}$.

Πρόταση 3. Για δοθέντα $a > 0, \varepsilon > 0$, ο εκτιμητής δ_ψ είναι καλύτερος από τον $\bar{X}$ εάν

$$r \leq \min\left\{\frac{2\,\varepsilon}{3}, \frac{2a}{n+1}\right\}.$$

Σκιαγράφηση απόδειξης. Ισχύει ότι εάν Y είναι τυχαία μεταβλητή με πεδίο τιμών $(0,\infty)$, και σε αυτό το διάστημα η $p(y)$ είναι μία θετική φθίνουσα συνάρτηση ενώ η $q(y)$ έχει μία αλλαγή προσήμου από αρνητική σε θετική στο σημείο $y_0 \in (0,\infty)$, τότε $E[p(Y)q(Y)] \leq p(y_0)Eq(Y)$ (βλ. και Kubokawa, 1994, Lemma 2.1).
Με εφαρμογή της Πρότασης 2 αρκεί

$$E\left[\frac{\psi^2(S)\,S^2}{(1+\varepsilon\,S)^2} - \frac{2a}{n\lambda+3a}\frac{\psi(S)\,S}{1+\varepsilon\,S}\right] = E\left[\frac{\psi^2(S)}{(1+\varepsilon\,S)^2(n\lambda+3a)}\left\{(n\lambda+3a)\,S^2 - 2a(1+\varepsilon\,S)\,S\frac{1}{\psi(S)}\right\}\right] \leq 0.$$

Επειδή $\psi(s) \le r$, ισοδύναμα αρκεί

$$E\left[\frac{\psi^2(S)}{(1+\varepsilon\, S)^2(n\lambda+3a)}\left\{(n\lambda+3a)\, S^2 - 2a(1+\varepsilon\, S)\, S\frac{1}{r}\right\}\right]$$
$$= E\left[\frac{\psi^2(S)}{(1+\varepsilon\, S)^2(n\lambda+3a)r}\{n\lambda r S^2 - 2aS + (3r-2\varepsilon)aS^2\}\right] \le 0$$

και συνεπώς, ισοδύναμα αρκεί

$$E\left[\frac{\psi^2(S)}{(1+\varepsilon\, S)^2(n\lambda+3a)r}\{n\lambda r S^2 - 2aS\}\right] \le 0,$$

ενώ επειδή $\psi^2(s)/(1+\varepsilon\, s)^2$ είναι φθίνουσα και $n\lambda r s^2 - 2as = s(n\lambda r s - 2a)$ έχει μία αλλαγή προσήμου στο $(0,\infty)$ από αρνητική σε θετική, ισοδύναμα αρκεί

$$E[n\lambda r S^2 - 2aS] \le 0 \Leftrightarrow r \le \frac{2aES}{n\lambda ES^2} = \frac{2a}{n+1}.$$

Σημειώνεται ότι για τους βελτιωμένους εκτιμητές δ_ψ της Πρότασης 3 ισχύει

$$h(s) = \frac{\psi(s)\, s}{1+\varepsilon\, s} \le \frac{r\, s}{1+\varepsilon\, s} \le \frac{(2/3)\,\varepsilon\, s}{1+\varepsilon\, s} \le 1.$$

Το ακόλουθο πόρισμα δίνει μία απλή μορφή βελτιωμένων εκτιμητών του $\bar{X}$.

Πόρισμα 4. Για δοθέντα $a > 0, \varepsilon > 0$, ο εκτιμητής

$$\delta_r = \left(1 - \frac{r\, S}{1+\varepsilon\, S} exp\left\{-\frac{a}{\bar{X}}\right\}\right)\bar{X},$$

είναι καλύτερος από τον $\bar{X}$ εάν $r \le \min\{2\varepsilon/3, 2a/(n+1)\}$.

Σημειώνεται ότι, μεταξύ των βελτιωμένων εκτιμητών της κλάσης του Πορίσματος 4, ο εκτιμητής με τη μεγαλύτερη συρρίκνωση προκύπτει για $r = \min\{2\varepsilon/3, 2a/(n+1)\}$, πράγμα που σημαίνει πως σε αυτή την περίπτωση έχουμε τους ακόλουθους καλύτερους εκτιμητές του $\bar{X}$,

- $\delta_{a,\varepsilon}^{(1)} = \left(1 - \frac{2\,\varepsilon\, S}{3(1+\varepsilon\, S)} exp\left\{-\frac{a}{\bar{X}}\right\}\right)\bar{X},\ για\ a > 0\ και\ 0 < \varepsilon < \frac{3\,a}{n+1}$,
- $\delta_{a,\varepsilon}^{(2)} = \left(1 - \frac{2\,a\, S}{(n+1)(1+\varepsilon\, S)} exp\left\{-\frac{a}{\bar{X}}\right\}\right)\bar{X},\ για\ a > 0\ και\ \varepsilon \ge \frac{3\,a}{n+1}$.

Παρατηρήσεις. Ισχύουν τα ακόλουθα.

1. $\hat{\mu}_a \equiv \delta_{a,\varepsilon}^{(2)}$ για $\varepsilon = 4a/n$, και αυτή η τιμή του ε είναι προφανώς μεγαλύτερη από $3a/(n+1)$.
2. $\{\delta_{a,\varepsilon}^{(2)} : a > 0, \varepsilon \ge 3a/(n+1)\} \supset \{\hat{\mu}_a : a > 0\}$.
3. Η κλάση $\{\delta_{a,\varepsilon}^{(2)} : a > 0, 3a/(n+1) \le \varepsilon < 4a/n\}$ περιέχει βελτιωμένους εκτιμητές με μεγαλύτερη συρρίκνωση από αυτούς της κλάσης $\{\hat{\mu}_a : a > 0\}$, ενώ η $\{\delta_{a,\varepsilon}^{(2)} : a > 0, \varepsilon > 4a/n\}$ με μικρότερη.

Στον Πίνακα 1 παρέχονται ενδεικτικά αριθμητικά αποτελέσματα από 50.000 Monte Carlo προσομοιώσεις σχετικά με τη μέση ζημία των εκτιμητών $\bar{X}$, $\hat{\mu}_1$, $\delta_{1,\varepsilon}^{(1)}$, $\delta_{1,\varepsilon}^{(2)}$, με μεγέθη δείγματος $n = 4,10$, $\mu = 1$, $\lambda = 0.1,0.4,1,2$, $\alpha = 1$, κατάλληλα επιλεγμένες τιμές για το ε, όπου $\varepsilon \approx 3a/(n+1)$ σημαίνει τιμή κοντά στο άνω φράγμα, $\varepsilon = 3a/(n+1)$ σημαίνει τιμή ακριβώς το άνω φράγμα, $\varepsilon > 4a/n$ σημαίνει τιμή ίση με $4a/n + 0.1$ που δίνει στον εκτιμητή $\delta_{1,\varepsilon}^{(2)}$ μικρότερη συρρίκνωση από τον $\hat{\mu}_1$. Από τη σύγκριση αυτών των αποτελεσμάτων φαίνεται ότι τη μεγαλύτερη βελτίωση της μέσης ζημίας επιτυγχάνει ο εκτιμητής $\delta_{1,\varepsilon}^{(2)}$ με $\varepsilon = 3a/(n+1)$ και ακολουθεί ο $\delta_{1,\varepsilon}^{(1)}$ με $\varepsilon \approx 3a/(n+1)$. Σημειώνεται ότι όταν το μέγεθος δείγματος n αυξάνεται, είναι αναμενόμενο να μειώνεται η βελτίωση της μέσης ζημίας καθενός από τους εκτιμητές $\hat{\mu}_1$, $\delta_{1,\varepsilon}^{(1)}$, $\delta_{1,\varepsilon}^{(2)}$ ως προς τον $\bar{X}$, διότι για μεγάλες τιμές του n αυτοί οι εκτιμητές τείνουν να ταυτιστούν με τον $\bar{X}$.

Πίνακας 1. *Αποτελέσματα προσομοίωσης μέσης ζημίας εκτιμητών του μ*

λ	$\bar{X}$	$\hat{\mu}_1$	$\delta_{1,\varepsilon}^{(1)}$ με $\varepsilon \approx \frac{3a}{n+1}$	$\delta_{1,\varepsilon}^{(2)}$ με $\varepsilon = \frac{3a}{n+1}$	$\delta_{1,\varepsilon}^{(2)}$ με $\varepsilon > \frac{4a}{n}$
	$n = 4$				
0.1	2.4080	1.2893	0.9613	0.9141	1.3579
0.4	0.6133	0.4249	0.3941	0.3785	0.4339
1	0.2522	0.2032	0.2004	0.1951	0.2048
2	0.1267	0.1109	0.1111	0.1092	0.1112
	$n = 10$				
0.1	1.0010	0.5470	0.4487	0.4475	0.6056
0.4	0.2518	0.1857	0.1775	0.1772	0.1913
1	0.0997	0.0853	0.0843	0.0842	0.0861
2	0.0504	0.0463	0.0461	0.0461	0.0464

3. ΒΕΛΤΙΩΜΕΝΗ ΕΚΤΙΜΗΣΗ ΓΙΑ ΤΟ σ

Στην ενότητα αυτή παράγονται βελτιωμένοι εκτιμητές τύπου Strawderman (1974) για το σ απλής μορφής. Μία γενική μορφή εκτιμητών τύπου Strawderman (1974) για το σ είναι

$$\delta_\varphi = \frac{n}{n-1}\big(1 - \varphi(W)\big)S, \qquad W = \frac{T}{nS},$$

όπου η φ ικανοποιεί, για κάποιο $\varepsilon > 0$,

- $(1+w)^\varepsilon \varphi(w)$ φθίνουσα συνάρτηση,
- $0 \le \varphi(w) \le B(\varepsilon)$, με $B(\varepsilon)$ κατάλληλο άνω φράγμα που πρέπει να προσδιορισθεί.

Μία απλή μορφή εκτιμητών τύπου Strawderman (1974) είναι να θεωρήσουμε

$$\delta^{(\varepsilon,r)} = \frac{n}{n-1}\left(1 - \frac{r}{(1+W)^{\varepsilon}}\right)S, \qquad W = \frac{T}{nS},$$

όπου $r \le B(\varepsilon)$. Το επόμενο αποτέλεσμα δίνει ικανές συνθήκες ώστε ο $\delta^{(\varepsilon,r)}$ να βελτιώνει τον $\delta_0 = [n/(n-1)]S$. Πρώτα όμως παρατίθεται ένα λήμμα (βλ. Lehmann and Romano, 2005, σελ.70), το οποίο είναι χρήσιμο για την απόδειξή του.

Λήμμα 5. Έστω $b_1(x)$ και $b_2(x)$ συναρτήσεις πυκνότητας ορισμένες στα διαστήματα S_1 και S_2 αντίστοιχα, όπου $S_1 \subseteq S_2$ και η $b_1(x)/b_2(x)$ είναι αύξουσα συνάρτηση για $x \in S_2$. Εάν X είναι τυχαία μεταβλητή με συνάρτηση πυκνότητας $b_1(x)$ ή $b_2(x)$ και $h(x)$, $x \in S_2$, είναι αύξουσα (αντίστοιχα, φθίνουσα) τότε $E_{b_1}h(X) \ge$ (αντίστοιχα, $\le$) $E_{b_2}h(X)$. Επιπλέον, εάν $h(x)$ και $b_1(x)/b_2(x)$ είναι γνησίως μονότονες, τότε $E_{b_1}h(X) >$ (αντίστοιχα, $<$) $E_{b_2}h(X)$.

Πρόταση 6. Για $\varepsilon > 0$, ο $\delta^{(\varepsilon,r)}$ είναι καλύτερος από τον δ_0 εάν $0 < r \le B(\varepsilon)$ με

$$B(\varepsilon) = \left\{1 + \frac{n-1}{2\varepsilon}\,\frac{\Gamma(\varepsilon + n/2 + 1)\Gamma(2\varepsilon + (n-1)/2)}{\Gamma(2\varepsilon + n/2)\Gamma(\varepsilon + (n-1)/2)}\right\}^{-1},$$

όπου $\Gamma(\alpha) = \int_0^{\infty} x^{a-1}e^{-x}dx$.

Σκιαγράφηση απόδειξης. Ισχύει ότι $\ln(1-x) = -\sum_{i=1}^{\infty} x^i/i \ge -x - x^2/2(1-x)$, $0 < x < 1$, ενώ η διαφορά των μέσων ζημιών των δ_0 και $\delta^{(\varepsilon,r)}$ δίνεται από

$$\begin{aligned}
R\left(\delta_0, \delta^{(\varepsilon,r)}\right) &= EL_{\sigma}(\delta_0, \sigma) - EL_{\sigma}\left(\delta^{(\varepsilon,r)}, \sigma\right) \\
&= E\left[\frac{1}{n-1}\frac{r}{(1+W)^{\varepsilon}}\,\frac{nS}{\sigma} + \ln\left(1 - \frac{r}{(1+W)^{\varepsilon}}\right)\right] \\
&= r\int_0^{\infty}\int_0^{\infty}\left\{\frac{1}{n-1}\frac{y}{(1+w)^{\varepsilon}}\right. \\
&\quad \left. + \frac{1}{r}\ln\left(1 - \frac{r}{(1+w)^{\varepsilon}}\right)\right\} yg(y)h(wy;\mu,\sigma)dydw \\
&\text{για} \qquad u = 1/(1+w), \qquad v = (1+w)y \\
&= r\int_0^{1}\int_0^{\infty}\left\{\frac{1}{n-1}u^{\varepsilon+1}v + \frac{1}{r}\ln(1 - ru^{\varepsilon})\right\} v\, g(uv)\, h\big((1 \\
&\quad - u)v;\mu,\sigma\big)dvdu \\
&\ge r\left\{\frac{1}{n-1}\int_0^{1} u^{\varepsilon+1}\int_0^{\infty} v^2\, g(uv)\, h\big((1-u)v;\mu,\sigma\big)dvdu\right. \\
&\quad \left. - \int_0^{1}\left(u^{\varepsilon} + \frac{ru^{2\varepsilon}}{2(1-r)}\right)\int_0^{\infty} v\, g(uv)\, h\big((1-u)v;\mu,\sigma\big)dvdu\right\}.
\end{aligned}$$

Για να είναι $R\left(\delta_0, \delta^{(\varepsilon,r)}\right) \ge 0$ αρκεί

$$r \le \left\{1 + \frac{1}{2}\int_0^1 u^{2\varepsilon}\int_0^\infty v\, g(uv)\, h\big((1-u)v;\mu,\sigma\big)dvdu \right.$$
$$\times \left[\frac{1}{n-1}\int_0^1 u^{\varepsilon+1}\int_0^\infty v^2\, g(uv)\, h\big((1-u)v;\mu,\sigma\big)dvdu \right.$$
$$\left.\left. - \int_0^1 u^{\varepsilon}\int_0^\infty v\, g(uv)\, h\big((1-u)v;\mu,\sigma\big)dvdu\right]^{-1}\right\}^{-1} = B(\varepsilon;\mu,\sigma),$$

δηλαδή αρκεί να δείξουμε ότι $B(\varepsilon;\mu,\sigma) \ge B(\varepsilon;\mu_0,\sigma) = B(\varepsilon)$. Η ισότητα στην τελευταία σχέση προκύπτει άμεσα με υπολογισμούς γάμμα και βήτα ολοκληρωμάτων, ενώ για τη διαδικασία απόδειξης της ανισότητας τα βασικά εργαλεία είναι η εφαρμογή του Λήμματος 6 και οι ιδιότητες (A1) και (A2).

Είναι

$$\frac{1}{B(\varepsilon;\mu,\sigma)} = 1 + \frac{1}{2\, I_1(\varepsilon;\mu,\sigma)[(1/(n-1))\, I_2(\varepsilon;\mu,\sigma) - 1]},$$

με $I_1(\varepsilon;\mu,\sigma)$ και $I_2(\varepsilon;\mu,\sigma)$ όπως αυτά αναλύονται στη συνέχεια.

Για το $I_1(\varepsilon;\mu,\sigma)$ έχουμε

$$I_1(\varepsilon;\mu,\sigma) = \frac{\int_0^1 u^{\varepsilon}\int_0^\infty v\, g(uv)\, h\big((1-u)v;\mu,\sigma\big)dvdu}{\int_0^1 u^{2\varepsilon}\int_0^\infty v\, g(uv)\, h\big((1-u)v;\mu,\sigma\big)dvdu} = E_\mu U^{-\varepsilon} \ge E_{\mu_0} U^{-\varepsilon}\,.$$

Η ανισότητα στην τελευταία σχέση ισχύει αφού για την πυκνότητα της U, $f(u;\mu,\sigma) \propto u^{2\varepsilon}\int_0^\infty v\, g(uv)\, h\big((1-u)v;\mu,\sigma\big)dv$, $0 < u < 1$, έχουμε ότι $f(u;\mu,\sigma)/\, f(u;\mu_0,\sigma)$ είναι φθίνουσα, διότι

$$\frac{f(u_1;\mu,\sigma)}{f(u_1;\mu_0,\sigma)} \propto \frac{\int_0^\infty v\, g(u_1 v)\, h\big((1-u_1)v;\mu,\sigma\big)dv}{\int_0^\infty v\, g(u_1 v)\, h\big((1-u_1)v\big)dv}$$
$$\propto \frac{\int_0^\infty y\, g\left(\frac{u_1}{1-u_1}y\right)\, h(y;\mu,\sigma)dy}{\int_0^\infty y\, g\left(\frac{u_1}{1-u_1}y\right)\, h(y)dy} \propto E_{u_1}\left[\frac{h(Y;\mu,\sigma)}{h(Y)}\right] \ge E_{u_2}\left[\frac{h(Y;\mu,\sigma)}{h(Y)}\right],$$

μιας και για την πυκνότητα της Y, $f_u(y) \propto y\, g\left(\frac{u}{1-u}y\right)\, h(y)$, $y > 0$, έχουμε ότι $f_{u_1}(y)/f_{u_2}(y)$ είναι αύξουσα για $0 < u_1 < u_2 < 1$.

Για το $I_2(\varepsilon;\mu,\sigma)$ έχουμε

$$I_2(\varepsilon;\mu,\sigma) = \frac{\int_0^1 u^{\varepsilon+1}\int_0^\infty v^2\, g(uv)\, h\big((1-u)v;\mu,\sigma\big)dvdu}{\int_0^1 u^{\varepsilon}\int_0^\infty v\, g(uv)\, h\big((1-u)v;\mu,\sigma\big)dvdu}$$
$$= \frac{\int_0^1 u^{\varepsilon-2}\int_0^\infty y^2\, g(y)\, h\left(\frac{(1-u)y}{u};\mu,\sigma\right)dydu}{\int_0^1 u^{\varepsilon-2}\int_0^\infty y\, g(y)\, h\left(\frac{(1-u)y}{u};\mu,\sigma\right)dydu}$$

$$= \frac{\int_0^\infty y^2\, g(y) \int_0^1 u^{\varepsilon-2}\, h\left(\frac{(1-u)y}{u}; \mu, \sigma\right) du dy}{\int_0^\infty y\, g(y) \int_0^1 u^{\varepsilon-2}\, h\left(\frac{(1-u)y}{u}; \mu, \sigma\right) du dy}$$
$$= \frac{\int_0^\infty y^{\varepsilon+1}\, g(y) \int_0^\infty (y+z)^{-\varepsilon}\, h(z; \mu, \sigma) dz dy}{\int_0^\infty y^{\varepsilon}\, g(y) \int_0^\infty (y+z)^{-\varepsilon}\, h(z; \mu, \sigma) dz dy} = E_\mu Y \geq E_{\mu_0} Y.$$

Η ανισότητα στην τελευταία σχέση ισχύει αφού για την πυκνότητα της Y, $f(y; \mu, \sigma) \propto y^\varepsilon\, g(y) \int_0^\infty (y+z)^{-\varepsilon}\, h(z; \mu, \sigma) dz,\ y > 0$, έχουμε ότι $f(y; \mu, \sigma) / f(y; \mu_0, \sigma)$ είναι αύξουσα, διότι

$$\frac{f(y_1; \mu, \sigma)}{f(y_1; \mu_0, \sigma)} \propto \frac{\int_0^\infty (y_1+z)^{-\varepsilon}\, h(z; \mu, \sigma) dz}{\int_0^\infty (y_1+z)^{-\varepsilon}\, h(z) dz} \propto E_{y_1}\left[\frac{h(Z; \mu, \sigma)}{h(Z)}\right] \leq E_{y_2}\left[\frac{h(Z; \mu, \sigma)}{h(Z)}\right]$$

μιας και για την πυκνότητα της Z, $f_y(z) \propto (y+z)^{-\varepsilon}\, h(z),\ z > 0$, έχουμε ότι $f_{y_2}(z) / f_{y_1}(z)$ είναι αύξουσα για $0 < y_1 < y_2$.

Τέλος, από την παραπάνω ανάλυση του $I_2(\varepsilon; \mu, \sigma)$ είδαμε ότι

$$I_2(\varepsilon; \mu_0, \sigma) = \frac{\int_0^\infty y^{\varepsilon+1}\, g(y) \int_0^\infty (y+z)^{-\varepsilon}\, h(z) dz dy}{\int_0^\infty y^{\varepsilon}\, g(y) \int_0^\infty (y+z)^{-\varepsilon}\, h(z) dz dy} = EY,$$

όπου Y έχει πυκνότητα $f(y; \mu_0, \sigma) \propto y^\varepsilon\, g(y) \int_0^\infty (y+z)^{-\varepsilon}\, h(z) dz, y > 0$, και επειδή $f(y; \mu_0, \sigma) / g(y) \propto \int_0^\infty (y/(y+z))^\varepsilon\, h(z) dz$ είναι προφανώς γνησίως αύξουσα ως προς $y > 0$, το Λήμμα 5 δίνει $I_2(\varepsilon; \mu_0, \sigma) = EY > E_g Y = n - 1$, και αυτό ολοκληρώνει την απόδειξη.

Στον Πίνακα 2 παρέχονται ενδεικτικά αριθμητικά αποτελέσματα από 10.000 Monte Carlo προσομοιώσεις σχετικά με τη μέση ζημία των εκτιμητών δ_0, δ_{φ_S}, $\delta_{\varphi_{BZ}}$, $\delta^{(\varepsilon,r)}$, με μεγέθη δείγματος $n = 4,10$, $\mu = 2$, $\sigma = 0.01, 0.02, 0.05, 0.1, 0.2, 0.4, 0.6,\ 1$, $\mu_0 = 1$, $\varepsilon = 1$, και $r = 0.2105, 0.0388$ ($\approx$ το άνω φράγμα $B(1)$) για τα αντίστοιχα δειγματικά μεγέθη. Από τη σύγκριση αυτών των αποτελεσμάτων φαίνεται ότι ο $\delta^{(1,r)}$ έχει τη μικρότερη μέση ζημία για μικρές τιμές του σ. Σημειώνεται ότι όταν το μέγεθος δείγματος n αυξάνεται, είναι αναμενόμενο να μειώνεται η βελτίωση της μέσης ζημίας καθενός από τους εκτιμητές δ_{φ_S}, $\delta_{\varphi_{BZ}}$, $\delta^{(\varepsilon,r)}$ ως προς τον δ_0, διότι για μεγάλες τιμές του n αυτοί οι εκτιμητές τείνουν να ταυτιστούν με τον δ_0. Επιπλέον, ανάλογο συμπέρασμα προκύπτει για τον εκτιμητή $\delta^{(\varepsilon,r)}$ και για μεγάλες τιμές του ε.

Πίνακας 2. *Αποτελέσματα προσομοίωσης μέσης ζημίας εκτιμητών του σ*

σ	δ_0	δ_{φ_S}	$\delta_{\varphi_{BZ}}$	$\delta^{(1,r)}$
	$n = 4$			
0.01	0.3690	0.3690	0.3681	0.3669

0.02	0.3793	0.3793	0.3766	0.3750
0.05	0.3725	0.3725	0.3641	0.3632
0.1	0.3626	0.3625	0.3466	0.3477
0.2	0.3743	0.3696	0.3538	0.3567
0.4	0.3763	0.3679	0.3601	0.3625
0.6	0.3720	0.3612	0.3579	0.3597
1	0.3667	0.3535	0.3545	0.3556
	$n = 10$			
0.01	0.1133	0.1133	0.1133	0.1131
0.02	0.1187	0.1187	0.1187	0.1184
0.05	0.1140	0.1140	0.1140	0.1134
0.1	0.1144	0.1144	0.1142	0.1133
0.2	0.1155	0.1155	0.1143	0.1139
0.4	0.1153	0.1152	0.1129	0.1145
0.6	0.1160	0.1152	0.1129	0.1145
1	0.1165	0.1152	0.1136	0.1153

ABSTRACT

This work deals with the estimation of the mean μ and the parameter $1/\lambda$ of an inverse Gaussian distribution $IG(\mu, \lambda)$ from a decision theoretic point of view. New classes of improved estimators are constructed and comparative simulation results regarding selected members of these classes and estimators available in the literature are presented.

ΑΝΑΦΟΡΕΣ

Bobotas, P. and Kourouklis, S. (2009). Strawderman-type estimators for a scale parameter with application to the exponential distribution. *Journal of Statistical Planning and Inference*, **139**, 3001--3012.

Brewster, J.F. and Zidek, J.V. (1974). Improving on equivariant estimators. *Annals of Statistics*, **2**, 21-38.

Chhikara, R.S. and Folks, J.L. (1989). *The inverse Gaussian distribution: Theory, Methodology and Applications,* New York: Marcel Dekker.

Hsieh, H.K., Korwar, R.M. and Rukhin, A.L. (1990). Inadmissibility of the maximum likelihood estimator of the inverse Gaussian mean. *Statistics and Probability Letters*, **9**, 83-90.

Jorgensen, B. (1982). *Statistical properties of the generalized inverse Gaussian distribution*. Lecture notes in Statistics, Vol. 9, New York: Springer.

Kourouklis, S. (1997). A new property of the inverse Gaussian distribution with applications. *Statistics and Probability Letters*, **32**, 161-166.

Kubokawa, T. (1994). A unified approach to improving equivariant estimators. *Annals of Statistics*, **22**, 290-299.

Lehmann, E.L. and Romano, J.P. (2005). *Testing Statistical Hypotheses*, New York: Springer.

Ma, T., Liu, S. and Ahmed, S.E. (2014). Shrinkage estimation for the mean of the inverse Gaussian population. *Metrika*, **77**, 733-752.

Schrödinger, V.E. (1915). Zur Theorie der Fall-und Steigversuche und Teilchen mit Brownscher Bewegung, *Physics Z,* **16**, 289-295.

Seshardi, V. (1993). *The inverse Gaussian distribution: A case study in exponential families*, Oxford: Oxford University Press.

Shuster, J. (1968). On the inverse Gaussian distribution function. *Journal of the American Statistical Association*, **63**, 1514-1516.

Stein, C. (1964). Inadmissibility of the usual estimator for the variance of a normal distribution with unknown mean. *Annals of the Institute of Statistical Mathematics*, **16**, 155-160.

Strawderman, W. (1974). Minimax estimation of powers of the variance of a normal population under squared error loss. *Annals of Statistics*, **2**, 190-198.

Tweedie, M.C.K. (1945). Inverse statistical variates. *Nature,* **155**, 543.

Tweedie, M.C.K. (1957a). Statistical properties of inverse Gaussian distributions I. *Annals of Mathematical Statistics*, **28**, 362-377.

Tweedie, M.C.K. (1957b). Statistical properties of inverse Gaussian distributions II. *Annals of Mathematical Statistics*, **28**, 696-705.

Wald, A. (1945). Sequential tests of statistical hypotheses, *Annals of Mathematical Statistics*, **16**, 117-186.

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 35ου Πανελληνίου & 1ου Διεθνούς Συνεδρίου Στατιστικής (2023), σελ.58-73

ΕΝΑ ΕΠΕΚΤΕΤΑΜΕΝΟ ΕΠΙΔΗΜΙΟΛΟΓΙΚΟ ΦΙΛΤΡΟ ΣΩΜΑΤΙΔΙΩΝ ΓΙΑ ΤΗΝ ΠΕΡΙΓΡΑΦΗ ΤΗΣ ΜΕΤΑΔΟΣΗΣ ΜΟΛΥΣΜΑΤΙΚΩΝ ΑΣΘΕΝΕΙΩΝ. ΕΦΑΡΜΟΓΗ ΣΤΑ ΔΕΔΟΜΕΝΑ COVID-19 ΣΤΗΝ ΙΤΑΛΙΑ

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

ΠΕΡΙΛΗΨΗ

Στο παρόν άρθρο, προτείνεται ένα επεκτεταμένο επιδημιολογικό φίλτρο σωματιδίων με χρονικά μεταβαλλόμενες παραμέτρους, με στόχο την περιγραφή της εξέλιξης πανδημιών. Η εγκυρότητα του μοντέλου εξετάζεται στις ημερήσιες καταγραφές COVID-19 στην Ιταλία για το διάστημα των 500 ημερών. Τα κύρια ευρήματα περιλαμβάνουν την εκτίμηση του ποσοστού ασυμπτωματικών φορέων, που αποτελεί γνωστό χαρακτηριστικό της COVID-19. Σε αντίθεση με άλλα μοντέλα που χρησιμοποιούν επιπλέον καταστάσεις για τους ασυμπτωματικούς φορείς, εξαναγκάζοντας την εκτίμηση αυτής της αναλογίας και αυξάνοντας την πολυπλοκότητα του μοντέλου, η προτεινόμενη προσέγγιση οδηγεί στην ανάδειξη της κρυφής δυναμικής της COVID-19 χωρίς πρόσθετη υπολογιστική επιβάρυνση. Επιπλέον ευρήματα που επιβεβαιώνουν την καταλληλότητα του μοντέλου, είναι η εξέλιξη των χρονικά μεταβαλλόμενων παραμέτρων καθώς και η προκύπτουσα διαφοροποίηση στα περιστατικά εισαγωγών σε ΜΕΘ συγκριτικά με τις επίσημες καταγραφές, κατά τη διάρκεια του ισχυρότερου κύματος κρουσμάτων τον Ιανουαρίου του 2022. Τέλος, προτείνεται στατιστικός αλγόριθμος για την εκτίμηση του πλήθους των επί του παρόντος αναρρωμένων και προστατευμένων μέσω του εμβολιασμού.

Λέξεις Κλειδιά: Στοχαστική Μοντελοποίηση, Φίλτρο Σωματιδίων, Δυναμική Εκτίμηση Παραμέτρων, Κατανομές Πιθανοτήτων, COVID-19

1. ΕΙΣΑΓΩΓΗ

Η εξάπλωση μολυσματικών ασθενειών περιγράφεται συχνά μέσω ντετερμινιστικών μοντέλων, με το μοντέλο SIR να αποτελεί τη δημοφιλέστερη προσέγγιση. Ωστόσο, οι προσεγγίσεις αυτές συχνά αποτυγχάνουν να

αντιμετωπίσουν τη μεταβαλλόμενη δυναμική των επιδημιών, ειδικά για μεγάλες χρονικές περιόδους. Δεδομένου ότι οι ασυμπτωματικές λοιμώξεις είναι δύσκολο να ανιχνευθούν, σε συνδυασμό με τα ποσοστά ψευδώς θετικών/αρνητικών διαγνωστικών τεστ, αντιλαμβανόμαστε την ύπαρξη θορύβου στις ημερήσιες καταγραφές. Επομένως, η μετάβαση από την ντετερμινιστική στη στοχαστική θεώρηση καθίσταται απαραίτητη.

Οι Costa et al. (2005) συνδυάζουν το μοντέλο SEIR με το εκτεταμένο φίλτρο Kalman (EKF) για την περιγραφή επιδημιών, ενώ οι Zhu et al. (2021) περιλαμβάνουν στο συνδυασμό SEIRD-EKF την εκτίμηση των παραμέτρων του επιδημιολογικού μοντέλου. Οι Papageorgiou & Tsaklidis (2023) προτείνουν ένα εκτεταμένο SEIRS μοντέλο, συνδυασμένο με το unscented φίλτρο Kalman για την πρόβλεψη της εξέλιξη της COVID-19 στη Γαλλία. Οι Calvetti et al. (2021), χρησιμοποιούν το μοντέλο SE(A)IR βασισμένο στο φίλτρο σωματιδίων για την εκτίμηση των πρώιμων σταδίων εξάπλωσης του COVID στο Οχάιο και το Μίσιγκαν.

Στο παρόν άρθρο παρουσιάζεται ένα νέο, επεκτεταμένο SEIRS μοντέλο με επαναλαμβανόμενους εμβολιασμούς, σε συνδυασμό με τη μέθοδο του φίλτρου σωματιδίων με δυναμική εκτίμηση παραμέτρων. Η στοχαστική προσέγγιση, μαζί με την εκτίμηση των χρονικά μεταβαλλόμενων παραμέτρων, ξεπερνά τους περιορισμούς των προσδιοριστικών προσεγγίσεων, ενώ παρέχει πολύτιμα συμπεράσματα αναφορικά με την κρυφή δυναμική των επιδημιών. Η εγκυρότητα των εκτιμήσεων εξετάστηκε στις ημερήσιες καταγραφές COVID-19 στην Ιταλία για περίοδο 500 ημερών από την έναρξη των εμβολιασμών.

Τα κυριότερα ευρήματα αφορούν την εκτίμηση υψηλότερου αριθμού κρουσμάτων συγκριτικά με τις ημερήσιες καταγραφές, γεγονός που μπορεί να αποδοθεί στις ασυμπτωματικές μολύνσεις που χαρακτηρίζουν την εξάπλωση του COVID. Το αντίστοιχο εκτιμώμενο ποσοστό κυμαίνεται μεταξύ 20 – 50%, εκτίμηση που υποστηρίζεται από την υπάρχουσα βιβλιογραφία. Εξίσου σημαντική είναι η εκτίμηση μεγαλύτερου πλήθους εισαχθέντων σε ΜΕΘ κατά τη διάρκεια του ισχυρότερου κύματος κρουσμάτων κατά τον Ιανουαρίου του 2022. Η έντονη μεταδοτικότητα σε σύντομο χρονικό διάστημα, πιθανότατα άσκησε έντονη πίεση στο εθνικό σύστημα υγείας, οδηγώντας σε πληρότητα στις ΜΕΘ που πιθανόν να μην επέτρεψε την εισαγωγή όλων των σοβαρών περιπτώσεων.

Το προτεινόμενο μοντέλο παρέχει μια πιο αντιπροσωπευτική εικόνα της εξέλιξης πανδημιών, λόγω του αυξημένου αριθμού καταστάσεων και μεταβάσεων. Τέλος, προτείνεται στατιστικός αλγόριθμος για την εκτίμηση των επί του παρόντος αναρρωμένων και προστατευμένων μέσω του εμβολιασμού, μέσω των χρονοσειρών του συνολικού πλήθους αναρρωμένων και εμβολιασμένων. Ο αλγόριθμος χρησιμοποιεί ένα σύνολο βαρών προερχόμενα

από κανονικές κατανομές, παρέχοντας μια διόρθωση για τις προαναφερθείσες χρονοσειρές.

2. ΜΕΘΟΔΟΛΟΓΙΑ

2.1. Το επεκτεταμένο SEIRS μοντέλο με επαναλαμβανόμενους εμβολιασμούς

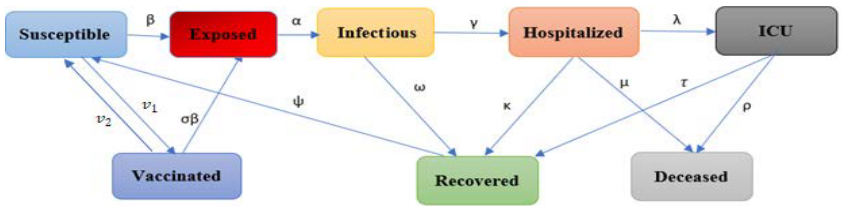
Σε αυτή την ενότητα, παρουσιάζεται η δομή του εκτεταμένου SEIRS μοντέλου, που πρόκειται να συνδυαστεί με τη μεθοδολογία του φίλτρου σωματιδίων για τη διερεύνηση της εξάπλωσης του COVID στην Ιταλία. Η προτεινόμενη επέκταση του τυπικού μοντέλου SEIRS, περιλαμβάνει επιπρόσθετα τις καταστάσεις (και τις σχετικές διαφορικές εξισώσεις) που αντιστοιχούν στους νοσηλευόμενους σε νοσοκομεία, ΜΕΘ, τους αποθανόντες και τους πλήρως εμβολιασμένους. Το προτεινόμενο σχήμα, επιτρέπει επίσης τη μετάβαση αναρρωμένων και εμβολιασμένων πίσω στην κατάσταση ευάλωτος, χαρακτηριστικό που είναι απαραίτητο για την περιγραφή επιδημιών με εξασθενούμενη ανοσία για μεγάλες χρονικές περιόδους. Στο Σχήμα 1 αναπαρίστανται οι μεταβάσεις μεταξύ των καταστάσεων του προτεινόμενου επιδημιολογικού μοντέλου. Η επιλογή των παρουσιαζόμενων καταστάσεων και μεταβάσεων έγινε με στόχο την περιγραφή του πλήρους φάσματος του εξεταζόμενου φαινομένου χωρίς να αυξηθεί σημαντικά η πολυπλοκότητα του μοντέλου.

$$
\begin{aligned}
\frac{dS(t)}{dt} &= \Lambda - \frac{\beta S(t) I(t)}{N} - v_1 S(t) + \psi R(t) + v_2 V(t) - \delta S(t) \\
\frac{dE(t)}{dt} &= \frac{\beta S(t) I(t)}{N} + \frac{\sigma \beta V(t) I(t)}{N} - aE(t) - \delta E(t) \\
\frac{dI(t)}{dt} &= aE(t) - \gamma I(t) - \omega I(t) - \delta I(t) \\
\frac{dH(t)}{dt} &= \gamma I(t) - \lambda H(t) - \kappa H(t) - \mu H(t) - \delta H(t) \\
\frac{dC(t)}{dt} &= \lambda H(t) - \tau C(t) - \rho C(t) - \delta C(t) \\
\frac{dR(t)}{dt} &= \omega I(t) + \kappa H(t) + \tau C(t) - \psi R(t) - \delta R(t) \\
\frac{dD(t)}{dt} &= \mu H(t) + \rho C(t) \\
\frac{dV(t)}{dt} &= v_1 S(t) - \frac{\sigma \beta V(t) I(t)}{N} - v_2 V(t) - \delta V(t)
\end{aligned}
\tag{1}
$$

Από το σύστημα (1), μπορεί να προσδιοριστεί το ισοζύγιο απουσίας της νόσου από τον πληθυσμό $\boldsymbol{Y}^0$. Για τον υπολογισμό του ισοζυγίου, θέτουμε $I^0 = 0$, εφόσον απαιτείται απουσία μολυσματικών φορών που να μεταδίδουν τη νόσο. Εξισώνοντας τις ροές του συστήματος (1) με το μηδέν, καταλήγουμε στο

$\boldsymbol{Y}^0 = (S^0, 0,0,0,0,0, V^0) = \left(\frac{\Lambda(v_2+\delta)}{\delta(v_1+v_2+\delta)}, 0,0,0,0,0, \frac{\Lambda v_1}{\delta(v_1+v_2+\delta)}\right)$. Για τον προσδιορισμό του $\boldsymbol{Y}^0$, μπορούμε να αφαιρέσουμε τον αριθμό των αποθανόντων $D(t)$, καθώς η ποσότητα αυτή δεν συμμετέχει στις υπόλοιπες εξισώσεις του συστήματος (1), ενώ μπορεί να περιγραφεί ως γραμμικός συνδυασμός των 7 υπολοίπων καταστάσεων. Στη συνέχεια, προχωράμε στην παρουσίαση 2 θεωρημάτων που σχετίζονται με τον βασικό αναπαραγωγικό ρυθμό R_0 και την ύπαρξη ενδημικού ισοζυγίου. Η παράθεση των σχετικών αποδείξεων παραλείπεται, ενώ ο ενδιαφερόμενος αναγνώστης παραπέμπεται στη μεθοδολογία των αποδείξεων των θεωρημάτων 2 και 4 των Papageorgiou & Tsaklidis (2023).

Θεώρημα 1. Ο βασικός αναπαραγωγικός ρυθμός με βάση το προτεινόμενο επεκτεταμένο επιδημιολογικό μοντέλο δίνεται από τον τύπο $R_0 = \frac{\beta\alpha(S^0+\sigma V^0)}{N(\alpha+\delta)(\gamma+\omega+\delta)} = \frac{\alpha\beta\Lambda(\sigma v_1+v_2+\delta)}{N\delta(\alpha+\delta)(\gamma+\omega+\delta)(v_1+v_2+\delta)}$.



Σχήμα 1. *Διαγραμματική αναπαράσταση του προτεινόμενου επιδημιολογικού μοντέλου*

Θεώρημα 2. Το επεκτεταμένο SEIRS έχει μοναδικό ενδημικό ισοζύγιο $\boldsymbol{Y}^*$ όταν $R_0 > 1$.

Πίνακας 1. *Ορισμός παραμέτρων και καταστάσεων του προτεινόμενου επεκτεταμένου SEIRS μοντέλου*

Σύμβολο	Ορισμός Παραμέτρων/Καταστάσεων
S	Ευάλωτοι (Susceptible)
E	Εκτεθειμένοι (Exposed)
I	Μολυσματικοί (Infectious)
H	Νοσηλευόμενοι σε Νοσοκομεία (Hospitalized)
C	Νοσηλευόμενοι σε ΜΕΘ (ICU admitted)
R	Αναρρωμένοι (Recovered)
D	Αποθανόντες (Deceased)
V	Εμβολιασμένοι (Vaccinated)
Λ	Ρυθμός Γεννήσεων και Μετανάστευσης
α	Ρυθμός Επώασης
β	Ρυθμός Μόλυνσης/Μετάδοσης
γ	Ρυθμός Εισαγωγών σε Νοσοκομεία

δ	Ρυθμός Θνησιμότητας από άλλα αίτια
λ	Ρυθμός Μετάβασης από το Νοσοκομείο στη ΜΕΘ
κ	Ρυθμός Ανάρρωσης μετά από Νοσηλεία σε Νοσοκομείο
μ	Ρυθμός Θνησιμότητας μετά από Νοσηλεία σε Νοσοκομείο
ν_1	Ρυθμός Εμφάνισης Νέων Πλήρως Εμβολιασμένων Ατόμων
ν_2	Ρυθμός Εξασθένισης Ανοσίας Προερχόμενης από Εμβολιασμό
ρ	Ρυθμός Θνησιμότητας μετά από Νοσηλεία σε ΜΕΘ
σ	Πιθανότητα Μόλυνσης Εμβολιασμένων
τ	Ρυθμός Ανάρρωσης μετά από Νοσηλεία σε ΜΕΘ
ψ	Ρυθμός Μετάβασης από Αναρρωμένος σε Ευάλωτος
ω	Ρυθμός Ανάρρωσης Κρουσμάτων με Ήπια Συμπτώματα

Θεώρημα 3. Το ισοζύγιο $\boldsymbol{Y}^0$ είναι ολικά ασυμπτωτικά ευσταθές αν-ν $R_0 < 1$.

Απόδειξη. Σύμφωνα με την αρχή αναλλοίωτου του LaSalle, επιλέγουμε τη συνάρτηση Lyapunov L, η οποία είναι θετικά ημιορισμένη στο σύνολο $\Omega = \{(S,E,I,H,C,R,D,V) \mid S,E,I,H,C,R,D,V \geq 0\} = \mathbb{R}_+^8$. Ορίζουμε τη συνάρτηση $L = \frac{1}{2}[(S-S^0)^2 + E^2 + I^2 + H^2 + C^2 + R^2 + (V-V^0)^2] = \frac{1}{2}\boldsymbol{X}^T\boldsymbol{X} > 0,$ με διάνυσμα $\boldsymbol{X} = [S-S^0\ E-E^0\ I-I^0\ H-H^0\ C-C^0\ R-R^0\ V-V^0]^T = [S-S^0\ E\ I\ H\ C\ R\ V-V^0]^T$, εκτός από την περίπτωση όπου $L(\boldsymbol{Y}^0) = 0.$ εκτός από την περίπτωση όπου $L\left(\boldsymbol{Y}^0\right) = 0.$ Παραγωγίζοντας την L ως προς t, παίρνουμε

$$\begin{aligned}
\frac{dL}{dt} &= (S-S^0)\frac{dS}{dt} + E\frac{dE}{dt} + I\frac{dI}{dt} + H\frac{dH}{dt} + C\frac{dC}{dt} + R\frac{dR}{dt} + (V-V^0)\frac{dV}{dt} \\
&= (S-S^0)[\Lambda - \beta_1(S-S^0)I - \beta_1 S^0 I + \psi R - (\nu_1+\delta)(S-S^0) + \nu_2(V-V^0) - (\nu_1+\delta)S^0 + \nu_2 V^0] \\
&\quad + E[-(\alpha+\delta)E + \beta_1(S-S^0)I + \beta S^0 I + \sigma\beta_1(V-V^0)I + \sigma\beta_1 V^0 I] \quad + I[\alpha E - (\gamma+\omega+\delta)I] \\
&\quad + H[\gamma I - (\lambda+\kappa+\mu+\delta)H] + C[\lambda H - (\tau+\rho+\delta)C] + R[\omega I + \kappa H + \tau C - (\psi+\delta)R] \\
&\quad + (V-V^0)[\nu_1(S-S^0) - (\nu_2+\delta)(V-V^0) - \sigma\beta_1 I(V-V^0) - \sigma\beta_1 V^0 I + \nu_1 S^0 - (\nu_2+\delta)V^0] \\
&= -\beta_1(S-S^0)^2 I - \beta_1 S^0(S-S^0)I + \psi(S-S^0)R - (\nu_1+\delta)(S-S^0)^2 + \nu_2(S-S^0)(V-V^0) \\
&\quad - (\alpha+\delta)E^2 + \beta_1 IE(S-S^0) + \beta S^0 EI + \sigma\beta_1 IE(V-V^0) + \sigma\beta_1 V^0 EI + aIE - (\gamma+\omega+\delta)I^2 + \gamma HI \\
&\quad - (\lambda+\kappa+\mu+\delta)H^2 + \lambda CH - (\tau+\rho+\delta)C^2 + \omega RI + \kappa RH + \tau RC - (\psi+\delta)R^2 + \nu_1(V-V^0)(S-S^0) \\
&\quad - (\nu_2+\delta)(V-V^0)^2 - \sigma\beta_1(V-V^0)^2 I - \sigma\beta_1 V^0(V-V^0)I = \boldsymbol{X^T Q X} = \boldsymbol{X}^T\boldsymbol{Q}_1\boldsymbol{X} + \boldsymbol{X^T Q}_2\boldsymbol{X}\,, \qquad (2)
\end{aligned}$$

με

$$\boldsymbol{Q}_1 = \begin{pmatrix}
-(\nu_1+\delta) & 0 & -\beta_1 S^0 & 0 & 0 & \psi & \nu_2 \\
0 & -(\alpha+\delta) & \beta_1(S^0+\sigma V^0) & 0 & 0 & 0 & 0 \\
0 & \alpha & -(\gamma+\omega+\delta) & 0 & 0 & 0 & 0 \\
0 & 0 & \gamma & -(\lambda+\kappa+\mu+\delta) & 0 & 0 & 0 \\
0 & 0 & 0 & \lambda & -(\tau+\rho+\delta) & 0 & 0 \\
0 & 0 & \omega & \kappa & \tau & -(\psi+\delta) & 0 \\
\nu_1 & 0 & -\sigma\beta_1 V^0 & 0 & 0 & 0 & -(\nu_2+\delta)
\end{pmatrix}, \qquad (3)$$

$$\boldsymbol{Q}_2(I) = \begin{pmatrix}
-\beta_1 I & 0 & 0 & 0 & 0 & 0 & 0 \\
\beta_1 I & 0 & 0 & 0 & 0 & 0 & \sigma\beta_1 I \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & -\sigma\beta_1 I
\end{pmatrix}, \qquad (4)$$

όπου ο $\boldsymbol{Q}_2(I)$ είναι πολυωνυμικός πίνακας. Σημειώνουμε ότι για το δεξί μέλος της 2ης εξίσωσης της σχέσης (2) ισχύει $\Lambda - (\nu_1+\delta)S^0 + \nu_2 V^0 = 0$ και

$v_1 S^0 - (v_2 + \delta)V^0 = 0$, ενώ $E^0 = I^0 = H^0 = C^0 = R^0 = 0$. Το χαρακτηριστικό πολυώνυμο του πίνακα $\boldsymbol{Q}_1$ είναι το

$$p(l) = \det(l\boldsymbol{I}_{7x7} - \boldsymbol{Q}_1) = (l+\delta)(l+\lambda+\kappa+\mu+\delta)(l+v_1+v_2+\delta)(l+\tau+\rho+\delta)(l+\delta+\psi)$$
$$(l^2 + (a+\gamma+\omega+2\delta)l + \delta(a+\gamma+\delta+\omega) + \alpha(\gamma+\omega-\beta_1 S^0 - \sigma\beta_1 V^0)) = 0, \quad (5)$$

οδηγώντας σε 5 αρνητικές ιδιοτιμές. Για το τετραγωνικό πολυώνυμο της σχέσης (5), εφαρμόζουμε το κριτήριο Routh-Hurwitz 2ου βαθμού, όπου οι ρίζες του πολυωνύμου λαμβάνουν χώρα στο αριστερό μιγαδικό ημιεπίπεδο όταν οι συντελεστές $(a+\gamma+\omega+2\delta)$ και $\delta(a+\gamma+\delta+\omega) + \alpha(\gamma+\omega-\beta_1 S^0 - \sigma\beta_1 V^0)$ είναι θετικοί. Καθώς $(a+\gamma+\omega+2\delta) > 0$, το κριτήριο ικανοποιείται αν και μόνο αν $\delta(a+\gamma+\delta+\omega) + \alpha(\gamma+\omega-\beta_1 S^0 - \sigma\beta_1 V^0) > 0$, όπου έπειτα από πράξεις καταλήγουμε στο ότι $R_0 < 1$. Οι ιδιοτιμές του πίνακα $\boldsymbol{Q}_2(I)$ είναι όλες μη θετικές καθώς ισχύει $I(t) > 0 \ \forall\, t \geq 0$. Επίσης, όλες οι ιδιοτιμές του $\boldsymbol{Q}_1$ έχουν αρνητικό πραγματικό μέρος αν και μόνο αν $R_0 < 1$. Σύμφωνα με τα παραπάνω έχουμε $\frac{dL}{dt} = \boldsymbol{X}^T \boldsymbol{Q} \boldsymbol{X} < 0$ αν και μόνο αν $R_0 < 1$. Με παρόμοιο τρόπο μπορεί να αποδειχτεί ότι το ενδημικό ισοζύγιο είναι ολικά ευσταθές αν-ν $R_0 > 1$. ■

2.2 Αλγόριθμος εκτίμησης των επί του παρόντος αναρρωμένων ή εμβολιασμένων

Η πλειοψηφία συνόλων δεδομένων περιέχει το αθροιστικό πλήθος αναρρωμένων. Ωστόσο, κατά την παρούσα ανάλυση, χρειαζόμαστε το επί του παρόντος πλήθος αναρρωμένων κρουσμάτων. Η διαφορά αυτή, οφείλεται στην εξασθένιση των αντισωμάτων που παράγονται από μια πιθανή μόλυνση με την πάροδο του χρόνου, ενώ ο ιός συνεχώς μεταλλάσσεται. Έτσι, δημιουργούμε τη χρονοσειρά των νέων αναρρώσεων $\{n_{rec_t}, t \geq 0\}$, παίρνοντας τις πρώτες διαφορές των αθροιστικών αναρρώσεων. Στη συνέχεια, επιλέγουμε τη στιγμή $t-k$ και προσαρμόζουμε κανονική κατανομή με μέσο όρο $\mu = t-k$. Λαμβάνοντας υπόψη ότι η ανοσία διαρκεί κατά κανόνα $k \pm c$ ημέρες, υπολογίζουμε την πιθανοφάνεια για κάθε $i = t-k-c, \dots, t-k+c$ με βάση την κανονική κατανομή, καταλήγοντας σε ένα σύνολο $2c+1$ πλήθους τιμών. Κανονικοποιούμε τις παραπάνω τιμές και κατασκευάζουμε ένα διάνυσμα $2c+1$ όρων στάθμισης $\boldsymbol{w}$. Στη συνέχεια, παίρνουμε το εσωτερικό γινόμενο του $\boldsymbol{w}$ με το τις νέες αναρρώσεις για $i = t-k-c, \dots, t-k+c$. Δεδομένου ότι πρέπει $Currently_{R_t} \in \mathbb{Z}^+ \ \ \forall t \in \mathbb{N}$, αφαιρούμε από τις αθροιστικές αναρρώσεις το παραπάνω εσωτερικό γινόμενο,

$$Currently_{R_t} = Cumulative_{R_t} - \left\lfloor < \boldsymbol{w}, \left(n_{rec_{t-k-c}}, \dots, n_{rec_{t-k+c}}\right) > \right\rfloor. \quad (6)$$

Η επιλογή της κανονικής κατανομής δεν είναι αυθαίρετη. Οι νέες αναρρώσεις σε κάθε χρονικό βήμα μπορούν να θεωρηθούν ως αφίξεις μιας διαδικασίας Poisson (Al-Dousari 2021). Για μεγάλο λ, η κατανομή Poisson μπορεί να προσεγγιστεί ικανοποιητικά από τη γκαουσιανή.

Ομοίως, μπορεί να υπολογιστεί ο αριθμός των ατόμων που είναι επί του παρόντος προστατευμένοι μέσω του εμβολιασμού. Η μόνη διαφορά στην εκτίμηση των επί του παρόντος προστατευμένων έγκειται στην προσθήκη του αθροιστικού πλήθους αναμνηστικών δόσεων, χαρακτηριστικό σε ασθένειες όπως ο COVID, η γρίπη, ο κίτρινος πυρετός κ.α. Έτσι έχουμε,

$$Currently_{V_t} = Cumulative_{V_t} - \left|< \boldsymbol{w}, \left(n_{vac_{t-k-c}}, \dots, n_{vac_{t-k+c}}\right) >\right| + booster_t\,, \qquad (7)$$

όπου με $Currently_{V_t}$ συμβολίζουμε τον πλήθος των ατόμων που εξακολουθούν να έχουν ανοσία λόγω εμβολιασμού, ενώ η χρονοσειρά $\{n_{vac_t},\ t \geq 0\}$ αντιστοιχεί στον ημερήσιο αριθμό νέων, πλήρως εμβολιασμένων ατόμων.

Αλγόριθμος 1 Εκτίμηση του πλήθους επί του παρόντος αναρρωμένων

Απαιτούνται: $\{Cumulative_{R_t}, t \geq 0\}, k, c, \sigma$
  Για $i = 0, \dots, N-1$ **κάνε**
    $n_{rec_i} = Cumulative_{R_{i+1}} - Cumulative_{R_i}$
  Τέλος
  Επέλεξε $t > k - c$
  Για $i = 0, \dots, 2c$ **κάνε**
    $p_i = P(i + (t-k-c) - 0.5 < X_i < i + (t-k-c) + 0.5)$ όπου $X_i \sim N(t-k, \sigma)$
  Τέλος
  Για $i = 0, \dots, 2c$ **κάνε**
    Κανονικοποίηση των βαρών $w_i = p_i / \sum_{i=0}^{2c} p_i$
  Τέλος
  Συγκέντρωσε τα w_i στο διάνυσμα $\boldsymbol{w}$
  Για $j = 0, \dots, N-1$ **κάνε**
    αν $j < t$
      $Currently_{R_j} = Cumulative_{R_j}$
    αλλιώς
      $Currently_{R_j} = Cumulative_{R_j} - \left|< \boldsymbol{w}, \left(n_{rec_{j-k-c}}, \dots, n_{rec_{j-k+c}}\right) >\right|$
  Τέλος

2.3 Το μεικτό επιδημιολογικό SEIRS φίλτρο σωματιδίων

Η εξάπλωση μολυσματικών ασθενειών συνήθως συνοδεύεται από παραμέτρους με υψηλή μεταβλητότητα, όπως ο ρυθμός μετάδοσης β που επηρεάζεται από εξωτερικούς παράγοντες. Άλλα παραδείγματα είναι τα ποσοστά θνησιμότητας και ανάρρωσης των κρουσμάτων. Τα προτεινόμενα ντετερμινιστικά μοντέλα δεν λαμβάνουν υπόψη τη στοχαστικότητα των

επιδημιών, ενώ εμφανίζουν δυσκολίες στη διαχείριση της μεταβαλλόμενης συμπεριφοράς τους. Στόχος είναι η ενσωμάτωση στο προτεινόμενο επεκτεταμένο SEIRS, οποιασδήποτε διαθέσιμης πληροφορίας σχετικά με την εξέλιξη της πανδημίας, όπως το πλήθος μολυσματικών, αναρρωμένων, εμβολιασμένων, αποθανόντων, νοσηλευόμενων σε νοσοκομεία και ΜΕΘ.

Στόχος είναι η προσαρμογή της μεθοδολογίας του φίλτρου σωματιδίων στις ιδιαιτερότητες της εξάπλωσης επιδημιών, κατασκευάζοντας ένα υβριδικό επιδημιολογικό σύστημα που μπορεί να συνδυάσει αποτελεσματικά τα στοιχεία του επεκτεταμένου SEIRS με τις διαθέσιμες παρατηρήσεις, αντιμετωπίζοντας τις αδυναμίες τόσο των παρατηρήσεων όσο και των επιδημιολογικών μοντέλων. Η συγχώνευση των δύο μεθοδολογιών απαιτεί την τροποποίηση του ντετερμινιστικού εκτεταμένου SEIRS μοντέλου καταλήγοντας σε μια ισοδύναμη στοχαστική προσέγγιση. Αρχικά, πρέπει να καθοριστεί η προτεινόμενη κατανομή μεταβάσεων ενός βήματος μεταξύ των κρυφών καταστάσεων $f_{\boldsymbol{\theta}}(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$. Εφόσον εργαζόμαστε με υποπληθυσμούς, χρησιμοποιούμε διακριτές κατανομές για τις μεταβάσεις μεταξύ καταστάσεων. Για τις καταστάσεις των εκτεθειμένων και των αναρρωμένων, επιλέγουμε

$$(d_{EI})_t \sim B(E_{t-1}, 1-e^{-a\,dt}) \qquad \kappa\alpha\iota \qquad (d_{RS})_t \sim B\left(R_{t-1}, 1-e^{-\psi\,dt}\right), \tag{8}$$

όπου το $1-e^{-a\,dt}$ αντιπροσωπεύει την πιθανότητα ενός ατόμου να μετακινηθεί από την κατάσταση εκτεθειμένος στην μολυσματικός, κατά τη διάρκεια του διαστήματος $(t, t+dt)$. Για καταστάσεις με περισσότερες από μία εξερχόμενες μεταβάσεις, χρησιμοποιούμε πολυωνυμικές κατανομές. Συγκεκριμένα,

$$\begin{pmatrix} d_{SE} \\ d_{SV} \end{pmatrix}_t \sim M\left(S_{t-1}, \begin{pmatrix} \frac{\beta I_{t-1}/N}{\beta I_{t-1}/N + v_1}\left(1-e^{-\left(\beta\frac{I_{t-1}}{N}+v_1\right)dt}\right) \\ \frac{v_1}{\beta I_{t-1}/N + v_1}\left(1-e^{-\left(\beta\frac{I_{t-1}}{N}+v_1\right)dt}\right) \end{pmatrix}\right), \qquad \begin{pmatrix} d_{IH} \\ d_{IR} \end{pmatrix}_t \sim M\left(I_{t-1}, \begin{pmatrix} \frac{\gamma}{\gamma+\omega}\left(1-e^{-(\gamma+\omega)\,dt}\right) \\ \frac{\omega}{\gamma+\omega}\left(1-e^{-(\gamma+\omega)\,dt}\right) \end{pmatrix}\right), \tag{9}$$

$$\begin{pmatrix} d_{HC} \\ d_{HR} \\ d_{HD} \end{pmatrix}_t \sim M\left(H_{t-1}, \begin{pmatrix} \frac{\lambda}{\lambda+\kappa+\mu}\left(1-e^{-(\lambda+\kappa+\mu)\,dt}\right) \\ \frac{\kappa}{\lambda+\kappa+\mu}\left(1-e^{-(\lambda+\kappa+\mu)\,dt}\right) \\ \frac{\mu}{\lambda+\kappa+\mu}\left(1-e^{-(\lambda+\kappa+\mu)\,dt}\right) \end{pmatrix}\right), \qquad \begin{pmatrix} d_{CR} \\ d_{CD} \end{pmatrix}_t \sim M\left(C_{t-1}, \begin{pmatrix} \frac{\tau}{\tau+\rho}\left(1-e^{-(\tau+\rho)\,dt}\right) \\ \frac{\rho}{\tau+\rho}\left(1-e^{-(\tau+\rho)\,dt}\right) \end{pmatrix}\right), \tag{10}$$

$$\begin{pmatrix} d_{VS} \\ d_{VE} \end{pmatrix}_t \sim M\left(V_{t-1}, \begin{pmatrix} \frac{v_2}{\sigma\beta I_{t-1}/N + v_2}\left(1-e^{-\left(\sigma\beta\frac{I_{t-1}}{N}+v_2\right)dt}\right) \\ \frac{\sigma\beta I_{t-1}/N}{\sigma\beta I_{t-1}/N + v_2}\left(1-e^{-\left(\sigma\beta\frac{I_{t-1}}{N}+v_2\right)dt}\right) \end{pmatrix}\right). \tag{11}$$

Η κατανομή που συνδέει τις κρυφές καταστάσεις με τις παρατηρήσεις $g_{\boldsymbol{\theta}}(\boldsymbol{y}_t|\boldsymbol{x}_t)$, καθώς έχουμε 6 παρατηρήσιμες χρονοσειρές μπορεί να βασιστεί στην μέση τιμή Poisson κατανομών με παραμέτρους τις τιμές των κρυφών καταστάσεων τη στιγμή t, δηλαδή

$$g_{\boldsymbol{\theta}}(\boldsymbol{y}_t|\boldsymbol{x}_t) = \frac{1}{6}\sum_{k=1}^{6} Poisson(y_{k,t}\,; x_{k,t}), \tag{12}$$

όπου $\boldsymbol{y}_t = \left(y_{1,t}, y_{2,t}, y_{3,t}, y_{4,t}, y_{5,t}, y_{6,t}\right)^T = (I_t^*, H_t^*, C_t^*, R_t^*, D_t^*, V_t^*)^T$ και $\boldsymbol{x}_t = (I_t, H_t, C_t, R_t, D_t, V_t)^T$. Το σύμβολο * χρησιμοποιείται για τη διάκριση μεταξύ παρατηρήσεων και κρυφών καταστάσεων.

Λαμβάνοντας υπόψη την κυμαινόμενη δυναμική των επιδημιών μεγάλης κλίμακας, αυξάνουμε τον χώρο καταστάσεων προσθέτοντας χρονικά μεταβαλλόμενες παραμέτρους που αντιστοιχούν στους ρυθμούς μεταδοτικότητας (β), ανάρρωσης (κ, τ), θνησιμότητας (μ, ρ), εισαγωγών σε νοσοκομεία (γ) και ΜΕΘ (λ), καταλήγοντας στο επαυξημένο διάνυσμα κρυφών καταστάσεων $\tilde{\boldsymbol{x}}_t$. Σε κάθε χρονικό βήμα, η μεταβολή των παραμέτρων βασίζεται σε κανονικές κατανομές μηδενικού μέσου όρου. Κατά αυτόν τον τρόπο, δεν προκαθορίζουμε την τάση των παραμέτρων, ενώ οι τροχιές τους καθορίζονται πλήρως από την πιθανοφάνεια των δεδομένων.

3. ΑΠΟΤΕΛΕΣΜΑΤΑ

3.1. Εφαρμογή στις ημερήσιες καταγραφές COVID-19 στην Ιταλία

Σε αυτό το σημείο θα διερευνήσουμε την προσαρμοστική/προβλεπτική ικανότητα του προτεινόμενου επιδημιολογικού φίλτρου σωματιδίων στις ημερήσιες καταγραφές COVID-19 στην Ιταλία. Εξετάζουμε ένα ευρύ χρονικό διάστημα 500 ημερών από τις 4 Ιανουαρίου 2021, που αντιπροσωπεύει την πρώτη ημέρα καταγραφής πλήρως εμβολιασμένων ατόμων. Στο διάστημα αυτό συναντάμε 4 σημαντικά κύματα κρουσμάτων, με το 3ο κύμα να είναι το επικρατέστερο. Ιδιαίτερα ενδιαφέρον είναι ότι η απότομη αύξηση μεταδοτικότητας σημειώθηκε σε μικρότερο χρονικό διάστημα (28 Δεκεμβρίου 2021 – 23 Ιανουαρίου 2022), σε σύγκριση με τα δύο προηγούμενα κύματα κρουσμάτων.

Αλγόριθμος 2 Αλγόριθμος φίλτρου σωματιδίων με δυναμική εκτίμηση παραμέτρων

Απαιτούνται: $N_p, f, T, threshold$
Αρχικοποίησε $\{\tilde{\boldsymbol{x}}_0^i = \left(\boldsymbol{x}_0^i, \boldsymbol{\theta}_0^i\right)^T, w_0^i\}$
Για $t = 1, 2, \dots, T$ **κάνε**
  Πάρε δείγμα $\tilde{\boldsymbol{x}}_t^i \sim f_{\boldsymbol{\theta}}(\tilde{\boldsymbol{x}}_t^i | \tilde{\boldsymbol{x}}_{t-1}^i, \boldsymbol{y}_t)$
  $w_t^i \propto w_{t-1}^i \frac{g_{\boldsymbol{\theta}}(y_t | \tilde{\boldsymbol{x}}_t^i) p(\tilde{\boldsymbol{x}}_t^i | \tilde{\boldsymbol{x}}_{t-1}^i)}{f_{\boldsymbol{\theta}}(\tilde{\boldsymbol{x}}_t^i | \tilde{\boldsymbol{x}}_{t-1}^i, y_t)}$
  Για $i = 1, 2, \dots, N_p$ **κάνε**
    Κανονικοποίηση $w_t^i = w_t^i / \sum_{i=1}^{N_p} w_t^i$
    Αν $\widehat{N}_{eff}(t) < threshold$
      Αναδειγματοληψία με επανάθεση N_p σωματιδίων με βάση την κατανομή βαρών
      $P(k(i) = l) = w_t^l, l = 1, \dots, N_p$
      $\tilde{\boldsymbol{x}}_{0:t}^i = \tilde{\boldsymbol{x}}_{0:t}^{k(i)}$
      Επαναφορά αρχικών βαρών $w_t^i = 1/N_p$
    Τέλος αν
  Τέλος

Επιπρόσθετες παρατηρήσεις για την εξέλιξη της πανδημίας, αφορούν το πλήθος των περιστατικών που εισήχθησαν σε νοσοκομεία και ΜΕΘ. Καθώς προχωράμε βαθύτερα στην περίοδο των εμβολιασμών, παρατηρείται μείωση στα ποσοστά εισαγωγών σε νοσοκομεία και ΜΕΘ, παρατήρηση που γίνεται ιδιαίτερα εμφανής κατά τα δύο πρώτα κύματα της εξεταζόμενης περιόδου. Το 1ο κύμα (19 Φεβρουαρίου 2021 – 21 Μαρτίου 2021) προκάλεσε περίπου 2 φορές περισσότερες μολύνσεις σε σχέση με το 2ο, ενώ οι εισαγωγές σε νοσοκομεία και ΜΕΘ του πρώτου κύματος ξεπέρασαν αυτές του 2ου κατά περίπου 5-6 φορές. Ακόμη, πτωτική τάση εμφανίζεται και στα ποσοστά θνησιμότητας από COVID-19 στην Ιταλία, ενώ τα αντίστοιχα ποσοστά αναρρωμένων παρουσιάζουν την ακριβώς αντίθετη τάση. Στον Πίνακα 2, περιλαμβάνονται οι εκ των προτέρων και εκ των υστέρων ($T = 500$) κατανομές των παραμέτρων και καταστάσεων του μοντέλου. Οι εκ των υστέρων κατανομές προκύπτουν μέσω του στοχαστικού SEIRS μοντέλου για $T = 500$ ημέρες και $N_p = 200$ σωματίδια.

Πίνακας 2. *Εκ των προτέρων/υστέρων κατανομές για τις παραμέτρους του μοντέλου. Με Unif και tN συμβολίζονται η ομοιόμορφη και περικομμένη κανονική κατανομή, αντίστοιχα.*

Initial Parameter	Prior Distribution	Posterior Mean (95% C.I.)	References
$\alpha,\ \sigma$	$1/5.6,\ 0.2$	$1/5.6\ ,\ 0.2$	Quesada et al. 2021; Ssentongo et al 2022
$\psi,\ v_2$	$tN(\frac{1}{240},\frac{1}{2400},\frac{1}{300},\frac{1}{180})$, $tN(\frac{1}{180},\frac{1}{1800},\frac{1}{210},\frac{1}{150})$	$tN(\frac{1}{240},\frac{1}{2400},\frac{1}{300},\frac{1}{180})$, $tN(\frac{1}{180},\frac{1}{1800},\frac{1}{210},\frac{1}{150})$	Feikin et al. 2022; Mahase et al. 2021
$\omega,\ v_1$	$1/14,\ 10^{-4}$	0.0668 (0.06411, 0.06966), 1.144×10^{-4} (1.1019×10^{-4}, 1.1708×10^{-4})	Voinsky et al. 2020
β	$Unif(0.5, 1.5)$	0.8144 (0.7570, 0.9421)	Based on Maximum Likelihood
λ	$Unif$ (0.036, 0.048)	0.0196 (0.0184, 0.02102)	Based on Maximum Likelihood
γ	tN(0.0085,0.00085, 0.0006,0.012)	0.000688 (0.000474, 0.000904)	Based on Maximum Likelihood
κ	$tN(\frac{1}{18},\frac{1}{180},\frac{1}{22},\frac{1}{14})$	0.0568 (0.0538, 0.0601)	Roomi et al. 2021
μ	$tN(\frac{1}{12},\frac{1}{120},\frac{1}{15},\frac{1}{9})$	0.0686 (0.0650, 0.0729)	Roomi et al. 2021
τ	$tN(\frac{1}{9},\frac{1}{90},\frac{1}{11},\frac{1}{7})$	0.1306 (0.1275, 0.1331)	Roomi et al. 2021
ρ	$tN(\frac{1}{8},\frac{1}{80},\frac{1}{10},\frac{1}{6})$	0.1481 (0.1456, 0.1514)	Roomi et al. 2021

Σύμφωνα με τη μετανάλυση των Queasada et al. (2021), η μέση περίοδος επώασης του COVID είναι 5.6 ημέρες. Επομένως, επιλέγουμε τη σημειακή εκτίμηση για το ρυθμό επώασης $a = 1/5.6$, ενώ οι εκτιμήσεις για τις υπόλοιπες παραμέτρους επιλέγονται με παρόμοιο τρόπο. Αυτή η παράμετρος, μαζί με την πιθανότητα μόλυνσης μετά τον εμβολιασμό (σ), διατηρούνται σταθερές λόγω

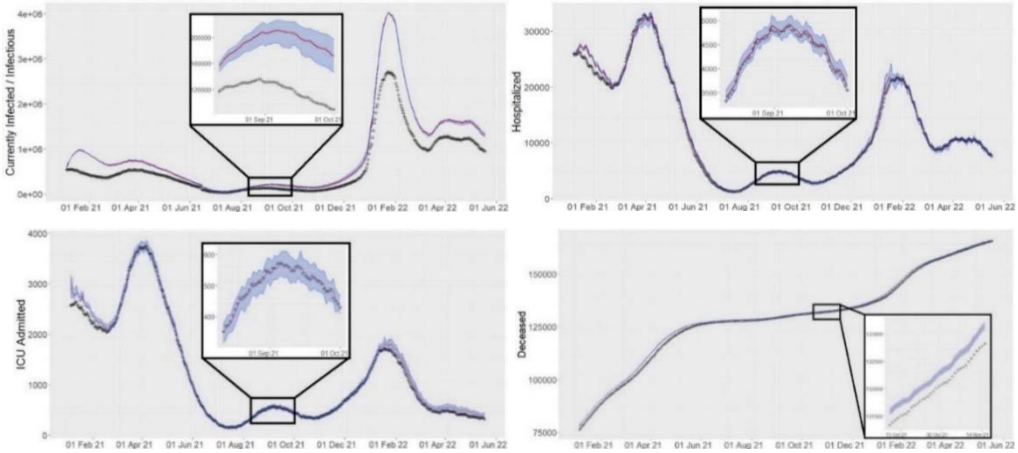
της χαμηλής τους μεταβλητότητας ως προς τον χρόνο. Οι ρυθμοί επανευαισθησίας μετά τη μόλυνση (ψ) ή τον εμβολιασμό (v_2), διατηρούνται επίσης σταθεροί. Χρησιμοποιούμε περικομμένες κανονικές κατανομές για τις εκ των προτέρων εκτιμήσεις λόγω των αβεβαιοτήτων που σχετίζονται με το ανοσοποιητικό σύστημα κάθε ατόμου.

3.2. Εξέταση απόδοσης του μοντέλου σύμφωνα με τις παραγόμενες εκτιμήσεις

Στο Σχήμα 2, παρουσιάζονται οι ημερήσιες παρατηρήσεις μολυσματικών, νοσηλευόμενων σε νοσοκομεία και ΜΕΘ και αποθανόντων, συνοδευόμενες από τις εκτιμήσεις που παράγονται από το προτεινόμενο στοχαστικό μοντέλο. Οι παραγόμενες εκτιμήσεις φαίνεται να ακολουθούν ικανοποιητικά τη δυναμική της πανδημίας ακόμη και για τη μακρά περίοδο των 500 ημερών. Συγκεκριμένα, οι εκτιμήσεις που παρέχονται για τους νοσηλευόμενους σε νοσοκομεία, ΜΕΘ, και αποθανόντες είναι πολύ κοντά στα πραγματικά δεδομένα, γεγονός εξαιρετικά ενθαρρυντικό, καθώς αυτές οι 3 χρονοσειρές αντιπροσωπεύουν τις ακριβέστερες καταγραφές. Σημαντικό εύρημα της παρούσας ανάλυσης αποτελεί η εκτίμηση μεγαλύτερου πλήθους μολυσματικών ατόμων συγκριτικά με τις ημερήσιες καταγραφές. Η υπερεκτίμηση αυτή, πιθανότατα μπορεί να αποδοθεί στις υπάρχουσες ασυμπτωματικές λοιμώξεις που χαρακτηρίζουν την μετάδοση του COVID-19. Το ποσοστό των ασυμπτωματικών κυμαίνεται μεταξύ 20.02 – 50.87%, γεγονός που υποστηρίζεται από την μετανάλυση των Oran & Topol.

Κατά τη διάρκεια του 3ου κύματος, παρατηρείται μικρή υπερεκτίμηση (περίπου 100 ατόμων) στους νοσηλευόμενους σε ΜΕΘ. Αυτό το κύμα μόλυνσης αντιστοιχεί στην εντονότερη διασπορά του ιού μέχρι σήμερα, ενώ ταυτόχρονα η διάρκειά του είναι μικρότερη σε σύγκριση με τα δύο προηγούμενα κύματα της εξεταζόμενης περιόδου. Αναμφίβολα, ο συνδυασμός αυτός άσκησε πίεση στο εθνικό σύστημα υγείας, οδηγώντας σε πληρότητα στις ΜΕΘ. Ως αποτέλεσμα, θεωρούμε ότι πιθανώς το μοντέλο οδηγεί στην αποκάλυψη αυτού του γεγονότος, παρέχοντας εκτίμηση για τον αριθμό των ατόμων που παρέμειναν εκτός ΜΕΘ λόγω μη διαθεσιμότητας κλινών. Ταυτόχρονα, παρατηρείται μικρή υπερεκτίμηση στον αριθμό των αποθανόντων (0.15%–0.95%). Η μικρή απόκλιση μεταξύ καταγραφών και εκτιμήσεων μπορεί να αποδοθεί σε ελλιπή καταγραφή θανάτων, χωρίς να μπορούμε να

είμαστε βέβαιοι για αυτό λόγω μη επαρκών στοιχείων στη βιβλιογραφία.

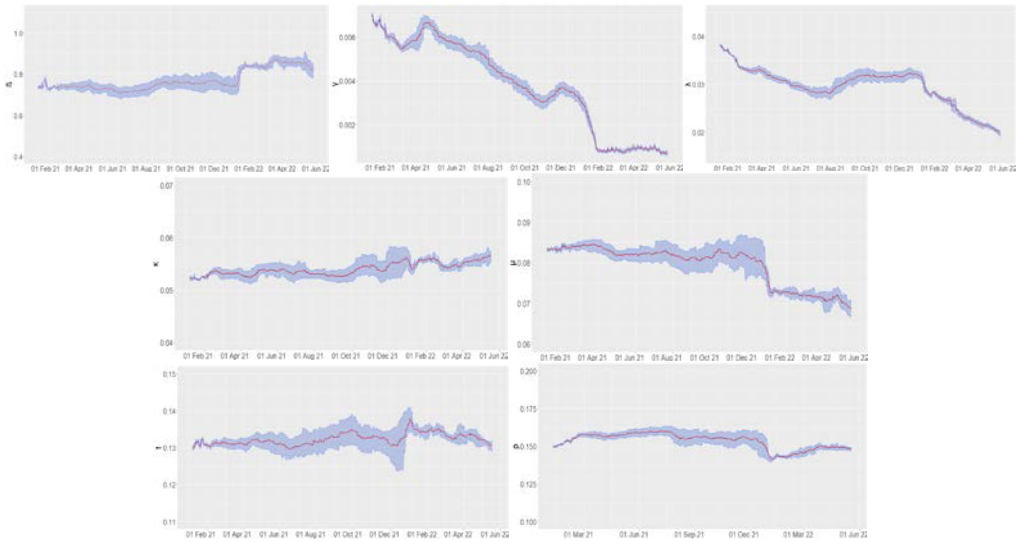


Σχήμα 2. *Ημερήσιες καταγραφές COVID και οι αντίστοιχες εκτιμήσεις για τους α) μολυσματικούς, β) νοσηλευόμενους σε νοσοκομεία, γ) και ΜΕΘ και δ) νεκρούς στην Ιταλία. Οι μαύρες κουκκίδες αντιστοιχούν στα πραγματικά δεδομένα, η κόκκινη γραμμή στις εκτιμήσεις του μοντέλου και οι μπλε περιοχές αντιστοιχούν στα 99% διαστήματα εμπιστοσύνης*

Στο Σχήμα 3, παρουσιάζεται η εξέλιξη των χρονικά μεταβαλλόμενων παραμέτρων κατά τη διάρκεια 500 ημερών πανδημίας στην Ιταλία. Εμφανίζεται αξιοσημείωτη πτωτική τάση στα ποσοστά εισαγωγών σε νοσοκομεία (γ) και ΜΕΘ (λ). Η τάση αυτή εντείνεται ιδιαίτερα κατά το 3ο κύμα, συμπεραίνοντας ότι παρά την αξιοσημείωτη αύξηση στη μεταδοτικότητα, τα υψηλά ποσοστά εμβολιασμών διατήρησαν το πλήθος των σοβαρών λοιμώξεων σε διαχειρίσιμα επίπεδα. Τα ποσοστά ανάρρωσης είτε σε νοσοκομεία (κ) είτε σε ΜΕΘ (τ) χαρακτηρίζονται από αυξητική τάση, ενώ τα ποσοστά θνησιμότητας (μ) και (ρ) εμφανίζουν αντίθετη συμπεριφορά. Η θνησιμότητα ατόμων που νοσηλεύθηκαν σε νοσοκομεία, μειώνεται εντονότερα από αυτή των περιστατικών στις ΜΕΘ, γεγονός αναμενόμενο καθώς ακόμη και μετά τον εμβολιασμό η κατάσταση της υγείας των εισαχθέντων σε ΜΕΘ παραμένει κρίσιμη.

Στον Πίνακα 3, παρουσιάζονται οι τιμές NRMSE της προσαρμοστικής ικανότητας 5 μοντέλων, όπου η τελευταία σειρά αντιστοιχεί στα σφάλματα της προτεινόμενης προσέγγισης. Συγκρίνουμε τα 5 μοντέλα με βάση τα ημερήσια περιστατικά νοσηλείας σε νοσοκομεία και ΜΕΘ, περιστατικά αποθανόντων και εμβολιασμένων. Το προτεινόμενο μοντέλο παρέχει τα καλύτερα αποτελέσματα προσαρμογής, δίνοντας τις μικρότερες τιμές NRMSE συγκριτικά με τα υπόλοιπα 4 μοντέλα, ενώ η ντετερμινιστική προσέγγιση παρέχει μακράν τις χειρότερες εκτιμήσεις. Σύγκριση γίνεται με μοντέλα που κάνουν χρήση του EKF, καθώς αποτελεί τη δημοφιλέστερη στοχαστική προσέγγιση στη βιβλιογραφία για τη μοντελοποίηση επιδημιών (Zhu et al. 2021). Το μοντέλο EKF-SEIRD επιλέχθηκε ώστε να γίνει σύγκριση με ένα από τα δημοφιλέστερα,

σύγχρονα μοντέλα της βιβλιογραφίας και εμφάνισε αυξημένες τιμές NRMSE για τα περιστατικά αποθανόντων σε σύγκριση με το προτεινόμενο φίλτρο σωματιδίων με χρονικά μεταβαλλόμενες παραμέτρους.



Σχήμα 3. *Αναπαράσταση της εξέλιξης των χρονικά μεταβαλλόμενων παραμέτρων για 500 ημέρες πανδημίας στην Ιταλία με τα αντίστοιχα 50% διαστήματα εμπιστοσύνης*

Πίνακας 3. *Τιμές NRMSE για το ντετερμινιστικό SEIHCRDV, EKF-SEIRD, EKF-SEIHCRDV με δυναμική εκτίμηση παραμέτρων και του επιδημιολογικού φίλτρου σωματιδίων με και χωρίς χρονικά μεταβαλλόμενες παραμέτρους*

	Hospitalized	ICU	Deceased	Vaccinated
Deterministic SEIHCRDV	0.9988	0.9991	0.9998	1.1509
Particle filtering SEIHCRDV (Constant parameters)	0.6106	0.1911	0.0874	0.0252
Dynamic EKF-SEIRD	-	-	0.1645	-
Dynamic EKF-SEIHCRDV	0.2082	0.0741	0.0795	0.0217
Particle filtering SEIHCRDV (Dynamic parameter estimation)	**0.0505**	**0.0550**	**0.0611**	**0.0240**

4. ΣΥΖΗΤΗΣΗ ΚΑΙ ΣΥΜΠΕΡΑΣΜΑΤΑ

Οι ντετερμινιστικές προσεγγίσεις αποτυγχάνουν να διαχειριστούν την περίπλοκη δυναμική επιδημιών, λόγω των σημαντικών διακυμάνσεων που τις συνοδεύουν. Οι μεταβολές στη δυναμική που οφείλονται σε εξωτερικούς παράγοντες, καθώς και η αβεβαιότητα στις καταγραφές, οδηγούν στην επιλογή πιο αντιπροσωπευτικών, στοχαστικών προσεγγίσεων. Τα υπάρχοντα ντετερμινιστικά μοντέλα δεν είναι τέλεια, ενώ η προσπάθεια συμπερίληψης όλων των πιθανών κρυφών καταστάσεων και μεταβάσεων μιας επιδημίας, θα

οδηγούσε σε περίπλοκες δομές συνοδευόμενες από περιπτώσεις υπερπροσαρμογής (Papageorgiou et al. 2023; 2022; 2021).

Στον παρόν άρθρο παρουσιάζουμε ένα νέο, επεκτεταμένο μοντέλο SEIRS συνοδευόμενο από ένα φίλτρο σωματιδίων με δυναμική εκτίμηση παραμέτρων. Με βάση αυτό, στόχος είναι η αντιμετώπιση των περιορισμών που συνοδεύουν τις ντετερμινιστικές προσεγγίσεις, ενώ παράλληλα ξεπερνάμε δραστικά την προγνωστική τους ικανότητα. Επεκτείνουμε το τυπικό μοντέλο SEIRS, αυξάνοντας επαρκώς τον αριθμό των διαφορικών εξισώσεων του συστήματος από 4 σε 8, λαμβάνοντας υπόψη τους υποπληθυσμούς των νοσηλευόμενων σε νοσοκομεία και ΜΕΘ, των αποθανόντων και των εμβολιασμένων. Μέσω του φίλτρου σωματιδίων, αξιοποιούμε τη διαθέσιμη πληροφορία που αφορά την εξέλιξη του COVID στην Ιταλία, βελτιώνοντας αισθητά τις παραγόμενες εκτιμήσεις ακόμη και για 500 ημέρες πανδημίας. Μέσω της δυναμικής εκτίμησης παραμέτρων, αποφεύγουμε τη συμπερίληψη διαφορετικών ρυθμών μετάβασης για εμβολιασμένους και ανεμβολίαστους (που θα οδηγούσε στην προσθήκη 7 έως 9 επιπλέον παραμέτρων).

Τα κυριότερα αποτελέσματα αφορούν την εκτίμηση υψηλότερου αριθμού κρουσμάτων σε σύγκριση με τις ημερήσιες παρατηρήσεις -αύξηση 20.02 – 50.87% (95% δ.ε.)-, που μπορεί να αποδοθεί σε ασυμπτωματικές λοιμώξεις. Με βάση την μετανάλυση των (Oran & Topol 2021), συμπεραίνεται ότι το ποσοστό των ατόμων θετικών στον COVID που δεν εκδήλωσαν συμπτώματα κυμαίνεται μεταξύ 21.7% και 47.3%, ενισχύοντας σημαντικά την εγκυρότητα των εκτιμήσεων του μοντέλου. Ακόμη, αξιοσημείωτο είναι ότι σε αντίθεση με άλλα επιδημιολογικά μοντέλα της βιβλιογραφίας που χρησιμοποιούν επιπλέον καταστάσεις για τις ασυμπτωματικές μολύνσεις, εξαναγκάζοντας την εκτίμηση αυτού του ποσοστού και αυξάνοντας σημαντικά την πολυπλοκότητα του μοντέλου, η παρούσα προσέγγιση αποκαλύπτει την κρυφή δυναμική της πανδημίας χωρίς πρόσθετο υπολογιστικό κόστος.

Το κύμα του Ιανουαρίου 2022 στην Ιταλία σηματοδοτεί όχι μόνο την εντονότερη μεταδοτικότητα του ιού έως τώρα, αλλά και την πιο σύντομη περίοδο κορύφωσης ενός κύματος κρουσμάτων. Ο συνδυασμός αυτός πιθανότατα οδήγησε σε συνωστισμό στις ΜΕΘ, με το μοντέλο να διαγιγνώσκει αυτό το γεγονός και να παρέχει εκτίμηση για το πλήθος των ατόμων που έμειναν εκτός των ΜΕΘ λόγω μη διαθεσιμότητας κλινών. Η παρούσα προσέγγισή υποστηρίζει την αποτελεσματικότητα του πλήρους εμβολιασμού κατά του COVID, καθώς τα δυναμικά εκτιμώμενα ποσοστά ανάρρωσης παρουσιάζουν αυξητική τάση καθώς προχωράμε βαθύτερα στην περίοδο των εμβολιασμών. Τα ποσοστά εισαγωγών σε νοσοκομεία, καθώς και οι ρυθμοί θνησιμότητας παρουσιάζουν φθίνουσα συμπεριφορά ιδιαίτερα κατά το 3ο – επικρατέστερο– κύμα κρουσμάτων. Το συμπέρασμα αυτό είναι σε συμφωνία

με πληθώρα μελετών της βιβλιογραφίας. Οι Watson et al. (2022) αναφέρουν ότι οι εμβολιασμοί οδήγησαν σε μείωση κατά 63% στους θανάτους, ενώ οι νοσηλείες σε νοσοκομεία και ΜΕΘ μειώθηκαν κατά 63.5% και 65.6%, σε μια περίοδο 300 ημερών στις Η.Π.Α. (Feikin et al. 2022).

Τέλος, η θεωρητική ανάλυση με βάση το επεκτεταμένο SEIRS, οδηγεί σε πολύτιμα συμπεράσματα για την εξάπλωση της πανδημίας. Ο τύπος για τον δείκτη R_0 (θεώρημα 1) παρέχει μια αντιπροσωπευτική εικόνα της εξέλιξης της πανδημίας, ενώ βοηθά στην απόδειξη της ύπαρξης και ολικής ευστάθειας επιδημικών ισοζυγίων. Με βάση τους Rosero-Bixby & Miller (2022), ο αναπαραγωγικός ρυθμός δεν διατηρείται μικρότερος της μονάδας μετά το τέλος των lockdowns, καθιστώντας την εξαφάνιση του ιού εξαιρετικά αμφίβολη. Δεδομένου λοιπόν ότι ο περιορισμός της μεταδοτικότητας του ιού φαίνεται αρκετά δύσκολος –τουλάχιστον στο άμεσο μέλλον– οι αρμόδιες αρχές θα πρέπει να επικεντρωθούν στη μείωση των επιπέδων σοβαρών λοιμώξεων και θνησιμότητας.

ABSTRACT

In this paper, an extended epidemiological particle filter with time-varying parameters is proposed, with the aim of describing the evolution of pandemics. The validity of the model is examined on daily records of COVID-19 in Italy over a period of 500 days. Key findings include estimating the rates of asymptomatic carriers, a well-known feature of COVID-19. Unlike other models that use additional states for asymptomatic carriers, forcing the estimation of this ratio and increasing the complexity of the model, the proposed approach leads to the emergence of the hidden dynamics of COVID-19 without extra computational burden. Additional findings that confirm the appropriateness of the model are the evolution of the time-varying parameters, as well as the resulting variation in the incidence of ICU admissions compared to official records, during the most prevalent infection wave of January 2022. Finally, a statistical algorithm is presented for the estimation of the currently recovered and vaccine-protected individuals.

ΑΝΑΦΟΡΕΣ

Al-Dousari A., Ellahi A., Hussain I. (2021). Use of non-homogeneous Poisson process for the analysis of new cases, deaths, and recoveries of COVID-19 patients: A case study of Kuwait. *Journal of King Saud University - Science*, **33**(8), 101614.

Calvetti D., Hoover A., Rose J., Somersalo E. (2021). Bayesian particle filter algorithm for learning epidemic dynamics, *Inverse Problems*, **37**, 115008.

Costa P.J., Dunyak J.P., Mohtashemi M. (2005). Models, prediction, and estimation of outbreaks of infectious disease, Proceedings. *IEEE SoutheastCon 2005*, 174-178.

Feikin D.R., et al. (2022). Duration of effectiveness of vaccines against SARS-CoV-2 infection and COVID-19 disease: results of a systematic review and meta-regression. *Lancet*, **399**, 924-944.

Mahase E. (2021). Covid-19: Antibody levels fall after second Pfizer dose, but protection against severe disease remains, studies indicate. *BMJ*, **375**, n2481.

Oran D.P., Topol A.J. (2021). The Proportion of SARS-CoV-2 Infections That Are Asymptomatic. *Annals of Internal Medicine*, **174**(5), 655-662.

Papageorgiou V.E., 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 and Fractals*, **166**, 112914.

Papageorgiou V.E. et al. (2023). A convolutional neural network of low complexity for tumor anomaly detection. *Proc. of 8th Inter. Cong, on Inform. and Comm. Technology*, London.

Papageorgiou V.E., Zegkos T., et al. (2022). Analysis of Digitalized ECG Signals Based on Artificial Intelligence and Spectral Analysis Methods Specialized in ARVC. *International Journal for Numerical Methods in Biomedical Engineering*, **38**(11), e3644.

Papageorgiou V. (2021). Brain Tumor Detection Based on Features Extracted and Classified Using a Low-Complexity Neural Network. *Traitement du Signal,* **38**(3):547-554.

Quesada J.A., López-Pineda A., Gil-Guillén V.F., et al. (2021). Período de incubación de la COVID-19: revisión sistemática y metaanálisis. *Rev Clin Esp.,* **221**, 109-117.

Roomi S., Shah S.O., et al. (2021). Declining Intensive Care Unit Mortality of COVID-19: A Multi-Center Study. *Journal of Clinical Medicine Research*, **13**(3), 184-190.

Rosero-Bixby L, Miller T. (2022). The mathematics of the reproduction number R for Covid-19: a primer for demographers. *Vienna Yearbook Popul Res*, **2022**(20), 1–24.

Ssentongo P., et al. (2022). SARS-CoV-2 vaccine effectiveness against infection, symptomatic and severe COVID-19: a systematic review and meta-analysis. *BMC Infect Dis*, **22**, 439.

Voinsky I., Baristaite G., Gurwitz D. (2020). Effects of age and sex on recovery from COVID-19: Analysis of 5769 Israeli patients. *Journal of Infection*, **81**(2), e102-e103.

Watson O.J., Barnsley G., et al. (2022). Global impact of the first year of COVID-19 vaccination: a mathematical modelling study. *Lancet Infect. Dis.* **22**(9), 1293–302.

Zhu X., Gao B., et al. (2021). Extended Kalman filter based on stochastic epidemiological model for COVID-19 modelling. *Computers in Biology and Medicine*, **137**, 104810.

Ελληνικό Στατιστικό Ινστιτούτο
Ιρακτικά 35ου Πανελληνίου & 1ου Διεθνούς Συνεδρίου Στατιστικής (2023), σελ.74-91

ΣΥΓΚΡΙΣΗ ΕΚΤΙΜΗΤΩΝ ΓΙΑ ΤΗΝ ΠΡΟΒΛΕΨΗ ΕΚΛΟΓΙΚΩΝ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

Γεώργιος Ξ. Παπαγεωργίου ***Μιλτιάδης Σ. Χαλικιάς***
gpapageorgiou@uniwa.gr mchalik@uniwa.gr

Τμήμα Λογιστικής και Χρηματοοικονομικής,
Πανεπιστήμιο Δυτικής Αττικής

ΠΕΡΙΛΗΨΗ

Στην εργασία αυτή συγκρίνουμε και προτείνουμε εκτιμητές με στόχο την ακριβή πρόβλεψη του αποτελέσματος (συνολικού αριθμού ψήφων των δύο υποψηφίων) της λαϊκής ψηφοφορίας των προεδρικών εκλογών του 2020 των Ηνωμένων Πολιτειών της Αμερικής, χρησιμοποιώντας μονοσταδιακή δειγματοληψία με τη μέθοδο των συστάδων. Στην περίπτωση που εξετάζουμε, επιλέγουμε αρχικά ως συστάδες τις πολιτείες και έπειτα επαναλαμβάνουμε την ίδια διαδικασία θέτοντας ως πρωτογενείς δειγματοληπτικές μονάδες τις επαρχίες. Για τη διεκπεραίωση του πειράματος χρησιμοποιήθηκαν τα δεδομένα των εκλογών του 2020, του 2016 και του 2012. Η μεθοδολογία που θα παρουσιάσουμε για τον υπολογισμό του τελικού αποτελέσματος μπορεί να εφαρμοστεί και στις εγχώριες βουλευτικές εκλογές, θέτοντας ως συστάδες είτε τους νομούς είτε τα εκλογικά τμήματα. Επιπροσθέτως, διερευνούμε τον ελάχιστο αριθμό συστάδων που είναι απαραίτητες για τον προσδιορισμό διαστημάτων εμπιστοσύνης επιπέδου 95% για κάθε διαμέριση των Η.Π.Α. που έχουμε επιλέξει.

Λέξεις κλειδιά: Δειγματοληψία κατά συστάδες, Εκτιμητής Λόγου, Εκτιμητής Horvitz - Thompson, Εκτιμητής Γραμμικής Παλινδρόμησης

1. ΕΙΣΑΓΩΓΗ

Η ανάπτυξη και η μελέτη μοντέλων για προβλέψεις εκλογικών αποτελεσμάτων συνδέεται στενά με την ανάπτυξη της εμπειρικής κοινωνικής έρευνας (Νικολακόπουλος 2003). Η διερεύνηση πολιτικών τάσεων και αντιλήψεων στον Ελλαδικό χώρο ξεκίνησε στα μέσα του 20ου αιώνα ως μία σύμπραξη των συμμαχικών δυνάμεων με σκοπό την επίβλεψη της διεξαγωγής της εκλογικής διαδικασίας (AMFOGE, 1946) αλλά δεν γνώρισε ιδιαίτερη ανάπτυξη μέχρι την εποχή της μεταπολίτευσης, κυρίως στις αρχές της δεκαετίας του ‘90, όπου τροφοδοτήθηκε από την ανταγωνιστικότητα των Μέσων Μαζικής Ενημέρωσης της νεοσύστατης ιδιωτικής τηλεόρασης.

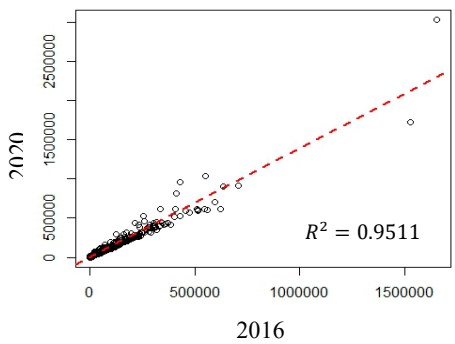
Συγκεκριμένα για την πρόβλεψη εκλογικών αποτελεσμάτων, οι μέθοδοι που χρησιμοποιούνται μπορούν να διακριθούν στις εξής κατηγορίες, τις προεκλογικές δημοσκοπήσεις για τη διάθεση ψήφου, τις δημοσκοπήσεις εξόδου

(exit polls) και τις εκτιμήσεις από προβολές που πραγματοποιούνται είτε πάνω στο σύνολο των αρχικών καταμετρημένων ψηφοδελτίων είτε πάνω σε ένα αντιπροσωπευτικό δείγμα εκλογικών τμημάτων. Η έγκαιρη πρόγνωση του εκλογικού αποτελέσματος δίνει τον απαραίτητο χρόνο στους εκπροσώπους των πολιτικών κομμάτων να προετοιμάσουν τις πρώτες τους κινήσεις μετά το πέρας της ψηφοφορίας. Σαφώς, μεγαλύτερη ακρίβεια ως προς τη μέτρηση του αποτελέσματος παρουσιάζουν οι μέθοδοι που βασίζονται στα δεδομένα της ημέρας διεξαγωγής των εκλογών και ο συνηθέστερος εμπειρικός κανόνας για τη μείωση του σφάλματος των εκτιμήσεων είναι η στάθμιση των αποτελεσμάτων σε αποτελέσματα προηγούμενων αναμετρήσεων. Μειονεκτήματα του κανόνα αυτού αποτελούν η αδυναμία του ψηφοφόρου να ανακαλέσει την προηγούμενη ψήφο του και το ποσοστό στο οποίο αντιπροσωπεύει την πολιτική του παράταξη. Η χρήση της μονοσταδιακής δειγματοληψίας με τη μέθοδο των συστάδων στην περίπτωση αυτή αποτελεί μία μερική λύση του άνωθεν προβλήματος επειδή χρησιμοποιεί τα εκλογικά τμήματα ως πρωτογενείς δειγματοληπτικές μονάδες αντί των ψηφοφόρων. Ο χαρακτηρισμός της μεθόδου ως μερική λύση οφείλεται κυρίως σε προβλήματα που αφορούν στην απόλυτη τήρηση της θεωρίας των στατιστικών μεθόδων που χρησιμοποιούνται και την ερμηνεία των αποτελεσμάτων. Η μέθοδος που παρουσιάζεται στην παρούσα μελέτη μας οδηγεί σε μία εκτίμηση του αποτελέσματος η οποία δεν πρέπει να συγχέεται με το αποτέλεσμα καθαυτό. Μελετώντας την ξένη βιβλιογραφία, παρατηρείται η τάση της εξαγωγής τελικής εκτίμησης από τη συνδυαστική μελέτη μοντέλων, όπως δημοσκοπήσεων, ποσοτικών μοντέλων και προβλέψεις ειδικών (Graefe, 2018). Οι κίνδυνοι που αντιμετωπίζουμε όταν επιλέγουμε να στηρίξουμε τα συμπεράσματά μας αποκλειστικά σε ένα μοντέλο πρόβλεψης μπορούν να περιγραφούν συνοπτικά σε τρία σημεία, την επιλογή μοντέλου με κριτήριο την ταύτιση με τις πολιτικές πεποιθήσεις, την ικανότητα του μοντέλου να δώσει έγκυρες προβλέψεις σε εφαρμογές του μεταξύ διαφορετικών χρονικών στιγμών και το είδος της πληροφορίας που περιέχει.

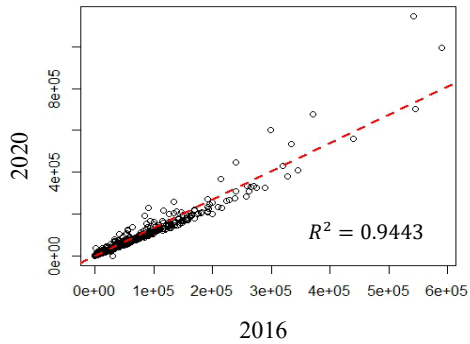
Στην εργασία αυτή επικεντρωνόμαστε στην εκτίμηση του εκλογικού αποτελέσματος χρησιμοποιώντας ως παράγοντα στάθμισης τα αποτελέσματα των προηγούμενων εκλογών. Ξεκινώντας από τη σύγκριση τριών εκτιμητών βάσει κριτηρίων που θα αναλύσουμε στη συνέχεια, σκοπός μας είναι να καταλήξουμε στην επιλογή του καταλληλότερου, πάνω στον οποίο θα βασίσουμε το μοντέλο πρόβλεψης μας. Οι τρεις προς σύγκριση εκτιμητές είναι ο εκτιμητής Horvitz – Thompson (Horvitz & Thompson, 1952), ο γραμμικός εκτιμητής για συστάδες με άνισα μεγέθη και ο εκτιμητής Λόγου (Cochran, 1977), οι ορισμοί των οποίων παρατίθενται στην επόμενη ενότητα. Κατά τη

σύγκριση τους, μελετώντας τα πλεονεκτήματα και τα μειονεκτήματα τους, θα αναδείξουμε τις βασικές προϋποθέσεις που πρέπει να πληροί ένας εκτιμητής για να είναι ακριβής καθώς και το απαραίτητο πλήθος συστάδων που χρειαζόμαστε ώστε να επιτευχθεί η επιθυμητή ακρίβεια. Οι συγκρίσεις θα πραγματοποιηθούν πάνω σε δύο διαφορετικές διαμερίσεις των Ηνωμένων Πολιτειών της Αμερικής. Σε πρώτο στάδιο, θα αναλύσουμε την επίδοση των εκτιμητών σε επίπεδο 51 πολιτειών και έπειτα θα πραγματοποιηθεί ανάλυση σε επίπεδο 3112 επαρχιών εκτελώντας μονοσταδιακή δειγματοληψία με τη μέθοδο των συστάδων. Η επιλογή της μονοσταδιακής μεθόδου έγινε με γνώμονα τον προσδιορισμό ενός αντιπροσωπευτικού δείγματος. Η χρήση ενός πολυσταδιακού μοντέλου κρίνεται ανεπαρκής για την επίτευξη του στόχου αυτού καθώς θα απέκλειε συστηματικά ένα μεγάλο κομμάτι του πληθυσμού σε κάθε επανάληψη του προγράμματος. Ολοκληρώνοντας την ενότητα των προκαταρκτικών συγκρίσεων, θα εισάγουμε έναν νέο εκτιμητή μέσω του οποίου λαμβάνουμε υπόψη μας και τα αποτελέσματα της προηγούμενης εκλογικής αναμέτρησης για κάθε συστάδα και δεδομένη διαμέριση που έχουμε επιλέξει να εξετάσουμε. Το εγχείρημα αυτό βασίζεται στην προϋπόθεση της ύπαρξης γραμμικής σχέσης μεταξύ των αποτελεσμάτων των εκλογών ανά συστάδα στην αρχή και στο τέλος της τετραετίας. Παραθέτουμε ενδεικτικά τα αντίστοιχα γραφήματα για τους δύο υποψηφίους.

***Εικόνα 1**. Ψήφοι Δημοκρατικών*



***Εικόνα 2.** Ψήφοι*



Η αντιμετώπιση της ετεροσκεδαστικότητας που παρατηρείται στα άνω γραφήματα θα συζητηθεί στην ενότητα 2, όπου θα παρουσιάσουμε τον εκτιμητή παλινδρόμησης που θα χρησιμοποιήσουμε για την πρόβλεψη του εκλογικού αποτελέσματος. Επιπροσθέτως, θα δούμε ότι γραμμική σχέση παρατηρείται και ανάμεσα στις εκτιμήσεις των εκλογικών αποτελεσμάτων

μεταξύ δύο διαδοχικών αναμετρήσεων. Σημειώνουμε ότι εκτιμητή παλινδρόμησης (διαφορετικό από αυτόν που εφαρμόσαμε) για εκτίμηση εκλογικών αποτελεσμάτων σε Ελληνικές εκλογές με συστάδες τα εκλογικά κέντρα παρουσίασαν πρώτη φορά οι Κουνιάς, Ε. & Συμεωνίδης, Γ. (1990).

2.ΜΕΘΟΔΟΛΟΓΙΑ ΚΑΙ ΟΡΙΣΜΟΙ

Κάνοντας εφαρμογή των προγραμματιστικών πακέτων της R, κατασκευάζουμε πρόγραμμα που πραγματοποιεί **n** ανεξάρτητες μονοσταδιακές δειγματοληπτικές διαδικασίες με τη μέθοδο των συστάδων. Στις επόμενες εφαρμογές οι τιμές του φυσικού αριθμού **n** κυμαίνονται στο διάστημα $[4 * 10^3, 10^4]$, αναλόγως της υπολογιστικής πολυπλοκότητας που συνδέεται με το πλήθος συστάδων δείγματος και της διαμέρισης του πληθυσμού. Στην περίπτωση του εκτιμητή Horvitz – Thompson η πιθανότητα κάθε συστάδας να επιλεγεί στο δείγμα είναι ανάλογη του μεγέθους της, ήτοι του πλήθους των έγκυρων ψηφοδελτίων εντός της συστάδας, ενώ για τους άλλους δύο εκτιμητές εφαρμόζεται απλή τυχαία δειγματοληψία. Για δεδομένο πλήθος συστάδων **m** λαμβάνουμε τρεις παραμέτρους. Ως πρώτη εξ αυτών ορίζουμε το ποσοστό **p (precision)** των διαστημάτων εμπιστοσύνης που περιείχαν τη μέση τιμή του εκτιμητή, δηλαδή τον προβλεπόμενο αριθμό ψήφων των δύο υποψηφίων, όπως προκύπτει από τις διαδοχικές δειγματοληψίες. Έπειτα, ορίζουμε μία νέα παράμετρο **c (accuracy)** που αποτελεί το αντίστοιχο ποσοστό για το πραγματικό εκλογικό αποτέλεσμα. Όταν ο εκτιμητής είναι αμερόληπτος και οι παράμετροι **p** και **c** ταυτίζονται. Το αντίστροφο εν γένει δεν ισχύει. Παρομοίως, καταγράφουμε την τετραγωνική ρίζα του μέσου τετραγωνικού σφάλματος του έκαστου εκτιμητή για τους δύο υποψήφιους, που συμβολίζουμε με $\boldsymbol{RMSE}$. Όλα τα διαστήματα εμπιστοσύνης είναι υπολογισμένα για επίπεδο σημαντικότητας α = 0.05. Ο συνδυασμός των παραπάνω τιμών μας επιτρέπει να αποφανθούμε για την καταλληλότητα του εκτιμητή.

Ορισμός 2.1. Ο αμερόληπτος εκτιμητής πληθυσμιακού ολικού X (συνόλου ψήφων υποψηφίου) για συστάδες με άνισα μεγέθη, $\mathbf{X_{Cl,u}}$, δίνεται από τον τύπο:

$$\hat{X} = X_{\mathrm{Cl,u}} = \frac{M}{m}\sum_{i=1}^{m} t_i \ , \qquad (2.1)$$

όπου **M** είναι ο συνολικός αριθμός συστάδων, **m** το πλήθος συστάδων στο δείγμα και $t_i,\ i = 1,2,\dots,m$, οι τιμές της τυχαίας μεταβλητής για το χαρακτηριστικό που εξετάζουμε, δηλαδή το σύνολο των ψήφων που αντιστοιχούν σε κάθε παράταξη στη συστάδα **i**.

Για τη διακύμανση και την εκτίμηση της διακύμανσης του εκτιμητή ισχύουν οι δύο ακόλουθες σχέσεις:

$$Var\left(X_{Cl,u}\right) = \frac{M(1-f)}{m}\sum_{i=1}^{M}\frac{\left(t_i - \overline{X}\right)^2}{M-1} \tag{2.2}$$

και

$$\widehat{Var}(X_{\mathrm{Cl,u}}) = \frac{M(1-f)}{m}\sum_{i=1}^{m}\frac{\left(t_i - \overline{X}_{Cl,u}\right)^2}{m-1}\ , \tag{2.3}$$

όπου $f = m/M$, $\overline{X} = (X/M) = (\sum_{i=1}^{M} t_i)/M$ και $\overline{X}_{Cl,u} = X_{Cl,u}/\mathrm{M}$.

Ορισμός 2.2. Ο εκτιμητής Λόγου πληθυσμιακού ολικου Χ, $\mathbf{X_{Cl,r}}$, ορίζεται από τη σχέση:

$$\hat{X} = X_{Cl,r} = N \cdot \left(\sum_{i=1}^{m} t_i \Big/ \sum_{i=1}^{m} y_i\right)\ , \qquad \mathrm{N} = \sum_{i=1}^{M} y_i\ , \tag{2.4}$$

όπου $\mathbf{y_i}$ είναι το μέγεθος κάθε συστάδας στον πληθυσμό ή ο συνολικός αριθμός των έγκυρων ψηφοδελτίων στη συστάδα i.

Η διακύμανση και η εκτίμηση της διακύμανσης του εκτιμητή Λόγου δίνονται προσεγγιστικά από τους παρακάτω τύπους:

$$Var\left(X_{Cl,r}\right) = \frac{M^2(1-f)}{m}\sum_{i=1}^{M}\frac{(t_i - Ry_i)^2}{M-1}\ , \tag{2.5}$$

$$\widehat{Var}(X_{Cl,\mathrm{r}}) = \frac{\widehat{\overline{M}}^2(1-f)}{m}\sum_{i=1}^{m}\frac{\left(t_i - \overline{X}_{Cl,r}y_i\right)^2}{m-1}\ , \tag{2.6}$$

όπου $R = \sum_{i=1}^{M} t_i / \sum_{i=1}^{M} y_i$, $\overline{X}_{Cl,r} = X_{Cl,r}/N$ και $\widehat{\overline{M}} = N/\left(\frac{1}{m}\sum_{i=1}^{m} y_i\right)$.

Ορισμός 2.3. Ο εκτιμητής Horvitz – Thompson (1952) για το πληθυσμιακό ολικό Χ ορίζεται ως:

$$\hat{X} = X_{\mathrm{HT}} = \sum_{i=1}^{m}\frac{t_i}{\pi_i}\ , \tag{2.7}$$

όπου $\pi_i = m \cdot y_i \,/\, \sum_{i=1}^{M} y_i = m \cdot p_i$ είναι η πιθανότητα εκλογής της συστάδας **i** στο δείγμα και **m** το πλήθος συστάδων δείγματος.

Σύμφωνα με τη βιβλιογραφία, η διακύμανση του εκτιμητή X_{HT} για **m** πλήθος συστάδων δίνεται από τους δύο παρακάτω τύπους:

$$Var(X_{\mathrm{HT}}) = \sum_{i=1}^{M}\frac{(1-\pi_i)}{\pi_i}{t_i}^2 + \sum_{i=1}^{M}\sum_{j\neq i}^{M}\frac{\left(\pi_{ij} - \pi_i\pi_j\right)}{\pi_i\pi_j}t_i t_j\ , \tag{2.8}$$

$$Var(X_{\mathrm{HT}}) = \frac{1}{2}\sum_{i=1}^{M}\sum_{j\neq i}^{M}\frac{(\pi_i\pi_j - \pi_{ij})}{\pi_i\pi_j}\left(\frac{t_i}{\pi_i} - \frac{t_j}{\pi_j}\right)^2 . \qquad (2.9)$$

Οι ποσότητες π_{ij} , $1 \leq i,j \leq M$, αποτελούν τις από κοινού πιθανότητες εκλογής των συστάδων στο δείγμα.

Η πρώτη εξίσωση (2.8) προτάθηκε από τους Horvitz και Thompson (1952), ενώ η δεύτερη (2.9) διατυπώθηκε από τους Yates, Grundy και Sen (1953). Για τις εκτιμήσεις τους ισχύουν οι επόμενες σχέσεις:

$$\widehat{Var}(X_{\mathrm{HT}}) = \sum_{i=1}^{m}\frac{(1-\pi_i)}{{\pi_i}^2}{t_i}^2 + \sum_{i=1}^{m}\sum_{j\neq i}^{m}\frac{(\pi_{ij} - \pi_i\pi_j)}{\pi_i\pi_j\pi_{ij}}t_it_j , \qquad (2.10)$$

$$\widehat{Var}(X_{\mathrm{HT}}) = \sum_{i=1}^{m}\sum_{j>i}^{m}\frac{(\pi_i\pi_j - \pi_{ij})}{\pi_i\pi_j\pi_{ij}}\left(\frac{t_i}{\pi_i} - \frac{t_j}{\pi_j}\right)^2 . \qquad (2.11)$$

Ο προσδιορισμός των πιθανοτήτων π_{ij} είναι εν γένει δύσκολος ακόμη και για μικρό πλήθος συστάδων. Αντικαθιστώντας τη σχέση:

$$\pi_{ij} = \pi_i\pi_j\big(1 - (1-\pi_i)(1-\pi_j)[\textstyle\sum_{k=1}^{M}\pi_k(1-\pi_k)]^{-1}\big),\ 1 \leq i,j \leq M, \qquad (2.12)$$

στην εκτίμηση της διακύμανσης των Yates, Grundy και Sen (2.11), προκύπτουν οι επόμενες εξισώσεις για τον υπολογισμό της διακύμανσης του εκτιμητή X_{HT} και της εκτίμησης της (Hájek, 1964).

$$Var(X_{\mathrm{HT}}) = \frac{m}{m-1}\sum_{i=1}^{M}\pi_i(1-\pi_i)\left(\frac{t_i}{\pi_i} - A_M\right)^2 , \qquad (2.13)$$

$$\widehat{Var}(X_{HT}) = \frac{m}{m-1}\sum_{i=1}^{m}(1-\pi_i)\left(\frac{t_i}{\pi_i} - A_m\right)^2 , \qquad (2.14)$$

όπου

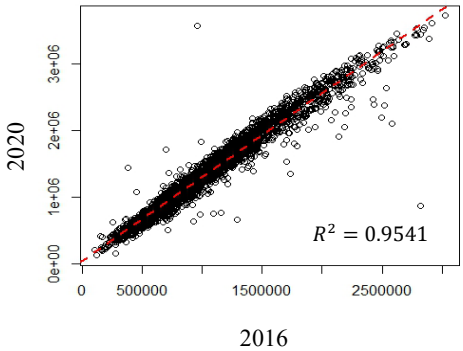
$$A_s = \sum_{k=1}^{s}\frac{1-\pi_k}{\sum_{j=1}^{s}(1-\pi_j)}\cdot\frac{t_k}{\pi_k} , \quad 1 \leq k \leq s \leq M . \qquad (2.15)$$

Οι τιμές των εκτιμήσεων της διακύμανσης του εκτιμητή X_{HT} που παρουσιάζονται στα αποτελέσματα της ενότητας 3 και 4 βασίζονται στις δύο προηγούμενες εξισώσεις.

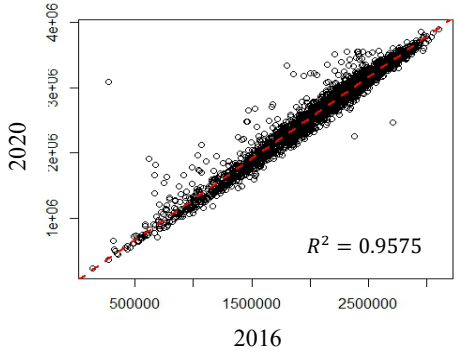
Όπως αναφέραμε στην εισαγωγή της εργασίας, γραμμική σχέση παρατηρούμε και μεταξύ των εκτιμήσεων των εκλογικών αποτελεσμάτων μεταξύ δύο διαδοχικών εκλογικών αναμετρήσεων. Προτού προβούμε στον ορισμό του εκτιμητή γραμμικής παλινδρόμησης, παρουσιάζουμε τα στικτογράμματα που προκύπτουν από τις τιμές της μεταβλητής

t_i/π_i , $1 \le i \le M$, για σταθερό $m = 30$ μεταξύ δύο διαδοχικών εκλογικών αναμετρήσεων.

Εικόνα 3. *Ψήφοι Δημοκρατικών*



Εικόνα 4. *Ψήφοι*



Παρατήρηση 2.1. Ο διορθωμένος συντελεστής Pearson (adjusted R squared) που αντιστοιχεί στα παραπάνω γραφήματα ισούται με 0.9541 και 0.9575 αντίστοιχα. Επομένως, η χρήση του ακόλουθου γραμμικού μοντέλου (Ορισμός 2.4) αναμένεται να συνεισφέρει στην ακρίβεια των εκτιμήσεων. Η ετεροσκεδαστικότητα των πληθυσμιακών καταλοίπων που παρατηρείται στα γραφήματα αντιμετωπίζεται με τον προσδιορισμό μίας νέας ευθείας με κλίση β_0 όπως θα δούμε στη συνέχεια (βλ. 2.18 και Παρατήρηση 2.2).

Ορισμός 2.4. Ο εκτιμητής γραμμικής παλινδρόμησης για δειγματοληψία ανάλογη του μεγέθους των συστάδων ορίζεται ως:

$$X_{REG}(s) = X_{HT}(s) + \hat{\beta}_0(s) \cdot \left(\mathrm{Y} - Y_{HT}(s)\right) , \quad s \in S^m . \tag{2.16}$$

Συμβολίζουμε ως S^m το δειγματικό χώρο για δείγματα πλήθους **m** και στοιχεία της μορφής $s = (x_1, x_2, x_3, \dots , x_m) \in S^m$. Ακόμη, ισχύει ότι:

$$\hat{\beta}_0(s) = \sum_{\mathrm{i}=1}^{\mathrm{m}} \left(\frac{\mathrm{x_i}}{\pi_\mathrm{i}} - \frac{\hat{X}}{\mathrm{m}}\right)\left(\frac{\mathrm{y_i}}{\pi'_\mathrm{i}} - \frac{\hat{Y}}{\mathrm{m}}\right) \Big/ \sum_{\mathrm{i}=1}^{\mathrm{m}} \left(\frac{\mathrm{y_i}}{\pi'_\mathrm{i}} - \frac{\hat{Y}}{\mathrm{m}}\right)^2 , \tag{2.17}$$

όπου $\hat{X} = X_{HT}(s)$ και $\hat{Y} = Y_{HT}(s)$, $s \in S^m$. Επιπλέον,

$$\beta_0 = \sum_{\mathrm{i}=1}^{M} \pi_i \left(\frac{\mathrm{x_i}}{\pi_\mathrm{i}} - \frac{\mathrm{X}}{m}\right)\left(\frac{\mathrm{y_i}}{\pi'_\mathrm{i}} - \frac{\mathrm{Y}}{\mathrm{m}}\right) \Big/ \sum_{\mathrm{i}=1}^{\mathrm{M}} \pi_i \left(\frac{\mathrm{y_i}}{\pi'_\mathrm{i}} - \frac{\mathrm{Y}}{\mathrm{m}}\right)^2 , \tag{2.18}$$

όπου π_i, π'_i $1 \le i \le M$, είναι οι πιθανότητες συμπερίληψης των συστάδων στις εκλογικές αναμετρήσεις του 2020 (ή 2016) και του 2016 (ή 2012) αντίστοιχα, ενώ ο συντελεστής παλινδρόμησης β_0 ελαχιστοποιεί την προτεινόμενη διακύμανση του εκτιμητή (2.16) (Cochran, 1977). Για τη

διακύμανση του συντελεστή $\hat{\beta}_0$ και την εκτίμηση της διακύμανσής του ισχύουν οι παρακάτω εξισώσεις (Draper, N. R., & Smith, H. (1998)):

$$\sigma^2(\hat{\beta}) = \sigma^2(\varepsilon_{i,m}) \cdot \left[\sum_{\mathrm{i}=1}^{\mathrm{m}} \left(\frac{\mathrm{y_i}}{\pi_\mathrm{i}'} - \frac{\hat{Y}}{\mathrm{m}}\right)^2\right]^{-1} \tag{2.19}$$

και

$$s^2(\hat{\beta}) = \frac{\sum_{i=1}^{m}(\hat{\varepsilon}_{i,m})^2}{m-2} \cdot \left[\sum_{\mathrm{i}=1}^{\mathrm{m}} \left(\frac{\mathrm{y_i}}{\pi_\mathrm{i}'} - \frac{\hat{Y}}{\mathrm{m}}\right)^2\right]^{-1}, \tag{2.20}$$

όπου

$$\varepsilon_{i,m} = \left(\frac{x_i}{\pi_i} - \frac{X}{m}\right) - \beta_0\left(\frac{y_i}{\pi_i'} - \frac{Y}{m}\right), \ 1 \le i \le M \tag{2.21}$$

και

$$\hat{\varepsilon}_{i,m} = \left(\frac{x_i}{\pi_i} - \frac{\hat{X}}{m}\right) - \hat{\beta}_0\left(\frac{y_i}{\pi_i'} - \frac{\hat{Y}}{m}\right), \ 1 \le i \le M, \tag{2.22}$$

αποτελούν τα κατάλοιπα (residuals) του πληθυσμού και τις εκτιμήσεις αυτών.

Παρατήρηση 2.2. Οι συντελεστές β_0 (2.18) και $\hat{\beta}_0$ (2.17) ελαχιστοποιούν τις συναρτήσεις (2.23) και (2.24) αντίστοιχα (Cochran, 1977). Επιλέον, κατά τη διεξαγωγή μιας σειράς **n** πειραμάτων παρατηρήθηκε ότι η μέση τιμή των συντελεστών $\hat{\beta}_0$ που προκύπτουν από τη μέθοδο ελαχίστων τετραγώνων στο κάθε δείγμα ταυτίζεται με τον συντελεστή (2.18). Ο συντελεστής β_0 (2.18) προκύπτει από την εφαρμογή της μεθόδου ελαχίστων τετραγώνων στον πληθυσμό χρησιμοποιώντας ως βάρη τις πιθανότητες συμπερίληψης $\pi_i, 1 \le i \le M$.

Ορισμός 2.5. Η συνάρτηση (2.23) (Cochran, 1977) ορίζεται ως η διακύμανση του εκτιμητή γραμμικής παλινδρόμησης υπό την προϋπόθεση ότι $\pi_i < 1$, $1 \le i \le M$. Δηλαδή, έχουμε ότι:

$$Var(X_{REG}) = \frac{1}{m-1}\sum_{\mathrm{i}\epsilon\mathrm{A}} \pi_\mathrm{i}(\varepsilon_{i,m})^2 \left[m + (\hat{Y} - Y)^2 \Big/ \sum_{\iota=1}^{m}\left(\frac{\hat{Y}}{m} - \frac{Y}{m}\right)^2\right], \tag{2.23}$$

οπου β_0 είναι ο συντελεστής που ελαχιστοποιεί την (2.23) και

$$A = \{\, i\,, 1 \le i \le M \mid \pi_i < 1 \,\}.$$

Για την εκτίμηση της (2.23) (Cochran W. 1977) ισχύει ότι,

$$\widehat{Var}(X_{REG}) = \frac{1}{m-2}\sum_{i=1}^{\mathrm{m}} (\hat{\varepsilon}_{i,m})^2 \left[m + (\hat{Y} - Y)^2 \Big/ \sum_{\iota=1}^{m}\left(\frac{\mathrm{y_i}}{\pi_\mathrm{i}'} - \frac{\hat{Y}}{m}\right)^2\right] \tag{2.24}$$

Προσαρμόζοντας τη θεωρία εκτιμητών παλινδρόμησης που ανέπτυξε ο W. G. Cochran (1977) στο δειγματοληπτικό μας πλάνο, προκειμένου να μπορεί να εφαρμοστεί το μοντέλο γραμμικής παλινδρόμησης πρέπει ο πληθυσμός μας να πληροί τις παρακάτω δύο προϋποθέσεις:

$$\sum_{i=1}^{M} \pi_i \varepsilon_{i,m} = 0 \quad \kappa\alpha\iota \quad \sum_{i=1}^{M} \pi_i \varepsilon_{i,m} \left(\frac{y_i}{\pi_i'} - \frac{Y}{m} \right) = 0 \,. \tag{2.25}$$

Εν γένει, στη θεωρία της γραμμικής παλινδρόμησης είναι απαραίτητο τα κατάλοιπα $\varepsilon_{i,m}$ (residuals 2.22) να είναι ανεξάρτητα μεταξύ τους και να μην παρατηρείται ετεροσκεδαστικότητα στο δείγμα. Για την επαλήθευση των δύο αυτών προϋποθέσεων χρησιμοποιήθηκαν οι έλεγχοι Breusch – Godfrey (autocorrelation) και Breusch – Pagan (homoscedasticity). Ενδεικτικά, αναφέρουμε ότι για $m = 30$, τα αποτελέσματα των ελέγχων για την πληρότητα των προϋποθέσεων πάνω σε 4000 δείγματα έδειξαν ότι για το 95% των δειγμάτων ικανοποιούνται οι μηδενικές υποθέσεις των ελέγχων αυτών.

Παρατήρηση 2.3. Για μεγάλες τιμές του $\boldsymbol{m}$ ($\boldsymbol{m \geq 30}$) ο 2ος όρος εντός της αγγύλης στις εξισώσεις (2.23) και (2.24) είναι αμελητέος. Κατά τη διεξαγωγή του πειράματος παρατηρήθηκε ότι η συμπερίληψη του δεύτερου όρου οδηγεί σε καλύτερη εκτίμηση της διακύμανσης όσον αφορά στα αποτελέσματα του κόμματος των Ρεπουμπλικάνων, δηλαδή επιτυγχάνεται ποσοστό ακρίβειας $\boldsymbol{p} = 95\%$ για μικρότερο πλήθος συστάδων σε σύγκριση με μία εκτίμηση που δεν περιείχε τον δεύτερο όρο. Το επίπεδο της μεροληψίας του εκτιμητή παλινδρόμησης διαφέρει αναλόγως με τη μεταβλητή που εξετάζουμε (ψήφοι κάθε υποψήφιου) και τη χρονική περίοδο που μελετούμε.

Παρατήρηση 2.4. Οι όροι στη σχέση (2.23) με πιθανότητα εκλογής στο δείγμα $\pi_i = 1$ δε συνεισφέρουν στον υπολογισμό της διακύμανσης, οπότε παραλείπονται. Σημαντική προϋπόθεση της τεχνικής που παρουσιάζουμε είναι ο περιορισμός του πλήθους δείγματος ώστε να μην υπάρχουν πολλές συστάδες με την ιδιότητα που αναφέραμε στον πληθυσμό. Υπενθυμίζουμε ότι ισχύει $\pi_i = mp_i \,, 1 \leq i \leq M$.

Ορισμός 2.5. Έστω $\hat{\theta}$ ο εκτιμητής της παραμέτρου θ. Ορίζεται η ρίζα του μέσου τετραγωνικού σφάλματος (**RMSE**) ως:

$$RMSE(\hat{\theta}) = \sqrt{E\left[\left(\theta - \hat{\theta}\right)^2\right]} \,. \tag{2.26}$$

3. ΣΥΓΚΡΙΣΗ ΕΚΤΙΜΗΤΩΝ

Σε πρώτο στάδιο, επιλέγουμε να συγκρίνουμε τους εκτιμητές $X_{\mathrm{Cl,u}}$ και $X_{\mathrm{Cl},r}$, καθότι αμφότεροι βασίζονται στη μέθοδο της απλής τυχαίας δειγματοληψίας. Στον επόμενο πίνακα παρουσιάζονται τα ποσοστά ακρίβειας $\overline{\boldsymbol{p}}$ και $\overline{\boldsymbol{c}}$ που ορίζουμε ως το ημιάθροισμα των αντίστοιχων ποσοστών $\boldsymbol{p}$ και $\boldsymbol{c}$ των δύο

υποψηφίων, όπως αυτά παρουσιάστηκαν στην ενότητα 2, καθώς και το ημιάθροισμα των ριζών των μέσων τετραγωνικών σφαλμάτων των δύο υποψηφίων ($\overline{RMSE}$) για $\boldsymbol{m}$ πλήθος συστάδων δείγματος πάνω στις δύο διαμερίσεις για μέγιστο αριθμό δειγμάτων $\mathbf{n} = 10000$. Όταν ο εκτιμητής είναι αμερόληπτος ισχύει ότι $\boldsymbol{p} = \boldsymbol{c}$ και το πρώτο ποσοστό παραλείπεται.

Πίνακας 3.1. *Σύγκριση Εκτιμητών* $X_{Cl,u} - X_{Cl,r}(Ratio)$

Διαμέριση	Εκτιμητής	$\bar{p}$	$\bar{c}$	$\overline{RMSE}$	Πλήθος συστάδων m	Πληθυσμός %
Πολιτείες	$X_{Cl,u}$	-	0.880	3.627e+07	5	9.87
	Ratio	0.875	0.853	6.109e+06	5	9.87
	$X_{Cl,u}$	-	0.881	2.978e+07	7	13.69
	Ratio	0.868	0.842	5.284e+06	7	13.69
	$X_{Cl,u}$	-	0.884	2.421e+07	10	19.67
	Ratio	0.864	0.849	4.510e+06	10	19.67
Επαρχίες	$X_{Cl,u}$	-	0.808	3.614e+07	40	1.28
	Ratio	0.801	0.769	8.547e+06	40	1.28
	$X_{Cl,u}$	-	0.880	1.913e+07	150	4.85
	Ratio	0.882	0.865	5.183e+06	150	4.85
	$X_{Cl,u}$	-	0.910	1.143e+07	380	12.24
	Ratio	0.909	0.901	3.454e+06	380	12.24

Παρατηρήσεις

1) Μειονέκτημα της διακύνμασης του πρώτου εκτιμητή αλλά και της εκτίμησης της διακύμανσής του αποτελεί η εξάρτηση τους από τη μεταβλητότητα των τιμών της τυχαίας μεταβλητής μας. Λόγω της α.τ.δ. είναι αρκετά πιθανό να παρουσιαστούν στο δείγμα μας συστάδες στις οποίες οι διαφορές μεταξύ των έγκυρων ψήφων είναι σημαντικές. Ως αποτέλεσμα, έχουμε ότι το εύρος των διαστημάτων εμπιστοσύνης να παρουσιάσει εξίσου μεγάλη μεταβλητότητα στις διαδοχικές επαναλήψεις της διαδικασίας. Συγκεκριμένα, παρατηρούμε ότι η ρίζα του μέσου τετραγωνικού σφάλματος του πρώτου εκτιμητή είναι σημαντικά μεγαλύτερη από την αντίστοιχη του εκτιμητή Λόγου. Η επιλογή αποκλειστικά πολιτειών με παρόμοια μεγέθη, αν και θα βοηθούσε στη συρρίκνωση του σφάλματος, δε βελτιώνει το επίπεδο ακρίβειας **c** των σημειακών εκτιμήσεων.

2) Το μέγεθος της εκτίμησης της διακύμανσης του εκτιμητή Λόγου συνδέεται άμεσα με τη γραμμική σχέση που υπάρχει ανάμεσα στο πλήθος των έγκυρων ψηφοδελτίων του έκαστου κόμματος και του συνολικού αριθμού των έγκυρων ψηφοδελτίων σε κάθε συστάδα. Παρόμοια γραμμική σχέση συναντάται επίσης ανάμεσα στο πλήθος των ψηφοδελτίων κάθε συστάδας, μεταξύ δύο διαδοχικών

εκλογικών αναμετρήσεων και για τους δύο υποψήφιους. Το γεγονός αυτό αποτελέσε τη βάση για τη δημιουργία του εκτιμητή (2.4).

3) Ως μειονεκτήματα για τον εκτιμητή Λόγου έχουμε ότι δεν είναι αμερόληπτος και πως για την εκτίμηση του ολικού Χ είναι απαραίτητη η γνώση του συνολικού πλήθους των έγκυρων ψηφοδελτίων.

Βάσει των άνωθεν παρατηρήσεων συμπεραίνουμε ότι ο εκτιμητής Λόγου, παρότι δεν είναι αμερόληπτος, υπερτερεί έναντι του εκτιμητή μονοσταδιακής δειγματοληψίας κατά συστάδες (2.1) λόγω της σημαντικής διαφοράς ανάμεσα στις εκτιμήσεις των μέσων τετραγωνικών σφαλμάτων τους.

Η επόμενη σύγκριση θα πραγματοποιηθεί ανάμεσα στον εκτιμητή Λόγου και τον εκτιμητή Horvitz – Thompson. Ακολουθούμε την ίδια μεθοδολογία και παραθέτουμε τα αποτελέσματα στον πίνακα που έπεται.

Πίνακας 3.2. *Σύγκριση Εκτιμητών* $X_{HT} - X_{Cl,r}(Ratio)$

Διαμέριση	Εκτιμητής	$\bar{p}$	$\bar{c}$	$\overline{RMSE}$	Πλήθος συστάδων m	Πληθυσμός %
Πολιτείες	X_{HT}	-	0.952	5.741e+06	5	20.69
	Ratio	0.875	0.853	6.109e+06	5	9.87
	X_{HT}	-	0.953	4.548e+06	7	29.07
	Ratio	0.868	0.842	5.284e+06	7	13.69
	X_{HT}	-	0.944	3.424e+06	10	40.55
	Ratio	0.864	0.849	4.510e+06	10	19.67
	X_{HT}	-	0.942	2.499e+06	15	55.95
	Ratio	0.874	0.860	3.713e+06	15	29.41
Επαρχίες	X_{HT}	-	0.949	6.669e+06	15	4.65
	Ratio	0.682	0.637	1.166e+07	15	0.49
	X_{HT}	-	0.957	3.908e+06	40	12.32
	Ratio	0.801	0.769	8.547e+06	40	1.28
	X_{HT}	-	0.959	1.688e+06	150	35.29
	Ratio	0.882	0.865	5.183e+06	150	4.85
	X_{HT}	-	0.959	6.795e+05	380	61.40
	Ratio	0.909	0.901	3.454e+06	380	12.24

Παρατηρήσεις

1) Όπως αναφέρθηκε στην αρχή της παρούσας ενότητας, ο εκτιμητής Λόγου βασίζεται σε απλή τυχαία δειγματοληψία, ενώ για τον εκτιμητή Horvitz – Thompson εφαρμόζουμε δειγματοληψία με πιθανότητες ανάλογες του μεγέθους των συστάδων. Στην περίπτωση που όλες οι συστάδες έχουν την ίδια πιθανότητα ο εκτιμητής Horvitz – Thompson ταυτίζεται με τον εκτιμητή του ορισμού 2.1. Κατά τη σύγκριση δύο εκτιμητών που βασίζονται σε διαφορετικές δειγματοληπτικές μεθόδους πρέπει να δοθεί ιδιαίτερη προσοχή στο ποσοστό

του πληθυσμού που αντιστοιχεί σε δεδομένο πλήθος συστάδων **m**. Όταν εφαρμόζουμε δειγματοληψία ανάλογη του μεγέθους των δειγματοληπτικών μονάδων είναι σύνηθες να εμφανίζονται με μεγαλύτερη συχνότητα στο δείγμα μας συστάδες μεγάλου μεγέθους.

2) Λαμβάνοντας υπόψη την παρατήρηση (1) διαπιστώνουμε πως για να πραγματοποιηθεί ορθή σύγκριση ανάμεσα στους δύο εκτιμητές, πρέπει να εξασφαλίσουμε ότι τα δείγματα που χρησιμοποιούμε αντιστοιχούν στο ίδιο ποσοστό του πληθυσμού σε κάθε περίπτωση.

3) Τα χαμηλά ποσοστά ***p*** και ***c*** του εκτιμητή Λόγου σε σχέση με τον εκτιμητή Horvitz – Thompson, αν και ο πρώτος παρουσιάζει μικρότερο μέσο τετραγωνικό σφάλμα, μας ωθούν στην επιλογή του δεύτερου ως πιο αξιόπιστου εκτιμητή.

Προφανώς, η επιλογή των επαρχιών ως συστάδων μας δίνει ακριβέστερα κατά τη διεξαγωγή των επαναλήψεων του πειράματος για $m \geq 40$. Στο σημείο αυτό επισημαίνουμε ότι για τις προβλέψεις αμφότερων των εκτιμητών είναι αναγκαία η γνώση του αριθμού (ή μιας εκτίμησης) των έγκυρων ψηφοδελτίων κάθε συστάδας με σκοπό τον υπολογισμό του πλήθους **N** στον εκτιμητή Λόγου αλλά και για τον υπολογισμό των πιθανοτήτων συπερίληψης π_iτου εκτιμητή Horvitz – Thompson.

4. ΕΚΤΙΜΗΤΗΣ ΓΡΑΜΜΙΚΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

Σκοπός μας στην ενότητα αυτή είναι να προβλέψουμε τα εκλογικά αποτελέσματα λαμβάνοντας υπόψη την έκβαση των εκλογών της προηγούμενης τετραετίας. Ο στόχος αυτός επιτυγχάνεται μέσω της εφαρμογής και αξιολόγησης του μοντέλου γραμμικής παλινδρόμησης που ορίστηκε στην ενότητα 2 πάνω στις δύο διαθέσιμες διαμερίσεις των Ηνωμένων Πολιτειών της Αμερικής. Προκειμένου να θεμελιώσουμε την αποτελεσματικότητα του εκτιμητή που έχουμε ορίσει, καθώς και της εκτίμησης της διακύμανσής του που προτείναμε, θα εξετάσουμε την ακρίβεια των προβλέψεων ανάμεσα σε τρεις διαδοχικές εκλογικές αναμετρήσεις. Αν και στη βιβλιογραφία υπάρχει διαθέσιμος εκτιμητής γραμμικής παλινδρόμησης (Cochran 1977), ο εκτιμητής που προτείνουμε καθώς και η εκτιμώμενη διακύμανσή του βασίζονται στη μέθοδο δειγματοληψίας με βάρη ανάλογα του μεγέθους των συστάδων.

Αρχικά, επιλέγουμε να συγκρίνουμε τον εκτιμητή Horvitz – Thompson με τον εκτιμητή γραμμικής παλινδρόμησης πάνω στις διαμερίσεις που έχουμε ορίσει για διαφορετικά πλήθη συστάδων. Ακολουθούν τα αποτελέσματα της πρώτης σύγκρισης.

Πίνακας 4.1. *Σύγκριση Εκτιμητών* $X_{HT} - X_{REG}$

Διαμέριση	Εκτιμητής	$\bar{p}$	$\bar{c}$	$\overline{RMSE}$	Πλήθος συστάδων m	Πληθυσμός %
Πολιτείες	X_{HT}	-	0.952	5.741e+06	5	20.69
	X_{REG}	0.937	0.937	1.433e+06	5	20.69
	X_{HT}	-	0.958	5.045e+06	6	24.74
	X_{REG}	0.953	0.949	1.233e+06	6	24.74
	X_{HT}	-	0.953	4.548e+06	7	29.07
	X_{REG}	0.958	0.952	1.097e+06	7	29.07
Επαρχίες	X_{HT}	-	0.949	6.669e+06	15	4.65
	X_{REG}	0.942	0.933	1.461e+06	15	4.65
	X_{HT}	-	0.954	5.079e+06	25	7.79
	X_{REG}	0.950	0.901	1.246e+06	25	7.79
	X_{HT}	-	0.954	4.585e+06	30	9.43
	X_{REG}	0.953	0.875	1.190e+06	30	9.43
	X_{HT}	-	0.951	4.221e+06	35	10.94
	X_{REG}	0.954	0.855	1.162e+06	35	10.94
	X_{HT}	-	0.957	3.908e+06	40	12.35
	X_{REG}	0.961	0.871	1.065e+06	40	12.35

Σύμφωνα με τα παραπάνω στοιχεία διακρίνουμε ότι το μέσο τετραγωνικό σφάλμα του εκτιμητή μας είναι σταθερά μικρότερο από το αντίστοιχο του X_{HT}. Επιπροσθέτως, παρατηρούμε τη μείωση του ποσοστού **c** καθώς ο αριθμός των συστάδων αυξάνεται. Το γεγονός αυτό συνδέεται με τη μεροληψία του εκτιμητή παλινδρόμησης, καθώς η αύξηση του πλήθους των συστάδων στο δείγμα έχει ως αποτέλεσμα τη συσσώρευση των σημειακών εκτιμήσεων περί της μέσης τιμής του και την ελάττωση του εύρους των διαστημάτων εμπιστοσύνης.

Η επόμενη εφαρμογή αποτελεί σύγκριση των προβλέψεων του εκτιμητή γραμμικής παλινδρόμησης ανάμεσα σε τρεις διαδοχικές εκλογικές αναμετρήσεις. Θα εξετάσουμε το επίπεδο ακρίβειας με το οποίο μπορούμε να εκτιμήσουμε τα αποτελέσματα της λαϊκής ψηφοφορίας (**popular vote**) του 2016, συνδυάζοντάς τα με τα εκλογικά δεδομένα του 2012. Έπειτα, κάνοντας σύγκριση με τις πληροφορίες που εξήχθησαν από την αντίστοιχη εφαρμογή για την τετραετία 2016-2020, θα λάβουμε τις πρώτες ενδείξεις για το πλήθος συστάδων που χρειαζόμαστε στο δείγμα μας ώστε να επιτύχουμε την επιθυμητή ακρίβεια.

Πίνακας 4.2. *Σύγκριση Εκτιμήσεων X_{REG} , 2016 - 2020*

Διαμέριση	Εκτιμητής	$\bar{p}$	$\bar{c}$	$\overline{RMSE}$	Πλήθος συστάδων m	Πληθυσμός %
Πολιτείες	$X_{REG}(2012 - 2016)$	0.947	0.945	2.261e+06	5	19.84

	$X_{REG}(2016-2020)$	0.937	0.937	1.433e+06	5	20.69
	$X_{REG}(2012-2016)$	0.945	0.940	1.779e+06	6	23.98
	$X_{REG}(2016-2020)$	0.953	0.949	1.233e+06	6	24.74
	$X_{REG}(2012-2016)$	0.952	0.950	1.601e+06	7	28.05
	$X_{REG}(2016-2020)$	0.958	0.952	1.097e+06	7	29.07
Επαρχίες	$X_{REG}(2012-2016)$	0.942	0.941	1.699e+06	15	3.85
	$X_{REG}(2016-2020)$	0.942	0.933	1.461e+06	15	4.65
	$X_{REG}(2012-2016)$	0.951	0.945	1.329e+06	25	6.40
	$X_{REG}(2016-2020)$	0.950	0.901	1.246e+06	25	7.79
	$X_{REG}(2012-2016)$	0.953	0.944	1.208e+06	30	7.71
	$X_{REG}(2016-2020)$	0.953	0.875	1.190e+06	30	9.43
	$X_{REG}(2012-2016)$	0.955	0.950	1.053e+06	40	10.28
	$X_{REG}(2016-2020)$	0.961	0.871	1.065e+06	40	12.35

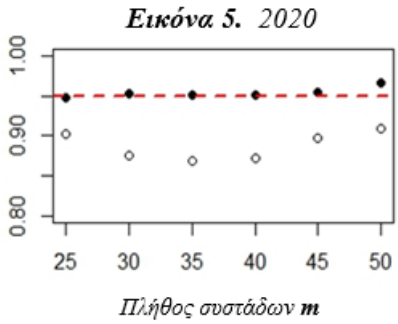
Για την επόμενη ανάλυση, εισάγουμε ένα νέο μέγεθος που σε συνδυασμό με τα ανωτέρω αποτελέσματα, θα μας δώσει μία πληρέστερη εικόνα για την ακρίβεια των εκτιμήσεων μας και θα συμβάλει καταλυτικά στον προσδιορισμό του πλήθους συστάδων που χρειαζόμαστε για την εξαγωγή εκτιμήσεων βάσει δεδομένης ακρίβειας. Συμβολίζουμε με $\boldsymbol{m'}$ το πλήθος συστάδων που χρειαζόμαστε, ώστε με πιθανότητα 95% η εκτίμηση του τυπικού σφάλματος που προκύπτει σε κάθε επανάληψη του προγράμματος να είναι μικρότερη από το αναγραφόμενο επίπεδο τυπικού σφάλματος.

Πίνακας 4.3. *Φράγματα Τυπικής Απόκλισης*

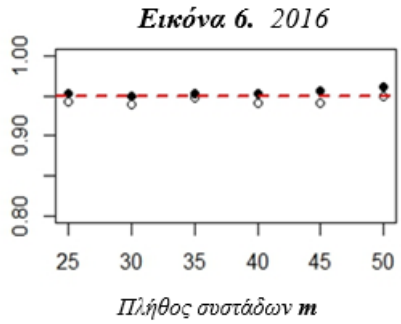
Εκτιμητής	95ο Ποσοστημόριο τυπικής απόκλισης 2020	$\boldsymbol{m'}(2020)$	95ο Ποσοστημόριο τυπικής απόκλισης 2016	$\boldsymbol{m'}(2016)$

ΔΗΜ	1.218e+06	27	1.184e+06	39
ΡΕΠ	1.219e+06	31	1.144e+06	63
ΔΗΜ	2.119e+06	13	2.087e+06	16
ΡΕΠ	2.030e+06	15	1.952e+06	29

Η ανάλυση του τυπικού σφάλματος μας παρέχει νέες πληροφορίες για την ακρίβεια των αναμενόμενων εκτιμήσεων που δε θα ήταν προφανείς αν βασιζόμασταν αποκλειστικά στην εκτίμηση του μέσου τετραγωνικού σφάλματος (πίνακας 4.2). Ειδικά στη μελέτη της διαμέρισης των επαρχιών για πλήθος συστάδων άνω των 25, οι εκτιμήσεις του μέσου τετραγωνικού σφάλματος σχεδόν ταυτίζονται στις δύο αναμετρήσεις αλλά τα επίπεδα μεροληψίας τους διαφέρουν σημαντικά. Στο σημείο αυτό ανακαλούμε από την εισαγωγή το δεύτερο από τα μειονεκτήματα της καθολικής εξάρτησης σε ένα μοντέλο πρόβλεψης, δηλαδή την ικανότητα του να δώσει έγκυρες προβλέψεις σε δύο διαδοχικές αναμετρήσεις. Στα παρακάτω γραφήματα αναπαρίστανται τα ποσοτά $\boldsymbol{p}$ και $\boldsymbol{c}$ (ως μαύρες και λευκές κουκκίδες), ξεχωριστά για τους δύο υποψηφίους, ως συναρτήσεις του πλήθους συστάδων δείγματος για τον εκτιμητή παλινδρόμησης. Ακολουθούν τα γραφήματα για το κόμμα των Δημοκρατικών.



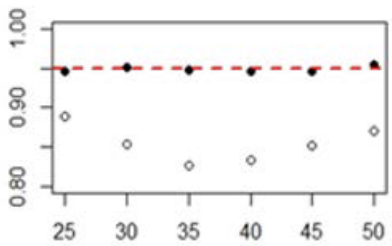
Εικόνα 5. *2020*



Εικόνα 6. *2016*

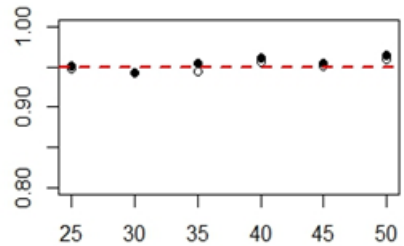
Στη συνέχεια παρουσιάζονται τα αντίστοιχα διαγράμματα για το κόμμα των Ρεπουμπλικάνων.

Εικόνα 7. *2020*



Πλήθος συστάδων ***m***

Εικόνα 8.



Πλήθος συστάδων ***m***

Η αύξηση της μεροληψίας που παρατηρείται κατά την τετραετία 2016-2020 οφείλεται στις διαφορές των πιθανοτήτων π_i και π_i'. Η φορά του σφάλματος που προέκυψε, δηλαδή η υποτίμηση ή υπερτίμηση των ψήφων του έκαστου υποψηφίου, μπορεί να εκτιμηθεί με τη χρήση συνδυαστικών μοντέλων (Graefe, 2018). Κλείνοντας, παραθέτουμε κάποιες από τις εκτιμήσεις που λάβαμε από τους εκτιμητές (2.2), (2.3) και (2.4) μαζί με τα πραγματικά αποτελέσματα των εκλογών για τα έτη 2016 και 2020. Τα δείγματα για τους εκτιμητές επιλέχθηκαν από $\mathbf{n} = 10000$ διαδοχικές επαναλήψεις με τρόπο τέτοιο ώστε να αντιπροσωπεύουν το μικρότερο δυνατό ποσοστό πληθυσμού από το οποίο μπορούμε να εξάγουμε τις πιο ακριβείς προβλέψεις.

Πίνακας 4.4. *Προβλέψεις τελικού αποτελέσματος 2020*

Εκτιμητής	*ΔΗΜ Ποσοστό %*	*ΔΗΜ Απόκλιση %*	*ΡΕΠ Ποσοστό %*	*ΡΕΠ Απόκλιση %*	Πληθυσμός % - συστάδες
X_{REG}	51.05	0.73	46.62	0.63	0.9 % - 15
X_{HT}	51.73	2.44	46.36	2.48	3.83 % - 25
$X_{Cl,r}$	52.28	2.31	45.91	2.28	6.47 % - 250
Αποτελέσματα 2020	*DEM*	51.32	*GOP*	46.83	

Πίνακας 4.5. *Προβλέψεις τελικού αποτελέσματος 2016*

Εκτιμητής	*ΔΗΜ Ποσοστό %*	*ΔΗΜ Απόκλιση %*	*ΡΕΠ Ποσοστό %*	*ΡΕΠ Απόκλιση %*	Πληθυσμός % - συστάδες
X_{REG}	48.03	0.97	47.14	1.09	3.51 % - 15
X_{HT}	48.06	2.73	46.83	2.85	4.27 % - 25
$X_{Cl,r}$	48.89	2.68	46.54	2.65	5.78 % - 250
Αποτελέσματα 2016	*DEM*	47.83	*GOP*	47.30	

5. ΣΥΜΠΕΡΑΣΜΑΤΑ

Λαμβάνοντας υπόψη τα αποτελέσματα των ενοτήτων 3 και 4 συμπεραίνουμε ότι η εφαρμογή της μεθόδου γραμμικής παλινδρόμησης στην πρόβλεψη εκλογικών αποτελεσμάτων συνεισφέρει σημαντικά στη μείωση του εύρους των διαστημάτων εμπιστοσύνης και στην ακρίβεια των σημειακών εκτιμήσεων συγκριτικά με τους τρεις πρώτους εκτιμητές που παρουσιάσαμε, για όλα τα πλήθη συστάδων δείγματος που ελέγχθησαν. Επιπλέον, η καθεαυτή επιλογή της σύγκρισης των εκτιμητών πάνω σε εκλογικά αποτελέσματα σε αντίθεση με την επιλογή ενός οποιοδήποτε άλλου αρχείου δεδομένων μας παρείχε έναν τρόπο επαλήθευσης της αποτελεσματικότητας τους σε πληθυσμιακό επίπεδο. Ωστόσο, ιδιαίτερη μνεία πρέπει να δοθεί στις προϋποθέσεις χρήσης των εκτιμητών και τη στατιστική θεωρία τους συνοδεύει, όπως για παράδειγμα τη γραμμική σχέση των δεδομένων στην περίπτωση του εκτιμητή παλινδρόμησης. Στα αρνητικά των μεθόδων που εφαρμόσαμε καταλογίζουμε ότι ο εκτιμητής παλινδρόμησης δεν είναι αμερόληπτος (Cochran 1977). Στις περιπτώσεις που εξετάσαμε όμως, το επίπεδο της μεροληψίας δεν υπερέβηκε το τυπικό σφάλμα του προτεινόμενου εκτιμητή (2.4) και ειδικά στη διαμέριση των πολιτειών η τάξη της μεροληψίας ήταν αμελητέα.

Κατά τη διεξαγωγή των συγκρίσεων της ενότητας 4 παρατηρήθηκε ότι η λεπτότερη διαμέριση του πληθυσμού οδηγεί σε ακριβέστερες προβλέψεις. Βασιζόμενοι στα στοιχεία του πίνακα 4.3, γνωρίζουμε ότι το 95% των εκτιμήσεων της τυπικής απόκλισης δεν θα ξεπερνά τα 1,2 εκατομμύρια όταν το $\boldsymbol{m'}$ ανήκει στο διάστημα [27, 39] για τον υπολογισμό των αποτελεσμάτων του δημοκρατικού κόμματος και στο διάστημα [31, 63] για το κόμμα των Ρεπουμπλικάνων. Βεβαίως, η εξαγωγή προβλέψεων μπορεί να πραγματοποιηθεί και για μικρότερο πλήθος συστάδων, όπως φαίνεται στον πίνακα 4.4, αλλά το μέσο τετραγωνικό σφάλμα θα αυξηθεί (βλ. πίνακα 4.2). Συγκεκριμένα, τα διαστήματα πλήθους συστάδων που προκύπτουν ώστε το 95% των εκτιμήσεων της τυπικής απόκλισης του εκτιμητή παλινδρόμησης να είναι μικρότερες ή ίσες των 2,1 εκατομμυρίων για τους υποψηφίους των δύο κομμάτων είναι [13, 16] και [15, 29] για τους Δημοκρατικούς και τους Ρεπουμπλικάνους αντίστοιχα (βλ. πίνακα 4.3). Σε συνδυασμό με τα στοιχεία του πίνακα 4.2., αν απαιτήσουμε επιπλέον το μέσο τετραγωνικό σφάλμα να μην υπερβαίνει τα 1,2 εκατομμύρια, διαπιστώνουμε ότι θα χρειαστούμε τουλάχιστον 30 συστάδες στο δείγμα μας. Ο περιορισμός που θέσαμε είναι ισοδύναμος με την απαίτηση ο συντελεστής μεταβολής (**cv**) που αντιστοιχεί στο δειγματοληπτικό μας πλάνο να είναι ≤ 0.01 για τις εκλογές του 2020 και ≤ 0.025 για τις εκλογές του 2016. Οι εκτιμήσεις για το απαιτούμενο πλήθος συστάδων

που παρουσιάζονται διασφαλίζουν ότι ακόμα και στην περίπτωση μεγάλης εκλογικής ανατροπής, όπως παρατηρήθηκε στις αμερικάνικες εκλογές του 2016, θα έχουμε προβλέψεις της ακρίβειας που απαιτήσαμε. Όσον αφορά στον ελάχιστο αριθμό συστάδων που χρειαζόμαστε στο δείγμα ώστε να προκύπτουν διαστήματα επιπέδου εμπιστοσύνης 95% ($\boldsymbol{p}$ = 0.95), από τους πίνακες των ενοτήτων 2 και 3 είναι προφανές ότι αρκούν 5 και 15 συστάδες για τις διαμερίσεις των πολιτειών και των επαρχιών αντίστοιχα. Το χαμηλό ποσοστο $\boldsymbol{c}$ που παρατηρείται στις προβλέψεις του 2020 για μεγάλο πλήθος συστάδων αντισταθμίζεται αυξάνοντας τον συντελεστή εμπιστοσύνης στο 99%, με τα παραγώμενα διαστήματα εμπιστοσύνης να μην παρουσιάζουν σημαντική μεταβολή του εύρους τους. Τελος, απορρίπτουμε τη χρήση της διαμέρισης των πολιτειών λόγω του ποσοστού του πληθυσμού που απαιτείται για να μας παρέχει ακριβείς προβλέψεις ($\geq$20%), του χαμηλότερου ποσοστού ακρίβειας $\boldsymbol{c}$ και του μεγαλύτερου μέσου τετραγωνικού σφάλματος σε σύγκριση με τη διαμέριση των επαρχιών.

ABSTRACT

In this paper, we compare and suggest estimators in order to accurately predict the outcome of the popular vote in the United States of America. Using the cluster sampling technique, we create two different partitions of the U.S.A., the first one consisting of states while in the second we choose the counties as our clusters. For this purpose, we used data containing the outcome of the past 3 presidential elections, from 2012 to 2020. The methodology presented for the prediction of the popular vote can also be applied to the Greek legislative elections. We also discuss the minimum number of clusters needed to achieve 95% confidence intervals for each partition chosen.

Keywords: Sampling with probabilities proportional to size, Cluster sampling, Ratio estimator, Horvitz – Thompson Estimator, Linear Regression.

ΑΝΑΦΟΡΕΣ

Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association,* 47(260), 663-685.

Yates, F., & Grundy, P. M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society: Series B (Methodological)*, 15(2), 253-261.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, *35*(4), 1491-1523.

Cochran, W. (1977), *Sampling Techniques*. *3rd ed. Wiley, New York*

Κουνιάς, Ε.&Συμεωνίδης Γ. (1990). Οι δήμοι και οι κοινότητες ως συστάδες στην πρόβλεψη εκλογικών αποτελεσμάτων, *10ο Πανελλήνιο Συνέδριο Στατιστικής*.

Draper, N. R., & Smith, H. (1998). Applied regression analysis (Vol. 326). *John Wiley & Sons.* 79-114

Νικολακόπουλος, Η. (2003), «Η ανάπτυξη των πολιτικών δημοσκοπήσεων στην Ελλάδα», www.sedea.gr

Graefe, A. (2018). Predicting elections: Experts, polls, and fundamentals. *Judgment and Decision making*, *13*(4), 334-344.

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 35ου Πανελληνίου & 1ουΔιεθνούς Συνεδρίου Στατιστικής (2023), σελ.92-105

ΑΝΩ ΜΟΝΟΠΛΕΥΡΑ ΔΙΑΓΡΑΜΜΑΤΑ ΕΛΕΓΧΟΥ ΜΕ ΕΚΤΙΜΗΜΕΝΕΣ ΠΑΡΑΜΕΤΡΟΥΣ ΓΙΑ ΜΗΔΕΝΟΔΙΟΓΚΩΜΕΝΕΣ ΔΙΕΡΓΑΣΙΕΣ

Αθανάσιος Ρακιτζής[1], Ευτυχία Μαμζερίδου[2]
[1] Πανεπιστήμιο Πειραιώς, Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης, 18534 Πειραιάς, Ελλάδα.
(arakitz@unipi.gr)
[2] Πανεπιστήμιο Αιγαίου, Τμήμα Στατιστικής και Αναλογιστικών-Χρηματοοικονομικών Μαθηματικών, 83200 Σάμος.
(emam@aegean.gr)

ΠΕΡΙΛΗΨΗ

Στην παρούσα εργασία μελετάται η απόδοση ενός άνω μονόπλευρου διαγράμματος ελέγχου τύπου Shewhart για την παρακολούθηση μιας μηδενοδιογκωμένης Poisson διεργασίας (*Zero-inflated Poisson process*), στην περίπτωση εκτιμημένων παραμέτρων. Συνήθως η απόδοση των διαγραμμάτων ελέγχου υπολογίζεται υπό την υπόθεση ότι οι παράμετροι της διεργασίας είναι γνωστές. Στην πράξη όμως οι τιμές αυτές είναι σπάνια γνωστές και θα πρέπει να εκτιμηθούν από ένα προκαταρκτικό δείγμα *m* εντός ελέγχου παρατηρήσεων (δείγμα Φάσης Ι). Για την εκτίμηση των παραμέτρων χρησιμοποιείται η μέθοδος της μέγιστης πιθανοφάνειας. Στόχος είναι να διερευνηθεί το πόσο επηρεάζεται η θεωρητική απόδοση του άνω μονόπλευρου διαγράμματος ελέγχου από την εκτίμηση των παραμέτρων. Επιπλέον, θα διερευνηθεί το ποια πρέπει να είναι η τιμή του *m* (μέγεθος του δείγματος Φάσης Ι), ώστε η απόδοση του διαγράμματος στην περίπτωση των εκτιμημένων παραμέτρων να είναι παρόμοια με τη θεωρητική απόδοση (περίπτωση γνωστών παραμέτρων). Χρησιμοποιώντας προσομοίωση Monte Carlo, μελετάται η εντός ελέγχου απόδοση του διαγράμματος με εκτιμημένες παραμέτρους, η οποία βασίζεται στη χρήση της δεσμευμένης κατανομής του μήκους ροής. Επιπλέον, προτείνεται και ένας τρόπος επιλογής του ορίου ελέγχου όταν το μέγεθος *m* του δείγματος Φάσης Ι είναι δεδομένο.

Λέξεις κλειδιά: Εκτιμητές Μέγιστης Πιθανοφάνειας, Πιθανότητα Εσφαλμένου Συναγερμού, Στατιστικός Έλεγχος Διεργασιών, Μηδενοδιογκωμένες Διεργασίες, Διαγράμματα Ελέγχου.

1. ΕΙΣΑΓΩΓΗ

Τα διαγράμματα ελέγχου αποτελούν το βασικότερο εργαλείο στον στατιστικό έλεγχο ποιότητας για την παρακολούθηση μιας παραγωγικής διεργασίας. Η χρήση των διαγραμμάτων ελέγχου επιτρέπει την ανίχνευση παρουσίας ειδικών αιτιών μεταβλητότητας σε μια διεργασία, οπότε και λέμε ότι σε αυτή την περίπτωση η

διεργασία είναι εκτός στατιστικού ελέγχου. Διαφορετικά, αν η διεργασία λειτουργεί μόνο παρουσία φυσικών αιτιών μεταβλητότητας, τότε λέμε ότι η διεργασία είναι εντός στατιστικού ελέγχου. Υπάρχουν δύο βασικές κατηγορίες διαγραμμάτων ελέγχου ανάλογα με το είδος των διαθέσιμων μετρήσεων από ένα χαρακτηριστικό X (τυχαία μεταβλητή, τ.μ.), το οποίο σχετίζεται με το επίπεδο της ποιότητας των προϊόντων που παράγει μια διεργασία. Αν οι μετρήσεις προέρχονται από μια συνεχή τ.μ. τότε χρησιμοποιούνται διαγράμματα ελέγχου για μεταβλητές (*control charts for variables*) ενώ αν οι μετρήσεις αποτελούν δεδομένα καταμετρήσεων (*count data*), τότε χρησιμοποιούνται διαγράμματα ελέγχου ιδιοτήτων (*control charts for attributes*).

Τα πιο συχνά χρησιμοποιούμενα διαγράμματα ελέγχου ιδιοτήτων (δείτε π.χ [Montgomery (2020)]) είναι τα p- και np-chart τα οποία χρησιμοποιούνται για την παρακολούθηση του ποσοστού ή του αριθμού των ελαττωματικών αντικειμένων ή τα c- και u-chart τα οποία χρησιμοποιούνται για την παρακολούθηση του αριθμού των ελαττωμάτων ή του μέσου αριθμού ελαττωμάτων ανά μονάδα ελέγχου. Στην πρώτη περίπτωση η X αντιστοιχεί στον αριθμό των ελαττωματικών αντικειμένων σε δείγμα μεγέθους n και συνήθως υποθέτουμε ότι ακολουθεί τη διωνυμική κατανομή $\mathcal{B}(n,p)$ ενώ στη δεύτερη εκφράζει τον αριθμό των ελαττωμάτων που εμφανίζονται σε μια μονάδα ελέγχου και συνήθως υποθέτουμε ότι ακολουθεί την κατανομή Poisson $\mathcal{P}(\lambda)$. Στην συνέχεια θα μας απασχολήσει η περίπτωση παρακολούθησης του αριθμού των ελαττωμάτων.

Λόγω της τεχνολογικής εξέλιξης, οι σύγχρονες παραγωγικές διαδικασίες εμφανίζουν πολύ χαμηλούς έως και μηδενικούς αριθμούς ελαττωμάτων στις επιθεωρούμενες μονάδες. Τέτοιες διεργασίες είναι γνωστές ως διεργασίες υψηλής ποιότητας. Έτσι, εμφανίζεται μεγάλος αριθμός μηδενικών τιμών σε ένα διάγραμμα c-chart με αποτέλεσμα η χρήση των παραδοσιακών ορίων 3σ να οδηγεί σε αυξημένο ποσοστό εσφαλμένων συναγερμών. Επιπλέον, λόγω της παρουσίας πολλών μηδενικών τιμών, στα δεδομένα παρατηρείται υπερ-μεταβλητότητα (*overdispersion*). Αυτό σημαίνει πως υπάρχει σημαντική απόκλιση από την υπόθεση ότι $X \sim \mathcal{P}(\lambda)$, αφού η μέση τιμή και η διασπορά της κατανομής Poisson είναι ίσες. Άρα θα πρέπει να αναζητηθεί ένα καταλληλότερο μοντέλο πιθανότητας (λόγω της παρουσίας μεγάλου αριθμού μηδενικών τιμών και της υπερ-μεταβλητότητας στα δεδομένα) για την κατανομή της X και με βάση αυτό να αναπτυχθεί ένα διάγραμμα ελέγχου.

Ένα τέτοιο μοντέλο είναι αυτό της μηδενο-διογκωμένης κατανομής Poisson (*zero-inflated Poisson*, ή ZIP, *distribution*) στο οποίο υπάρχει μια επιπλέον παράμετρος η οποία επηρεάζει την πιθανότητα εμφάνισης μηδενικών τιμών, έναντι της αντίστοιχης πιθανότητας υπό την κατανομή $\mathcal{P}(\lambda)$. Τα διαγράμματα ελέγχου για την κατανομή ZIP έχουν μελετηθεί εκτενώς τα τελευταία χρόνια και ο ενδιαφερόμενος αναγνώστης παραπέμπεται στην εργασία των [Mahmood and Xie (2019)] και τις εκεί αναφορές.

Συνήθως η απόδοση των διαγραμμάτων ελέγχου υπολογίζεται υπό την υπόθεση ότι οι παράμετροι της διεργασίας είναι γνωστές (*Περίπτωση γνωστών παραμέτρων* ΠΓΠ ή Case K). Στην πράξη όμως οι τιμές αυτές είναι σπάνια γνωστές (*Περίπτωση άγνωστών*

παραμέτρων ΠΑΠ ή Case U) και θα πρέπει να εκτιμηθούν από ένα εντός ελέγχου δείγμα Φάσης Ι, μεγέθους m. Είναι όμως γνωστό (δείτε π.χ. [Montgomery (2020)]) ότι υπάρχει σημαντική διαφορά στην απόδοση του διαγράμματος μεταξύ της ΠΓΠ (θεωρητική απόδοση) και της ΠΑΠ (πραγματική απόδοση). Στόχος της παρούσας εργασίας είναι να διερευνήσει το πόσο επηρεάζεται η θεωρητική απόδοση ενός άνω μονόπλευρου διαγράμματος ελέγχου τύπου Shewhart για την παρακολούθηση μιας μηδενο-διογκωμένης Poisson διεργασίας (*ZIP process*) όταν οι παράμετροί της είναι άγνωστες και πρέπει να εκτιμηθούν.

Η διάρθρωση της εργασίας είναι η εξής: Στην Ενότητα 2 παρουσιάζονται οι βασικές ιδιότητες του διαγράμματος υπό την ΠΓΠ. Η απόδοση του διαγράμματος υπό την ΠΑΠ αξιολογείται στην Ενότητα 3, μέσω μελέτης προσομοίωσης. Στην Ενότητα 4 προτείνονται τροποποιήσεις στον αρχικό σχεδιασμό του διαγράμματος ώστε να μετριαστεί η επίδραση των εκτιμημένων παραμέτρων στην πραγματική εντός ελέγχου απόδοση του διαγράμματος. Τέλος, στην Ενότητα 5 δίνονται τα συμπεράσματα της μελέτης καθώς και προτάσεις για μελλοντική έρευνα.

2. ΤΟ ΔΙΑΓΡΑΜΜΑ ΕΛΕΓΧΟΥ ZIP-SH

2.1 Περίπτωση Γνωστών Παραμέτρων

Σε αυτήν την ενότητα παρουσιάζονται οι βασικές ιδιότητες ενός άνω μονόπλευρου διαγράμματος ελέγχου για την παρακολούθηση μιας διεργασίας ZIP (ZIP-SH *chart*) στην περίπτωση γνωστών παραμέτρων λ_0 και ϕ_0 (ΠΓΠ).

Έστω X η τυχαία μεταβλητή με στήριγμα $\mathcal{S} = \{0,1,\dots\} = \mathbb{N}_0$. Τότε η X ακολουθεί την κατανομή ZIP με παραμέτρους λ και ϕ αν η συνάρτηση πιθανότητας της είναι (δείτε π.χ. [Johnson et al. (2005)])

$$f_{ZIP}(x|\phi,\lambda) = \begin{cases} \phi + (1-\phi) f_{\mathcal{P}}(x|\lambda), & x = 0 \\ f_{\mathcal{P}}(x|\lambda), & x = 1,2,\dots \end{cases}$$

όπου $\lambda > 0$ και $\phi \in [0,1]$. Η παράμετρος ϕ καθορίζει το βαθμό μηδενοδιόγκωσης της κατανομής (*zero-inflation parameter*) ενώ $f_{\mathcal{P}}(x|\lambda)$, $x \in \mathbb{N}_0$ είναι η συνάρτηση πιθανότητας της $\mathcal{P}(\lambda)$. Η αθροιστική συνάρτηση κατανομής της κατανομής ZIP ορίζεται ως

$$F_{ZIP}(x|\phi,\lambda) = \phi + (1-\phi) F_{\mathcal{P}}(x|\lambda)$$

όπου $F_{\mathcal{P}}(x|\lambda)$, $x \in \mathbb{R}$ είναι η αθροιστική συνάρτηση κατανομής της $\mathcal{P}(\lambda)$.

Έστω ότι πρόκειται να παρακολουθήσουμε μια διεργασία ZIP χρησιμοποιώντας ένα άνω μονόπλευρο διάγραμμα ελέγχου τύπου Shewhart με ένα άνω όριο ελέγχου UCL. Θα αναφερόμαστε στο διάγραμμα αυτό ως το διάγραμμα ZIP-SH, το οποίο είναι κατάλληλο μόνο για την ανίχνευση αυξήσεων στο μέσο επίπεδο της διεργασίας. Αυτού του είδους η αλλαγή συνδέεται με τη χειροτέρευση της διεργασίας αφού π.χ. ένδειξη εκτός ελέγχου διεργασίας σημαίνει πως έχει αυξηθεί ο μέσος αριθμός ελαττωμάτων ανά μονάδα ελέγχου.

Υποθέτουμε επίσης πως όταν η διεργασία είναι εντός ελέγχου τότε $Y_i \sim ZIP(\phi_0, \lambda_0)$, $i = 1,2, \ldots$, όπου $Y_1, Y_2, \ldots$ είναι διαδοχικές παρατηρήσεις από τη διεργασία. Έστω $\alpha \in (0,1)$ η επιθυμητή (ή ονομαστική, *nominal*) πιθανότητα εσφαλμένου συναγερμού (*false alarm rate*, FAR). Τότε, για δεδομένες τιμές α και UCL η πιθανότητα το διάγραμμα να δώσει ένδειξη εκτός ελέγχου διεργασίας, ενώ στην πραγματικότητα η διεργασία είναι εντός ελέγχου, ικανοποιεί την παρακάτω ανισότητα

$$P(Y > UCL|IC) \leq \alpha \Leftrightarrow P(Y \leq UCL|IC) = F_{ZIP}(UCL|\phi_0, \lambda_0) \geq 1 - \alpha \quad (1)$$

Για δεδομένη τιμή α και λόγω της διακριτής φύσης της κατανομής ZIP, μπορούμε να προσδιορίσουμε το UCL ως τον μικρότερο ακέραιο που ικανοποιεί την (1). Κατά συνέπεια, για την πραγματική τιμή α^* της πιθανότητας εσφαλμένου συναγερμού θα ισχύει ότι $\alpha^* \leq \alpha$. Προφανώς $\alpha^* = F_{ZIP}(UCL|\phi_0, \lambda_0)$

Όταν η διεργασία είναι εκτός ελέγχου (OOC), τότε οι τιμές των παραμέτρων της είναι $\phi_1 \neq \phi_0$ και $\lambda_1 \neq \lambda_0$. Άρα, η πιθανότητα το διάγραμμα να μη δώσει ένδειξη εκτός ελέγχου διεργασίας ενώ στην πραγματικότητα η διεργασία είναι εκτός ελέγχου, ισούται με

$$\beta = P(Y \leq UCL|\text{OOC}) = F_{ZIP}(UCL|\phi_1, \lambda_1) \quad (2)$$

όπου $Y \sim ZIP(\phi_1, \lambda_1)$.

Έστω N η τυχαία μεταβλητή που εκφράζει το πλήθος των σημείων που απεικονίζονται στο διάγραμμα μέχρι αυτό να δώσει για πρώτη φορά ένδειξη εκτός ελέγχου διεργασίας. Τότε, στην περίπτωση γνωστών παραμέτρων, η $N \sim \mathcal{G}e(1 - \beta)$, όπου $\mathcal{G}e(p)$ είναι η Γεωμετρική κατανομή με παράμετρο $p \in (0,1)$. Η κατανομή της N είναι γνωστή ως κατανομή του μήκους ροής (*run length distribution*) με συνάρτηση πιθανότητας $f_N(u)$ και αθροιστική συνάρτηση κατανομής $F_N(u)$, όπου

$$f_N(u) = P(N = u) = (1 - \beta)\beta^{u-1}, \quad F_N(u) = P(N \leq u) = 1 - \beta^u$$

για $u = 1,2, \ldots$. Η $E(N)$ εκφράζει το αναμενόμενο πλήθος σημείων που απεικονίζονται στο διάγραμμα μέχρι αυτό να δώσει για πρώτη φορά ένδειξη εκτός ελέγχου διεργασίας, γνωστό και ως μέσο μήκος ροής (*Average Run Length*, ARL). Το ARL αποτελεί το πιο συχνά χρησιμοποιούμενο μέτρο αξιολόγησης της απόδοσης ενός διαγράμματος ελέγχου. Εκτός του ARL χρησιμοποιείται συχνά και η τυπική απόκλιση της κατανομής της N (*standard deviation of the run length*, $SDRL$), δηλαδή η $\sqrt{V(N)} = SDRL$, όπου $V(N)$ είναι η διασπορά της N. Κατά συνέπεια, τα ARL και $SDRL$ του διαγράμματος ZIP-SH στην ΠΓΠ υπολογίζονται από τις σχέσεις

$$ARL = (1 - \beta)^{-1}, \quad SDRL = \sqrt{\beta}/(1 - \beta)$$

όπου το β δίνεται στην εξίσωση (2).

2.2 Περίπτωση Άγνωστων Παραμέτρων

Σε αυτήν την ενότητα παρουσιάζονται οι βασικές ιδιότητες του διαγράμματος ελέγχου ZIP-SH στην περίπτωση άγνωστων παραμέτρων λ_0 και ϕ_0 (ΠΑΠ). Στην πράξη, οι

τιμές των παραμέτρων της διεργασίας είναι άγνωστες και θα πρέπει να εκτιμηθούν πριν την εφαρμογή του διαγράμματος για την παρακολούθησή της. Στις εργασίες των [He et al. (2003)], [Rakitzis and Castagliola (2016)] θεωρήθηκε άγνωστη μόνο μια παράμετρος, και συγκεκριμένα το λ_0. Στη συνέχεια, υποθέτουμε πως οι τιμές των ϕ_0, λ_0 είναι άγνωστες και πρέπει να εκτιμηθούν. Έστω $X_1, X_2, \dots, X_m$ ένα δείγμα Φάσης I από μια εντός ελέγχου διεργασία ZIP με παραμέτρους ϕ_0, λ_0. Χρησιμοποιώντας τη μέθοδο της Μέγιστης Πιθανοφάνειας, οι εκτιμητές μέγιστης πιθανοφάνειας $\hat{\phi}$, $\hat{\lambda}$ των ϕ_0, λ_0 υπολογίζονται αριθμητικά ως οι λύσεις του παρακάτω μη γραμμικού συστήματος εξισώσεων ([Cohen (1991)], [Xie and Goh (1993)])

$$\hat{\lambda} = \bar{X}^{+}\left(1 - e^{-\hat{\lambda}}\right), \quad \hat{\phi} = 1 - \bar{X}/\hat{\lambda},$$

όπου $\bar{X}^{+}$ είναι ο μέσος όρος των μη-μηδενικών τιμών στο δείγμα των $X_1, X_2, \dots, X_m$ και $\bar{X}$ είναι ο συνήθης μέσος όρος των τιμών στο δείγμα Φάσης I.

Στην ΠΑΠ, το άνω όριο ελέγχου εξαρτάται από την πιθανότητα εσφαλμένου συναγερμού α καθώς και από τις εκτιμήσεις των παραμέτρων της διεργασίας. Συνεπώς είναι μια τυχαία μεταβλητή, έστω $\widehat{UCL}$. Επιπλέον, έστω $\boldsymbol{\theta}_0 = (\phi_0, \lambda_0)$ και $\hat{\boldsymbol{\theta}} = \left(\hat{\phi}, \hat{\lambda}\right)$. Τότε για δεδομένες τιμές των α και $\hat{\phi}, \hat{\lambda}$, το $\widehat{UCL}$ προκύπτει ως ο ελάχιστος θετικός ακέραιος που ικανοποιεί την παρακάτω ανισότητα

$$F_{ZIP}\left(\widehat{UCL}\middle|\hat{\phi}, \hat{\lambda}\right) = \hat{\phi} + \left(1 - \hat{\phi}\right)F_{\mathcal{P}}\left(\widehat{UCL}\middle|\hat{\lambda}\right) \geq 1 - \alpha. \tag{3}$$

Θα συμβολίζουμε ως $\hat{\alpha} = 1 - F_{ZIP}\left(\widehat{UCL}\middle|\hat{\phi}, \hat{\lambda}\right)$ και θα αναφερόμαστε σε αυτή την πιθανότητα ως η υπό συνθήκη πιθανότητα εσφαλμένου συναγερμού (*conditional false alarm rate* ή CFAR, [Goedhart (2021)]).

Άμεσα έπεται ότι κατά την ΠΑΠ, το εντός ελέγχου ARL είναι επίσης τυχαία μεταβλητή αφού διαφορετικά δείγματα Φάσης I από την ίδια εντός ελέγχου διεργασία θα δώσουν διαφορετικές εκτιμήσεις $\hat{\phi}, \hat{\lambda}$ και άρα διαφορετικές τιμές για το άνω όριο ελέγχου $\widehat{UCL}$. Η μεταβλητότητα αυτή στις τιμές του ARL είναι γνωστή ως *μεταβλητότητα από χρήστη-σε-χρήστη* (*practitioner-to-practitioner variability*). Στη συνέχεια, χρησιμοποιώντας Monte Carlo προσομοίωση, θα διερευνήσουμε την επίδραση των εκτιμημένων παραμέτρων στην απόδοση του διαγράμματος. Τα βήματα της μελέτης προσομοίωσης είναι τα εξής:

Βήμα 1: Επιλέγουμε δείγμα Φάσης I μεγέθους m από μια $ZIP(\phi_0, \lambda_0)$ διεργασία.

Βήμα 2: Από το δείγμα Φάσης I εκτιμούμε τις παραμέτρους ϕ, λ και έστω $\hat{\phi}$, $\hat{\lambda}$ οι εκτιμήσεις μέγιστης πιθανοφάνειας.

Βήμα 3: Για δεδομένες τιμές των α και $\hat{\phi}$, $\hat{\lambda}$ προσδιορίζουμε την τιμή $\widehat{UCL}$ από την εξίσωση (3).

Βήμα 4: Υπολογίζουμε την τιμή του εντός ελέγχου ARL μέσω της σχέσης

$$ARL_{\hat{\theta}|\theta_0} = \left(1 - F_{ZIP}\left(\widehat{UCL}\middle|\phi_0, \lambda_0\right)\right)^{-1}$$

όπου $\alpha_{\hat{\theta}|\theta_0} = 1 - F_{ZIP}\left(\widehat{UCL}\middle|\phi_0, \lambda_0\right)$.

Βήμα 5: Επαναλαμβάνουμε T φορές (π.χ. $T = 25000$) τα βήματα 1-4 και υπολογίζουμε τη δειγματική μέση τιμή $AARL_{\hat{\theta}|\theta_0}$ και τη δειγματική τυπική απόκλιση $SDARL_{\hat{\theta}|\theta_0}$ των T τιμών $ARL_{\hat{\theta}|\theta_0}$.

Σημειώνεται πως το $ARL_{\hat{\theta}|\theta_0}$ είναι γνωστό και ως εντός ελέγχου δεσμευμένο ARL αφού υπολογίζεται για δεδομένη τιμή του $\widehat{UCL}$. Γενικά, το $\widehat{UCL}$ θεωρείται τυχαία μεταβλητή αφού εξαρτάται από τα $\hat{\phi}$, $\hat{\lambda}$. Άρα και το $ARL_{\hat{\theta}|\theta_0}$ είναι τυχαία μεταβλητή. Επιπλέον, τα $AARL_{\hat{\theta}|\theta_0}$ και $SDARL_{\hat{\theta}|\theta_0}$ αποτελούν εκτιμήσεις της μέσης τιμής και της τυπικής απόκλισης της κατανομής του $ARL_{\hat{\theta}|\theta_0}$. Επίσης, η πιθανότητα εσφαλμένου συναγερμού $\alpha_{\hat{\theta}|\theta_0}$ υπολογίζεται στην περίπτωση που $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, για δεδομένη τιμή του $\widehat{UCL}$. Η συγκεκριμένη πιθανότητα δεν μπορεί να υπολογιστεί στην πράξη (αφού τα ϕ_0, λ_0 είναι άγνωστα) όμως περιέχει πληροφορία σχετικά με την απόκλιση στην απόδοση του διαγράμματος μεταξύ ΠΓΠ και ΠΑΠ. Δείτε επίσης [Zhao and Driscoll (2016)]. Για το λόγο αυτό είναι και το βασικό μέτρο απόδοσης στην παρούσα μελέτη αφού $ARL_{\hat{\theta}|\theta_0} = \left(\alpha_{\hat{\theta}|\theta_0}\right)^{-1}$.

3. ΑΡΙΘΜΗΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

Στην ενότητα αυτή παρουσιάζονται τα αποτελέσματα μιας μελέτης προσομοίωσης σχετικά με την απόδοση του διαγράμματος ελέγχου ZIP-SH στην περίπτωση εκτιμημένων παραμέτρων. 25000 δείγματα Φάσης Ι προσομοιώθηκαν από μια $ZIP(\phi_0, \lambda_0)$ διαδικασία, για διάφορες τιμές ϕ_0, λ_0 και για διάφορα δείγματα Φάσης Ι μεγέθους m. Καθένα από αυτά τα δείγματα χρησιμοποιήθηκε για την εκτίμηση των παραμέτρων ϕ_0, λ_0.

Συγκεκριμένα το $m \in \{100,200,500,1000,2000,5000,\infty\}$, όπου με $m = \infty$ συμβολίζουμε την περίπτωση γνωστών παραμέτρων (ΠΓΠ), το $\phi_0 \in \{0.9,0.8,0.7\}$ και το $\lambda_0 \in \{1,2,5,6\}$. Ο λόγος που επιλέχθηκαν αυτές οι τιμές για το ϕ_0 αποδίδεται στην πρόθεση να μελετηθεί η απόδοση του διαγράμματος στην περίπτωση παρακολούθησης μιας διεργασίας υψηλής ποιότητας, όπου η εμφάνιση μηδενικής τιμής συνεπάγεται την απουσία ελαττώματος στο υπό επιθεώρηση αντικείμενο. Όσο πιο μεγάλη η τιμή ϕ_0 τόσο πιο μεγάλη η πιθανότητα μη εμφάνισης ελαττώματος στο προϊόν. Για τους παραπάνω συνδυασμούς (ϕ_0, λ_0) προκύπτουν διεργασίες με εντός ελέγχου μέσο επίπεδο από 0.1έως 2.4. Επίσης η επιθυμητή πιθανότητα εσφαλμένου συναγερμού α του διαγράμματος είναι 0.0027. Άρα, η πραγματική τιμή του FAR δεν ξεπερνά το 0.0027 οπότε, υπό την ΠΓΠ, το εντός ελέγχου ARL, έστω ARL_0, είναι τουλάχιστον 370.4. Στους Πίνακες 1 και 2 δίνονται οι τιμές των μέτρων $AARL_{\hat{\theta}|\theta_0}$ και $SDARL_{\hat{\theta}|\theta_0}$, αντίστοιχα, για τις διάφορες τιμές των ϕ_0, λ_0 και m. Επιπλέον, στον Πίνακα 3 δίνεται το εκτιμημένο ποσοστό των περιπτώσεων $\{ARL_{\hat{\theta}|\theta_0} < ARL_0\}$.

Στη στήλη «∞» του Πίνακα 1 δίνονται οι τιμές του ARL_0 στην ΠΓΠ. Καθώς το μέγεθος m του δείγματος Φάσης Ι αυξάνει, το $AARL_{\hat{\theta}|\theta_0}$ τείνει στη θεωρητική τιμή ARL_0 για (σχεδόν) όλα τα ζεύγη τιμών (ϕ_0, λ_0). Δεν είναι δύσκολο να διαπιστώσουμε

ότι απαιτούνται πολύ μεγάλα μεγέθη δείγματος Φάσης Ι (π.χ. $m \geq 2000$) ώστε να μειωθεί σημαντικά η διαφορά μεταξύ $AARL_{\hat{\theta}|\theta_0}$ και ARL_0, κυρίως στις περιπτώσεις που το $\phi_0 = 0.9$ ή 0.8. Για τις περιπτώσεις που το $\phi_0 = 0.7$, ενδέχεται να χρειάζεται και μικρότερο δείγμα Φάσης Ι (π.χ. $m = 1000$ ή 500), κάτι που όμως εξαρτάται και από την τιμή του λ_0. Για τις περιπτώσεις $(\phi_0, \lambda_0) \in \{(0.7,1), (0.7,2), (0.7,5)\}$ η διαφορά του $AARL_{\hat{\theta}|\theta_0}$ με το ARL_0 είναι περίπου 2%-3% όταν $m = 500$ ενώ για $m = 1000$ είναι περίπου 0.1%.

Αξίζει επίσης να σημειωθεί πως σε αντίθεση με το τι συμβαίνει στην περίπτωση των διαγραμμάτων ελέγχου για μεταβλητές με εκτιμημένες παραμέτρους, η σύγκλιση του $AARL_{\hat{\theta}|\theta_0}$ στο ARL_0 δεν είναι μονότονη. Δείτε π.χ. τις περιπτώσεις (0.9,5), (0.8,1), (0.8,2) και (0.8,5). Το φαινόμενο αυτό είχε επισημανθεί επίσης από τους [Rakitzis and Castagliola (2016)].

Πίνακας 1. *Τιμές $AARL_{\hat{\theta}|\theta_0}$ για το διάγραμμα ZIP-SH με Εκτιμημένες Παραμέτρους*

ϕ_0	λ_0	m						
		100	200	500	1000	2000	5000	∞
0.9	1	1522.94	1249.49	1066.34	922.73	747.99	575.88	526.64
	2	1388.20	973.26	809.55	707.30	629.49	604.82	603.73
	5	930.43	737.70	637.62	619.95	644.06	691.21	730.18
	6	882.21	699.11	612.25	578.16	538.14	504.30	497.71
0.8	1	1518.49	1224.89	1128.86	1202.76	1288.54	1356.34	1366.18
	2	1108.81	911.33	870.26	924.19	997.74	1071.65	1102.83
	5	806.82	691.68	648.02	654.35	657.58	674.22	916.91
	6	751.40	664.47	603.54	578.08	567.52	566.41	566.41
0.7	1	1479.89	1165.64	945.44	911.73	910.67	910.78	910.78
	2	1029.84	875.05	759.72	736.42	735.22	735.22	735.22
	5	760.10	678.86	623.44	610.84	611.03	611.27	611.27
	6	711.59	653.80	632.58	625.65	618.20	596.47	377.61

Στον Πίνακα 2 δίνονται οι τιμές του $SDARL_{\hat{\theta}|\theta_0}$ το οποίο αποτελεί εκτίμηση της τυπικής απόκλισης (και άρα της μεταβλητότητας) της κατανομής του $ARL_{\hat{\theta}|\theta_0}$. Συγκεκριμένα, δίνει μια εικόνα του πόσο διαφορετικές τιμές $ARL_{\hat{\theta}|\theta_0}$ θα υπάρχουν μεταξύ των διαφορετικών χρηστών του διαγράμματος, λόγω της εκτίμησης των τιμών των ϕ_0, λ_0 από τα διαφορετικά δείγματα Φάσης Ι. Όσο μεγαλύτερη είναι η τιμή του $SDARL_{\hat{\theta}|\theta_0}$ τόσο πιο μεγάλη είναι η *μεταβλητότητα από χρήστη-σε-χρήστη* (*practitioner-to-practitioner variability*). Παρατηρούμε ότι καθώς το m αυξάνει το $SDARL_{\hat{\theta}|\theta_0}$ μειώνεται, το οποίο είναι επιθυμητό αλλά και αναμενόμενο. Όμως για συγκεκριμένα ζεύγη τιμών (ϕ_0, λ_0) οι τιμές του μπορεί να εξακολουθούν να είναι υψηλές, ακόμη και για $m \geq 2000$. Αντίστοιχη συμπεριφορά παρατηρήθηκε και από τους [Zhao και Driscoll (2016)] στην περίπτωση του c-chart και αποδίδεται στη διακριτή φύση της

κατανομής ZIP. Οι [Saleh et al. (2015)] πρότειναν τον προσδιορισμό του m με βάση την τιμή του $SDARL_{\hat{\theta}|\theta_0}$. Συγκεκριμένα, το m πρέπει να είναι τέτοιο ώστε η τιμή του $SDARL_{\hat{\theta}|\theta_0}$ να είναι περίπου ίση με το 5%-10% της επιθυμητής τιμής ARL_0 υπό την ΠΓΠ. Στα αποτελέσματα του Πίνακα 2 αυτό συμβαίνει σε 4 μόνο περιπτώσεις, όταν $m = 2000$ και σε 6 περιπτώσεις, όταν $m = 5000$ (τιμές με έντονη γραμματοσειρά).

Πίνακας 2. *Τιμές $SDARL_{\hat{\theta}|\theta_0}$ για το διάγραμμα ZIP-SH με Εκτιμημένες Παραμέτρους*

ϕ_0	λ_0	m					
		100	200	500	1000	2000	5000
0.9	1	2511.93	1620.21	987.54	848.22	662.79	325.84
	2	3591.45	1184.67	605.49	415.13	206.25	**41.76**
	5	1406.72	654.51	329.56	222.67	169.55	121.22
	6	1299.92	592.57	319.86	239.07	161.80	64.65
0.8	1	2267.65	1452.89	607.14	403.39	282.14	**103.70**
	2	1622.64	796.34	395.41	334.31	275.79	154.91
	5	792.67	450.30	288.84	275.60	275.42	273.91
	6	692.9	423.15	252.03	144.19	**50.78**	**0.00**
0.7	1	2064.49	1281.68	538.51	144.06	**9.30**	**0.00**
	2	1119.27	709.02	316.81	87.92	**0.00**	**0.00**
	5	626.69	400.71	205.19	79.78	**14.34**	**0.00**
	6	542.24	357.57	272.71	269.59	268.87	265.55

Στις περισσότερες από τις εξεταζόμενες περιπτώσεις, καθώς το m αυξάνει το ποσοστό αυτό μειώνεται και για $m = 5000$ τείνει να μηδενιστεί. Το εκτιμημένο ποσοστό των περιπτώσεων $\{ARL_{\hat{\theta}|\theta_0} < ARL_0\}$ δίνεται στον Πίνακα 3. Μικρή τιμή του συγκεκριμένου ποσοστού σημαίνει ότι αναμένεται μόλις σε ένα μικρό ποσοστό χρηστών να παρατηρηθούν διαγράμματα με τιμή μικρότερη από ARL_0, και άρα με αυξημένη πιθανότητα εσφαλμένου συναγερμού, έναντι της επιθυμητής.

Όμως, για συγκεκριμένα ζεύγη τιμών (π.χ. για $(\phi_0, \lambda_0) = (0.9,5)$ ή $(0.8,5)$) το ποσοστό των περιπτώσεων $\{ARL_{\hat{\theta}|\theta_0} < ARL_0\}$ παραμένει σημαντικά μεγαλύτερο του 0%, ακόμη και για $m = 5000$. Η συμπεριφορά αυτή αποδίδεται επίσης στη διακριτή φύση της κατανομής ZIP.

Πίνακας 3. *Εκτιμημένο Ποσοστό Περιπτώσεων* $\{ARL_{\hat{\theta}|\theta_0} < ARL_0\}$ *για το άνω μονόπλευρο ZIP-SH διάγραμμα*

ϕ_0	λ_0	m					
		100	200	500	1000	2000	5000
0.9	1	17.50%	12.12%	2.86%	0.26%	0.04%	0.00%
	2	29.37%	19.37%	8.09%	2.06%	0.23%	0.00%
	5	43.35%	39.82%	**34.14%**	**28.98%**	**20.76%**	**9.37%**
	6	32.28%	23.74%	12.48%	5.05%	0.90%	0.02%
0.8	1	40.43%	34.63%	23.33%	14.92%	7.04%	0.89%
	2	42.30%	37.96%	29.74%	22.32%	13.74%	3.89%
	5	49.66%	49.51%	49.20%	**47.58%**	**47.00%**	**43.98%**
	6	28.79%	20.92%	9.69%	3.20%	0.47%	0.00%
0.7	1	21.99%	12.76%	3.21%	0.41%	0.01%	0.00%
	2	20.58%	12.02%	2.94%	0.33%	0.00%	0.00%
	5	24.95%	16.64%	5.72%	1.32%	0.09%	0.00%
	6	8.90%	2.58%	1.60%	0.00%	0.00%	0.00%

4. ΔΙΟΡΘΩΜΕΝΑ ΟΡΙΑ ΕΛΕΓΧΟΥ

Τα έως τώρα αποτελέσματα έδειξαν ότι είναι απαραίτητη η λήψη ενός δείγματος Φάσης Ι μεγάλου μεγέθους (π.χ. $m > 1000$ ή ακόμη και $m > 5000$) ώστε η πραγματική απόδοση του διαγράμματος να είναι κοντά στη θεωρητική. Επειδή στην πράξη δεν είναι πάντοτε εφικτό να συλλέγονται τόσες πολλές προκαταρκτικές μετρήσεις, στη συνέχεια προτείνονται δύο τροποποιήσεις (*Σχεδιασμός Α* και *Σχεδιασμός Β*) στη διαδικασία σχεδιασμού του διαγράμματος, όταν το μέγεθος m του δείγματος Φάσης Ι είναι δεδομένο. Πριν προχωρήσουμε, πρέπει να σημειωθεί πως για δεδομένο m και αφού το άνω όριο ελέγχου εξαρτάται από τις εκτιμήσεις $\hat{\phi}$, $\hat{\lambda}$ και την πιθανότητα α, η μόνη τροποποίηση που μπορεί να γίνει είναι στην τιμή α. Έτσι, αντί της επιθυμητής τιμής α για την πιθανότητα εσφαλμένου συναγερμού, να χρησιμοποιηθεί μια άλλη τιμή.

Σύμφωνα με το Σχεδιασμό Α προτείνεται η χρήση πιθανότητας εσφαλμένου συναγερμού $\alpha' \neq \alpha$ ώστε το $AARL_{\hat{\theta}|\theta_0}$ να είναι όσο γίνεται πιο κοντά στην επιθυμητή τιμή του εντός ελέγχου μέσου μήκους ροής ARL (έστω αυτή c_0). Η διαδικασία εύρεσης του α' είναι αντίστοιχη αυτής που περιγράφεται στην Ενότητα 2 (Βήματα 1-5) μόνο που αντί να εκτελείται για μια τιμή α, εκτελείται για διάφορες τιμές $\alpha' \neq \alpha$ μέχρι να βρεθεί η τιμή α' που θα δώσει $AARL_{\hat{\theta}|\theta_0} \approx c_0$.

Σύμφωνα με το Σχεδιασμό Β προτείνεται η χρήση πιθανότητας εσφαλμένου συναγερμού $\alpha'' \neq \alpha$ ώστε το εκτιμημένο ποσοστό των περιπτώσεων $\{ARL_{\hat{\theta}|\theta_0} < c_0\}$ να είναι όσο γίνεται πιο κοντά σε μια προκαθορισμένη τιμή (έστω αυτή γ^*). Η

διαδικασία είναι όμοια με αυτή στο Σχεδιασμό Α με τη μόνη διαφορά να είναι στο κριτήριο που χρησιμοποιείται για την εύρεση της τιμής α''.

Στον Πίνακα 4 δίνονται τα αποτελέσματα των διορθωμένων τιμών α' και α'' (με ακρίβεια 4 δεκαδικών ψηφίων) για την πιθανότητα εσφαλμένου συναγερμού. Ως c_0 χρησιμοποιήθηκε η τιμή ARL_0 (δείτε στήλη «∞» στον Πίνακα 1) ενώ ως γ^* χρησιμοποιήθηκε η τιμή 0.05. Για λόγους οικονομίας χώρου παρουσιάζονται τα αποτελέσματα στις περιπτώσεις $\phi_0 \in \{0.9, 0.8, 07\}$, $\lambda_0 \in \{2,6\}$ και $m \in \{200,500\}$.

Αξίζει να σημειωθεί πως το μέγεθος m του δείγματος Φάσης Ι εξαρτάται από την υπό παρακολούθηση διεργασία και ενδέχεται ακόμη και η περίπτωση $m = 200$ να μην είναι εφικτή στην πράξη. Αν όμως το m δεν είναι αρκετά μεγάλο ενδέχεται να προκληθεί πρόβλημα στην εύρεση ευσταθών εκτιμήσεων των άγνωστων παραμέτρων, λόγω της υψηλής πιθανότητας εμφάνισης μηδενικής τιμής ενώ είναι πολύ πιθανό να ληφθεί δείγμα το οποίο να αποτελείται μόνο από μηδενικές τιμές. Παρά ταύτα η προτεινόμενη μεθοδολογία δουλεύει και για $m < 200$.

Πίνακας 4. *Αποτελέσματα Διορθωμένων Σχεδιασμών Α και Β*

			Σχεδιασμός Α				*Σχεδιασμός Β*			
ϕ_0	λ_0	m	α'	$AARL_{\hat{\theta}\|\theta_0}$	$SDARL_{\hat{\theta}\|\theta_0}$	%	α''	$AARL_{\hat{\theta}\|\theta_0}$	$SDARL_{\hat{\theta}\|\theta_0}$	%
0.9	2	200	0.0041	608.65	623.39	36.29%	0.0013	2285.50	3102.06	5.38%
		500	0.0035	599.19	387.87	18.89%	0.0024	911.46	686.77	5.41%
		∞	0.0027	603.73	-	-	0.0027	603.73	-	-
0.9	6	200	0.0036	502.56	391.59	39.97%	0.0014	1436.82	1348.55	5.48%
		500	0.0032	506.70	252.10	23.24%	0.0022	765.28	404.15	5.14%
		∞	0.0027	497.71	-	-	0.0027	497.71	-	-
0.8	2	200	0.0023	1082.15	968.82	25.94%	0.0011	2470.99	2422.22	5.62%
		500	0.0020	1090.61	544.85	8.18%	0.0018	1183.51	698.58	6.32%
		∞	0.0027	1102.83	-	-	0.0027	1102.83	-	-
0.8	6	200	0.0031	571.52	353.84	26.92%	0.0018	1026.40	695.18	6.00%
		500	0.0029	562.94	222.53	11.96%	0.0029	678.06	311.23	5.07%
		∞	0.0027	566.41	-	-	0.0027	566.41	-	-
0.7	2	200	0.0032	736.01	545.76	19.51%	0.0021	1148.41	943.28	5.61%
		500	0.0028	733.08	250.22	4.45%	0.0030	719.62	218.15	5.53%
		∞	0.0027	735.22	-	-	0.0027	735.22	-	-
0.7	6	200	0.0045	376.99	192.49	23.00%	0.0031	563.33	307.29	5.41%
		500	0.0042	380.59	114.72	7.23%	0.0041	386.74	119.75	5.95%
		∞	0.0027	377.61	-	-	0.0027	377.61	-	-

Από τα αποτελέσματα στον Πίνακα 4 έπεται ότι οι διορθωμένες τιμές α' και α'' δεν είναι απαραίτητο ότι ταυτίζονται. Με βάση το Σχεδιασμό Α, το $AARL_{\hat{\theta}|\theta_0}$ είναι πολύ κοντά στην επιθυμητή τιμή ARL_0 ακόμα και για $m = 200$ για όλα τα υπό εξέταση ζεύγη τιμών (ϕ_0, λ_0). Όμως το εκτιμημένο ποσοστό των περιπτώσεων $\{ARL_{\hat{\theta}|\theta_0} < ARL_0\}$ (στήλη «%» υπό το Σχεδιασμό Α) έχει αυξηθεί, σε σχέση με τις αντίστοιχες

τιμές στον Πίνακα 3. Τα αποτελέσματα αντιστρέφονται με βάση το Σχεδιασμό Β. Πλέον το εκτιμημένο ποσοστό περιπτώσεων $\{ARL_{\widehat{\theta}|\theta_0} < ARL_0\}$ είναι περίπου στο 5% για όλες σχεδόν τις εξεταζόμενες περιπτώσεις αλλά οι τιμές των $AARL_{\widehat{\theta}|\theta_0}$ και $SDRL_{\widehat{\theta}|\theta_0}$ είναι μεγαλύτερες έναντι των αντίστοιχων τιμών υπό το Σχεδιασμό Α.

Στον Πίνακα 5 παρουσιάζεται μια συνοπτική εικόνα των τιμών μέτρων απόδοσης $AARL_{\widehat{\theta}|\theta_0}$, $SDRL_{\widehat{\theta}|\theta_0}$ και του ποσοστού των περιπτώσεων $\{ARL_{\widehat{\theta}|\theta_0} < ARL_0\}$ όταν η πιθανότητα εσφαλμένου συναγερμού είναι α, α' ή α'', για $m \in \{200,500\}$, $\lambda_0 \in \{2,6\}$ και $\phi_0 \in \{0.9,0.8,0.7\}$.

Πίνακας 5. *Συγκριτικός Πίνακας Μέτρων Απόδοσης για το Διάγραμμα ZIP-SH*

Σχεδιασμός	ϕ_0	λ_0	m	FAR	$AARL_{\widehat{\theta}\|\theta_0}$	$SDARL_{\widehat{\theta}\|\theta_0}$	*perc*
	0.9	2	200	0.0027	973.26	1184.67	19.37%
A			200	0.0041	608.65	623.39	36.29%
B			200	0.0013	2285.50	3102.06	5.38%
Case K			∞	0.0027	603.73	-	-
	0.9	6	500	0.0027	612.25	319.86	12.48%
A			500	0.0032	506.70	252.10	23.24%
B			500	0.0022	765.28	404.15	5.14%
Case K			∞	0.0027	497.71	-	-
	0.8	2	200	0.0027	870.26	395.41	29.74%
A			200	0.0023	1082.15	968.82	25.94%
B			200	0.0011	2470,99	2422.22	5.62%
Case K			∞	0.0027	1102.83	-	-
	0.8	6	500	0.0027	603.54	252.03	9.69%
A			500	0.0029	562.94	222.53	11.96%
B			500	0.0029	678.06	311.23	5.07%
Case K			∞	0.0027	566.41	-	-
	0.7	2	200	0.0027	875.05	709.02	12.02%
A			200	0.0032	736.01	545.76	19.51%
B			200	0.0021	1148.41	943.28	5.61%
Case K			∞	0.0027	735.22	-	-
	0.7	6	500	0.0027	632.58	272.71	1.60%
A			500	0.0042	380.59	114.72	7.23%
B			500	0.0041	386.74	119.75	5.95%
Case K			∞	0.0027	377.61	-	-

Για παράδειγμα για $(\phi_0, \lambda_0) = (0.9,2)$ και $\alpha = 0.0027$, το $ARL_0 = 603.73$ (ΠΓΠ). Όταν το μέγεθος του δείγματος Φάσης Ι είναι $m = 200$, για $\alpha = 0.0027$, τα $AARL_{\widehat{\theta}|\theta_0}$ και $SDRL_{\widehat{\theta}|\theta_0}$ είναι, αντίστοιχα, 973.26 και 1184.67 ενώ το ποσοστό των περιπτώσεων $\{ARL_{\widehat{\theta}|\theta_0} < ARL_0\}$ είναι 19.37%. Υπό το Σχεδιασμό Α, προτείνεται η χρήση FAR

$\alpha' = 0.0041$ η οποία δίνει $AARL_{\hat{\theta}|\theta_0} = 608.65$ και $SDRL_{\hat{\theta}|\theta_0} = 623.39$. Δηλαδή το $AARL_{\hat{\theta}|\theta_0}$ είναι πολύ κοντά στη (θεωρητική) τιμή ARL_0 και υπάρχει μείωση στο $SDARL_{\hat{\theta}|\theta_0}$, συγκριτικά με την περίπτωση χωρίς τροποποίηση. Όμως το ποσοστό των περιπτώσεων $\{ARL_{\hat{\theta}|\theta_0} < ARL_0\}$ έχει αυξηθεί και πλέον είναι 36.29%. Υπό το Σχεδιασμό Β, παρατηρούνται τα αντίστροφα αποτελέσματα αφού για FAR $\alpha'' = 0.0013$ το $AARL_{\hat{\theta}|\theta_0} = 2285.50$, το $SDARL_{\hat{\theta}|\theta_0} = 3102.06$ αλλά το ποσοστό των περιπτώσεων $\{ARL_{\hat{\theta}|\theta_0} < ARL_0\}$ είναι πλέον κοντά στο 5% και συγκεκριμένα ίσο με 5.38%.

Τα παραπάνω αποτελέσματα αποδίδονται στο ότι προκειμένου να επιτευχθεί χαμηλό ποσοστό διαγραμμάτων με εντός ελέγχου ARL μικρότερο του ARL_0 (*Σχεδιασμός Β*) μειώνεται η πιθανότητα εσφαλμένου συναγερμού (σε σύγκριση με την αρχική τιμή 0.0027) το οποίο οδηγεί σε μεγαλύτερες τιμές για το UCL. Η αύξηση αυτή στην τιμή του UCL έχει ως αποτέλεσμα κάποιοι από τους χρήστες να καταλήξουν σε ένα διάγραμμα ελέγχου με εντός ελέγχου ARL πολύ μεγαλύτερο από ARL_0. Κατά συνέπεια, αυξάνεται και η μέση τιμή και η τυπική απόκλιση της κατανομής του $ARL_{\hat{\theta}|\theta_0}$. Αντίθετα, υπό το Σχεδιασμό Α, για να είναι η μέση τιμή της κατανομής του $ARL_{\hat{\theta}|\theta_0}$ κοντά στο ARL_0, αυξάνεται η πιθανότητα εσφαλμένου συναγερμού (σε σύγκριση με την αρχική τιμή 0.0027) το οποίο έχει ως αποτέλεσμα κάποιοι από τους χρήστες να καταλήξουν σε ένα διάγραμμα με τιμή UCL μικρότερη από αυτή που προκύπτει όταν το $\alpha = 0.0027$ και άρα με εντός ελέγχου ARL μικρότερο από ARL_0.

Με αυτόν τον τρόπο, για δεδομένο μέγεθος m του δείγματος Φάσης Ι, το $AARL_{\hat{\theta}|\theta_0} \approx ARL_0$ ενώ ταυτόχρονα μειώνεται και το $SDARL$. Αξίζει πάντως να αναφερθεί πως καθώς το m αυξάνει, οι διαφορές στις τιμές των α' και α'' μειώνονται και τελικά οι δύο σχεδιασμοί καταλήγουν σε παρόμοια απόδοση. Από τα αποτελέσματα του Πίνακα 5, εκτός από τις περιπτώσεις που το $\phi_0 = 0.9$, η διαφορά $|\alpha' - \alpha''|$ δεν ξεπερνά το 0.0002 όταν το $m = 500$. Αυξάνοντας κι άλλο το m (π.χ. για $m = 1000$) η διαφορά $|\alpha' - \alpha''|$ μειώνεται περαιτέρω, ακόμη και για $\phi_0 = 0.9$.

5. ΣΥΜΠΕΡΑΣΜΑΤΑ

Στην εργασία αυτή μελετήθηκε η απόδοση ενός άνω μονόπλευρου διαγράμματος ελέγχου τύπου Shewhart για την παρακολούθηση μιας μηδενοδιογκωμένης Poisson διεργασίας, στην περίπτωση που και οι δύο παράμετροι της διεργασίας είναι άγνωστες και πρέπει να εκτιμηθούν. Σε αντίστοιχες εργασίες στην έως τώρα βιβλιογραφία, η βασική υπόθεση ήταν ότι η παράμετρος ϕ_0 είναι γνωστή και πρέπει να εκτιμηθεί μόνο το λ_0. Για την εκτίμηση των ϕ_0, λ_0 χρησιμοποιήθηκε η μέθοδος της μέγιστης πιθανοφάνειας λόγω της δημοφιλίας της, αλλά και των ασυμπτωτικών ιδιοτήτων που έχουν οι αντίστοιχοι εκτιμητές.

Χρησιμοποιώντας μια σειρά από μέτρα αξιολογήθηκε η απόδοση του διαγράμματος στην περίπτωση των εκτιμημένων παραμέτρων. Τα μέτρα απόδοσης βασίστηκαν στην υπό συνθήκη κατανομή του εντός ελέγχου μέσου μήκους ροής και συγκεκριμένα στη

μέση τιμή ($AARL_{\hat{\theta}|\theta_0}$) και στην τυπική απόκλιση ($SDARL_{\hat{\theta}|\theta_0}$) αυτής. Η τυπική απόκλιση $SDARL_{\hat{\theta}|\theta_0}$ δίνει μια εικόνα της μεταβλητότητας μεταξύ των χρηστών (*practitioner-to-practitioner variability*) και όσο πιο μικρή είναι, τόσο περισσότερο τα διαφορετικά διαγράμματα (που προκύπτουν από τους διαφορετικούς χρήστες μετά από χρήση των διαφορετικών δειγμάτων Φάσης Ι) θα έχουν παρόμοια εντός ελέγχου απόδοση. Επιπλέον, μελετήθηκε και το ποσοστό των περιπτώσεων $\{ARL_{\hat{\theta}|\theta_0} < ARL_0\}$ το οποίο δίνει μια εικόνα του πόσοι χρήστες θα καταλήξουν τελικά σε ένα διάγραμμα ZIP-SH με αυξημένο ποσοστό εσφαλμένων συναγερμών.

Τα αποτελέσματα της μελέτης προσομοίωσης έδειξαν ότι γενικά απαιτούνται μεγάλα μεγέθη δείγματος ώστε να μην υπάρχουν μεγάλες διαφορές στην απόδοση του διαγράμματος μεταξύ της ΠΓΠ και της ΠΑΠ. Στις περισσότερες από τις περιπτώσεις που εξετάστηκαν, προτείνεται το μέγεθος m του δείγματος Φάσης Ι να είναι μεγαλύτερο του 2000. Επιπλέον, αν οι διαθέσιμοι πόροι είναι δεδομένοι και το μέγεθος του δείγματος Φάσης Ι είναι προκαθορισμένο, προτάθηκε η τροποποίηση του ονομαστικού επιπέδου εσφαλμένου συναγερμού α του διαγράμματος, χρησιμοποιώντας δύο διαφορετικά κριτήρια. Το ένα απαιτεί η τιμή του $AARL_{\hat{\theta}|\theta_0}$ να είναι κοντά στο ARL_0 ενώ το δεύτερο απαιτεί μόνο ένα μικρό ποσοστό διαγραμμάτων να έχει εντός ελέγχου ARL μικρότερο από ARL_0.

Για την επιβεβαίωση των αποτελεσμάτων στους Πίνακες 1-5 έχουν αναπτυχθεί προγράμματα στην R ([R Core Team (2023)]) τα οποία είναι διαθέσιμα από τους συγγραφείς κατόπιν σχετικού αιτήματος. Τέλος, ως προτάσεις για μελλοντική έρευνα προτείνεται η μελέτη της απόδοσης των αντίστοιχων διαγραμμάτων ελέγχου τύπου CUSUM ([He et al. (2012)]) και EWMA ([Alevizakos and Koukouvinos (2020)]) στην περίπτωση εκτιμημένων παραμέτρων.

ABSTRACT

In this work we study the performance of an upper one-sided Shewhart-type control charts for monitoring a zero-inflated Poisson process, in the case of estimated parameters (Case U). In real problems, process parameters are unknown and must be estimated from a Phase I dataset. However, this affects the theoretical performance (case of known parameters, Case K) of the chart and, thus, further investigation is needed to assess these differences in chart's performance. In addition, practical guidelines are necessary to determine the size of the Phase I sample to reduce the difference in chart's performance between Case K and Case U. Using simulation, we evaluate the performance of the chart, focusing on the (conditional) distribution of the in-control average run length. The main aim is study the practitioner-to-practitioner variability and suggest practical guidance regarding its reduction, in order to provide charts that does not demonstrate an excessive number of false alarms. In addition, two approaches for adjusting the usually employed design are discussed, when the size of the Phase I dataset is pre-specified. The results show that either approaches provide charts that meet the design requirements, even for Phase I samples of medium size.

Ευχαριστίες: This work has been partly supported by the University of Piraeus Research Center.

ΑΝΑΦΟΡΕΣ

Alevizakos V. and Koukouvinos C. (2020). Monitoring of zero-inflated Poisson processes with EWMA and DEWMA control charts. *Quality and Reliability Engineering International*, **36**, 88-111.

Cohen A. C. (1991). *Truncated and Censored Samples: Theory and Applications*. CRC Press.

Goedhart R. (2021). Design considerations and trade-offs for Shewhart control charts. In: *Frontiers in Statistical Quality Control 13* (pp. 13-23). Cham: Springer International Publishing.

He B., Xie M., Goh T.N. and Ranjan P. (2003). On the estimation error in zero-inflated Poisson model for process control. *International Journal of Reliability, Quality and Safety Engineering*, **10**, 159-169.

He S., Huang W. and Woodall W. H. (2012). CUSUM charts for monitoring a zero-inflated poisson process. *Quality and Reliability Engineering International*, **28**, 181-192.

Johnson N. L., Kemp A. W. and Kotz S. (2005). *Univariate Discrete Distributions*, John Wiley & Sons.

Mahmood T. and Xie M. (2019). Models and monitoring of zero-inflated processes: The past and current trends. *Quality and Reliability Engineering International*, **35**, 2540-2557.

Montgomery D. C. (2020). *Introduction to Statistical Quality Control*, 8^{th} *edn*. John Wiley & Sons.

R Core Team. (2022). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.r-project.org/

Rakitzis A. C. and Castagliola P. (2016). The effect of parameter estimation on the performance of one-sided Shewhart control charts for zero-inflated processes. *Communications in Statistics - Theory and Methods*, **45**, 4194-4214.

Saleh N. A., Mahmoud M. A., Keefe M. J. and Woodall W. H. (2015). The difficulty in designing Shewhart $\bar{X}$ and X control charts with estimated parameters. *Journal of Quality Technology*, **47**, 127-138.

Xie M. and Goh T. N. (1993). SPC of a near zero-defect process subject to random shocks. *Quality and Reliability Engineering International*, **9**, 89-95.

Zhao M. J. and Driscoll A. R. (2016). The c-chart with bootstrap adjusted control limits to improve conditional performance. *Quality and Reliability Engineering International*, **32**, 2871-2881.

Εργασίες στα Αγγλικά

Papers in English

Greek Statistical Institute
Proceedings of the 35th Panhellenic and 1st International Statistics Conference (2023), p.107-121

Optimal Threshold Selection on a Two-Cutpoint Optimal ROC curve

***A. Anastasiou*[1], *J. Tsimikas*[1]**
[1]Laboratory of Statistics and Data Analysis, Department of Statistics and Actuarial-Financial Mathematics, University of the Aegean, Greece
sasd20003@aegean.gr, tsimikas@aegean.gr

ABSTRACT

This work aims to explore techniques that integrate misclassification cost analysis into optimal Receiver Operating Characteristics (ROC) curves. Specifically, assuming a family of distributions with the unimodality property of the Likelihood ratio the optimal ROC curve is a two cutpoint ROC curve. We obtain results concerning the estimation of the optimal threshold/s and the corresponding False Positive Rate (FPR) given the two misclassification costs and the disease prevalence. We present details for some common parametric families that exhibit the unimodality property of the Likelihood ratio. We combine maximum likelihood etimation with resampling methods for statistical inference on the optimal threshold.

Keywords: Optimal ROC, Unimodality, Misclassification Cost, Optimal threshold.

1. Introduction

A well-established and widely used tool to evaluate the diagnostic ability of biomarker measurements is the Receiver Operating Characteristics (ROC) curve. Its application was initially introduced by Green and Swets (1966). Later, Lusted (1971) demonstrated that medical diagnostic testing holds great potential in situations where a decisive determination is needed regarding the presence or absence of a disease. Since then numerous studies have been conducted to assess the discriminating potential of medical diagnostic tests with most studies focusing on continuous biomarker measurements or ordinal ratings of the suspicion of a disease. In this paper we focus on continuous biomarker measurements. We denote the continuous biomarker measurement by Y and the presence of disease indicator by D. The two Cumulative Distributions Functions (CDF) for diseased and healthy individuals are denoted by $F_1(y) = P(Y \leq y|D = 1) = 1 - S_1(y)$ and $F_0(y) = P(Y \leq y|D = 0) = 1 - S_0(y)$ respectively and have common support, where $S_0(y), S_1(y)$ are the Survival Functions.The corresponding densities are denoted by $f_1(y)$ and $f_0(y)$. In the ROC curve literature it is commonly assumed

that higher measurements are more indicative of the presence of the disease. Under this assumption the decision rule is to declare a case positive when $Y \geq c$, c being the decision cutoff. For a given threshold, c, we define the True Positive Rate, $TPR(c) = P(Y \geq c|D = 1) = S_1(c) = Sensitivity(c)$, and the False Positive Rate $FPR(c) = P(Y \geq c|D = 0) = S_0(c) = 1 - Specificity(c)$. In general the ROC curves provide a comprehensive graphical representation of the trade-off between the sensitivity and the false positive rate of a diagnostic test as one varies decision cutoff(s). The ROC curve is an increasing function in the FPR,

$$\{(t, ROC(t)), t \in (0, 1)\},$$

where t denotes the FPR and $ROC(t)$ is the corresponding sensitivity. More formally (see Pepe (2003) and Zhou et al. (2002)), the ROC curve is given by:

$$ROC(t) = S_1(S_0^{-1}(t)), t \in (0, 1), \tag{1}$$

where $c = S_0^{-1}(t)$ gives the connection between a given $FPR = t$ and the corresponding value of the cutoff. The selection of the optimal cutoff, (c_{opt}), should be based on the cost due to misclassification. However, since agreement on misclassification costs can be elusive, other criteria have been investigated (see Zou et al. (2013)). We refer to the ROC curve in ((1)) as the right directional ROC curve. It is easily shown that the right directional ROC curve is concave when the density (likelihood) ratio $r(y) = \dfrac{f_1(y)}{f_0(y)}$ is increasing in y. In such cases the ROC curve is called proper (Egan (1975), pp 19, 37). When the density (likelihood) ratio is not increasing in y the optimal decision rule needs to be modified to 'classify as positive if $r(y) > c$'. We refer to the ROC curve obtained by using the optimal decision rule as the Optimal ROC curve.

The selection of an optimal threshold in Optimal ROC curves is crucial as it directly impacts the reliability and reproducibility of diagnostic tests. A lower threshold may increase the sensitivity of the test, enabling the detection of a greater number of true positive cases. However, this may also lead to a higher false positive rate, which can result in unnecessary interventions or treatments for individuals who do not have the condition of interest and that lead to an unnecessary cost. On the other hand, a higher threshold can enhance the specificity of the test, reducing the false positive rate but potentially missing true positive cases. The Optimal decision cutoff(s) minimize the misclassification cost.

When we assume that the distributions of cases and controls follow a family of distributions with the unimodal density (likelihood) ratio property, the resulting Optimal ROC curve is a two-cutpoint ROC curve. Martinez-Camblor et. al. (2017) and Martinez-Camblor et. al. (2019) thoroughly discuss the two cutoff ROC curve. Its optimality, under the assumption of unimodality of the density ratio, is discussed in Bantis

et.al. (2021). Due to the concavity of Optimal ROC and its increasing nature, a unique optimal cutoff can be determined, providing a path for estimating the corresponding Optimal FPR.

The structure of the paper is as follows: In section 2 we discuss how one could select the threshold of a continuous diagnostic test using the classical directional ROC curve. In addition, we study the optimal ROC curve under a unimodal Likelihood ratio family and the performance to some common distributions. In section 3 we conduct a small simulation study and in section 4 we apply our methods to a real dataset concerning genes associated with ovarian cancer. We specifically focus on the beta distribution and its relevance to the diagnostic testing of these genes.

2. Threshold selection in optimal ROC curves

2.1 Obtaining thresholds using cost analysis

The Cost when operating at given $FPR = t$ is,

$$\text{Cost}(t) = k + t(1-\pi)C_{01} + (1 - ROC(t))\pi C_{10} + ROC(t)\pi C_{11}$$

where k is the overall operating cost, $\pi = P(D = 1)$ denotes the known disease prevalence in the target population, C_{01} is the cost of the unnecessary treatment of a false positive result, C_{10} is the cost of a false negative result and C_{11} is the cost of necessary treatment of a correctly classified diseased individual.

Let $\psi = \dfrac{\pi}{1-\pi}\dfrac{C_{10}}{C_{01}}$. The threshold that potentially minimizes the expected cost is the threshold that corresponds to an $FPR = t$ that solves,

$$\frac{\partial \text{Cost}(t)}{\partial t} = (1-\pi)C_{01} - \frac{\partial ROC(t)}{\partial t}\pi(C_{10} - C_{11}) = 0 \Leftrightarrow$$

$$\frac{\partial ROC(t)}{\partial t} = \left(\frac{1-\pi}{\pi}\right)\left(\frac{C_{01}}{C_{10} - C_{11}}\right) = \frac{1}{\psi}$$

For ease we consider $k = 0, C_{11} = 0$. If $C_{11} > 0$ then simply replace C_{10} by $(C_{10} - C_{11})$ in what follows. As for k it is just a constant cost unrelated to misclassification. We have,

$$Cost(t) = (1-\pi)C_{01}Q(t)$$

where,

$$Q(t) = [t + \psi(1 - ROC(t))],$$

It is obvious that the optimal t minimizes $Q(t)$.

We have $Q(0) = \psi$, $Q(1) = 1$ and

$$Q^{'}(t) = 1 - \psi \frac{\partial ROC(t)}{\partial t}$$

$$Q^{''}(t) = -\psi \frac{\partial^2 ROC(t)}{\partial t^2}.$$

A potential minimum of $Q(t)$ corresponds to,

$$\frac{\partial ROC(t)}{\partial t} = \frac{1}{\psi} \tag{2}$$

According to equation (2) optimal thresholds and corresponding optimal FPRs are determined by the "relative importance" of the misclassification costs adjusted for disease prevalence as summarized by ψ.

When $\psi = 1$ we have $\pi C_{10} = (1 - \pi)C_{01}$ and the two prevalence adjusted misclassification costs are equal. Then

$$Q(t) = t + (1 - ROC(t)).$$

In this special case the optimal threshold obtained by minimizing the expression above is called the Youden index based threshold. This can be easily seen by noting that $1 - Q(t) = (1 - t) + ROC(t) - 1 = specificity + sensitivity - 1$, the Youden index.

2.2 Optimal ROC curve

The optimal classification decision rule based on a biomarker value, Y is: classify as diseased if $W = r(Y) = \frac{f_1(Y)}{f_0(Y)} > c$. This fact is established in the following theorem.

Let $W|(D = 0) \sim G_0$, $W|(D = 1) \sim G_1$ and denote the corresponding densities and Survival functions by g_0, g_1 and by $\bar{G}_0, \bar{G}_1$.

Let $ROC_W(t) = \bar{G}_1(\bar{G}_0^{-1}(t)) = ROC_W(t)$. Then, following Pepe (2003) and Green and Swets (1966) we have,

1. If $W^* = h(Y)$ then $ROC_{W^*}(t) \leq ROC_W(t)\ \forall t \in (0, 1)$, with equality holding $(\forall t)$ iff $h(\cdot)$ is an increasing function of $r(\cdot)$.

2. $\frac{g_1(w)}{g_0(w)} = w$. Thus, $\frac{\partial ROC_W(t)}{\partial t} = \frac{g_1(\bar{G}_0^{-1}(t))}{g_0(\bar{G}_0^{-1}(t))} = \bar{G}_0^{-1}(t)$, which implies $ROC_W(t) = \int_0^t \bar{G}_0^{-1}(v)\, dv$

3. $ROC_W(t)$ is concave.

We will refer to $ROC_W(t)$ as the optimal ROC curve. That is $ROC_{opt}(t) = ROC_W(t)$.

The optimal ROC curve is simply the right directional ROC curve based on the random variable $W = r(Y)$. Since the optimal ROC is increasing and concave there exists a unique optimal decision threshold (which may translate to multiple thresholds in the original Y scale). We minimize $Q(t) = [t + \psi(1 - ROC_{opt}(t))]$ which results in the estimating equation $\frac{\partial ROC_{opt}(t)}{\partial t} = \frac{1}{\psi}$, or equivalently $\bar{G}_0^{-1}(t) = \frac{1}{\psi}$. Thus the optimal operating false positive rate equals $t^*_{opt} = \bar{G}_0\left(\frac{1}{\psi}\right)$. We note that t^*_{opt} is different than 0 or 1 when $\psi \in \left(inf\frac{f_1}{f_0}, sup\frac{f_1}{f_0}\right) = (inf\ r(y), sup\, r(y))$. Otherwise, $t^*_{opt} = 0$ for $\psi < inf\frac{f_1}{f_0}$ and $t^*_{opt} = 1$ for $\psi > sup\frac{f_1}{f_0}$.

2.3 Optimal ROC curve for a unimodal Likelihood ratio family

Let $\mathcal{F}$ be a family of Distributions that has the unimodality property of the Likelihood ratio . That is if we denote the support of $\mathcal{F}$ by the interval $\mathcal{S}$, we have for any $F_0, F_1 \in \mathcal{F}$, $r(y) = \frac{f_1(y)}{f_0(y)}$ is either a strictly monotone function of y or a strictly unimodal function of y ($r(y)$ attains a unique maximum or minimum in the interior of $\mathcal{S}$, thus $r(y)$ is either Decreasing - Increasing or Increasing-Decreasing). If $r(y)$ is strictly increasing then $ROC_{opt}(t) = S_1(S_0^{-1}(t))$. If $r(y)$ is strictly decreasing then $ROC_{opt}(t) = F_1(F_0^{-1}(t))$.

If $r(y)$ is Decreasing - Increasing then $r(y)$ has a minimum at $y_{(1)}$ in $\mathcal{S}^0$, the interior of the interval $\mathcal{S}$. That is $inf_{y\in\mathcal{S}}\ r(y) = min_{y\in\mathcal{S}^0}\ r(y) = r(y_{(1)}) = r_{min}$. Assume $sup_{y\in\mathcal{S}}\ r(y) = +\infty$ and is attained at the two borders of $\mathcal{S}$. The support of $W = r(Y)$ is the interval $(r(y_{(1)}), +\infty) = (r_{min}, +\infty)$. For $w \in (r_{min}, +\infty)$ the equation $r(y) = w$ (equivalently $log(r(y)) = log(w)$ has one root $c_1(w) < y_{(1)}$ and one root $c_2(w) > y_{(1)}$). We have, $G_0(w) = F_0(c_2(w)) - F_0(c_1(w))$. If $\psi < \frac{1}{r_{min}}$ then $t^*_{opt} = \bar{G}_0\left(\frac{1}{\psi}\right)$, whereas if $\psi > \frac{1}{r_{min}}$ then $t^*_{opt} = 1$. We observe that when ψ exceeds a certain threshold the optimal decision is to classify everybody as diseased.

If $r(y)$ is Increasing-Decreasing then $r(y)$ has a maximum at $y_{(1)}$ in $\mathcal{S}^0$. That is $sup_{y\in\mathcal{S}}\ r(y) = sup_{y\in\mathcal{S}^0}\ r(y) = r(y_{(1)}) = r_{sup}$. Assume $inf_{y\in\mathcal{S}}\ r(y) = -\infty$ and is attained at the two borders of $\mathcal{S}$. The support of $W = r(Y)$ is the interval $(0, r(y_{(1)})) = (0, r_{sup})$. For $w \in (0, r_{sup})$, the equation $r(y) = w$. We have, $G_0(w) = 1 - F_0(c_2(w)) + F_0(c_1(w))$. If $\psi > \dfrac{1}{r_{sup}}$ then $t^*_{opt} = \bar{G}_0\left(\dfrac{1}{\psi}\right)$, whereas $\psi < \dfrac{1}{r_{sup}}$ then $t^*_{opt} = 0$. We observe that when ψ falls below a certain threshold the optimal decision is to classify everybody as healthy.

2.4 Some results for commonly used parametric distributions

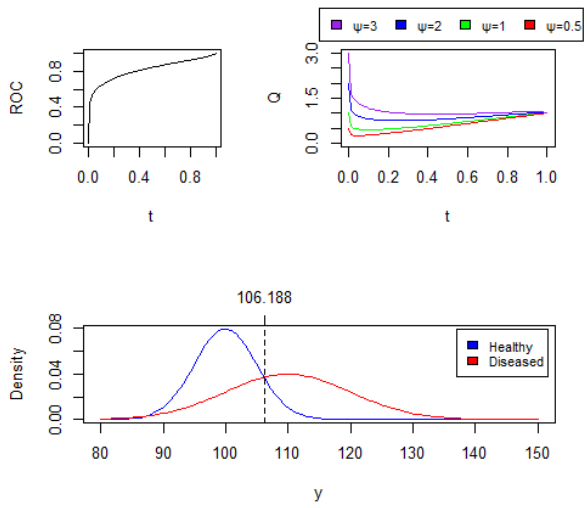
2.4.1 Threshold selection when using the Binormal ROC curve

Suppose, Y is a normal random variable such that $Y|D = 0 \sim N(\mu_0, \sigma_0^2)$ $Y|D = 1 \sim N(\mu_1, \sigma_1^2)$. Since the ROC curve is invariant to strictly increasing transformations of Y we can standardize according to the diseased distribution. That is, without loss of generality we assume $Y|D = 0 \sim N(-a, b^2)$, $Y|D = 1 \sim N(0, 1)$, where $-\infty < a = \dfrac{\mu_1 - \mu_0}{\sigma_1} < +\infty$, $b = \dfrac{\sigma_0}{\sigma_1} > 0$. Then the classical right directional ROC is:

$$ROC(t) = \Phi(a + b\Phi^{-1}(t)),$$

Example: Assume that $Y|D = 0 \sim N(100, 5^2)$, $Y|D = 1 \sim N(110, 10^2)$. In figure 1 we present the two normal densities with a specific cutpoint value (for $\psi = 1$), the right directional ROC curve and $Q(t)$ for multiple ψ.

Figure 1: *Densities $ROC(t)$ and $Q(t)$*



After some algebra we obtain the survival functions of the optimal transform $W = r(Y) = \dfrac{f_1(Y)}{f_0(Y)}$ for the two populations as follows:

Case 1: $b < 1$. For $w < \inf_y r(y) = \exp\left[\ln b - \dfrac{a^2}{2(1-b^2)}\right]$ we have $\bar{G}_0(w) = 1 - G_0(w) = 1$ and $\bar{G}_1(w) = 1 - G_1(w) = 1$. For $w \geq \exp\left[\ln b - \dfrac{a^2}{2(1-b^2)}\right]$ we have,

$$\bar{G}_0(w) = \bar{\Phi}\left(k_0(w) - \frac{ba}{1-b^2}\right) + \Phi\left(-k_0(w) - \frac{ba}{1-b^2}\right)$$

$$\bar{G}_1(w) = \bar{\Phi}\left(k_1(w) - \frac{a}{1-b^2}\right) + \Phi\left(-k_1(w) - \frac{a}{1-b^2}\right)$$

where,

$$k_0(w) = \sqrt{\frac{a^2}{(1-b^2)^2} + \frac{2}{1-b^2}\ln\left(\frac{w}{b}\right)}$$

$$k_1(w) = \sqrt{\frac{a^2b^2}{(1-b^2)^2} + \frac{2b^2}{1-b^2}\ln\left(\frac{w}{b}\right)}$$

Case 2: $b > 1$. For $w > \sup_y r(y) = \exp\left[\ln b - \dfrac{a^2}{2(1-b^2)}\right]$ we have $G_0(w) = G_1(w) = 1$. For $w \leq \exp\left[\ln b - \dfrac{a^2}{2(1-b^2)}\right]$ we have ,

$$\bar{G}_0(w) = \Phi\left(k_0(w) - \frac{ba}{1-b^2}\right) - \Phi\left(-k_0(w) - \frac{ba}{1-b^2}\right)$$

$$\bar{G}_1(w) = \Phi\left(k_1(w) - \frac{a}{1-b^2}\right) - \Phi\left(-k_1(w) - \frac{a}{1-b^2}\right)$$

In the case where $b = 1$ the optimal ROC curve is simply a classical directional one. It is either the right directional ROC curve (when $a > 0$) or a left directional one (when $a < 0$).

Example 1 (continues):

In figure 2 we can visually compare the Optimal $ROC_{opt}(t)$ (green color) compared to the Directional $ROC(t)$.

Figure 2: *Optimal and Directional ROC*

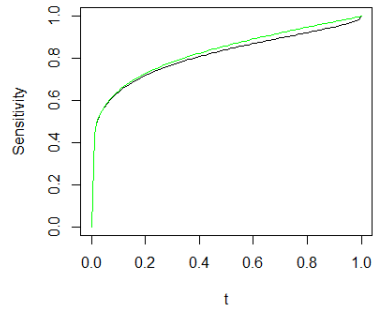
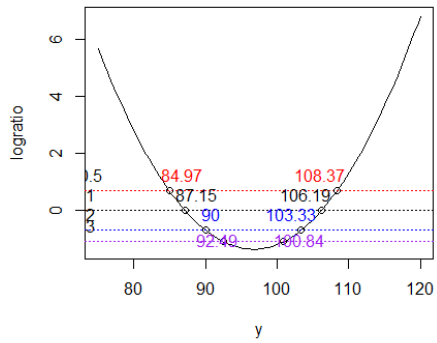


Figure 3 shows the logarithm of the likelihood ratio versus marker values. Optimal cutoffs in the original scale are indicated in the plot.

Figure 3: *Logratio vs* $\psi = 1(black), 2(blue), 3(purple), 0.5(red)$



2.4.2 Gamma Distributions

Assume that $Y|D = 0 \sim Gamma(a_0, b_0)$ and $Y|D = 0 \sim Gamma(a_1, b_1)$ (using the shape-rate parametrization). Notice that $b_0 Y|D = 0 \sim Gam(a_0, 1)$ and $b_0 Y|D = 1 \sim Gamma(a_1, \gamma = \frac{b_1}{b_0})$. Hence, without loss of generality we consider $Gamma(a_0, 1)$ and $Gamma(a_1, \gamma)$ as the control and case distributions respectively. The logarithm of the density ratio is then,

$$\ln(r(y)) = \ln(\delta) + \kappa \ln(y) + \lambda y,$$

where $\gamma = \frac{b_1}{b_0}$, $\delta = \frac{\Gamma(a_0)\gamma^{a_1}}{\Gamma(a_1)} > 0$, $\kappa = a_1 - a_0$, $\lambda = 1 - \gamma$.

The derivative of the logarithm of the density ratio is

$$\left(\ln(r(y))\right)' = \frac{\kappa}{y} + \lambda.$$

We thus have the following :

Case 1: $\kappa > 0, \lambda > 0$. The $\ln(r(y))$ function is strictly increasing with $\ln(r(0)) = -\infty$, $\ln(r(\infty)) = +\infty$. In this case the optimal ROC curve is the classical right directional ROC curve. Case 2: $\kappa < 0, \lambda < 0$. The function $\ln(r(y))$ is strictly decreasing with $\ln(r(0)) = +\infty$, $\ln(r(\infty)) = -\infty$. The optimal ROC curve is the left directional ROC curve. Case 3: $\kappa < 0, \lambda > 0$. The function $\ln(r(y))$ is decreasing in the interval $\left(0, \frac{-\kappa}{\lambda}\right)$ and increasing in the interval $\left(\frac{-\kappa}{\lambda}, \infty\right)$. We also have that $\ln(r(0)) = +\infty$, $\ln(r(\infty)) = +\infty$. Hence there exists a unique minimum at $min(\ln(r(y))) = \ln(r(\frac{-\kappa}{\lambda}))$ and we conclude that the optimal ROC curve is a two cut point ROC. Case 4: $\kappa > 0, \lambda < 0$. $\ln(r(y))$ is increasing in the interval $\left(0, \frac{-\kappa}{\lambda}\right)$ and decreasing in the interval $\left(\frac{-\kappa}{\lambda}, \infty\right)$. We have that $\ln(r(0)) = -\infty$, $\ln(r(\infty)) = -\infty$. Hence We a unique maximum exists at $max(\ln(r(y))) = \ln(r(\frac{-\kappa}{\lambda}))$ and we conclude that the optimal ROC curve is a two cut point ROC.

Note that there are four special cases that occur. In the first special case, where $\kappa < 0$ and $\lambda = 0$, $r(y)$ is increasing with $r(y) \underset{y\to 0}{\to} 0$ and $r(y) \underset{y\to+\infty}{\to} +\infty$. In the second special case, where $\kappa > 0$ and $\lambda = 0$, $r(y)$ is decreasing with $r(y) \underset{y\to 0}{\to} +\infty$ and $r(y) \underset{y\to+\infty}{\to} 0$. In the third special case, where $\lambda > 0$ and $\kappa = 0$, $r(y)$ is decreasing with $r(y) \underset{y\to 0}{\to} \delta_0$ and $r(y) \underset{y\to+\infty}{\to} 0$, where $\delta_0 = \gamma^{a_0}$. In the fourth special case, where $\lambda < 0$, $\kappa = 0$, $r(y)$ is increasing with $r(y) \underset{y\to 0}{\to} \delta_0$ and $r(y) \underset{y\to+\infty}{\to} +\infty$.

By letting $W = r(Y)$ we can easily derive the distribution function of W, for controls (and cases). Since the optimal ROC curve is determined by the CDF of W for controls ($ROC_W(t) = \int_0^t \bar{G}_0^{-1}(v)\, dv$) we provide the CDF of W for controls in the two cases where the optimal ROC is the two cutpoint ROC on the original scale.

We have for $w > 0$ that $G_0(w) = P[W < w|D = 0]P[\ln(W) < \ln(w)|D = 0] = P[\kappa \ln(Y) + \lambda Y < \ln(\frac{w}{\delta})|D = 0] = P[h(Y) < \ln(\frac{w}{\delta})|D = 0]$, where, $h(y) = \kappa \ln(y) + \lambda y$.

So, for Case 3 ($\kappa < 0$, $\lambda > 0$), we have

$$G_0(w) = \begin{cases} 0, & w < \delta e^{h\left(\frac{-\kappa}{\lambda}\right)} \\ \frac{\gamma(a_0, c_2(w)) - \gamma(a_0, c_1(w))}{\Gamma(a_0)}, & w \geq \delta e^{h\left(\frac{-\kappa}{\lambda}\right)}, \end{cases}$$

where $\gamma(a_0, \cdot)$ is the lower incomplete gamma function and $c_1(w) < c_2(w)$ denote the

two roots of the equation $h(y) - \ln(\frac{w}{\delta}) = 0$.

For Case 4 ($\kappa > 0$, $\lambda < 0$), we have

$$G_0(w) = \begin{cases} 0, & w \leq 0 \\ 1 - \frac{\gamma(a_0, c_2(w)) - \gamma(a_0, c_1(w))}{\Gamma(a_0)}, & 0 < w \leq \delta e^{h\left(\frac{-\kappa}{\lambda}\right)} \\ 1, & w > \delta e^{h\left(\frac{-\kappa}{\lambda}\right)} \end{cases}$$

2.4.3 Beta Distribution

Let assume that cases and controls follow $Beta(a_0, b_0)$ and $Beta(a_1, b_1)$ distributions respectively. We have,

$$\ln(r(y)) = \ln(\delta) + \kappa \ln(y) + \lambda \ln(1 - y),$$

where, $\delta = \dfrac{B(a_0, b_0)}{B(a_1, b_1)}$, $\kappa = a_1 - a_0$, $\lambda = b_1 - b_0$.

Hence

$$\left(\ln(r(y))\right)' = \frac{\kappa}{y} - \frac{\lambda}{1 - y}.$$

We have the general four cases to consider, Case 1: $\kappa > 0, \lambda > 0$.. We have that $\ln(r(y))$ is increasing in the interval $\left(0, \frac{\kappa}{\kappa+\lambda}\right)$ and $\ln(r(y))$ is decreasing in the interval $\left(\frac{\kappa}{\kappa+\lambda}, 1\right)$. Also we observe that at the borders of the interval $\mathcal{S}$, $-\infty$ is attained. We have that $\ln(r(0)) = -\infty \quad \ln(r(1)) = -\infty$. We obtain a unique maximum $max(\ln(r(y))) = \ln(r(\frac{\kappa}{\kappa+\lambda}))$. Thus, the optimal ROC curve is a two cut point ROC. Case 2: $\kappa < 0, \lambda < 0$. We have $\ln(r(y))$ is decreasing in the interval $\left(0, \frac{\kappa}{\kappa+\lambda}\right)$ and $\ln(r(y))$ is increasing in the interval $\left(\frac{\kappa}{\kappa+\lambda}, 1\right)$. Also we observe that at the borders of the interval $\mathcal{S}$, $+\infty$ is attained. We have that $\ln(r(0)) = +\infty, \quad \ln(r(1)) = +\infty$. We obtain a unique minimum $min(\ln(r(y))) = \ln(r(\frac{\kappa}{\kappa+\lambda}))$. We have that the optimal ROC curve is a two cut point ROC. Case 3: $\kappa < 0, \lambda > 0$. $\ln(r(y))$ is strictly decreasing with $\ln(r(0)) = +\infty, \ln(r(1)) = -\infty$. In this case the optimal ROC curve is the left directional one. Case 4 for $\kappa > 0, \lambda < 0$. $\ln(r(y))$ is strictly increasing with $\ln(r(0)) = -\infty, \ln(r(1)) = +\infty$. So in this case the optimal ROC curve is the right directional one.

Note that there are four special cases to consider. In the first special case, where $\kappa < 0$ and $\lambda = 0$, $r(y)$ is increasing and we have $r(y) \underset{y \to 0}{\to} 0$ and $r(y) \underset{y \to 1}{\to} \delta_1$, where $\delta_1 = \dfrac{B(a_0, b_0)}{B(a_1, b_0)}$. The optimal ROC curve is the right directional one. In the second

special case, where $\kappa > 0$ and $\lambda = 0$, we have $r(y)$ is decreasing and $r(y) \underset{y\to 0}{\to} +\infty$ and $r(y) \underset{y\to 1}{\to} \delta_1$. The optimal ROC curve is the left directional one. In the third special case, where $\lambda > 0$ and $\kappa = 0$ we have $r(y)$ is decreasing and $r(y) \underset{y\to 0}{\to} \delta_2$ and $r(y) \underset{y\to 1}{\to} 0$, where $\delta_2 = \dfrac{B(a_0, b_0)}{B(a_0, b_1)}$. The optimal ROC curve is the left directional one. Lastly, in the fourth special case, where $\lambda < 0$ and $\kappa = 0$ we have $r(y)$ is increasing and $r(y) \underset{y\to 0}{\to} \delta_2$ and $r(y) \underset{y\to 1}{\to} +\infty$. The optimal ROC curve is the right directional one.

The CDF of $W = r(Y)$ for controls is given below for the two scenarios where a two cutpoint ROC curve is optimal.

We have for $w > 0$, $G_0(w) = P[\ln(W) < \ln(w)|D = 0] = P[\kappa \ln(Y) + \lambda \ln(1 - Y) < \ln(\frac{w}{\delta})|D = 0] = P[h(Y) < \ln(\frac{w}{\delta})|D = 0]$, where $h(y) = \kappa \ln(y) + \lambda \ln(1 - y)$.

Hence, for Case 1 ($\kappa > 0$, $\lambda > 0$),

$$G_0(w) = \begin{cases} 0, & w \le 0 \\ 1 - [I_{c_2(w)}(a_0, b_0) - I_{c_1(w)}(a_0, b_0)], & 0 < w \le \delta e^{h(\frac{\kappa}{\kappa+\lambda})} \\ 1, & w > \delta e^{h(\frac{\kappa}{\kappa+\lambda})} \end{cases} \tag{3}$$

where, $I_y(a_0, b_0)$ is the regularized incomplete beta function (the CDF of $Y|D = 0$) and $c_1(w) < c_2(w)$ are the two roots of the equation $h(y) - \ln(\frac{w}{\delta}) = 0$

For Case 2 ($\kappa < 0$, $\lambda < 0$) we have

$$G_0(w) = \begin{cases} 0, & w < \delta e^{h(\frac{\kappa}{\kappa+\lambda})} \\ I_{c_2(w)}(a_0, b_0) - I_{c_1(w)}(a_0, b_0), & w \ge \delta e^{h(\frac{\kappa}{\kappa+\lambda})} \end{cases} \tag{4}$$

As mentioned before the optimal ROC curve for these two scenarios can be obtained by $ROC_{opt}(t) = \int_0^t \bar{G}_0^{-1}(v)\, dv$.

2.5 Statistical inference

Since in this paper we deal with standard parametric families, estimation is straightforward and is based on the maximum likelihood estimates of the parameters for the two distributions involved. Suppose we have $Y_{01}, Y_{02}, ..., Y_{0n_1}$, a sample from the healthy population ($D = 0$) and $Y_{11}, Y_{12}, ..., Y_{1n_2}$, a sample from the diseased population

$(D = 1)$. Given ψ, we estimate the $t^*_{opt} = \bar{G}_0\left(\frac{1}{\psi}\right)(FPR)$ by $\hat{t}^*_{opt} = \hat{\bar{G}}_0\left(\frac{1}{\psi}\right)$. Where $\hat{\bar{G}}_0(y)$ is the maximum likelihood estimate of $\bar{G}_0$, the CDF of $W = r(Y)|D = 0$. For example, if we have two Beta-distributed populations we can obtain an estimate of $\hat{t}^*_{opt}$ by first obtaining the parameter MLEs $\hat{\kappa}$, $\hat{\lambda}$ and then estimating $\hat{\bar{G}}_{0,MLE}\left(\frac{1}{\psi}\right)$ using (3) if $\hat{\kappa} > 0$, $\hat{\lambda} > 0$ or equation (4) if $\hat{\kappa} < 0$, $\hat{\lambda} < 0$. We obtain a $(1 - \alpha)$ confidence interval by using a bootstrap procedure. We acquire B bootstrap samples from the two original samples by sampling with replacement and we compute $\hat{t}^*_{opt,i}$, $i = 1, ..., B$. The endpoints of the $100(1 - \alpha)\%$ confidence interval are the $\left(\frac{\alpha}{2}\right)$ and $\left(1 - \frac{\alpha}{2}\right)$ quantiles from the bootstrap procedure. We note that our confidence intervals are closed intervals.

3. Application

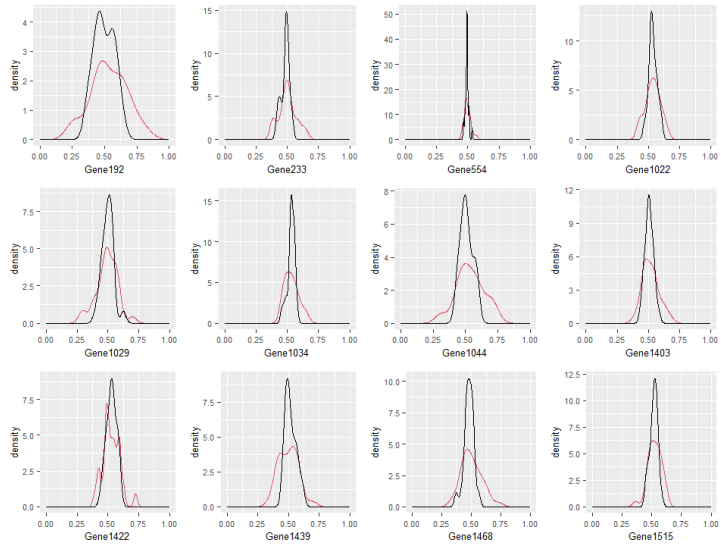
In this application, we analyze a real dataset derived from expression arrays of 30 ovarian cancer tissues and 23 tissues without cancer. The dataset consists of mRNA expression measurements for 1536 gene clones (Pepe 2003). The expression levels, denoted as $Y_{i,g}$, represent the relative mRNA expression of the gth gene in the ith tissue sample compared to a control tissue. It is important to note that the same control tissue was used for all experiments.

Dudoit (2002) and Newton (2000) provide comprehensive explanations of this technology, including a technical overview of how the expression values $Y_{i,g}$ are calculated. This technology involves using glass arrays spotted with gene clones to measure mRNA expression levels in the tissue samples. These measurements provide valuable insights into the differences in gene expression between ovarian cancer tissues and healthy tissues.

In this application, a transformation is applied to each gene, given by the formula: $y_* = \dfrac{y}{y+1}$. We note that if we assume beta-prime distributions for the two populations in the original scale, this transformation leads to two beta distributions. After applying this transformation to all 1536 genes, a goodness-of-fit test for beta distributions was conducted separately for the cases and controls. A total of 915 of the genes exhibit p-values greater than or equal to 0.05 for both cases and controls. This indicates that for these 915 genes, there is no significant evidence to suggest a departure from the assumed Beta distribution.

We provide the density plots for 12 specific genes, namely: g192, g233, g554, g1022, g1029, g1034, g1044, g1403, g1422, g1439, g1468 and g1515. These genes were selected because they suggest a two cutpoint ROC curve as appropriate and, in addition, have an empirical Area Under the Curve (AUC) close to 0.5, indicating that they are individually useless when considering the classical ROC framework.

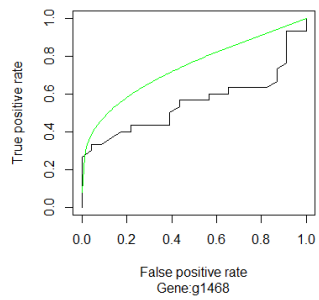
Figure 4: *Kernel density estimates for the 12 selected Genes (black:Controls , red:Cases)*



We examined Gene g1468 in more detail. It has an empirical AUC$= 0.5521739$ and s.e.$= 0.08147481$, yielding a 95% C.I. equal to $(0.3924833, 0.7118645)$. By using Beta Distribution Assumptions as we have saw at section 2.4.4 we obtain the MLE estimate for the Optimal AUC, which is equal to 0.7425499. Using the bootstrap method we obtain a 95% C.I. $(0.6364681, 0.8415323)$, (with 1000 bootstraps).

In the following plot we present with black color the Empirical ROC and with green color the Optimal ROC of Gene:g1468.

Figure 5: *Empirical and Optimal ROC*



Finally, in Table 2 we present estimates and 95% confidence intervals for the optimal FPR, Sensitivity and Cost for various values of ψ.

Table 1: *Optimal $FPRs$, Sensitivities and Costs*

ψ	t^*_{opt}	ROC_{opt} (t^*_{opt})	Q (t^*_{opt})	C.I t^*_{opt}	C.I. ROC_{opt} (t^*_{opt})	C.I. $Q(t^*_{opt})$
0.5	0.052	0.387	0.359	(0.037, 0.054)	(0.314, 0.466)	(0.305, 0.396)
1	0.149	0.519	0.631	(0.089, 0.179)	(0.479, 0.556)	(0.552, 0.697)
1.5	0.299	0.638	0.849	(0.151, 0.419)	(0.577, 0.686)	(0.677, 0.91)
2	0.541	0.775	0.991	(0.227, 0.874)	(0.626, 0.938)	(0.933, 1.005)
$\hat{\kappa} = -62.7$, $\hat{\lambda} = -69.6$, $bootsize = 1000$						

Following the application presented we conducted a small simulation by considering a scenario where the number of cases is 30 and the number of controls is 23. We assumed that both samples were drawn from beta distributions identical to the estimated ones in the application. We evaluated the coverage performance of our confidence interval for t^*_{opt} and for multiple values of ψ. We performed 1000 simulations with 5000 bootstrap samples for each. The resulting coverage of our 95% C.I. are summarized in the following table:

Table 2: *Coverage of Confidence Intervals*

ψ	90%	95%	99%
0.5	0.897	0.956	0.986
1	0.905	0.946	0.985
1.5	0.912	0.938	0.989
2	0.911	0.965	0.986
$\kappa = -62.7$, $\lambda = -69.6$, $bootsize = 5000$, $nsim = 1000$			

Conclusion: In this paper we presented methods for obtaining the optimal operating points on the ROC curve given unimodality of the likelihood ratio and in the presence of known disease prevalence and misclassification costs. Our approach was fully parametric. In the future we will pursuit nonparametric inference under the same assumptions.

ΠΕΡΙΛΗΨΗ:

Αυτή η εργασία αποσκοπεί στην εξερεύνηση τεχνικών που ενσωματώνουν ανάλυση κόστους λανθασμένης ταξινόμησης στις βέλτιστες καμπύλες ROC. Ειδικότερα, υποθέτοντας μια οικογένεια κατανομών με την ιδιότητα του μονότονου λόγου πιθανοφάνειας, η βέλτιστη καμπύλη ROC είναι μια καμπύλη ROC με δύο κατωφλιακά σημεία. Λαμβάνουμε αποτελέσματα σχετικά με την εκτίμηση του βέλτιστου κατωφλιακού σημείου/σημείων και του αντίστοιχου ψευδώς θετικού ποσοστού (FPR), λαμβάνοντας υπόψη τα δύο κόστη λανθασμένης ταξινόμησης και του επιπολασμού της νόσου. Παρουσιάζουμε λεπτομέρειες για ορισμένες κοινές παραμετρικές οικογένειες που εκδηλώνουν την ιδιότητα του μονοτονικού λόγου πιθανοφάνειας. Συνδυάζουμε τη μέθοδο της μέγιστης πιθανοφάνειας με μεθόδους αναδειγματοληψίας

για στατιστική εκτίμηση του βέλτιστου κατωφλιακού σημείου.

REFERENCES

Bantis, L., Tsimikas, J., Chambers, G., Capello, M., Hanash, S., Feng, Z., (2021). The length of the receiver operating characteristic curve and the two cutoff Youden index within a robust framework for discovery, evaluation, and cutoff estimation in biomarker studies involving improper receiver operating characteristic curves, *Statistics in Medicine*, **40**(7) , 1767–1789.

Dudoit, S., Yang, Y. H., Speed, T. P., Callow, M. J., (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments, *Statistica Sinica* , **12**(1), 111–139.

Egan JP. (1975). *Signal Detection Theory and ROC Analysis*, Academic Press, New York.

Green, D.M., Swets, J.A. (1966). *Signal detection theory and psychophysics*, Wiley, New York.

Hillis, L., (2016). Equivalence of binormal likelihood-ratio and bi-chi-squared ROC curve models, *Statistics in Medicine* , **35**(12), 2031–2057.

Lusted, L., B., (1971). Signal detectability and medical decision making. *Science*, **171**, 1217–1219.

Martínez-Camblor, P., Corral, N., Rey, C., Pascual, J., Cernuda-Morollón, E., (2017). Receiver operating characteristic curve generalization for non-monotone relationships, *Stat Methods Med Res*, **26**(1) , 113–123.

Martínez-Camblor, P., Pardo-Fernandez, JC., (2019. Parametric estimates for the receiver operating characteristic curve generalization for non-monotone relationships, *Stat Methods Med Res*, **28**(7) , 2032–2048.

Newton, M. A., Kendziorski, C. M., Richmond, C. S., Blattner, F. R., Tsui, K. W., (2000). On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data, *Journal of Computer Biology* , **81**, 37–52.

Pepe, M., S., (2003). *The statistical evaluation of medical tests for classification and prediction*, Oxford University Press.

Pepe, M., S., Longton, G., Anderson, G., Schummer, M., (2003). Selecting Differentially Expressed Genes from Microarray Experiments *Biometrics* , **59**(1), 133–142.

https://research.fredhutch.org/diagnostic-biomarkers-center/en/datasets.html.

Zhou, X.H., Obuchowski, N.A. and McClish, D.K. (2002). Statistical methods in diagnostic medicine Wiley, New York.

Zou, K. H., Yu, C. R., Liu, K., Carlsson, M. O., Cabrera, J. (2013). Optimal Thresholds by Maximizing or Minimizing Various Metrics via ROC-Type Analysis, *Academic Radiology*, **20**(7), 807–815.


Greek Statistical Institute
Proceedings of the 35th Panhellenic Statistics conference 2023),p. 122−135

SUBDATA SELECTION FOR BIG DATA REGRESSION BASED ON LEVERAGE SCORES

V. Chasiotis, D. Karlis
Department of Statistics, Athens University of Economics and Business, Greece
{chasiotisv, karlis}@aueb.gr

ABSTRACT

Regression can be really difficult in case of big datasets, since we have to dealt with huge volumes of data. The demand of computational resources for the modeling process increases as the scale of the datasets does, since traditional approaches for regression involve inverting huge data matrices. The main problem relies on the large data size, and so a standard approach is subsampling that aims at obtaining the most informative portion of the big data. In the current paper we provide a new algorithm based on leverages scores that improves existing approaches that select subdata using properties of orthogonal arrays as well as information criteria. A simulation experiment and a real data application are also provided.

Keywords: Designs of experiments; Optimality; Influential data points; Linear regression; Subsampling.

1. INTRODUCTION

Due to the size and complexity of big datasets, they can easily exceed the capacity of traditional computing resources. This may lead to the inability to perform certain analyses altogether. Also, more memory or processing power may be required by some statistical models. However, due to unavailability on standard machines, the analysis process can be really complicating.

An approach, in order to address these challenges, is the selection of subdata from the big data to conduct the analysis. Data reduction performs the analysis on a smaller sample that have been selected from the full data. Such an approach leads to reduction of the computational resources that are required for analysis, and so a targeted analysis on the smaller dataset is possible.

The selection of the subdata should be carefully done, in order to ensure that it is representative of the big data, and so to conclusions from the analysis of the subdata can be extrapolated to the big data. Overall, due to computational resource limitations, significant challenges for statistical analyses can be posed by big data, data reduction can be a useful technique for overcoming these challenges. However, before the im-

plementation of the subsampling, it is really important to take into consideration any potential drawbacks and limitations.

Drineas et al. (2011) proposed to select randomly a portion of the data. Their idea based on a randomized Hadamard transform on data and then to take subdata at random using uniform subsampling. Their goal was the approximation of the ordinary least squares estimators in linear regression models. However, their approach suffers from the inherent randomness.

As a consequence, an alternative approach, rapidly developing in recent years, focuses on selecting data points deterministically, so that a small portion of the full data preserves most of the information contained in the full data. Since optimal-design problems relies on data selection, such approaches are connected with the concept of the design of experiments. Therefore, the theory of optimal designs can be very useful in establishing a framework to select the most informative subdata from the full data.

As a first attempt, Wang et al. (2019) proposed the information-based optimal subdata selection (IBOSS) approach, which is motivated by the concept of optimal experimental designs. They focused on the selection of the most informative subdata from the full data for the estimation of unknown parameters. Their idea was the "maximization" of an information matrix, that is the central goal in the theory of optimal experimental designs. Overall, they concluded that a subdata that maximizes the determinant of the inverse of the covariance matrix of the unknown parameters is D-optimal, and so it contains the most informative data points from the full data. Also, they developed an algorithm based on a result for an upper bound of the determinant of the inverse of the covariance matrix of the unknown parameters given the subdata. 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, and its time complexity is $O(np + kp^2)$, or $O(np)$ when $n > kp$. It is important to mention that the IBOSS algorithm outperforms both the uniform and the leverage-based subsampling approaches (random subsampling-based approaches) for the selection of informative subdata from the full data.

Wang et al. (2021) proposed the orthogonal subsampling (OSS) approach to select subdata, that is their approach is based on the optimality of two-level orthogonal arrays. We need to mention that a two-level orthogonal array minimizes the average variance of the estimated parameters as well as provides the best predictions (Dey and Mukerjee, 1999), and so it represents an optimal design for linear regression. The sequential addition algorithm developed by Wang et al. (2021) is based on the combinatorial orthogonality of a two-level orthogonal array. Also, they prevent the algorithm to be time-consuming, by eliminating data points, and so its computational complexity is $O(np\log k)$. The algorithm is based on a discrepancy function that measures the distortion of data points on keeping two features simultaneously that are connected with the optimality of orthogonal arrays. The first feature is the selection of extreme data points and the second one is that the signs of the selected data points are as dissimilar as possible (combinatorial orthogonality). Moreover, the OSS approach outperforms the

IBOSS approach for the selection of informative subdata from the full data.

The approach of Ren and Zhao (2021), motivated by Wang et al. (2021), focuses on selecting subdata that approach a $k \times p$ two-level orthogonal array of strength 2. However, Chasiotis and Karlis (2023) mentioned an issue about implementation of Algorithm 3 of Ren and Zhao (2021), and so their approach is not taken into consideration.

Furthermore, the approach of Chasiotis and Karlis (2023) aims at identifying and interchanging data points that were not selected with those that have already been selected by an approach, i.e. the OSS or the IBOSS one, in order to improve the value of the D-optimality criterion. The two proposed algorithms in Chasiotis and Karlis (2023) are considered as extensions to existing ones, and so they should be evaluated considering a trade-off between improving the value of the D-optimality criterion at the cost of some additional computational time.

The approach of Wang et al. (2019) has been extended to other cases, e.g. multinomial logistic regression (Yao and Wang, 2019), quantile regression (Wang and Ma, 2021), and for logistic regression (Cheng et al., 2020). Also, for further related work we refer the reader to Wang (2019), Lee et al. (2021), and Deldossi and Tommasi (2022). More information on subdata selection or subsampling from big data based on designs can be found in the review papers by Yao and Wang (2021) and Yu et al. (2023), which provides a comprehensive overview of the current state of research in this area.

The approach in the current paper, which is a deterministic selection of the most informative data points from the full data based on leverages scores (LEVSS), has already been proposed to select subdata for linear model discrimination by Yu and Wang (2022). In the review paper by Yu et al. (2023), the authors have evaluated a numerous of algorithms for their ability to select the most informative subdata from big data providing simulation experiments as well as a real data application. One of the evaluated algorithms is the one by Yu and Wang (2022). However, there are some motivations that drive us to further investigate the role of leverage scores in the selection of the most informative data points. Also, we are interested in comparing the subdata selected based on leverage scores with the IBOSS and OSS approaches, in case of linear models, that is we assume predictors and responses follow a postulated model. This means that we do not take into consideration model-free subsampling methods, with the aim of elucidating the underlying processes in case of assuming a postulated model.

At first, since the IBOSS and OSS approaches obtain subdata that are D-optimal in some sense, Theorem 2 in Yu and Wang (2022) motivates us that the selection of the most informative data points based on leverages scores should be further investigated. We are also motivated by Chasiotis and Karlis (2023), who focused on selecting data points with large convex hull as close as possible to the one generated by the full data, that is, under the subdata, the determinant of the information matrix will be large. Also, they proved that the maximization of the determinant of the information matrix can be addressed as the maximization of the generalized variance of covariates under the selected subdata. It is important to note that the volume of space occupied by the cloud of the selected data points is proportional to the square root of the generalized variance.

The remaining of the paper is organized as follows. In Section 2., we briefly review the best linear unbiased estimator based on the subdata under a linear regression model. In Section 3., we describe the algorithm of the LEVSS approach. In Section 4., we provide simulation evidence to support the LEVSS approach, including a comparison with existing approaches (IBOSS and OSS). In Section 5., we use a real dataset for illustration, and in Section 6. this article is concluded with some discussions.

2. PRELIMINARIES

Assume that the full data are denoted by $(\mathbf{x}_i, y_i), \ldots, (\mathbf{x}_n, y_n)$. Let the 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 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, y_i is a response, and ϵ_i is an error term. The y_i's are uncorrelated given the covariates $\mathbf{x}_i$, $i = 1, 2, \ldots, n$ and ϵ_i's satisfy $\mathrm{E}(\epsilon_i) = 0$ and $\mathrm{V}(\epsilon_i) = \sigma^2$.

Taking into consideration the full data under model (1), the least-square estimator of $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1^{\mathrm{T}})^{\mathrm{T}}$, that is its best linear unbiased estimator, 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 inverse of

$$\mathbf{Q}_{\text{Full}} = \frac{1}{\sigma^2} \sum_{i=1}^{n} \mathbf{z}_i \mathbf{z}_i^{\mathrm{T}}$$

is equal to the covariance matrix of $\hat{\boldsymbol{\beta}}_{\text{Full}}$, where $\mathbf{Q}_{\text{Full}}$ is the observed Fisher information matrix of $\boldsymbol{\beta}$ for the full data in case that the error terms ϵ_i's are normally distributed. $\mathbf{Q}_{\text{Full}}$ is still called the information matrix, even though the normality assumption is not required.

If the sample size n of the full data is too large, then a fully analysis of the whole data may be infeasible. Therefore, based on limitations of the computational resources, we are interested in gaining useful information from the full data by the selection of a subset of the full data.

Let δ_i be a indicator variable about the inclusion of $(\mathbf{x}_i, y_i)$ in the subdata. Therefore, $\delta_i = 0$ if $(\mathbf{x}_i, y_i)$ is not included in the subdata and $\delta_i = 1$ otherwise. Also, we assume that we want to select subdata of size k, that is $\sum_{i=1}^{n} \delta_i = k$. Thus, the least-square estimator of $\boldsymbol{\beta}$ is still the best linear unbiased estimator based on the subdata, that is,

$$\hat{\boldsymbol{\beta}}_{\text{Sub}} = \left(\sum_{i=1}^{n} \delta_i \mathbf{z}_i \mathbf{z}_i^{\mathrm{T}}\right)^{-1} \sum_{i=1}^{n} \delta_i \mathbf{z}_i y_i.$$

The information matrix under the 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 D-optimality criterion, a subdata of size k is D-optimal in some sense if the determinant of the corresponding $\mathbf{Q}_{\text{Sub}}$ is maximized.

Chasiotis and Karlis (2023) proved that the determinant of $\mathbf{Q}_{\text{Sub}}$ is the generalized variance (Wilks, 1932) of covariates $\mathbf{x}_i$, $i = 1, 2, \ldots, p$ under the selected subdata, and so they addressed the problem of maximizing the determinant of $\mathbf{Q}_{\text{Sub}}$ as a problem of maximizing the generalized variance of covariates under the selected subdata.

3. LEVERAGE SCORE BASED ALGORITHM

In this section, we provide the deterministic leverage score selection (LEVSS) algorithm proposed by Yu and Wang (2022) for linear model discrimination.

Following the notations given by Yu and Wang (2022), let $\#(\Gamma)$ denote the cardinal number of a set Γ, and $\kappa(\mathbf{B}) := \lambda_{max}(\mathbf{B}) / \lambda_{min}(\mathbf{B})$ denote the condition number of a matrix $\mathbf{B}$, where $\lambda_{max}(\mathbf{B})$ and $\lambda_{min}(\mathbf{B})$ are the maximum and minimum eigenvalues of the squared matrix $\mathbf{B}$, respectively. Also, when $\mathbf{B}$ is a singular matrix, then $\kappa(\mathbf{B}) = \infty$.

The LEVSS algorithm is provided in Algorithm 1.

Algorithm 1 LEVSS

Input: The design matrix $\mathbf{X} = (\mathbf{x}_i^{\mathrm{T}}), i = 1, 2, \ldots, n$, the target sample size ($k > p$), and the threshold T (≥ 1).

Output: The selected index set Γ and the design matrix under the subdata.

Initialization: $\Gamma = \emptyset$, $\mathbf{U}_\Gamma = \emptyset$, $\kappa\left(\mathbf{U}_\Gamma^{\mathrm{T}} \mathbf{U}_\Gamma\right) = \infty$.

Perform a singular value decomposition of $\mathbf{X}$ as $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}$, calculate the leverage scores $h_{ii} := \|U_{i\cdot}\|^2$, where $U_{i\cdot}$ denotes the ith row of $\mathbf{U}$, and sort h_{ii}'s to have $h_{(11)} \geq \ldots \geq h_{(nn)}$.

for i in $1, \ldots, n$ **do**

  if $\#(\Gamma) \leq k$ or $\kappa\left(\mathbf{U}_\Gamma^{\mathrm{T}} \mathbf{U}_\Gamma\right) \geq T$ **then**

    Add the index of the data point corresponding to $h_{(ii)}$ to set Γ.

    Update the $\mathbf{U}_\Gamma$ as the selected rows of $\mathbf{U}$ in Γ.

  else

    break

  end if

end for

Yu and Wang (2022) mentioned that the stopping criterion for the LEVSS algorithm on the condition number is in order to ensure that the design matrix under the subdata is not ill-conditioned, that is to prevent multicollinearity. Also, they mentioned that, from

geometrical perspective, the threshold T on the condition number prevents subdata to be lied in a low-rank subspace. Moreover, in case that LEVSS algorithm selects more than k data points, say k^*, then they suggested a simple random sampling selecting k out of k^*.

Furthermore, they remarked that the stopping criterion for the LEVSS algorithm on the condition number is not crucial when the covariates are from the family of elliptically contoured distributions (Fang et al., 1990), since the space of covariates of the subdata expands to the space of covariates of the full data in a quick way.

The time complexity of LEVSS algorithm is $O(np^2)$.

4. SIMULATION EXPERIMENTS

In this section, we evaluate the performance of LEVSS algorithm based on simulated data, presenting the results of the algorithms of the approaches of IBOSS and OSS as well, in order to make a comparison.

4.1 First-order linear model

Under model (1), the observations $\mathbf{x}_i$'s, $i = 1, 2, \ldots, p$:

- Case 1. are independent and follow a multivariate uniform distribution on $[0, 1]^p$ with all covariates independent.
- Case 2. follow a multivariate normal distribution, that is, $\mathbf{x}_i \sim N(\mathbf{0}, \mathbf{\Sigma})$, with

$$\mathbf{\Sigma} = \left(0.5^{I(i,j)}\right), i, j = 1, 2, \ldots, p, \tag{3}$$

  where $I(i, j) = 0$ for $i = j = 1, 2, \ldots, p$ and $I(i, j) = 1$ for $i \neq j = 1, 2, \ldots, p$.
- Case 3. follow a truncated multivariate normal distribution on $[-5, 5]^p$, that is, $\mathbf{x}_i \sim N(\mathbf{0}, \mathbf{\Sigma})$, with covariance matrix $\mathbf{\Sigma}$ in (3).

The response data are generated from the linear model in (1) with the true value of $\boldsymbol{\beta}$ being a 51 dimensional vector with all elements equal to 1 and $\sigma^2 = 9$. We include an intercept, and so $p = 50$.

The simulation is repeated 1000 times. We calculate the mean squared error (MSE) of the subdata selected by the approaches of IBOSS, OSS and LEVSS. We estimate the intercept with the adjusted estimator $\hat{\beta}_0 = \bar{y} - \bar{\mathbf{x}}^{\mathrm{T}} \hat{\boldsymbol{\beta}}_1^{Sub}$ (Wang et al., 2019), 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, respectively, 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 investigate the cases that the full data sizes are $n = 5 \times 10^3, 10^4, 10^5$ and 10^6, and the subdata size is fixed at $k = 1000$. Figures 1, 2 and 3 show the MSEs of the

estimated slope parameters for the subdata selected by different approaches for Cases 1, 2 and 3, respectively. We also provide the mean values (⧫).

Figure 1: *The MSEs of the estimated slope parameters for the subdata selected by different approaches for the covariates of Case 1, when the full data size is* $n = 5 \times 10^3, 10^4, 10^5$ *and* 10^6 *and the subdata size is* $k = 1000$*.*

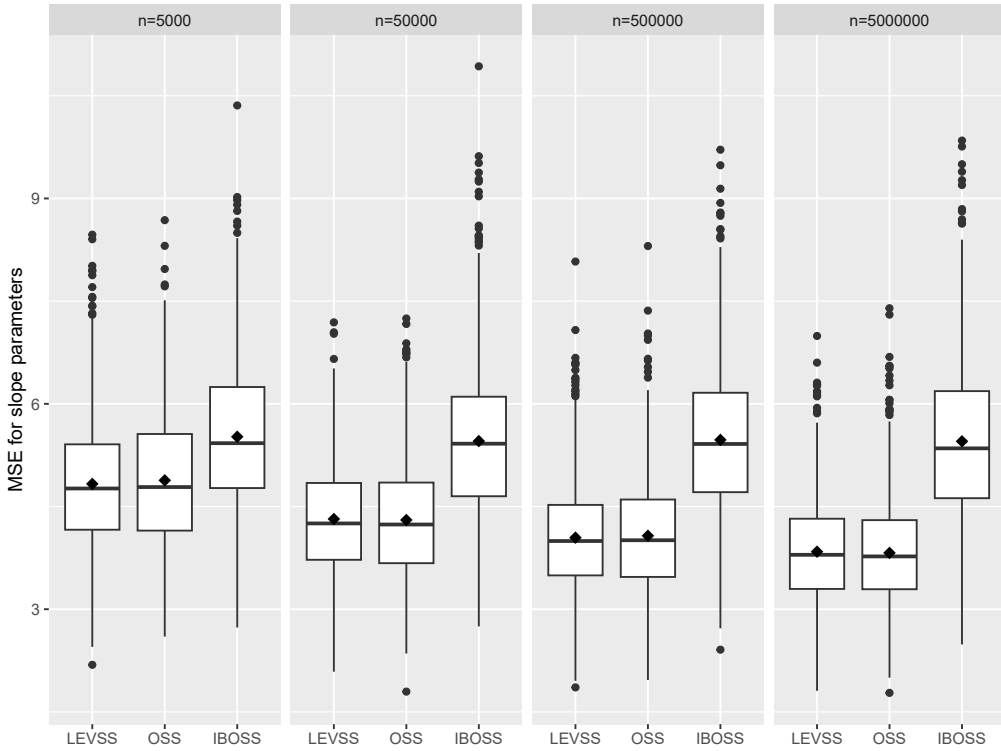


Figure 2: *The MSEs of the estimated slope parameters for the subdata selected by different approaches for the covariates of Case 2, when the full data size is* $n = 5 \times 10^3, 10^4, 10^5$ *and* 10^6 *and the subdata size is* $k = 1000$*.*

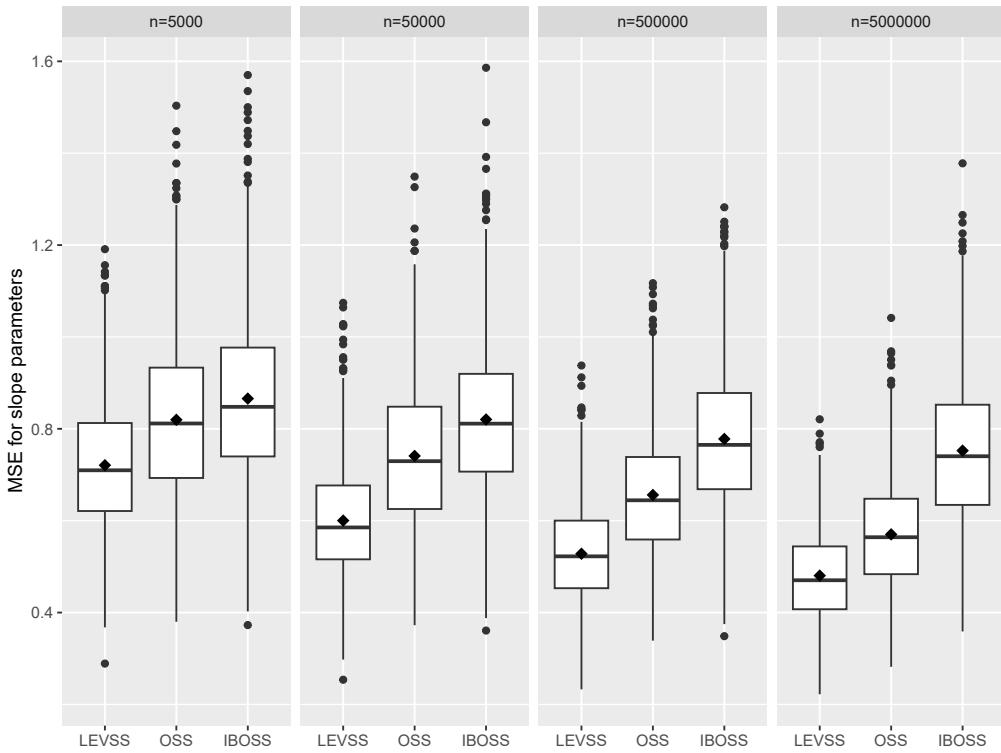
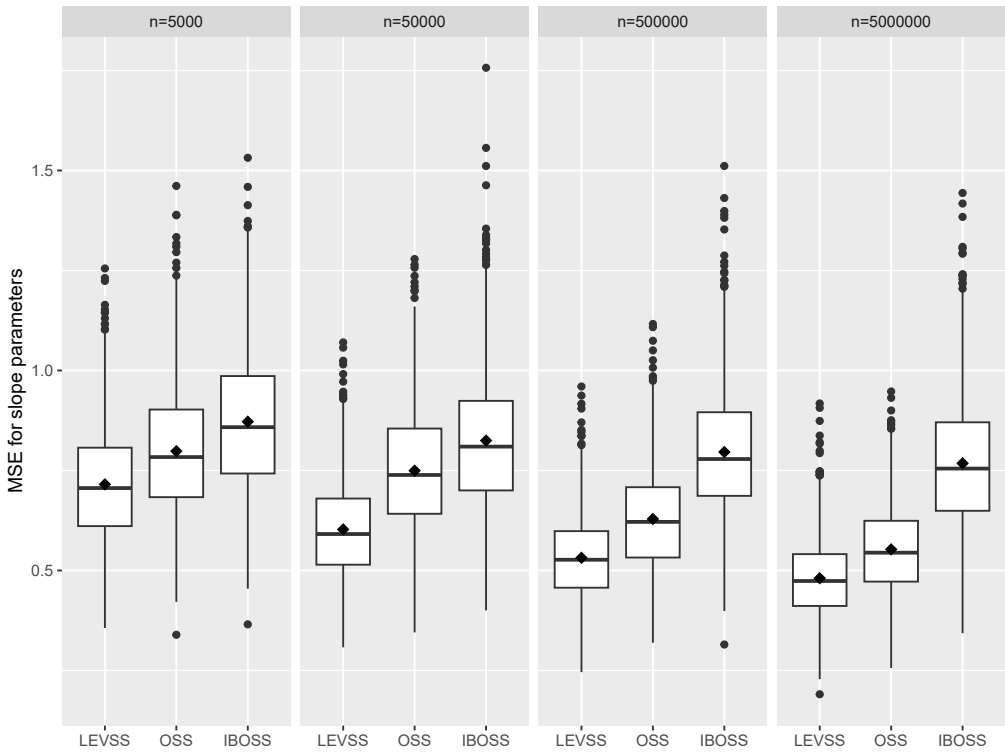


Figure 3: *The MSEs of the estimated slope parameters for the subdata selected by different approaches for the covariates of Case 3, when the full data size is* $n = 5 \times 10^3, 10^4, 10^5$ *and* 10^6 *and the subdata size is* $k = 1000$*.*



The LEVSS algorithm consistently outperforms the IBOSS and OSS ones in Cases 2 and 3. Also, in Case 1, LEVSS algorithm consistently outperforms the IBOSS one, and it is slightly better than the OSS one. This happens because the structure of the selected subdata by the LEVSS and OSS algorithm are very similar, when the observations $\mathbf{x}_i$'s, $i = 1, 2, \ldots, 50$ are independent and follow a multivariate uniform distribution with all covariates independent. Moreover, MSE of the estimated slope parameters by LEVSS approach decreases as the full data size n increases, even though the subdata size is fixed at $k = 1000$. This indicates that LEVSS approach identifies more informative data points from the full data as the full data size increases. Also, either the covariates are unbounded or not, as n increases, the MSE of the estimated slope parameters decreases fast in the LEVSS approach, as in the OSS one. Overall, LEVSS approach provide more accurate estimates for the model parameters compared with the IBOSS and OSS approaches, as one can see in Figures 1, 2 and 3. The MSE for intercept is immutable among the three approaches, and so the results are omitted for brevity. For the results from the IBOSS and OSS approaches, we refer the reader to Figure 1 in the supplementary material by Wang et al. (2021).

4.2 Computing time

We focus on the computing time of the approaches IBOSS, OSS and LEVSS for Case 1 for different full data sizes n, when the subdata size is equal to $k = 1000$ and the number of covariates is equal to $p = 50$. All computations are carried out on a PC with 3.6 GHz Intel 8-Core I7 processor and 16GB memory.

In Table 1, we present the mean computing times (in seconds) of the approaches IBOSS, OSS and LEVSS.

Table 1: *The mean execution time (in seconds) of the approaches IBOSS, OSS and LEVSS for different full data sizes* $n = 5 \times 10^3, 10^4, 10^5$ *and* 10^6*, when the subdata size is equal to* $k = 1000$ *and the number of covariates is equal to* $p = 50$*.*

n	5×10^3	5×10^4	5×10^5	5×10^6
IBOSS	0.175	0.758	6.858	73.14
OSS	3.205	6.886	18.351	150.06
LEVSS	1.205	1.812	7.911	83.16

For any full data size n, the algorithm of the LEVSS approach is faster than the one of the OSS approach. Also, noting that the difference in the computing time between the algorithms of the LEVSS and the IBOSS approach is very small.

5. REAL DATA APPLICATION

In this section, we evaluate the performance of the LEVSS approach approach on a real data example, examining the accuracy of the ordinary least squares estimates of slope parameters in model (1).

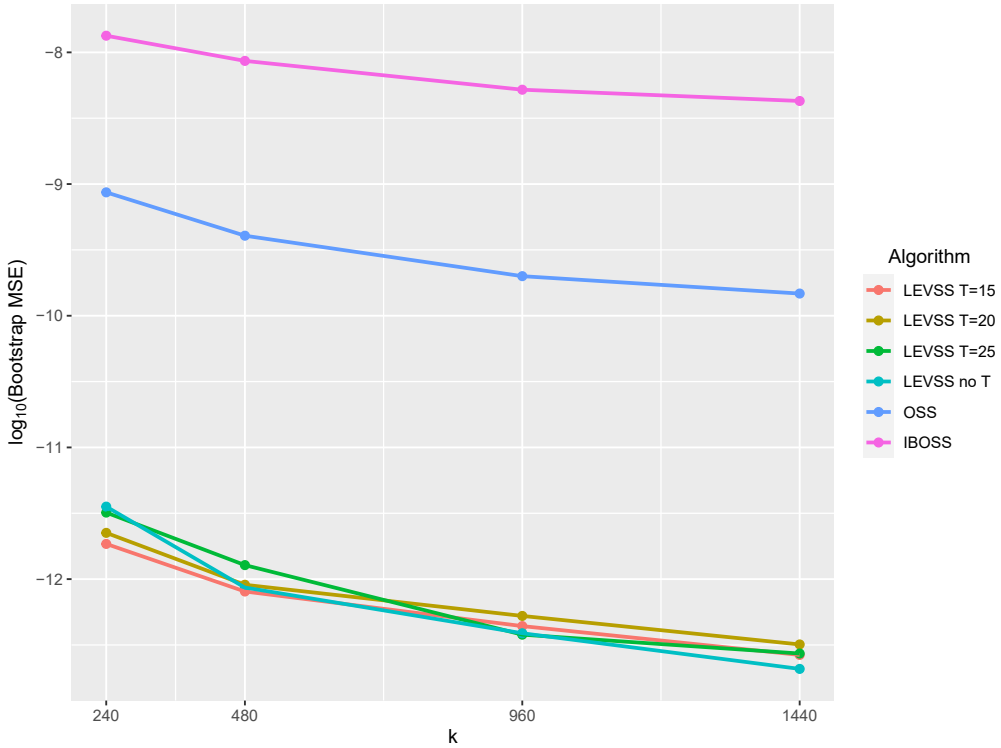
The dataset of the real data example consists of locations and absorbed power of wave energy converters in four real wave scenarios from the southern coast of Australia (Sydney, Adelaide, Perth and Tasmania). The full data consists of $n = 288,000$ data points and contains readings of 32 location variables and 16 absorbed power variables, so the number of covariates in the model is $p = 48$. The response variable is the total power output of the farm, and in the analysis we work with its log-transformation. Further information about the dataset can be found in "UCI Machine Learning Repository" (Dua and Graff, 2019).

We compare the performance of the algorithm of the LEVSS approach with the algorithms of the approaches of IBOSS and OSS, considering the MSE for the vector of slope parameters for each algorithm by using 100 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, each algorithm is implemented to obtain the subdata and then from the selected subdata the parameters of the model are estimated. The algorithm of the LEVSS approach is implemented under different values for the threshold T on the condition number, that is for

$T = 25$, 20 and 15, and when the stopping criterion on the condition number does not exist. These modifications on the algorithm of the LEVSS approach take place in the real data example, but not in the simulation experiments in Section 4., since the condition number was too small when the k have already been selected, that is the space of covariates of the subdata expanded quickly to the space of covariates of the full data.

Figure 4 shows the bootstrap MSEs by different approaches for $k = 5p, 10p, 20p$ and $30p$. Also, we take logarithm with base 10 of each MSE for a better presentation of the Figure 4. Moreover, we provide the mean values (⧫). All modifications on the algorithm of the LEVSS approach outperform the IBOSS and OSS algorithms in minimizing the bootstrap MSEs. Also, as Wang et al. (2021) stated, the IBOSS approach performs poorly in this real data example, because probably not all variables are important. On the other hand, one could say that the LEVSS approach approximates as close as possible the convex hull generated by the full data, and so performs very well. It is important to make a discussion on the modifications of the algorithm of the LEVSS approach. As the subdata size k is getting bigger, the bootstrap MSE is the smallest one for the case that the stopping criterion on the condition number is ignored. Also, consider that the LEVSS approach acts like simple random subsampling on the full data in some sense when the T is getting smaller, since then more than k data points are selected. This consideration seems to make sense as the subdata size k is getting larger, since for a small k, say $k = 240$, the bootstrap MSE is the smallest for the case that the T is the smallest one among others used. However, another value of T, which is lower that 15, will lead to a bigger bootstrap MSE for any subdata size k. Moreover, we should note that the algorithm of the LEVSS approach is faster when the value of T is getting larger. The fastest modification of the algorithm of the LEVSS approach is when we ignore the stopping criterion on the condition number.

Figure 4: *The bootstrap MSEs for the subdata selected by different approaches, when the subdata size is $k = 5p, 10p, 20p$ and $30p$. The different values for the threshold T on the condition number are equal to 25 (LEVSS $T = 25$), 20 (LEVSS $T = 20$) and 15 (LEVSS $T = 15$). Also, the algorithm of the LEVSS approach is implemented without the stopping criterion on the condition number (LEVSS no T).*



6. CONCLUDING REMARKS

We have evaluated the algorithm of the LEVSS approach for the selection of data points in an optimal way from a big dataset, in order to be able to run regression and derive the most informative coefficients as possible. Also, the algorithm of the LEVSS approach was compared with these of the IBOSS and the OSS approaches in order to show the improvement gained.

Also, the modifications of the algorithm of the LEVSS approach provide very useful information about the later. It seems that a larger value on threshold T, or even more the absence of the stopping criterion on the condition number, can lead to the selection of more informative data points. However, one should consider the level of multicollinearity caused when the algorithm of the LEVSS approach is applied under such considerations, since as Yu and Wang (2022) stated, a large value on the condition number may lead to a ill-conditioned matrix and thus cause multicollinearity.

Moreover, according to the results provided in Yu et al. (2023), some model-free

subsampling approaches (SPARTAN, SP) perform better than the LEVSS approach in some cases of the simulation experiments. Therefore, not only a further and a more comprehensive investigation but also the development of new methods on the accommodation of real data in the big data era is required.

We need to mention that we did not optimize the R used in anyway, and so further time savings could be possible.

ΠΕΡΙΛΗΨΗ

Η παλινδρόμηση μπορεί να είναι πολύ δύσκολη στην περίπτωση μεγάλων συνόλων δεδομένων, καθώς πρέπει να αντιμετωπίσουμε τεράστιους όγκους δεδομένων. Η ζήτηση υπολογιστικών πόρων για τη διαδικασία μοντελοποίησης αυξάνεται όσο αυξάνεται η κλίμακα των συνόλων δεδομένων, καθώς οι παραδοσιακές προσεγγίσεις για παλινδρόμηση περιλαμβάνουν την αντιστροφή τεράστιων πινάκων δεδομένων. Το κύριο πρόβλημα βασίζεται στο μεγάλο μέγεθος δεδομένων, και έτσι μια τυπική προσέγγιση είναι η επιλογή ενός αντιπροσωπευτικού δείγματος από τα μεγάλα δεδομένα. Στην παρούσα εργασία παρέχουμε έναν νέο αλγόριθμο που βασίζεται σε βαθμολογίες μόχλευσης που βελτιώνει τις υπάρχουσες προσεγγίσεις που επιλέγουν υποδεδομένα χρησιμοποιώντας ιδιότητες ορθογώνιων πινάκων καθώς και κριτήρια πληροφοριών. Παρέχεται επίσης ένα πείραμα προσομοίωσης και μια εφαρμογή σε πραγματικά δεδομένα.

REFERENCES

Chasiotis V. and Karlis D. (2023). Subdata selection for big data regression: an improved approach. doi.org/10.48550/arXiv.2305.00218.

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 Sarlos T. (2011). Faster least squares approximation. *Numerische mathematik*, **117**(2), 219–249.

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository. http://archive.ics.uci.edu/ml.

Fang, K.-T., Kotz, S., and Ng, K. W. (1990). *Symmetric multivariate and related distributions*. Springer.

Lee J., Schifano E. D. and Wang H. (2021). Fast optimal subsampling probability approximation for generalized linear models. doi.org/10.1016/j.ecosta.2021.02.007.

Ren M. and Zhao S.-L. (2021). Subdata selection based on orthogonal array for big data. doi.org/10.1080/03610926.2021.2012196.

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.

Yu J., Ai M. and Ye, Z. (2023). A review on design inspired subsampling for big data. doi.org/10.1007/s00362-022-01386-w.

Yu J. and Wang H. (2022). Subdata selection algorithm for linear model discrimination. *Statistical Papers*, **63**, 1883–1906.

Greek Statistical Institute
Proceedings of t he 35th Panhellenic Statistics Conference 2023), Ap. 136−150

APPLYING RENYI'S PSEUDODISTANCE FOR ROBUST MODEL SELECTION FOR INDEPENDENT NOT IDENTICALLY DISTRIBUTED OBSERVATIONS

A. Felipe, M. Jaenada, P. Miranda, L. Pardo
Dept. of Statistics and Operations Research, Complutense University of Madrid, Spain
{afelipe, mjaenada, pmiranda, lpardo}@ucm.es

ABSTRACT

We propose a new model selection criterion based on Rényi's pseudodistance (RP) for the case of independent and not identically distributed observations. This new criterion recovers as a particular case the classical Akaike's criterion. The choice of RP responds to the goal of obtaining a procedure aimed to be robust to outliers. We show that this new criterion is unbiased. Next, we develop the case of comparing a model to a restricted model, establishing the asymptotic distribution of the restricted estimator. We also derive a formula to decide if the restricted model better fits the data and obtain its asymptotic distribution. A simulation study shows that this new criterion seems to have a good behavior in the presence of outliers.

Keywords: Rényi's pseudodistance, robustness, restricted model.

1. INTRODUCTION

Consider a set of real-life observations coming from an unknown distribution to be statistically modeled. Then, it could be the case that different candidate models may be considered. Hence, a natural question arises as how to choose the model that best fits the data. Note that if the assumed model is too simple, with a reduced number of parameters, it may not capture some important patterns and relationships in the data. In contrast, if the assumed model is too complex with a large number of parameters, the estimated model parameters may overfit the observed data (including possible sample noise), then resulting in a poor performance when the model is applied to new data.

A model selection criterion is a rule used to select a statistical model among a set of candidates based on the observed data. It consists in an objective criterion function quantifying the compromise between goodness of fit and model complexity, measured via an expected dissimilarity or divergence. Consequently, the dissimilarity measure needs to be minimized to select the model with the best trade-off. In other words, model selection criteria rely on a measure of fairness between a candidate model and the true model (i.e., the probability distribution generating the data).

The Akaike information criterion (AIC) developed in Akaike (1973) and Akaike (1974) is the first model selection criterion proposed in the literature, and it is one of the most widely known and used in statistical practice as model selection criterion. The AIC estimates the expected Kullback-Leibler divergence defined in Kullback and Liebler (1951) between the true model underlying the data and a fitted candidate model, and selects as the best appropriate model the one with minimum AIC. Of course, the true model underlying the data is generally unknown and so an empirical estimate obtained from the observed data is used instead.

Following the same approach as AIC, several other model selection criteria have been proposed in the literature, as the "Bayesian information criterion" (BIC) defined in Schwarz (1978), the "Corrected Akaike information criterion" (AIC_C) developed in Hurvich and Tsai (1989), Hurvich and Tsai (1993) and Hurvich and Tsai (1995), the "Generalized Information Criterion" (GIC) defined in Konishi and Kitagawa (1996), and many ohters (see Rao and Wu (2001) and Cavanaugh and Neath (2011) for interesting surveys about model selection criteria).

Most of the previous procedures measure the fairness in terms of the Kullback-Leibler divergence. However, some other divergence measures have been explored, extending the methods with better robustness properties. For example, in Mattheou et al. (2009) it is considered the density power divergence (DPD) defined in Basu et al. (1998) to define a robust model selection criterion. Similarly, in Toma et al. (2020) it is introduced another robust criterion for model selection based on the Rényi pseudodistance (RP) defined in Jones et al. (2001).

All the previous criteria assume that the observations are independent and identically distributed. A new problem appears if the observations are independent but not identically distributed (i.n.i.d.o.). In this context, in Kurata and Hamada (2018) it is considered a criterion based on DPD, extending the theory of Mattheou et al. (2009). In this paper we introduce a new robust model selection criterion in the context of i.n.i.d.o. based on RP extending the methods of Toma et al. (2020).

The rest of the paper goes as follows. In Section 2 we introduce RP for i.n.i.d.o. and we present some theoretical results necessary for next sections. The criterion based on RP is considered in Section 3. Section 4 studies the restricted case, where some additional conditions on the parameter space are imposed. In Section 5 a simulation study illustrates the robustness of the proposed criterion and compare it with other model selection criteria. Some final conclusions are presented in Section 6.

2. Minimum Rényi's pseudodistance estimators for independent but not identically distributed observations

Let $Y_1, ..., Y_n$ be i.n.i.d.o., where each Y_i has true probability distribution function $G_i, i = 1, ..., n$, and probability density function $g_i, i = 1, ..., n$, respectively. We assume that the true density function g_i belongs to a parametric family of densities, $f_i(y, \boldsymbol{\theta}), i = 1, ..., n$, with $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^p$ being a common parameter for all the density

functions. We shall denote by $F_i(y, \boldsymbol{\theta})$ the distribution function associated to the density function $f_i(y, \boldsymbol{\theta}), i = 1, ..., n$.

The value of $\boldsymbol{\theta}$ that best fits the original distributions $g_1, ..., g_n$, would naturally minimize some kind of distance between the true and assumed densities, $(g_1(y), ..., g_n(y))$ and $(f_1(y, \boldsymbol{\theta}), ..., f_n(y, \boldsymbol{\theta}))$. Many distances can be considered, but we will use the family of RP divergence measures.

The definition of RP is given below.

Definition 01 *Consider $f(\cdot, \boldsymbol{\theta}), g(\cdot)$ two probability density functions. The* **Rényi's pseudodistance** *(RP) between f and g of tuning parameter $\alpha > 0$ is defined by*

$$R_\alpha\left(f(\cdot, \boldsymbol{\theta}), g(\cdot)\right) = \frac{1}{\alpha+1}\log\left(\int f(y, \boldsymbol{\theta})^{\alpha+1} dy\right) - \frac{1}{\alpha}\log\left(\int f(y, \boldsymbol{\theta})^{\alpha} g(y) dy\right) + \frac{1}{\alpha\left(\alpha+1\right)}\log\left(\int g(y)^{\alpha+1} dy\right). \tag{1}$$

The tuning parameter α controls the trade-off between efficiency and robustness. Thus, small values of α lead to more efficient results while less robust. On the other hand, for large values of α, the results will lead to robustness but with a loss of efficiency.

The best model parameter value approximating the underlying distribution would naturally minimize Eq. (1) in $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

At $\alpha = 0$, the corresponding **Rényi's pseudodistance** between f and g can be defined by taking continuous limits so that we obtain

$$\begin{aligned} R_0\left(f(\cdot, \boldsymbol{\theta}), g(\cdot)\right) &= \lim_{\alpha \downarrow 0} R_\alpha\left(f(y, \boldsymbol{\theta}), g(y)\right) = \int g(y) \log \frac{g(y)}{f(y, \boldsymbol{\theta})} dy \\ &= \int g(y) \log g(y) dy - \int g(y) log f(y, \boldsymbol{\theta}) dy, \end{aligned} \tag{2}$$

i.e. the Kullback-Leibler divergence measure between g and f.

Note that the last term in Eq. (1) does not depend on $\boldsymbol{\theta}$. Hence, the minimizer of the RP measure can be obtained, for $\alpha > 0$, by minimizing the surrogate function

$$\frac{1}{\alpha+1}\log\left(\int f_i(y, \boldsymbol{\theta})^{\alpha+1} dy\right) - \frac{1}{\alpha}\log\left(\int f(y, \boldsymbol{\theta})^{\alpha} g(y) dy\right).$$

The above expression can be rewritten using logarithm properties as

$$-\frac{1}{\alpha}\log\frac{\int f(y, \boldsymbol{\theta})^{\alpha} g(y) dy}{\left(\int f(y, \boldsymbol{\theta})^{\alpha+1} dy\right)^{\frac{\alpha}{\alpha+1}}},$$

and thus minimizing $R_\alpha(f(\cdot,\boldsymbol{\theta}),g(\cdot))$ in $\boldsymbol{\theta}$, for $\alpha>0$, is equivalent to minimize

$$V_\alpha^*(\boldsymbol{\theta}) = -\frac{\int f(y,\boldsymbol{\theta})^\alpha g(y)dy}{\left(\int f(y,\boldsymbol{\theta})^{\alpha+1}dy\right)^{\frac{\alpha}{\alpha+1}}}. \tag{3}$$

Similarly, for $\alpha=0$, minimizing $R_0\left(f(\cdot,\boldsymbol{\theta}),g(\cdot)\right)$ is equivalent to minimize

$$V_0^*(\boldsymbol{\theta}) = -\int g(y)logf(y,\boldsymbol{\theta})dy. \tag{4}$$

However, Expression (3) does not tend to Expression (4) when $\alpha\to 0$. In order to recover such convergence, we slightly modify Expression (3) as

$$V_\alpha(\boldsymbol{\theta}) = -\frac{\int f(y,\boldsymbol{\theta})^\alpha g(y)dy}{\alpha\left(\int f(y,\boldsymbol{\theta})^{\alpha+1}dy\right)^{\frac{\alpha}{\alpha+1}}} + \frac{1}{\alpha}, \tag{5}$$

where of course the value of $\boldsymbol{\theta}$ minimizing (3) is the same as for minimizing (5).

Now, let us denote by $V_{i,\alpha}(\boldsymbol{\theta})$ the corresponding objective functions for each pair of distributions $(f_i(y,\boldsymbol{\theta}),g_i(y)), i=1,...,n$, as given in (5). As all densities $f_i(y,\boldsymbol{\theta})$ share a common parameter, the model parameter that best approximates the different underlying densities should minimize the weighted objective function, giving equal weights to all functions $V_{i,\alpha}(\boldsymbol{\theta})$. Hence, we consider as objective function

$$H_\alpha(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n V_{i,\alpha}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n\left[-\frac{\int f_i(y,\boldsymbol{\theta})^\alpha g_i(y)dy}{\alpha\left(\int f_i(y,\boldsymbol{\theta})^{\alpha+1}dy\right)^{\frac{\alpha}{\alpha+1}}} + \frac{1}{\alpha}\right]. \tag{6}$$

Hence, the following definition arises.

Definition 02 *Consider $(g_1(y),...,g_n(y))$ and $(f_1(y,\boldsymbol{\theta}),...,f_n(y,\boldsymbol{\theta}))$, n pairs of true and assumed densities for i.n.i.d.o. random variables $Y_i, i=1,...,n$. For any $\alpha\geq 0$, the value $\boldsymbol{\theta}_{\boldsymbol{g},\alpha}$ satisfying*

$$\boldsymbol{\theta}_{\boldsymbol{g},\alpha} = \arg\min_{\boldsymbol{\theta}}\frac{1}{n}\sum_{i=1}^n\left[-\frac{\int f_i(y,\boldsymbol{\theta})^\alpha g_i(y)dy}{\alpha\left(\int f_i(y,\boldsymbol{\theta})^{\alpha+1}dy\right)^{\frac{\alpha}{\alpha+1}}} + \frac{1}{\alpha}\right] = \arg\min_{\boldsymbol{\theta}}\frac{1}{n}\sum_{i=1}^n V_{i,\alpha}(\boldsymbol{\theta}).$$

is called the **best-fitting parameter according to RP**.

For any fixed $i=1,...,n$, the true distribution g_i of the random variable Y_i is usually unknown in practice and thus $\boldsymbol{\theta}_{\boldsymbol{g},\alpha}$ should be estimated. As we only have one observation of each variable Y_i, the best way to estimate g_i is assuming that the distribution is degenerate in y_i. We will denote this degenerate distribution by $\widehat{g}_i$. Therefore, the empirical estimate of the RP divergence with $\alpha>0$, given in Eq. (1) is

$$R_\alpha\left(f_i(Y_i,\boldsymbol{\theta}),\widehat{g}_i\right) = \frac{1}{\alpha+1}\log\left(\int f_i(y,\boldsymbol{\theta})^{\alpha+1}dy\right) - \frac{1}{\alpha}\log f_i(Y_i,\boldsymbol{\theta})^\alpha + k, \tag{7}$$

and similarly the empirical estimate of the RP for $\alpha = 0$, stated in (2), yields to

$$R_0\left(f_i(Y_i,\boldsymbol{\theta}),\widehat{g}_i\right) = -\log f_i(Y_i,\boldsymbol{\theta}) + k, \tag{8}$$

where k in (7) and (8) denotes a constant that does not depend on $\boldsymbol{\theta}$. As discussed earlier, the best estimator of the model parameter $\boldsymbol{\theta}$, based on the RP divergence should minimize its empirical estimate. And again, this is equivalent to

$$\widehat{V}_{i,\alpha}\left(Y_i,\boldsymbol{\theta}\right) = -\frac{f_i(Y_i,\boldsymbol{\theta})^{\alpha}}{\alpha\left(\int f_i(y,\boldsymbol{\theta})^{\alpha+1}dy\right)^{\frac{\alpha}{\alpha+1}}} + \frac{1}{\alpha}.$$

As we have n observations, it makes sense to consider the arithmetic mean of expressions $\widehat{V}_{i,\alpha}\left(Y_i,\boldsymbol{\theta}\right)$, i.e.

$$H_{n,\alpha}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\widehat{V}_{i,\alpha}(Y_i,\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\left[-\frac{f_i(Y_i,\boldsymbol{\theta})^{\alpha}}{\alpha L_{\alpha}^{i}\left(\boldsymbol{\theta}\right)} + \frac{1}{\alpha}\right], \tag{9}$$

with

$$L_{\alpha}^{i}\left(\boldsymbol{\theta}\right) = \left(\int f_i(y,\boldsymbol{\theta})^{\alpha+1}dy\right)^{\frac{\alpha}{\alpha+1}}.$$

and correspondingly,

$$H_{n,0}(\boldsymbol{\theta}) = \lim_{\alpha\to 0} H_{n,\alpha}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\widehat{V}_{i,0}(Y_i,\boldsymbol{\theta}). \tag{10}$$

Remark at this point that the expected values of the estimates are indeed the theoretical objective functions

$$V_{i,\alpha}(\boldsymbol{\theta}) = E_{Y_i}\left[\widehat{V}_{i,\alpha}(Y_i,\boldsymbol{\theta})\right], \quad H_{\alpha}(\boldsymbol{\theta}) = E_{Y_1,\ldots,Y_n}\left[H_{n,\alpha}(\boldsymbol{\theta})\right].$$

This leads to the following definition.

Definition 03 *Given $Y_1,...,Y_n$ be i.n.i.d.o. and $\alpha > 0$, the* **minimum RP estimator (MRPE)***, $\widehat{\boldsymbol{\theta}}_{\alpha}$, is given by*

$$\widehat{\boldsymbol{\theta}}_{\alpha} = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} H_{n,\alpha}(\boldsymbol{\theta}), \tag{11}$$

with $H_{n,\alpha}(\boldsymbol{\theta})$ defined in (9) for $\alpha > 0$ and in (10) for $\alpha = 0$.

As the MRPE, $\widehat{\boldsymbol{\theta}}_{\alpha}$, is a minimum of a differentiable function, it must annul the first derivatives of the function $H_{n,\alpha}(\boldsymbol{\theta})$. This leads to a system of equations that allows to obtain $\widehat{\boldsymbol{\theta}}_{\alpha}$.

We next study the asymptotic distribution of the MRPE, $\widehat{\boldsymbol{\theta}}_\alpha$. Let us define the matrices $\boldsymbol{\Psi}_{n,\alpha}\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)$ and $\boldsymbol{\Omega}_{n,\alpha}\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)$ as follows:

$$\boldsymbol{\Psi}_{n,\alpha}\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{J}_{\alpha}^{(i)}\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right),$$

with

$$\boldsymbol{J}_{\alpha}^{(i)}\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)=\left(E_{Y_i}\left[\frac{\partial^2\widehat{V}_{i,\alpha}(Y_i;\boldsymbol{\theta})}{\partial\theta_j\partial\theta_k}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\boldsymbol{g},\alpha}}\right)_{j,k=1,...,p},i=1,...,n,$$

and

$$\boldsymbol{\Omega}_{n,\alpha}\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)=\frac{1}{n}\sum_{i=1}^{n}Var_{Y_i}\left[\left(\frac{\partial\widehat{V}_{i,\alpha}(Y_i;\boldsymbol{\theta})}{\partial\theta_j}\right)_{j=1,..,p}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\boldsymbol{g},\alpha}}.$$

Let $\lambda_1,...,\lambda_n$ be the eigenvalues of $\boldsymbol{\Omega}_{n,\alpha}\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)$. We assume that $\inf_n\lambda_n>0$, so that $\boldsymbol{\Omega}_{n,\alpha}\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)$ can be inverted.

We consider the following regularity conditions, that are the usual considered for maximum likelihood estimators:

C1. The support, $\mathcal{X}$, of the density functions $f_i(y,\boldsymbol{\theta})$ is the same for all i and it does not depend on $\boldsymbol{\theta}$. Besides, the true probability density functions $g_1,...,g_n$ have the same support $\mathcal{X}$.

C2. For almost all $y\in\mathcal{X}$ the density $f_i(y,\boldsymbol{\theta})$ admits all third derivatives with respect to $\boldsymbol{\theta}\in\boldsymbol{\Theta}$ and $i=1,...,n$.

C3. For $i=1,2,...,n$ the integrals

$$\int f_i(y,\boldsymbol{\theta})^{1+\alpha}dy$$

can be differentiated thrice with respect to $\boldsymbol{\theta}$ and we can interchange integration and differentiation. As a consequence of this condition, it follows that

$$\frac{\partial V_{i,\alpha}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}=E_{Y_i}\left[\frac{\partial\widehat{V}_{i,\alpha}(Y_i,\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right],\quad\frac{\partial^2V_{i,\alpha}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}=E_{Y_i}\left[\frac{\partial^2\widehat{V}_{i,\alpha}(Y_i,\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right]=\boldsymbol{J}_{\alpha}^{(i)}(\boldsymbol{\theta}).$$

C4. For $i=1,2,...,n$ the matrices $\boldsymbol{J}_{\alpha}^{(i)}\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)$ are positive definite.

C5. There exist functions $M_{jkl}^{(i)}$ and constants m_{jkl} such that

$$\left|\frac{\partial^3\widehat{V}_{i,\alpha}(y;\boldsymbol{\theta})}{\partial\theta_j\partial\theta_k\partial\theta_l}\right|\leq M_{jkl}^{(i)}\left(y\right),\qquad\forall\boldsymbol{\theta}\in\boldsymbol{\Theta},\ \forall j,k,l$$

and

$$E_Y\left[M_{jkl}^{(i)}(Y)\right] = m_{jkl} < \infty, \qquad \forall \boldsymbol{\theta} \in \boldsymbol{\Theta},\ \forall j,k,l.$$

C6. For all indices j,k,l and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, the sequences given by $\left\{\frac{\partial \widehat{V}_{i,\alpha}(Y_i,\boldsymbol{\theta})}{\partial \theta_j}\right\}_{j=1,...,p}$, by $\left\{\frac{\partial^2 \widehat{V}_{i,\alpha}(Y_i,\boldsymbol{\theta})}{\partial \theta_j \partial \theta_k}\right\}_{j,k=1,..,p}$ and by $\left\{\frac{\partial^3 \widehat{V}_{i,\alpha}(Y_i,\boldsymbol{\theta})}{\partial \theta_j \partial \theta_k \partial l}\right\}_{j,k,l=1,..,p}$ are uniformly integrable in the Cesàro sense, i.e.

$$\lim_{n\to\infty}\left(\sup_{n>1}\frac{1}{n}\sum_{i=1}^{n} E_{Y_i}\left[\left|\frac{\partial \widehat{V}_{i,\alpha}(Y_i,\boldsymbol{\theta})}{\partial \theta_j}\right| I_{\left\{\frac{\partial V_{i,\alpha}(Y_i,\boldsymbol{\theta})}{\partial \theta_j} > n\right\}}(Y_i)\right]\right) = 0,$$

$$\lim_{n\to\infty}\left(\sup_{n>1}\frac{1}{n}\sum_{i=1}^{n} E_{Y_i}\left[\left|\frac{\partial^2 \widehat{V}_{i,\alpha}(Y_i,\boldsymbol{\theta})}{\partial \theta_j \partial \theta_k}\right| I_{\left\{\frac{\partial^2 V_{i,\alpha}(Y_i,\boldsymbol{\theta})}{\partial \theta_j \partial \theta_k} > n\right\}}(Y_i)\right]\right) = 0,$$

$$\lim_{n\to\infty}\left(\sup_{n>1}\frac{1}{n}\sum_{i=1}^{n} E_{Y_i}\left[\left|\frac{\partial^3 \widehat{V}_{i,\alpha}(Y_i,\boldsymbol{\theta})}{\partial \theta_j \partial \theta_k \partial \theta_l}\right| I_{\left\{\frac{\partial^3 V_{i,\alpha}(Y_i,\boldsymbol{\theta})}{\partial \theta_j \partial \theta_k \partial \theta_l} > n\right\}}(Y_i)\right]\right) = 0.$$

C7. For all $\varepsilon > 0$

$$\lim_{n\to\infty}\left\{\frac{1}{n}\sum_{i=1}^{n} E_{Y_i}\left[\left\|\boldsymbol{\Omega}_n^{-\frac{1}{2}}(\boldsymbol{\theta})\frac{\partial \widehat{V}_{i,\alpha}(Y_i,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right\|_2^2 I_{\left\{\left\|\boldsymbol{\Omega}_n^{-\frac{1}{2}}(\boldsymbol{\theta})\frac{\partial \widehat{V}_{i,\alpha}(Y_i,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right\|_2^2 > \varepsilon\sqrt{n}\right\}}(Y_i)\right]\right\} = 0.$$

Now, the following result, whose proof can be seen in Castilla et al. (2022), holds.

Theorem 01 *Suppose the previous regularity conditions* **C1- C7** *hold. Then,*

$$\sqrt{n}\boldsymbol{\Omega}_{n,\alpha}(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})^{-\frac{1}{2}}\boldsymbol{\Psi}_{n,\alpha}(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})\left(\widehat{\boldsymbol{\theta}}_\alpha - \boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right) \underset{n\to\infty}{\overset{L}{\to}} N(\mathbf{0}_p, \boldsymbol{I}_p),$$

being $\boldsymbol{I}_p$ the p-dimensional identity matrix.

3. Model selection criterion based on RP

Let us consider a collection of l candidate models

$$\left\{\boldsymbol{M}^{(s)} = \left(M_1^{(s)}, ..., M_n^{(s)}\right)\right\}_{s\in\{1,...,l\}}$$

such that each $\boldsymbol{M}^{(s)}$ is characterized by the parametric density functions

$$\boldsymbol{f}(\cdot,\boldsymbol{\theta}_s) = (f_1(\cdot,\boldsymbol{\theta}_s), ..., f_n(\cdot,\boldsymbol{\theta}_s)),\ \boldsymbol{\theta}_s \in \boldsymbol{\Theta}_s \subset \mathbb{R}^{p_s},$$

with associated distribution functions $\boldsymbol{F}(.,\boldsymbol{\theta}_s) = (F_1(\boldsymbol{\theta}_s), ..., F_n(.,\boldsymbol{\theta}_s))$, where we assume that $\boldsymbol{\theta}_s$ is common for density functions in model s. That is, each candidate model would represent a parametric family defined by a common parameter, which may contain different number of parameters for different models. We aim to select the

best model from the collection $\{\boldsymbol{M}^{(s)}\}_{s\in\{1,...,l\}}$ according to some suitable selection criterion. Thus, for each model $\boldsymbol{M}^{(s)}$, we first determine the best parameter $\boldsymbol{\theta}_s$ fitting the sample and next we select the best fitted model from the collection. Then, for a set of observations, the model selection is performed in two steps: we first fit all the candidate models to the data, and then select the model with best trade-off between goodness of fit and complexity in terms of RP.

We next describe the first step of the model selection algorithm. Let us consider a fixed parametric model $\boldsymbol{M}^{(s)}$ modeling the true distribution underlying the data. If the true distribution of data were known, the parameter that best fits the model $\boldsymbol{M}^{(s)}$, denoted by $\boldsymbol{\theta}_{\boldsymbol{g}}^{s}$, can be obtained by maximizing the theoretical averaged objective function $H_\alpha(\boldsymbol{\theta})$ defined in Eq. (6) under the s-model, as proposed in Def. 02.

Following the discussion in the previous section, if the true distribution underlying is unknown but we have a random sample $Y_1, ..., Y_n$, the best estimate of the true parameter based on the sample from the RP approach is the MRPE $\widehat{\boldsymbol{\theta}}_{\alpha}^{s}$ defined in (11), as proposed in Def. 03.

Let us turn to the second step. Once all candidate models are fitted to the observed data, we should select the model with the best trade-off between fitness and complexity. Therefore, we need a measure of fairness between the best candidate for each model and the true distribution. The goodness of fit of a certain model $\boldsymbol{M}^{(s)}$ with associated densities $\boldsymbol{f}(\cdot, \boldsymbol{\theta}_g^s)$ and the best-fitting parameter $\boldsymbol{\theta}_g^s$ based on the RP can be quantified by the averaged objective function $H_\alpha(\boldsymbol{\theta}_g^s)$ given in Eq. (6).

As the true distribution is generally unknown, $\boldsymbol{\theta}_g^s$ is estimated by $\widehat{\boldsymbol{\theta}}_{\alpha}^{s}$. Hence, we can estimate $H_\alpha(\boldsymbol{\theta}_g^s)$ by $H_\alpha(\widehat{\boldsymbol{\theta}}_{\alpha}^{s})$. But again $H_\alpha(\widehat{\boldsymbol{\theta}}_{\alpha}^{s})$ needs to be estimated, and the natural estimator is $H_{n,\alpha}(\widehat{\boldsymbol{\theta}}_{\alpha}^{s})$ given in Eq. (9). But proceeding this way, it comes up that the sample is used both for estimating the parameter and for estimating H_α, and hence

$$E_{Y_1,...,Y_n}\left[H_{n,\alpha}\left(\widehat{\boldsymbol{\theta}}_{\alpha}^{s}\right)\right] \neq E_{Y_1,...,Y_n}\left[H_{\alpha}\left(\widehat{\boldsymbol{\theta}}_{\alpha}^{s}\right)\right].$$

Even worse, the estimation bias is important for determining the best model because it could depend on the model. Hence, we neeed to add a term correcting such bias.

The AIC criterion selects the model that minimizes

$$-2\sum_{i=1}^{n}\log f_i(y_i, \boldsymbol{\theta}) + 2p = 2H_{n,0}\left(\boldsymbol{\theta}\right) + 2p,$$

where $2p$ is the term correcting the bias. Following the same idea, we define the $RP_{NH}-$criterion as follows:

Definition 04 *Let* $\left\{\left(M_1^{(s)}, ..., M_n^{(s)}\right)\right\}_{s\in\{1,...,l\}}$ *be* l *candidate models for the i.n.i.d.o.* $Y_1, ..., Y_n$. *We define the* $RP_{NH}-$**Criterion** *as the criterion that selects the model satisfying*

$$(M_1^*, ..., M_n^*) = \min_{s\in\{1,...,l\}} RP_{NH}\left(M_1^{(s)}, ..., M_n^{(s)}, \widehat{\boldsymbol{\theta}}_\alpha^s\right),$$

where

$$RP_{NH}\left(M_1^{(s)}, ..., M_n^{(s)}, \widehat{\boldsymbol{\theta}}_\alpha^s\right) = H_{n,\alpha}\left(\widehat{\boldsymbol{\theta}}_\alpha^s\right) + \frac{1}{n} trace\left(\boldsymbol{\Omega_n}\left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{\alpha}}^{\boldsymbol{s}}\right)\boldsymbol{\Psi}_n^{-1}\left(\widehat{\boldsymbol{\theta}}_\alpha^s\right)\right).$$

We can observe that

$$\lim_{\alpha\to 0} RP_{NH}\left(M_1^{(s)}, ..., M_n^{(s)}, \widehat{\boldsymbol{\theta}}_\alpha^s\right) = -\frac{1}{n}\sum_{i=1}^{n}\log f_i(Y_i, \boldsymbol{\theta}) + \frac{p}{n},$$

and hence we recover AIC criterion up to the multiplicative constant $2n$.

In order to justify the $RP_{NH}-$Criterion in terms of the bias, the following can be shown. For this purpose, we shall assume the following additional regularity condition:

C8. The matrices $\boldsymbol{\Psi}_n^{-1}(\boldsymbol{\theta})$ and $\boldsymbol{\Omega}_n(\boldsymbol{\theta})$ are continuous for arbitrary $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

Theorem 02 [1] *Suppose the regularity conditions* **C1- C8** *hold. Then,*

$$E_{Y_1,...,Y_n}\left[RP_{NH}\left(M_1^{(s)}, ..., M_n^{(s)}, \widehat{\boldsymbol{\theta}}_\alpha^s\right)\right] = E_{Y_1,...,Y_n}\left[H_\alpha\left(\widehat{\boldsymbol{\theta}}_\alpha^s\right)\right], \forall s = 1, ..., l.$$

This means that there is no bias in this procedure for selecting the model. This is important because the bias might depend on the model and hence, a model that is not approppiate could be selected due to an unbiased estimation.

4. The restricted model

Let us study a particular case of the model selection problem. In some situations, it is interesting to compare a full model based on $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^p$, having p parameters with other restricted models where the parameter has to satisfy additionally linear constraints of the form

$$\{\boldsymbol{\theta} \in \boldsymbol{\Theta} / \ \boldsymbol{m}(\boldsymbol{\theta}) = \mathbf{0}_r\},$$

where $\mathbf{0}_r$ denotes the null vector of dimension r with $r < p$ and $\boldsymbol{m} : \mathbb{R}^p \to \mathbb{R}^r$ is a vector-valued function such that the $p \times r$ matrix

$$\mathbf{M}(\boldsymbol{\theta}) = \frac{\partial \boldsymbol{m}^T(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

[1] The detailed proofs of all results in the paper can be found in A. Felipe, M. Jaenada, P. Miranda, L. Pardo. Model Selection for independent not identically distributed observations based on Rényi's pseudodistances. To appear in Journal of Computational and Applied Mathematics.

exists and is continuous in $\boldsymbol{\theta}$, and $\text{rank}(\mathbf{M}(\boldsymbol{\theta})) = r, \forall \boldsymbol{\theta} \in \boldsymbol{\Theta}$.

We have already given a definition of the best fitting parameter for the full model based on RP in the previous section. Applying the same criterion for the restricted model, we obtain that the best-fitting parameter for the restricted model is given by

$$\boldsymbol{\theta}_{\boldsymbol{g},\alpha}^{R} = \arg \min_{\boldsymbol{\theta} \in \Theta / \boldsymbol{m}(\boldsymbol{\theta}) = \mathbf{0}_r} H_\alpha(\boldsymbol{\theta}).$$

Following similar arguments than in Section 2, we define the restricted MRPE as follows.

Definition 05 *Given $Y_1, ..., Y_n$ be i.n.i.d.o., the* **restricted MRPE** *(RMRPE), $\widetilde{\boldsymbol{\theta}}_\alpha$, is given by*

$$\widetilde{\boldsymbol{\theta}}_\alpha = \arg \min_{\boldsymbol{\theta} \in \Theta / \boldsymbol{m}(\boldsymbol{\theta}) = \mathbf{0}_r} H_{n,\alpha}(\boldsymbol{\theta}),$$

with $H_{n,\alpha}(\boldsymbol{\theta})$ defined in (9) for $\alpha > 0$ and in (10) for $\alpha = 0$.

Note that as the full model is more general, it follows that

$$H_{n,\alpha}(\widehat{\boldsymbol{\theta}}_\alpha) \leq H_{n,\alpha}(\widetilde{\boldsymbol{\theta}}_\alpha).$$

We next present a representation of the RMPRE.

Theorem 03 *Suppose the regularity conditions* **C1- C8** *hold and assume that the best fitting parameter, $\boldsymbol{\theta}_{\boldsymbol{g},\alpha}$, belongs to the restricted model. Then,*

$$n^{1/2}(\widetilde{\boldsymbol{\theta}}_\alpha - \boldsymbol{\theta}_{\boldsymbol{g},\alpha}) = \boldsymbol{P}^*(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}) n^{1/2} \left(\frac{\partial H_{n,\alpha}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)_{\boldsymbol{\theta} = \boldsymbol{\theta}_{\boldsymbol{g},\alpha}} + o_p(1),$$

being

$$\boldsymbol{P}^*(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}) = \boldsymbol{Q}_\alpha(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}) \boldsymbol{M}(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})^T \boldsymbol{\Psi}_n(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})^{-1} - \boldsymbol{\Psi}_n(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})^{-1},$$

with

$$\boldsymbol{Q}_\alpha(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}) = \boldsymbol{\Psi}_n(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})^{-1} \boldsymbol{M}(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}) \left[\boldsymbol{M}(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})^T \boldsymbol{\Psi}_n(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})^{-1} \boldsymbol{M}(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}) \right]^{-1}.$$

Suppose now that we have chosen a model as the best fitting model and we wonder if this model overfits the data and a restricted model is more accurate. Then, we can pose this problem as a model selection problem with two models, the full one and a restricted model, and apply the results of the previous section. Hence, it suffices to compute $RP_{NH}((M_1^{(s)}, ..., M_n^{(s)}, \boldsymbol{\theta})$ for both models and select the one attaining the minimum. For this, the following result provides the asymptotic distribution of a test statistic linked to this problem.

Theorem 04 *Suppose the regularity conditions* **C1- C8** *hold and assume that the best fitting parameter, $\boldsymbol{\theta}_{\boldsymbol{g},\alpha}$, belongs to the restricted model. Then, it follows that*

$$2n\left(RP_{NH}\left((M_1^{(s)},...,M_n^{(s)},\widehat{\boldsymbol{\theta}}_\alpha\right)-RP_{NH}\left((M_1^{(s)},...,M_n^{(s)},\widetilde{\boldsymbol{\theta}}_\alpha\right)\right)$$

coincides with the distribution of the random variable

$$\sum_{j=1}^{r}\lambda_j(\boldsymbol{\theta}_{g,\alpha})\left(\boldsymbol{\theta}\right)Z_j^2+2trace(\boldsymbol{\Omega}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)\boldsymbol{\Psi}_n^{-1}\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right))-2trace(\boldsymbol{\Omega}_n^R\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)\left(\boldsymbol{\Psi}_n^R\right)^{-1}\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)),$$

where $Z_1,\ldots,Z_k$ are independent standard normal variables, $\lambda_1(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}),\ldots,\lambda_r(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})$ are the nonzero eigenvalues of $-\boldsymbol{Q}_\alpha(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})\boldsymbol{M}(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})^T\boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)^{-1}\boldsymbol{\Omega}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)$ and

$$r=rank\left(\boldsymbol{\Omega}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)\boldsymbol{Q}_\alpha(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})\boldsymbol{M}(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})^T\boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)^{-1}\boldsymbol{\Omega}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)\right).$$

5. Simulation Study

To evaluate the performance of the RP_{NH}-criterion introduced in the previous sections, we consider the situation of a polynomial regression model. We consider $n=100$ and take the model

$$Y_i=X_i+2X_i^2-X_i^3+X_i^4+\epsilon_i, i=1,...,100,$$

where $\epsilon_i\sim\mathcal{N}(0,1)$ and the variables X_i are fixed and given by

$$X_i=-2+\frac{4}{101}(i), i=1,...,100.$$

We consider several theoretical models aiming to fit this data depending on the degree p of the polynomial. We take 1000 different sample data $(Y^s,X^s), s=1,...,1000$ and for each sample, we select the best fitting model according to several criteria. We have considered AIC, BIC, AIC_c and the RP_{NH}-criterion for different values of the tuning parameter, namely $\alpha=0.01,0.02,0.04,0.1,0.2,0.4,0.5,0.7$ and 1.

In Table 1 we present the number of times a model is selected for each model selection criterion. From these results, it can be seen that BIC seems to be the best fitting selection criterion.

Table 1: *Number of times each model is selected for uncontaminated data.*

p	0	1	2	3	4	5
AIC	0	0	0	0	836	164
BIC	0	0	0	0	967	33
AIC_c	0	0	0	0	864	136
$RPNH_{0.01}$	0	0	0	0	822	178
$RPNH_{0.02}$	0	0	0	0	822	178
$RPNH_{0.04}$	0	0	0	0	826	174
$RPNH_{0.1}$	0	0	0	0	822	178
$RPNH_{0.2}$	0	0	0	0	834	166
$RPNH_{0.4}$	0	0	0	0	842	158
$RPNH_{0.5}$	0	0	0	0	841	159
$RPNH_{0.7}$	0	0	0	0	838	162
$RPNH_{1.0}$	0	0	0	0	837	163

As it was explained in the motivation of the paper, we expect RP_{NH} to be a robust selection criterion. To check this, we have considered the previous model but we introduce contamination in some of the data. More concretely, we define

$$\epsilon_i \sim \mathcal{U}(\min_i(X_i + 2X_i^2 - X_i^3 + X_i^4) - r, \max_i(X_i + 2X_i^2 - X_i^3 + X_i^4) + r),$$

for some of the data chosen at random. Here, r is a constant measuring the strength of contamination, in the sense that the bigger r, the strongest the contamination. We have considered three values $r = 1, 5, 10$. Moreover, we have varied the proportion of data affected by contamination. In this study, we have chosen the proportion of contamination as $0.05, 0.10, 0.20, 0.30$.

Again, we have obtained the best fitting model according different model selection criteria, and we have conducted this experiment 1000 times. The number of times that each model is selected for each combination of contamination and strength of contamination r is given in Tables 2, 3, 4 and 5.

From the results of Tables 1-5, it can be seen that the performance of AIC, BIC and AIC_c dramatically decreases when we include contamination. As expected, the bigger the rate of contaminated data, the poorer the performance. Note however that they are not very affected for different values of r.

On the other hand, the results are quite similar to the uncontaminated case if we consider RP_{NH} and big values of the tuning parameter. This was the expected result and it follows the same behavior as other situations where RP has been considered. The best behavior appears for $\alpha = 0.4$ and $\alpha = 0.5$, where the efficiency is good and the performance in terms of robustness is very good. Hence, we conclude that the RP_{NH}-criterion seems to be a robust criterion.

Table 2: *Number of times each model is selected for a contamination degree of* 5%

	$r=1$						$r=5$						$r=10$					
p	0	1	2	3	4	5	0	1	2	3	4	5	0	1	2	3	4	5
AIC	0	0	1	19	659	321	0	0	10	27	622	341	0	0	9	38	616	337
BIC	0	0	16	64	802	118	0	0	39	84	751	126	0	0	57	96	715	132
AIC_c	0	0	1	22	694	283	0	0	12	33	651	304	0	0	12	43	634	311
$RPNH_{0.01}$	0	0	3	14	844	139	0	0	5	18	830	147	0	0	9	22	812	157
$RPNH_{0.02}$	0	0	0	13	866	121	0	0	2	47	833	118	0	0	6	62	810	122
$RPNH_{0.04}$	0	0	2	20	844	134	0	0	1	18	833	148	0	0	1	18	833	148
$RPNH_{0.1}$	0	0	0	0	835	165	0	0	0	0	836	164	0	0	0	0	830	170
$RPNH_{0.2}$	0	0	0	0	829	171	0	0	0	0	833	167	0	0	0	0	833	167
$RPNH_{0.4}$	0	0	0	0	837	163	0	0	0	0	835	165	0	0	0	0	839	161
$RPNH_{0.5}$	0	0	0	0	837	163	0	0	0	0	848	152	0	0	0	0	834	166
$RPNH_{0.7}$	0	0	0	0	842	158	0	0	0	0	836	164	0	0	0	0	831	169
$RPNH_{1.0}$	0	0	0	0	838	162	0	0	0	0	836	164	0	0	0	0	830	170

Table 3: *Number of times each model is selected for a contamination degree of* 10%.

	$r=1$						$r=5$						$r=10$					
p	0	1	2	3	4	5	0	1	2	3	4	5	0	1	2	3	4	5
AIC	0	0	17	64	591	328	0	0	18	82	575	325	0	0	47	106	538	309
BIC	0	0	94	155	629	122	0	0	95	180	611	114	0	0	153	178	558	111
AIC_c	0	0	21	75	621	283	0	0	24	93	601	282	0	0	55	123	556	266
$RPNH_{0.01}$	0	0	24	37	770	169	0	0	26	48	750	176	0	0	37	61	705	197
$RPNH_{0.02}$	0	0	19	30	800	151	0	0	24	40	780	156	0	0	30	60	747	163
$RPNH_{0.04}$	0	0	16	60	809	115	0	0	19	100	764	117	0	0	23	94	770	113
$RPNH_{0.1}$	0	0	0	5	845	150	0	0	0	1	839	160	0	0	0	1	851	148
$RPNH_{0.2}$	0	0	0	0	829	171	0	0	0	0	835	165	0	0	0	0	844	156
$RPNH_{0.4}$	0	0	0	0	829	171	0	0	0	0	840	160	0	0	0	0	850	150
$RPNH_{0.5}$	0	0	0	0	824	176	0	0	0	0	845	155	0	0	0	0	841	159
$RPNH_{0.7}$	0	0	0	0	831	169	0	0	0	0	833	167	0	0	0	0	834	166
$RPNH_{1.0}$	0	0	0	0	841	159	0	0	0	0	835	165	0	0	0	0	833	167

Table 4: *Number of times each model is selected for a contamination degree of* 20%.

	$r=1$						$r=5$						$r=10$					
p	0	1	2	3	4	5	0	1	2	3	4	5	0	1	2	3	4	5
AIC	0	0	52	134	509	305	0	0	76	176	464	284	0	0	148	169	397	286
BIC	0	0	238	210	440	112	0	0	278	234	398	90	0	0	367	223	325	85
AIC_c	0	0	64	154	512	270	0	0	91	191	476	242	0	0	179	180	400	241
$RPNH_{0.01}$	0	0	41	92	596	271	0	0	43	93	561	303	0	0	52	95	514	339
$RPNH_{0.02}$	0	0	37	85	625	253	0	0	39	92	589	280	0	0	47	90	561	302
$RPNH_{0.04}$	0	0	29	75	676	220	0	0	32	88	661	219	0	0	43	93	648	216
$RPNH_{0.1}$	0	0	42	214	631	113	0	0	41	138	693	128	0	0	20	64	810	106
$RPNH_{0.2}$	0	0	0	0	836	164	0	0	0	0	831	169	0	0	0	0	849	151
$RPNH_{0.4}$	0	0	0	0	837	163	0	0	0	0	833	167	0	0	0	0	840	160
$RPNH_{0.5}$	0	0	0	0	836	164	0	0	0	0	829	171	0	0	0	0	840	160
$RPNH_{0.7}$	0	0	0	0	845	155	0	0	0	0	827	173	0	0	0	0	849	151
$RPNH_{1.0}$	0	0	0	0	834	166	0	0	0	0	823	177	0	0	0	0	836	164

Table 5: *Number of times each model is selected for a contamination degree of* 30%.

	$r=1$						$r=5$						$r=10$					
p	0	1	2	3	4	5	0	1	2	3	4	5	0	1	2	3	4	5
AIC	0	0	112	178	433	277	0	0	114	209	373	304	0	0	192	183	374	251
BIC	0	0	327	256	327	90	0	0	385	240	276	99	0	0	457	212	253	78
AIC_c	0	0	136	191	436	237	0	0	137	233	368	262	0	0	219	189	371	221
$RPNH_{0.01}$	0	0	51	79	519	351	0	0	55	90	488	367	0	0	58	90	428	424
$RPNH_{0.02}$	0	0	46	77	540	337	0	0	48	84	520	348	0	0	52	87	472	389
$RPNH_{0.04}$	0	0	44	81	573	302	0	0	46	80	555	319	0	0	53	78	533	336
$RPNH_{0.1}$	0	0	70	187	550	193	0	0	63	221	537	179	0	0	55	139	628	178
$RPNH_{0.2}$	0	0	17	68	774	141	0	0	11	13	817	159	0	0	2	8	854	136
$RPNH_{0.4}$	0	0	0	0	856	144	0	0	0	0	833	167	0	0	0	0	841	159
$RPNH_{0.5}$	0	0	0	0	845	155	0	0	0	0	830	170	0	0	0	0	832	168
$RPNH_{0.7}$	0	0	0	0	834	166	0	0	0	0	815	185	0	0	0	0	826	174
$RPNH_{1.0}$	0	0	0	0	828	172	0	0	1	1	813	185	0	0	0	0	841	159

6. Conclusions

In this paper we present a new procedure for model selection for independent but not identically distributed observations aiming to compete with other methods based on maximum likelihood in terms of efficiency but being more robust in the presence of outliers. For this purpose, we have chosen RP, a tool that has proved itself to lead to robust estimations in other statistical problems. Based on RP, we have developed a model selection criterion, the RP_{NH}-criterion, that extends the well-known AIC. Besides, we have shown that the sample estimator is an unbiased estimator. As a special case, we have studied the case of having a restricted model and we have developed a procedure to decide if the restricted model is more appropriate for modeling the data. A simulation study seems to show that, as expected, this new procedure works very well under contamination, i.e. simulations suggest that the procedure is robust. Besides, it seems that the cost in terms of efficiency is reduced for several values of the tuning parameter α that are also robust.

Acknowledgements

This work was supported by the Spanish Grant PID2021-124933NB-I00.

ΠΕΡΙΛΗΨΗ

Στην παρούσα εργασία προτείνουμε ένα νέο κριτήριο επιλογής μοντέλου με βάση την Rényi's pseudodistance (RP) για την περίπτωση ανεξάρτητων αλλά όχι ισόνομα κατανεμημένων παρατηρήσεων. Το νέο κριτήριο περιλαμβάνει ως ειδική περίπτωση το κλασσικό κριτήριο του Akaike. Η επιλογή του RP ανταποκρίνεται στον στόχο της δημιουργίας μιας διαδικασίας ανθεκτικής σε ακραίες τιμές. Για το νέο κριτήριο αποδεικνύεται ότι είναι αμερόληπτο ενώ εξετάζεται η περίπτωση σύγκρισης ενός μοντέλου με ένα περιορισμένο μοντέλο, καθορίζοντας την ασυμπτωτική κατανομή του περιορισμένου εκτιμητή. Προτείνουμε επίσης μια στατιστική συνάρτηση για να

αποφασίσουμε εάν το περιορισμένο μοντέλο ταιριάζει καλύτερα στα δεδομένα και αποδεικνύουμε την ασυμπτωτική κατανομή της. Μια μελέτη προσομοίωσης δείχνει ότι το προτεινόμενο κριτήριο έχει καλή συμπεριφορά παρουσία ακραίων τιμών.

REFERENCES

Akaike, H. (1973). *Information theory and an extension of the maximum likelihood principle*. *Akadémia Kiadó*, Budapest, Hungary.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716–723.

Basu, A., Harris, I.R., Hjort, N.L. and Jones, M.C. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika*, **85**(3), 549–559.

Castilla, E., Jaenada, M. and Pardo, L. (2022). Estimation and testing on independent not identically distributed observations based on Rényi's pseudodistances. *IEEE Transactions on Information Theory*, **68**(7), 4588–4609.

Cavanaugh, J.E. and Neath, A.A. (2011). Akaike's information criterion: Background, derivation, properties, and refinements. In *International Encyclopedia of Statistical Science*, pages 26–29.

Hurvich, C.M. and Tsai, C.L. (1989). Regression and time series model selection in small samples. *Biometrika*, **76**, 297–307.

Hurvich, C.M. and Tsai, C.L. (1993). A corrected Akaike information criterion for vector autoregressive model selection. *J. of Time Series Analysis*, **14**, 271–279.

Hurvich, C.M. and Tsai, C.L. (1995). Model selection for extended quasi–likelihood models in small samples. *Biometrics*, **51**, 1077–1084.

Jones, M.C., Hjort, N.L., Harris, I.R. and Basu, A. (2001). A comparison of related density-based minimum divergence estimators. *Biometrika*, **88**, 865–873.

Konishi, S. and Kitagawa, G. (1996). Generalised information criteria in model selection. *Biometrika*, **83**, 875–890.

Kurata, S. and Hamada, E. (2018). A robust generalization and asymptotic properties of the model selection criterion family. *Communication in Statistics (Theory and Methods)*, **47**(3), 532–547.

Kullback, S. and Leibler, R.A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, **22**, 79–86.

Mattheou, K., Lee, S. and Karagrigoriou, A. (2009). A model selection criterion based on the BHHJ measure of divergence. *Journal of Statististical Planning and Inference*, **139**, 228–235.

Rao, C.R. and Wu, Y. (2001). On model selection. In *IMS Lecture Notes–Monograph Series*, volume 31, pages 1–57.

Schwarz, G.E. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**(2), 461–464.

Toma, A., Karagrigoriou, A. and Trentou, P. (2020). Robust model selection criteria based on pseudodistances. *Entropy*, 22(3):304.

Greek Statistical Institute
Proc. of the 35th Panhellenic & 1st International Statistics Conf. (2023), pp.151-164

A NEW ADAPTIVE RUN SUM MAX CHART

K. G. Fountoukidis, D. L. Antzoulakos, and A. C. Rakitzis
University of Piraeus, Department of Statistics and Insurance Science, 18534, Piraeus, Greece
k.fountoukidis@unipi.gr, dantz@unipi.gr, arakitz@unipi.gr

ABSTRACT

A variable sample size and sampling interval run sum Max chart (VSSI-RSMax) is proposed to efficiently monitor the mean and/or variability of a process. The VSSI-RSMax chart varies the sample size and sampling interval of the RSMax chart according to the current cumulative score. A Markov chain method is used to evaluate the performance of the proposed chart in terms of the average time to signal (ATS), the adjusted average time to signal (AATS), and the expected AATS (EAATS). The VSSI-RSMax chart is compared with other competitive single control charts, such as the standard fixed sample size and sampling interval (FSSI) Max chart, the VSSI-Max chart, the FSSI-RSMax chart and the FSSI Max-EWMA chart.

Keywords: Average Time to Signal, Average Sample Size, Average Sampling Interval, Markov Chains, Single Control Charts, Statistical Process Monitoring.

1. INTRODUCTION

Control chart is the most widely used technique for monitoring a process and detect changes in it. There are several types of control charts and among them, the most popular ones are the Shewhart, EWMA and CUSUM charts. We refer to [Montgomery (2020)] for a gentle introduction on the fundamentals, the properties, and the use of control charts. Shewhart-type charts are mainly used for detecting sudden and sustained shifts in process parameters while EWMA- and CUSUM-type charts are better than Shewhart charts in the detection of transient shifts. In addition, Shewhart charts are control charts with no memory and they are better in the detection of shifts of large size in process parameters. On the other hand, for small or medium sized shifts, control charts with memory such as the EWMA and CUSUM charts, are preferable than Shewhart charts.

However, EWMA and CUSUM charts are not widely used in practice due to the difficulty in choosing properly the values for their design parameters. Therefore, there are plenty of charts that can be viewed as intermediate solutions between the Shewhart charts and the EWMA and CUSUM charts. These charts aim to combine the simplicity in implementation and interpretation of the usual Shewhart chart with an improved performance in the detection of small and moderate shifts in process parameters.

One of these intermediate solutions is the run sum chart ([Roberts (1996)], [Reynolds (1971)]) which has been applied successfully in various industries and, according to [Jaehn (1991)], practitioners find it appealing, due to its ease in implementation and interpretation. A run sum chart is set up by dividing the zone above and the zone below the chart's center line into separate (distinct) regions. Next, in each of these regions we assign non-negative (or non-positive) scores. Thus, for each point on the chart a score is assigned, which is accumulated as more points are plotted on it. This cumulative score is monitored and when it reaches or surpasses a critical value, then an out-of-control (OOC) signal is triggered. The theoretical properties of the run sum chart can be studied via the Markov chain method of [Champ and Rigdon (1997)]. [Davis et al. (1990)] and [Davis et al. (1994)] considered a special case of the run sum chart, known as *zone chart*, which consists of four regions above (below) the center line. They studied the performance of the zone chart for various score combinations with or without a head-start feature. The literature on run sum charts is quite extensive and there are schemes for monitoring a wide variety of processes. See, for example, [Abubakar et al. (2022)], [Antzoulakos et al. (2023)] and references therein.

The majority of the available control charts is based on the assumption that the distribution of the quality characteristic X is the normal distribution. In this case, in order to detect changes in either or in both process parameters, two control charts need to be applied (e.g., the $\bar{X}$ and S charts), one for detecting changes in process mean and the other for detecting changes in process variability. However, instead of using two different control charts for process monitoring it is possible to use a *single* control chart. Single control charts are control charts that use *one* plotting statistic for detecting changes in either or in both process parameters. Hence, only one chart (instead of two) can be used.

The Max chart ([Chen and Cheng (1998)]) is the most popular single control chart. It is a Shewhart-type chart in which the plotting statistic is the maximum of the absolute values of the transformed sample mean and sample variance. These transformed statistics follow the standard normal distribution when the process is in-control (IC). Therefore, any change in process parameters is reflected on the values of this statistic and the chart triggers an OOC signal. For a recent overview on single control charts see [Thaga and Sivasamy (2015)] and [Jalilibal et al. (2022)]. Since the Max chart is a Shewhart-type chart, it is not very sensitive in the detection of small and moderate shifts in process parameters. Recently, [Antzoulakos et al. (2023)] introduced and studied the

run sum Max chart (RSMax) which improves the performance of the usual Max chart in detecting changes in process mean and/or variability.

Apart from developing charts based on statistics that incorporate memory on them, another method to improve their performance in the detection of an OOC situation, is to use variable sample sizes and/or intervals instead of fixed ones. If at least one of the sample sizes and sampling intervals is allowed to vary, the control chart is considered as adaptive. See [Tagaras (1998)] and [Psarakis (2015)] for comprehensive surveys on adaptive control charts. Recently, [Chew et al. (2015)], [Mim et al. (2022)], [Saha et al. (2019)], [Yeong et al. (2022a)], [Yeong et al. (2022b)] and [Yeong et al. (2023)] studied the statistical design and performance of various adaptive run sum charts. Motivated by these works, we propose and study a single run sum chart with variable sample size and variable sampling intervals, based on the Max statistic. We will refer to this chart as the VSSI-RSMax chart. To the best of our knowledge, single adaptive run sum charts have not been studied in the literature so far.

The remaining part of this manuscript is organized as follows: In Section 2, we provide the operation of the proposed VSSI run sum Max chart. In Section 3 we introduce the appropriate measures for evaluating chart's performance. Section 4 consists of the results of a numerical study regarding the statistical design and performance of several VSSI run sum Max charts, under various scenarios. Numerical comparisons between the proposed charts and other competitive single charts, adaptive or not, are given in Section 5. Finally, conclusions are provided in Section 6.

2. THE PROPOSED CONTROL CHART

Let us assume that the *critical-to-quality* characteristic (CTQ) X is a random variable (r.v.) that follows a normal distribution with mean μ and variance σ^2, i.e. $X \sim N(\mu, \sigma^2)$. For $i = 1,2, \ldots$, let $\boldsymbol{X}_i = \left(X_{i1}, X_{i2}, \ldots, X_{in_i}\right)$ be independent random samples of size $n_i \geq 2$ from the process. Let also μ_0 and σ_0 be the in-control (IC) process mean and standard deviation. Furthermore, we assume that, in general, the process parameters μ and σ can be expressed as $\mu = \mu_0 + \mathrm{a} \cdot \sigma_0$ and $\sigma = \mathrm{b} \cdot \sigma_0$ ($\mathrm{a} \in \mathbb{R}$, $b > 0$). For an IC process we have that $\mathrm{a} = 0$ and $\mathrm{b} = 1$, otherwise the process has changed due to some assignable causes. We will denote the OOC values of the process parameters by μ_1 and σ_1. The constants a and b represent the shifts in process mean and standard deviation, respectively.

Let $\bar{X}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}$ and $S_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2$ be the sample mean and the sample variance of the i^{th} sample, respectively. We define

$$U_i = \frac{(\bar{X}_i - \mu_0)}{\sigma_0/\sqrt{n_i}}, \quad V_i = \Phi^{-1}\left\{H_{n_i-1}\left(\frac{(n_i - 1)S_i^2}{\sigma_0^2}\right)\right\},$$

where $\Phi(\cdot)$ denotes the cumulative distribution function (cdf) of the $N(0,1)$ distribution, and $H_m(\cdot)$ is the cdf of the chi-square distribution with m degrees of freedom. Statistics U_i and V_i are independent, and they both have $N(0,1)$ distribution when the process is IC. Without loss of generality, we assume that the random samples have common (*fixed*) sample size n (i.e., $n_i = n$, $i = 1,2,\ldots$) while the intervals between the successive samplings are fixed, too. Both settings (fixed sample sizes and sampling intervals) are common practices in applications of control charts.

The Max control chart of [Chen and Cheng (1998)] is a Shewhart-type single chart, based on the statistic

$$M_i = \max\{|U_i|, |V_i|\}. \tag{1}$$

The values of M_i are plotted on the chart against the sample number i ($i = 1,2,\ldots$). A large (positive) value of M_i results from a shift in the IC process mean μ_0 and/or IC process standard deviation σ_0. Otherwise, the value of M_i will be small and the process is IC. When $X_{ij} \sim N(\mu_0 + \mathrm{a}\cdot\sigma_0,\ (\mathrm{b}\cdot\sigma_0)^2)$, the cdf of M_i is given by

$$F_{M_i}(t|\mathrm{a},\mathrm{b},n) = \left\{\Phi\left(\frac{t}{\mathrm{b}} - \sqrt{n}\frac{\mathrm{a}}{\mathrm{b}}\right) - \Phi\left(-\frac{t}{\mathrm{b}} - \sqrt{n}\frac{\mathrm{a}}{\mathrm{b}}\right)\right\} \times \left\{H_{n-1}\left(\frac{\chi^2_{n-1;\Phi(t)}}{\mathrm{b}^2}\right) - H_{n-1}\left(\frac{\chi^2_{n-1;\Phi(-t)}}{\mathrm{b}^2}\right)\right\}.$$

The upper control limit (UCL) of the Max chart satisfies the equation $F_{M_i}(UCL|0,1,n) = 1 - a$ and it is given by

$$UCL = \left(\chi^2_{1;\sqrt{1-a}}\right)^{1/2}$$

where a is the pre-specified value of the false alarm rate of the chart, and $\chi^2_{1;p}$ denotes the lower *p*-percentage point of the chi-square distribution with one degree of freedom. Also, the center line of the Max chart is $CL = \left(\chi^2_{1;\sqrt{0.5}}\right)^{1/2}$. We will denote as n_F, h_F the fixed sample size n and the fixed sampling interval h of the Max chart. In the sequel we will refer to this chart as the fixed sampling size and sampling interval Max chart (FSSI-Max chart).

[Antzoulakos et al., (2023)] introduced the RSMax chart by dividing the charting region above the *CL* of the FSSI-Max chart into four regions and assigned scores to each one of them. Specifically they assigned the nonnegative scores a_1, a_2, a_3, a_4 to the regions $R_1 = [CL, UCL_1)$, $R_2 = [UCL_1, UCL_2)$, $R_3 = [UCL_2, UCL_3)$ and $R_4 = [UCL_3, \infty)$, respectively, where $a_1 \le a_2 \le a_3 \le a_4$, $0 < UCL_0 < UCL_1 < UCL_2 < UCL_3 < UCL_4$, $UCL_0 = CL$ and $UCL_4 = \infty$. As long as successive values of the plotted statistic M_i fall above the CL, the scores are accumulated. However, if the plotted statistic M_i falls below the CL the cumulative scoring process resets to zero. The cumulative score CS_i at sampling stage *i* is defined as follows

$$CS_i = \begin{cases} CS_{i-1} + a_{j+1}, & \text{if } UCL_j \le M_i < UCL_{j+1} \\ 0, & \text{if } M_i < CL \end{cases} \quad (2)$$

where $= 1,2, \ldots, j = 0,1,2,3$ and $CS_0 \ge 0$. Usually, $CS_0 = 0$, otherwise a head-start feature ([Lucas and Crosier, (1982)]) is used on the chart. The RSMax chart gives an OOC signal at the i^{th} sample if $CS_i \ge H$, where H is an appropriate critical value. In the sequel we will refer to this chart as the fixed sampling size and sampling interval run sum Max chart (FSSI-RSMax chart).

The VSSI-RSMax results from the FSSI-RSMax chart when we allow both sample size and sampling interval to vary. As already stated, the use of adaptive features on a control chart improves its sensitivity and OOC performance. The operation of the VSSI-RSMax chart requires a small sample size n_S, a large sample size n_L ($2 \le n_S < n_L$), a short sampling interval h_S and a long sampling interval h_L ($0 < h_S < h_L$). The sample size n_{i+1} of the $(i+1)^{th}$ sample, and the sampling interval h_{i+1} between the i^{th} and the $(i+1)^{th}$ sample depend on CS_i and they are determined according to the following rule:

$$(n_{i+1}, h_{i+1}) = \begin{cases} (n_L, h_S), & \text{if } H/G \le CS_i < H \\ (n_S, h_L), & \text{if } 0 \le CS_i < H/G \end{cases}. \quad (3)$$

In Equation (3), G is a positive integer ($G \ge 2$) that controls the rate of switching between the pairs (n_L, h_S) and (n_S, h_L).

From Equation (3) we deduce that a small sample size n_S and a long sampling interval h_L should be used for the next sample, if there is not an indication of a change in the process. On the other hand, if such indications are present, a large sample size n_L and a short sampling interval h_S should be used. If the process really operates with the presence of assignable causes, this must be detected as soon as possible.

Therefore, the idea behind the operation of the VSSI-RSMax is the following: If the value of the current cumulative score is large but less than H (i.e. $H/G \le CS_i < H$) we assume that this is an indication that a change has occurred in the process and the next sample might be from an OOC process. Thus, to reduce the number of samples and the time to detect the possible presence of assignable causes, a large sample size n_L and a short sampling interval h_S should be used for the next sample. On the other hand, if the value of the current cumulative score is small (i.e., $0 \le CS_i < H/G$) we may reasonably assume that there are no indications for a change in the process. Consequently, a small sample size n_S and long sampling interval h_L should be used for the next sample.

Table 1 gives a tabular representation of the VSSI-RSMax chart. Region R_0 extends below CL and no score is assigned to it. The probability p_j ($j = 1,2,3,4$) that the i^{th} plotted point falls in region R_j is given by

$$p_j = P\left(UCL_{j-1} \le M_i < UCL_j | \text{a,b,} n_i\right) = F_{M_i}\left(UCL_j | \text{a,b,} n_i\right) - F_{M_i}(UCL_{j-1} | \text{a,b,} n_i)$$

and

$$p_0 = P(M_i < CL|\text{a,b},n_i) = F_{M_i}(CL|\text{a,b},n_i)$$

where $n_i = n_S$ or n_L. For an IC process the $p_0 = 0.5$.

Table 1. *Regions of the VSSI-RSMax chart with their associated control limits, probabilities, and scores*

Region	Interval	Probability	Score
R_0	$[0, CL)$	p_0	$-$
R_1	$[CL, UCL_1)$	p_1	a_1
R_2	$[UCL_1, UCL_2)$	p_2	a_2
R_3	$[UCL_2, UCL_3)$	p_3	a_3
R_4	$[UCL_3, \infty)$	p_4	a_4

The upper control limits UCL_j of the VSSI-RSMax chart are given by

$$UCL_j = L\sqrt{\chi^2_{1;\sqrt{\Phi(j)}}}, \quad j = 1,2,3 \tag{4}$$

(see also [Antzoulakos et al., (2023)] and [Khoo et al., (2013)]). Parameter L is a multiplication factor that needs to be determined in order to achieve the desired IC performance. Given the set of non-negative scores (a_1, a_2, a_3, a_4), the critical value H and the pairs (n_L, h_S), (n_S, h_L), the proposed VSSI-RSMax chart will be denoted as VSSI-RSMax$_H(a_1, a_2, a_3, a_4)$.

3. PERFORMANCE MEASURES

Different performance measures are used when adaptive features are enabled on a control chart or not. The most common performance measure for a VSSI control chart is the average time to signal (ATS) which is defined as the time needed, on average, from the beginning of process monitoring until the chart produces for the first time an OOC signal. Using a Markov chain method (see e.g., [Saccucci et al. (1992)]), ATS is computed as

$$\text{ATS} = \mathbf{q}^{\mathrm{T}}(\mathbf{I} - \mathbf{Q})^{-1}\mathbf{h} \tag{5}$$

where $\mathbf{Q}$ is the $m \times m$ transition probability matrix of the transient states of the chain, $\mathbf{q}$ is the $m \times 1$ initial probability vector, $\mathbf{I}$ is the $m \times m$ identity matrix, and $\mathbf{h}$ is the $m \times 1$ vector whose i^{th} element h_i is the sampling interval corresponding to the transient state i, $i = 1, 2, \ldots, m$ (h_i is either h_S or h_L). See [Antzoulakos et al. (2023)] regarding the construction of matrix $\mathbf{Q}$. [Chew et al. (2015)] suggested a modification on the computation of ATS and specifically, not to measure the time to signal from the start of the process but from the time the first sample is taken. In this case, the ATS is computed as

$$\mathrm{ATS} = \mathbf{q}^{\mathrm{T}}(\mathbf{I}-\mathbf{Q})^{-1}\mathbf{h} - \mathbf{q}^{\mathrm{T}}\mathbf{h}. \tag{6}$$

However, a more realistic performance measure is the adjusted average time to signal (AATS, see [Reynolds et al. (1988)]) which is the expected time from the occurrence of the shift in the process until the chart gives an OOC signal. The AATS can measure the average time needed by the chart to produce an OOC signal when the shift(s) in process parameter(s) does not occur at the beginning of process monitoring but after the process has operated in the IC state for some time. The AATS is computed as

$$\mathrm{AATS} = \boldsymbol{v}^{\mathrm{T}}[(\mathbf{I}-\mathbf{Q})^{-1} - \mathbf{I}/2]\mathbf{h} \tag{7}$$

where $\boldsymbol{v}$ is the $m \times 1$ vector whose i^{th} element $(i = 1,2,\dots,m)$ is $v_i = g_i h_i/(\mathbf{g}^{\mathrm{T}}\mathbf{h})$, and g_i is the i^{th} element of the $m \times 1$ cyclical steady state probability vector $\mathbf{g}$. For the determination of $\mathbf{g}$ we used the methodology as described by [Champ (1992)].

Even though the performance of an adaptive chart is mostly evaluated in terms of ATS and AATS, their computation requires the knowledge of the shift sizes a and b. However, in practice, these shift sizes are usually unknown. For this reason, we may assume that a and b are random variables and thus proceed with measures of the overall performance of the chart such as the expected ATS (EATS) and the expected AATS (EAATS). See, for example, [Khaw et al. (2017)]. These measures are given respectively by

$$\mathrm{EATS} = \int_{\mathrm{a}_{min}}^{\mathrm{a}_{max}} \int_{\mathrm{b}_{min}}^{\mathrm{b}_{max}} f(\mathrm{a},\mathrm{b})\mathrm{ATS}(\mathrm{a},\mathrm{b})d\mathrm{b}d\mathrm{a} \tag{8}$$

and

$$\mathrm{EAATS} = \int_{\mathrm{a}_{min}}^{\mathrm{a}_{max}} \int_{\mathrm{b}_{min}}^{\mathrm{b}_{max}} f(\mathrm{a},\mathrm{b})\mathrm{AATS}(\mathrm{a},\mathrm{b})d\mathrm{b}d\mathrm{a} \tag{9}$$

where $f(\mathrm{a},\mathrm{b})$ is the joint probability density function of (a,b) and a_{min}, a_{max} (resp. b_{min}, b_{max}) are the minimum and maximum shifts of a (resp. b). A common assumption made in practice is that the random variables a and b are independent with marginal uniform distributions over the intervals $[\mathrm{a}_{min},\mathrm{a}_{max}]$ and $[\mathrm{b}_{min},\mathrm{b}_{max}]$, respectively.

Before closing this section, we should notice that there must be a fair comparison in the performance of VSSI and FSSI charts. For FSSI control charts the ATS is the product of the expected number of samples to signal and the fixed sampling interval. The expected number of samples to signal is the so-called *average run length* (ARL) and thus, $\mathrm{ATS} = h_F \cdot \mathrm{ARL}$. Let ASS_0 and ASI_0 be the IC average sample size and average sampling interval, respectively. Then (see, e.g. [Lee and Lin (2012)]),

$$\mathrm{ASS}_0 = \frac{\mathrm{ANOS}_0}{\mathrm{ARL}_0}, \quad \mathrm{ASI}_0 = \frac{\mathrm{ATS}_0}{\mathrm{ARL}_0},$$

where ARL_0 is the IC average number of samples to signal, and ANOS_0 is the IC average number of observations to signal. The ARL_0, ANOS_0 are computed by

$$\mathrm{ARL}_0 = \mathbf{q}^{\mathrm{T}}(\mathbf{I}-\mathbf{Q}_0)^{-1}\mathbf{1}, \quad \mathrm{ANOS}_0 = \mathbf{q}^{\mathrm{T}}(\mathbf{I}-\mathbf{Q}_0)^{-1}\boldsymbol{n},$$

where $\mathbf{Q}_0$ is the IC $m \times m$ transition probability matrix of the transient states of the chain, $\mathbf{1}$ is the $m \times 1$ vector with all its elements being 1, and $\boldsymbol{n}$ is the $m \times 1$ vector whose i^{th} element n_i is the sample size corresponding to the transient state i, $i = 1, 2, \ldots, m$ (thus, n_i is either n_S or n_L). To ensure a fair comparison, the ASS_0 and ASI_0 must be the same for all charts. For FSSI schemes we have that $\mathrm{ASS}_0 = n_F$ and $\mathrm{ASI}_0 = h_F$.

4. DESIGN AND PERFORMANCE OF THE PROPOSED CHART

In this section we present the statistical design of the VSSI-RSMax$_H(a_1, a_2, a_3, a_4)$ chart and assess its performance using the performance measures presented in Section 3. The parameters of the VSSI-RSMax$_H(a_1, a_2, a_3, a_4)$ chart are: the critical value H, the set of scores (a_1, a_2, a_3, a_4), the parameter L of the control limits UCL_j ($j = 1,2,3$), the sample sizes n_S and n_L, the sampling intervals h_S and h_L, and the parameter G controlling the switching between the two sets of sample sizes and intervals (n_S, h_L) and (n_L, h_S). Below, we provide the steps of the algorithmic procedure for the statistical design of the VSSI-RSMax chart:

Step 1: Choose the score vector (a_1, a_2, a_3, a_4), the critical value H, the small sample size n_S ($n_S < n_F$), the short sampling interval h_S ($h_S < h_F$), and the value of G.

Step 2: Set the in-control ARL equal to the desired value c, i.e., $\mathrm{ARL}_0 = c$.

Step 3: Determine the value of L as the unique solution of the equation $\mathrm{ARL}_0 = c$.

Step 4: Set $UCL_0 = CL = \left(\chi^2_{1;\sqrt{0.5}}\right)^{1/2}$ and $UCL_j = L\sqrt{\chi^2_{1;\sqrt{\Phi(j)}}}$, $j = 1, 2, 3$.

Step 5: Determine the n_L value so that $\mathrm{ASS}_0 = n_F$ subject to the constraint $n_L > n_F$.

Step 6: Determine the h_L value so that $\mathrm{ASI}_0 = h_F$ subject to the constraint $h_L > h_F$.

Step 7: Set up the VSSI-RSMax$_H(a_1, a_2, a_3, a_4)$ chart and declare the process as OOC at the i^{th} sample if $CS_i \geq H$.

In the sequel, we present the results of an extensive numerical study, regarding the design and performance of various VSSI-RSMax control charts. The score sets and the critical values for each of these schemes are (see also [Antzoulakos et al., (2023)] and references therein):

Table 2. *Scores and Critical Values for the VSSI-RSMax control charts*

Scheme	(a_1, a_2, a_3, a_4)	H	Scheme	(a_1, a_2, a_3, a_4)	H
C_1	(0,1,2,3)	5	C_6	(1,2,7,14)	14
C_2	(0,1,2,4)	4	C_7	(1,3,11,19)	19
C_3	(0,1,6,12)	12	C_8	(1,3,8,15)	15
C_4	(0,3,4,12)	12	C_9	(1,2,7,13)	13
C_5	(0,2,3,8)	8			

For the IC settings, we chose $\mathrm{ARL}_0 = 250$, $\mathrm{ASS}_0 = 5$, $\mathrm{ASI}_0 = 1$, $(n_S, h_S) = (4, 0.1)$ and $G = 3$. The OOC performance of the VSSI-RSMax$_H(a_1, a_2, a_3, a_4)$ chart is evaluated in terms of AATS while the ATS is used for the IC performance. Thus, all schemes have the same IC ATS.

Table 3. *Best VSSI-RSMax chart based on AATS for* $n_S = 4$, $h_S = 0.1$, $G = 3$, $ASI_0 = 1$, $ASS_0 = 5$, $ARL_0 = 250$

a	b	h_L	n_L	L	AATS	Chart
0	0.25	1.1270	12	0.9196	1.35	C_1
	0.5	1.1270	12	0.9196	7.34	C_1
	1.5	1.1270	12	0.9196	3.45	C_1
	2.0	1.1270	12	0.9196	1.36	C_1
0.5	0.25	1.1270	12	0.9196	1.35	C_1
	0.5	1.1189	12	1.0312	4.59	C_8
	1.0	1.0682	19	1.0246	9.80	C_2
	1.5	1.1270	12	0.9196	2.49	C_1
	2.0	1.0930	15	0.9743	1.26	C_4
1.0	0.25	1.1270	12	0.9196	1.15	C_1
	0.5	1.1270	12	0.9196	1.53	C_1
	1.0	1.1270	12	0.9196	1.82	C_1
	1.5	1.1270	12	0.9196	1.47	C_1
	2.0	1.0930	15	0.9743	1.05	C_4
1.5	0.25	1.0568	22	0.9516	0.61	C_3
	0.5	1.1270	12	0.9196	0.71	C_1
	1.0	1.0930	15	0.9743	0.91	C_4
	1.5	1.0930	15	0.9743	0.96	C_4
	2.0	1.0930	15	0.9743	0.85	C_4

In Table 3 we provide the best scheme among the considered schemes C_1-C_9, i.e., the scheme achieving the minimum OOC AATS. For the shift sizes we chose a $\in \{0, 0.5, 1, 1.5\}$ and b $\in \{0.25, 0.5, 1, 1.5, 2\}$. The results in Table 3 show that the C_1 scheme has the best performance for most of the considered shifts. In addition, Table 3

can be used by the practitioners who would like to implement a VSSI-RSMax chart in practice, since it provides the best chart to use, for the examined n_S, h_S, G, a and b values.

Next, in Table 4, we provide the EAATS values of the schemes C_1-C_9 for $\text{ARL}_0 = 250$, $(n_S, h_S) = (4, 0.1)$, $G = 3$, $\text{ASS}_0 = 5$ and $\text{ASI}_0 = 1$. The boldfaced entries indicate the chart attaining the lowest EAATS value. The lesser the value of EAATS, the better is the overall performance of the chart. For the distribution of the independent random shifts a and b we assume that $\text{a}, \text{b} \sim U(0.3,2)$, where $U(\theta_1, \theta_2)$ is the Uniform distribution with support the interval $[\theta_1, \theta_2]$. Note also that a large difference $\theta_2 - \theta_1$ indicates a high degree of uncertainty. The results of Table 4 show that the scheme C_1 is the one with minimum EAATS value among the nine schemes.

Table 4. *EAATS values of VSSI-RSMax charts*

Chart	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9
EAATS	**2.4146**	2.6813	3.1687	2.5523	2.5730	2.6082	2.6175	2.5534	2.6306

5. NUMERICAL COMPARISONS

In this section we provide numerical comparisons between the proposed VSSI-RSMax chart, the VSSI-Max chart of [Lee and Lin (2012)] and the (FSSI) Max-EWMA chart of [Chen et al., (2001)]. Due to space economy, we do not present results from the comparisons between the proposed chart and the (FSSI) Max and RSMax charts since the latter have worse performance.

For all the examined adaptive schemes we use $(n_S, h_S) = (4,0.1)$ and $G = 3$ (for the VSSI-RSMax chart). To ensure a fair comparison between adaptive and fixed control charts we set $\text{ARL}_0 = 250$, $n_F = \text{ASS}_0 = 5$ and $h_F = \text{ASI}_0 = 1$. Also, for the shift sizes we choose $\text{a} \in \{0, 0.5, 1, 1.5\}$ and $\text{b} \in \{0.25, 0.5, 1, 1.5, 2\}$. The values of the design parameters of the VSSI-Max chart when $\text{ASS}_0 = 5$, $\text{ASI}_0 = 1$ and $(n_S, h_S) = (4,0.1)$ are $(n_L, h_L, UCL_1, UWL_1) = (27, 1.0451, 3.089, 2.2581)$. These values have been derived using the procedure described in [Lee and Lin (2012)].

Additionally, for the (FSSI) Max-EWMA control charts we use the four pairs for the design parameters (λ, K) provided in [Chen et al. (2001)], which are $(0.05, 2.45)$, $(0.10, 2.78)$, $(0.20, 3.03)$ and $(0.30, 3.14)$. According to [Chen et al. (2001)], the value $M_i' = \max\{|Y_i|, |Z_i|\}$, $i = 1,2, \ldots$, is plotted on the Max-EWMA chart where

$$Y_i = (1-\lambda)Y_{i-1} + \lambda U_i, \quad Z_i = (1-\lambda)Z_{i-1} + \lambda V_i,$$

and the statistics U_i, V_i are those defined in the Max chart. Also, $Y_0 = Z_0 = 0$. Since $M_i' \geq 0$, only an upper control limit is needed (say $UCL_{MaxEWMA}$), which is given by

$$UCL = (1.12838 + 0.60281 \cdot K)\sqrt{\lambda/(2-\lambda)}.$$

The parameter $\lambda \in (0,1]$ is the smoothing parameter of the chart while parameter K controls the limits width.

In Table 5 we provide numerical comparisons in terms of AATS for the examined control charts. The bold-faced entries correspond to the minimum AATS value of each row. The control chart among $C_1 - C_9$ with the best performance for every pair of shifts (a, b) is given in parenthesis (see in column labeled VSSI-RSMax). For the (FSSI) Max-EWMA chart the AATS values are calculated as $\mathrm{ARL} - 0.5$ (see, e.g. [Lee and Lin, (2012)]).

The results from Table 5 show that the VSSI-RSMax chart is more efficient than the VSSI-Max chart for small and moderate shift sizes (a, b). However, for large shifts in the IC process standard deviation σ_0 (i.e., $\mathrm{b} = 2$) and/or large shifts in the IC process mean μ_0 (i.e. $\mathrm{a} = 1.5$), the VSSI-Max chart outperforms the proposed chart. This fact was also mentioned by [Lee and Lin (2012)] and it is attributed to the ability of Shewhart-type charts to react more quickly than control charts with memory to large shifts in process parameters. Also, the VSSI-RSMax chart outperforms the (FSSI) Max-EWMA chart for almost all shifts, except for the cases $(\mathrm{a}, \mathrm{b}) = (0.5,1)$ and $(0,0.5)$.

Table 5. *AATS comparison between VSSI-Max, Max-EWMA, and VSSI-RSMax charts for* $G = 3$, $ASI_0 = 1$, $ASS_0 = 5$, $ARL_0 = 250$

a	b	VSSI-Max	Max-EWMA	(λ,K)	VSSI-RSMax
0.0	0.25	1.91	1.90	(0.3,3.14)	**1.38** (C_1)
	0.5	11.80	**4.89**	(0.2,3.03)	7.28 (C_1)
	1.5	3.77	5.80	(0.3,3.14)	**3.40** (C_1)
	2.0	**1.31**	2.17	(0.3,3.14)	1.36 (C_1)
0.5	0.25	1.91	1.91	(0.3,3.14)	**1.37** (C_1)
	0.5	11.04	4.73	(0.2,3.03)	**4.59** (C_8)
	1	14.68	**8.09**	(0.2,3.03)	9.37 (C_1)
	1.5	2.63	4.11	(0.3,3.14)	**2.48** (C_1)
	2.0	**1.21**	1.99	(0.3,3.14)	1.26 (C_1)
1.0	0.25	1.54	1.86	(0.3,3.14)	**1.18** (C_1)
	0.5	2.44	2.37	(0.3,3.14)	**1.53** (C_1)
	1	2.07	2.53	(0.3,3.14)	**1.82** (C_1)
	1.5	**1.46**	2.30	(0.3,3.14)	1.47 (C_1)
	2.0	**1.00**	1.61	(0.3,3.14)	1.05 (C_4)
1.5	0.25	**0.58**	1.49	(0.3,3.14)	0.61 (C_3)
	0.5	**0.64**	1.47	(0.3,3.14)	0.71 (C_3)
	1	**0.86**	1.43	(0.3,3.14)	0.91 (C_4)
	1.5	**0.90**	1.41	(0.3,3.14)	0.96 (C_4)
	2.0	**0.80**	1.23	(0.3,3.14)	0.85 (C_4)

Finally, in Table 6 we provide EAATS values for all the examined control charts in the scenario for the support of the random shifts a and b introduced in Section 4. The VSSI-RSMax chart consistently attains the lowest EAATS value according to Table 6 (case $(n_S, h_S) = (4,0.1)$). Especially, the C_1 scheme is recommended when there is further information regarding the shift sizes in process parameters.

Before closing this section, it is worth noting that the best Max-EWMA chart either in terms of AATS or EAATS is the one with $\lambda = 0.3$. Such a value is not very typical since smaller values (e.g., $\lambda = 0.05$ or 0.10) are suggested for better performance in the detection of small shifts in process parameters.

Table 6. *EAATS values of the VSSI-Max, Max-EWMA and VSSI-RSMax charts* $ASI_0 = 1$, $ASS_0 = 5$, $ARL_0 = 250$

VSSI-Max	**Max-EWMA**				**VSSI-RSMax**								
	0.05	0.10	0.20	0.30	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9
3.90	4.09	3.31	2.84	2.74	**2.41**	2.68	3.17	2.55	2.57	2.61	2.62	2.55	2.63

6. CONCLUSIONS

In this work, we studied the statistical design and performance of a variable sample size and sampling interval single run sum chart, namely the VSSI-RSMax chart, which it is suitable for the monitoring of a normally distributed process. The run sum chart offers an improved sensitivity in the detection of small and moderate shifts in process parameters, compared to Shewhart-type charts while it is considered as simpler in implementation and interpretation.

The results of an extensive numerical study showed that the proposed VSSI-RSMax chart outperforms the VSSI-Max chart for almost all the shift sizes. Also, under certain cases, the proposed chart outperforms the fixed sample size and sampling interval Max-EWMA chart. The latter is better when there is a small decreasing shift in standard deviation (e.g., for $b = 0.5$) and a small shift in mean (e.g., $a = 0.5$). In addition, among nine commonly used run sum charts, the C_1 scheme, is the one with the best performance in most of the examined OOC situations while it outperforms the VSSI-Max chart and the Max-EWMA in terms of overall performance. Therefore, we recommend the use of this scheme in practice, especially when is no prior information regarding the shift(s) in process parameters.

Finally, we should mention that the results presented in this work are valid only under the assumption of normality for the distribution of X. A separate study is needed to assess how robust is the proposed VSSI-RSMax chart, as well as the Max-EWMA chart, when this assumption is violated.

ΠΕΡΙΛΗΨΗ

Στην παρούσα εργασία προτείνεται η εφαρμογή της τεχνικής της ροής αθροίσματος σε ένα Max διάγραμμα ελέγχου εφοδιασμένο με τις διαδικασίες του μεταβλητού μεγέθους δείγματος και διαστήματος δειγματοληψίας (VSSI-RSMax διάγραμμα). Το VSSI-RSMax διάγραμμα μεταβάλει το μέγεθος του επόμενου δείγματος και το διάστημα δειγματοληψίας σύμφωνα με την τρέχουσα τιμή του αθροιστικού σκορ. Χρησιμοποιείται η μέθοδος της Μαρκοβιανής αλυσίδας για τον υπολογισμό διαφόρων μέτρων απόδοσης του προτεινόμενου διαγράμματος ελέγχου. Επίσης, το VSSI-RSMax διάγραμμα συγκρίνεται με άλλα ανταγωνιστικά διαγράμματα ελέγχου όπως το VSSI-Max διάγραμμα και το FSSI Max-EWMA διάγραμμα.

REFERENCES

Abubakar S.S., Khoo M.B.C., Saha S., and Teoh W.L. (2022). Run sum control chart for monitoring the ratio of population means of a bivariate normal distribution. *Communications in Statistics-Theory and Methods*, **51**, 4559-4588.

Antzoulakos D.L., Fountoukidis K.G., and Rakitzis A.C. (2023). A phase II run rum chart for monitoring process mean and variability. *Journal of Statistical Computation and Simulation*, https://doi.org/10.1080/00949655.2023.2178652 .

Champ C. W. (1992). Steady-state run length analysis of a Shewhart quality control chart with supplementary runs rules. *Communications in Statistics-Theory and Methods*, **21**, 765-777.

Champ C.W., and Rigdon S.E. (1997). An analysis of the run sum control chart. *Journal of Quality Technology*, **29**, 407-417.

Chen G., and Cheng S.W. (1998). MAX chart: Combining X-bar chart and S chart. *Statistica Sinica*, **8**, 263-271.

Chew X.Y., Khoo M.B.C., Teh S.Y., and Castagliola P. (2015). The variable sampling interval run rum $\bar{X}$ control chart. *Computers & Industrial Engineering*, **90**, 25-38.

Davis R.B., Homer A., and Woodall W.H. (1990). Performance of the zone control chart. *Communications in Statistics-Theory and Methods*, **19**, 1581-1587.

Davis R.B., Jin C., and Guo Y. (1994). Improving the performance of the zone control chart. *Communications in Statistics-Theory and Methods*, **23**, 3557-3565.

Jaehn A.H. (1991).The zone control chart. *Quality Progress*, **24**(7), 65-68.

Jalilibal Z., Amiri A., and Khoo M.B.C. (2022). A literature review on joint control schemes in statistical process monitoring. *Quality and Reliability Engineering International*, **38**, 3270-3289.

Khaw K. W., Khoo M. B., Yeong W. C., and Wu Z. (2017). Monitoring the coefficient of variation using a variable sample size and sampling interval control chart. *Communications in Statistics-Simulation and Computation*, **46**, 5772-5794.

Khoo M.B.C., Sitt C.K., Wu Z., and Castagliola P. (2013). A run sum Hotelling's χ^2 control chart. *Computers & Industrial Engineering*, **64**, 686-695.

Lee P.H., and Lin C.S. (2012). Adaptive Max charts for monitoring process mean and variability. *Journal of Chinese Institute of Industrial Engineers*, **29**, 193-205.

Lucas J.M., and Crosier R.B. (1982). Fast initial response for CUSUM quality control schemes. *Technometrics.* **24**, 199-205.

Mim F.N., Khoo M.B.C., Saha S., and Haq, A. (2022). New run sum t charts with variable sampling intervals for process mean. *Communications in Statistics - Simulation and Computation*, **51**, 350-5372.

Montgomery, D. C. (2020). Introduction to statistical quality control. *John Wiley & Sons*.

Psarakis S. (2015). Adaptive control charts: Recent developments and extensions. *Quality and Reliability Engineering International*, **31**, 1265-1280.

Roberts S.W. (1966). A comparison of some control chart procedures. *Technometrics*, **8**, 411–430.

Reynolds J.H. (1971). The run sum control chart procedure. *Journal of Quality Technology*, **3**, 23-27.

Reynolds Jr. M.R., Amin R.W., Arnold J.C., and Nachlas J.A. (1988). X charts with variable sampling intervals. *Technometrics*, **30**, 181-192.

Saccucci M.S., Amin R.W., Lucas J.M. (1992). Exponentially weighted moving average control schemes with variable sampling intervals. *Communications in Statistics-Simulation and Computation*, **21**, 627-657.

Saha S., Khoo M.B.C., Ng P.S., and Chong Z.L. (2019). Variable sampling interval run sum median charts with known and estimated process parameters. *Computers & Industrial Engineering*, **127**, 571-587.

Tagaras G. (1998). A survey of recent developments in the design of adaptive control charts. *Journal of Quality Technology*, **30**, 212-231.

Thaga K., and Sivasamy R. (2015). Single variables control charts: A further overview. *Indian Journal of Science and Technology*, **8**, 18-528.

Yeong W.C., Lim S.L., Khoo M.B.C., Ng P.S., and Chong Z.L. (2022a). A variable sampling interval run sum chart for the coefficient of variation. *Journal of Statistical Computation and Simulation*, **92**, 3150-3166.

Yeong W.C., Tan Y.Y., Lim S.L., Khaw K.W., and Khoo M.B.C. (2022b). A variable sample size run sum coefficient of variation chart. *Quality and Reliability Engineering International*, **38**, 1869-1885.

Yeong W.C., Tan Y.Y., Lim S.L., Khaw K.W., and Khoo M.B.C. (2023). Variable sample size and sampling interval (VSSI) and variable parameters (VP) run sum charts for the coefficient of variation. *Quality Technology & Quantitative Management*.

Greek Statistical Institute
Proc. of the 35th Panhellenic & 1st International Statistics Conf.(2023) ,pp.165-180

THEMATIC ANALYSIS AND RESEARCH TRENDS ON MEDICAL DIAGNOSIS PATENTS USING STATE-OF-THE-ART TOPIC MODELING

Konstantinos Georgiou, Konstantinos Charmanas, Lefteris Angelis
School of Informatics, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece
Correspondence email (K.G.): konsgeor@csd.auth.gr

ABSTRACT

Medical research and technologies play a crucial role in the welfare of humanity, as they ensure that people have access to effective healthcare. The diagnosis of diseases for the application of preventive measures along with the timely detection and containment of pandemics and epidemics have been brought to the spotlight in recent years, particularly during the COVID-19 pandemic. Medical companies and technological giants strive to provide Information and communications technology (ICT) solutions for medical diagnosis, to be leveraged in a plethora of cases. As competition is high and each company has benefits and profits from these technologies, many organizations secure their inventions by patenting them. Patents are a highly efficient way of protecting the commercial rights of an invention while they are also proven to be excellent indicators of technological innovation and progress. The information hidden in patent applications, which is frequently in textual form, can uncover valuable insights regarding the thematic trends and topics in granted patents as well as pinpoint the organizations and companies that have the highest patent shares in medical diagnosis technologies. In this study, a patent dataset related to ICT technologies for medical diagnosis and epidemic monitoring from the United States Patent and Trademark Office (USPTO) is collected and the topic modeling algorithms Pseudo-document based Topic Modeling (PTM), Correlated Topic Models (CTM), Hierarchical Pachinko Allocation (HPAM) and Latent Dirichlet Allocation (LDA) algorithms are applied. The optimal algorithm is then chosen for topic analysis, in combination with a specialized indicator for topic growth throughout the years. The results from the topic analysis indicate that medical diagnosis patents cover multiple thematic axes, ranging from disease prevention (cancer, diabetes, neural stimulation) and human welfare (circadian rhythm, dental practices, orthopedics) to expanded practices (sleeping and eating disorders, bullying) and innovative technologies (wearables, biometric augmented reality, microbiome diagnostics).

Keywords: Topic Modeling, CAGR, Patent Analysis, Medical Diagnosis

1. INTRODUCTION

The advanced capabilities of computer science have contributed to the development of significant technologies that provide efficient solutions in different domains and applications, including medicine and diagnostics. According to Tian et al. (2019), the process of disease diagnosis and potential treatment is currently assisted by artificial intelligence, and often achieves more accurate results than the existing approaches. In recent research, deep learning architectures constitute a common practice in providing solutions and outcomes for medical diagnostics by uncovering patterns from image and sound data (Bakator and Radosav, 2018).

Thus, the various medical diagnoses ICT are of great importance since the healthcare industry has active implementations that leverage these technologies daily. In this study, at the main goal is to demystify the primary investors and thematic axis related to medical diagnosis ICT by analyzing patents from the United States Patent and Trademark Office ($USPTO$). Patent data are usually considered as a knowledge basis in identifying the main industrial trends, across the years, and competitors of a specific technological domain (Abbas et al., 2014).

The process of analyzing and extracting knowledge from patent data is described as patent analysis. In this study, the framework introduced by Choi et al. (2018) which leverages the different patent properties to detect the main topics and assignees from a patent dataset is employed, which leads to providing insights about the emerging and dominant technologies of medical diagnosis ICT along with their main investors/competitors. Overall, the findings of this study can help researchers and industries, which are relevant to medical diagnosis ICT, in important objectives like strategic planning, competitor analysis, trending forecasting, technology scouting as well as developing new technologies.

2. RELATED WORK

In this Section, the main existing research that is relevant to medical technologies is discussed, especially covered by patents. Among the many applications of machine learning in diagnostics, cardiovascular disease classification or heartbeat classification, diabetes mellitus, tuberculosis, and Alzheimer's disease are frequently considered as research areas of interest (Chui et al., 2017). Overall, the primary contributions of Artificial Intelligence (AI) applications in the healthcare industry are related to advancing disease treatments, enhancing patient engagement, improving accuracy, reducing errors as well as optimizing costs and efficiency (Lee and Yoon, 2021). Although the advances of AI offer numerous benefits, He et al. (2019) suggest that there are some practical issues concerning clinical workflows, data sharing, privacy, and patient safety that should be addressed before implementing such methods. Thus,

AI technologies have not yet been completely adopted in the healthcare sector and associated industries.

In the field of cardiovascular diseases, Flores et al. (2022) review several patents to address the current practices aiding the process of patient well-caring and treatment for different types of relevant diseases. The objective that is covered by this research is to provide insights on some critical technologies that should be studied in order to develop new sufficient inventions that would provide in time and empowered solutions.

Moreover, the rise of Covid-19 urged the development of new technologies and methods that would help the mitigation and treatment of this disease. Soni et al (2022) investigate relevant patent technological areas, including medical diagnostics, and further discuss the content, importance and development of relevant inventions which have recently emerged and carry a vital role in this crisis, hence providing valuable insights on significant inventions that would enhance the public awareness and knowledge sharing.

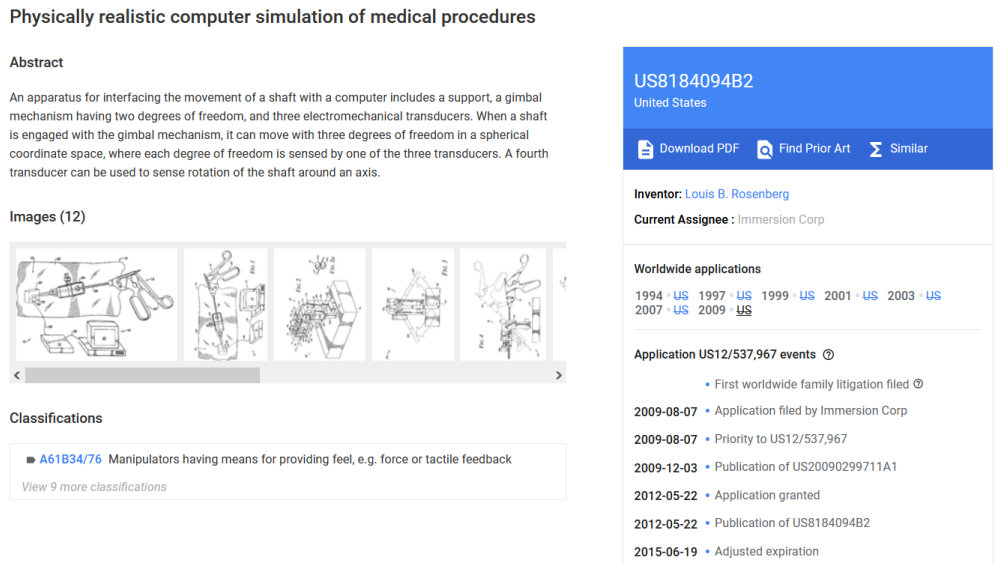
From a different perspective, Wang and Hsu (2023) analyze smart health patent data aiming at assessing and forecasting multiple technology convergence patterns, i.e., innovations that are formed by multiple existing technologies. Their outcomes uncovered some new core technologies implementing machine learning algorithms and manufacturing wearable devices. At the same time, some emerging technological areas are distinguished which are associated with communication networks, data transmission, speech recognition, and head-up displays. Finally, this study proposes that the related patent data and outcomes offer valuable knowledge that could benefit both inventors and firms.

In general, the various applications of patent analysis have contributed to extracting valuable findings from many areas of interests. Briefly, patent analysis has been notably used in the domain of Augmented Reality (Evangelista et al., 2020) to discover prominent technologies and the industrial landscape. In addition, various ICT domains such as the Internet of Things (IoT) and AI have been studied under the spectrum of patent analysis (Ardito et al., 2018, Liu et al., 2021) with the purpose of mapping the primary technologies, the innovations that patents entail as well as pinpointing primary assignees and organizations. In other domains, patent analysis has been used in low-carbon energy technologies (Albino et al., 2014) for the purpose of establishing policies via the profiling of patent commercialization, logistics (Choi et al., 2018, Kwon et al., 2022) with the purpose of extracting thematic axes and defining topics that drive innovation and photovoltaics (Sampaio et al., 2018) for identifying areas of knowledge and country specific competitors.

3. METHODOLOGY

In this Section, the main framework of this study which leads to the identification of the main topics and assignees of medical diagnosis *ICT* patents are presented. A patent entry provides information concerning several patent properties and characteristics such as assignees, titles, inventors, countries and more (Figure 1).

Figure 1. *Example of a patent entry*



The first step of the employed framework includes all the necessary steps that help in forming a representative dataset with relevant patented technologies. To do so, the Cooperative Patent Classification (*CPC*) scheme was manually searched in order to detect an appropriate classification that is relevant to this study. After a thorough examination, the subgroup *G16H50/00*, entitled *"ICT specially adapted for medical diagnosis, medical simulation or medical data mining; ICT specially adapted for detecting, monitoring or modeling epidemics or pandemics"*, was selected as a representative classification of medical diagnosis *ICT*. Thus, after collecting all the *USPTO* patent data that fall under this classification, a dataset of 15734 entries is established.

Afterwards, the methodology proposed by Choi et al. (2018) is employed to identify and classify the underlying topics of patent titles, through topic models, while also assess the main assignees per topic. In the process, multiple text preprocessing techniques are applied to transform the initial text into an appropriate data format that can be pipelined into a topic model (Maier et al., 2021). These techniques are lower-case transformation, punctuation and stopwords removal, tokenization and stemming.

In addition, a Document-Term Matrix is formed where only the stemmed words that occurred in more than 5 patent titles are included. This matrix constitutes the input to the topic models.

To reveal the main topics of the retrieved patents, four different algorithms are implemented to train topic models in the range from two to twenty topics. These algorithms are the following: i) Pseudo-document based Topic Model (Zuo et al., 2016), ii) Correlated Topic Models (Blei and Lafferty, 2006), iii) Hierarchical Pachinko Allocation (Mimmo et al., 2007), iv) Latent Dirichlet Allocation (Blei et al., 2003). To select the *optimal* model, two evaluation metrics that are relevant to topic coherence (Röder et al., 2015) and topic divergence (Dieng et al., 2020) are aggregated. Thus, for the rest of the employed methodology, the model that obtains the maximum cumulative evaluation is utilized.

In general, the main outputs of a topic model are the *Document – Topic Matrix*, usually referred to as *theta* or *gamma*, and the *Term – Topic Matrix*, usually referred to as *phi* (Blei et al., 2003). Briefly, *theta* stores the distributions of topics over documents, i.e., the probability of a document belonging to a topic, and *phi* stores the distributions of words over topics, the probability that a randomly selected token that belongs to a topic is a specific word. It should be noted that the rows of these two matrices sum to one. Also, these two matrices constitute the basis in the interpretation of the representative titles of each topic, where the concepts of the most probable words and documents (patent titles) are considered more.

Furthermore, the dominant topic for a document is denoted as the topic that obtains the highest probability. This information helps in evaluating the patent share of a topic (Equation 1), in different time periods, and in assessing the ten-year Compound Annual Growth Rate ($CAGR$) of each topic (Equation 2).

$$patent\ share\ in\ a\ topic = \frac{\#patents\ in\ the\ topic}{total\ \#\ patents} \times 100 \quad (1)$$

$$CAGR\ per\ Topic = \left[\left(\frac{Patent\ share\ of\ topic\ in\ 2023}{Patent\ share\ of\ topic\ in\ 2013}\right)^{\frac{1}{2023-2013}} - 1\right] \times 100 \quad (2)$$

where *#patents in the topic* denotes the number of patents in which the topic is dominant.

Finally, each topic is classified into one of the four classes, as proposed by Choi et al., 2018, describing *Emerging*, *Declining*, *Dominant* and *Saturated* topics. These classes are assessed by evaluating the $CAGR$ (y axis) and the 2023 patent share (x axis) of each topic in a two-dimensional scatter plot. This plot is split into four quartiles, one for each class, which are formed using two thresholds denoting the neutral $CAGR$ ($CAGR$ equal to 0) and the median 2023 patent share. Thus, the aforementioned classes can be discriminated as follows:

- Emerging: High $CAGR$ – Low 2023 patent share
- Dominant: High $CAGR$ – High 2023 patent share
- Saturated: Low $CAGR$ – High 2023 patent share
- Declining: Low $CAGR$ – Low 2023 patent share

The final step towards this approach is to identify the representative assignees per topic where the technology share of an assignee into a topic is measured by the number of patents that are owned-assigned by the assignee and belong to this topic (Equation 3).

$$technology\ share = \frac{\#\ patents\ that\ the\ assignee\ has\ in\ the\ topic}{total\ \#\ patents\ in\ the\ topic} \quad (3)$$

The motivation of this study is to provide answers for two Research Questions (RQs) encompassing concepts that are relevant to the general objectives of patent analysis:

RQ_1: Which are the primary thematic axes derived from medical diagnosis *ICT* patents and which are driving innovation? Uncovering and assessing the main themes of a patent dataset is a usual choice in tasks like technology forecasting and scouting as well as trending analysis. To provide insights regarding these general goals, a methodology that both uncovers and classifies the main topics of medical diagnosis ICT patents is employed.

RQ_2: Which are the leading organizations in medical diagnosis patenting? The identification of the main competitors, i.e., competitor analysis, in a technology domain is very important task that can help organizations in decision making and strategic planning. To satisfy this goal, an approach that distributes the various organizations-assignees in the classified topics is applied and leads to the extraction of valuable information depending on the goals and interests of each researcher-organization.

4. RESULTS

This Section discusses the main results of this study, with respect to the two **RQs**. In Section 4.1 the main topics of medical diagnosis ICT patents, as extracted by patent titles, are presented and classified according to the employed methodology. Also, Section 4.2 provides insights on the main investors of the different technologies, expressed by topics, belonging to the investigated domain. In summary, the interpretation of the topics will help in assessing the overall main areas of interest that are related to medical diagnosis while the topic classifications will help in providing information on trending and emerging themes (**RQ_1**). Regarding **RQ_2**, the evaluation of the technology shares for each assignee and topic will distinguish the main stakeholders of dominant and emerging technologies, thus showing which assignees carry a beneficiary positioning in the investigated technological area.

4.1 RQ_1: Which are the primary thematic axes derived from medical diagnosis *ICT* patents and which are driving innovation?

By finalizing the text preprocessing steps, a Document-Term Matrix is established and consists of 1815 words. By pipelining this matrix into the different topic models, overall 76 topic models, for 4 algorithms and 19 different parameters for numbers of topics, are evaluated. Figure 2 presents the topic coherence, Figure 3 presents the topic divergence and Figure 4 shows the aggregated scores of the trained topic models.

Figure 2. *Topic coherence of the models*

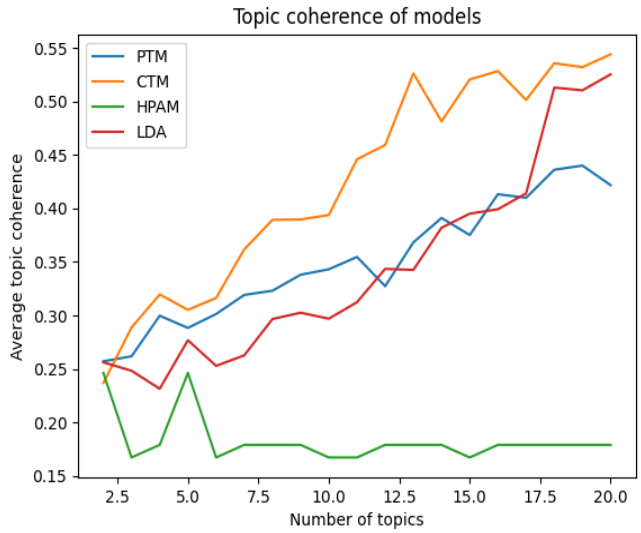


Figure 3. *Topic divergence of the models*

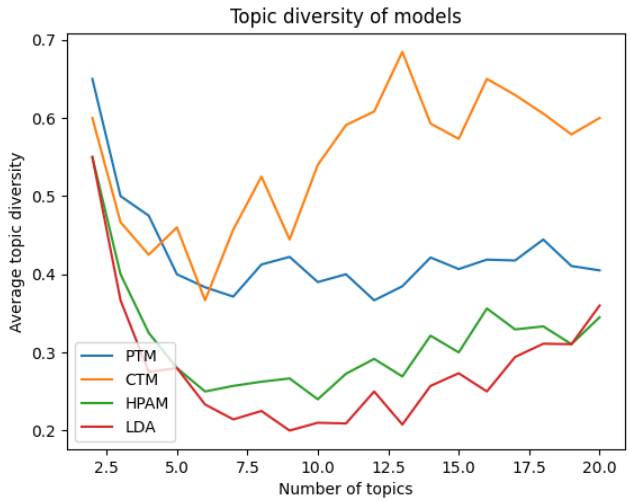
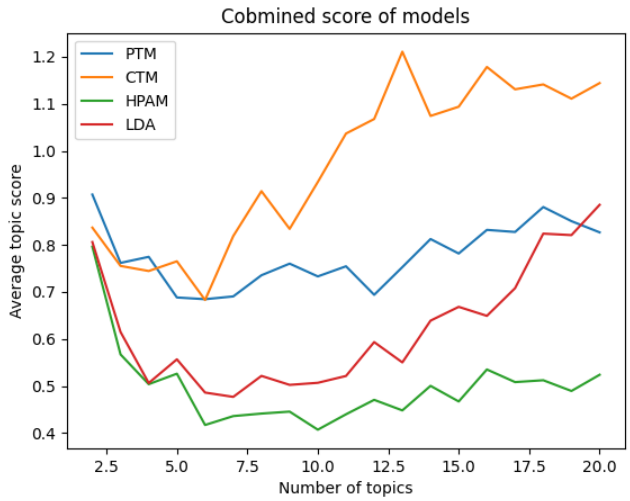


Figure 4. *Combined score of the models*



The above figures indicate that the overall best performing topic modeling algorithm for the utilized Document-Term Matrix is the Correlated Topic Models, since the majority of the models that are trained using this algorithm outperform the rest models for the same number of topics. At the same time, the aggregated/combined score shows that the best performing model is trained under the aforementioned algorithm for 13

topics. As discussed previously, from now the respective model will be used for the rest of this study. Thus, Table 1 presents the representative topic titles and the current patent share along with the $CAGR$ of each topic.

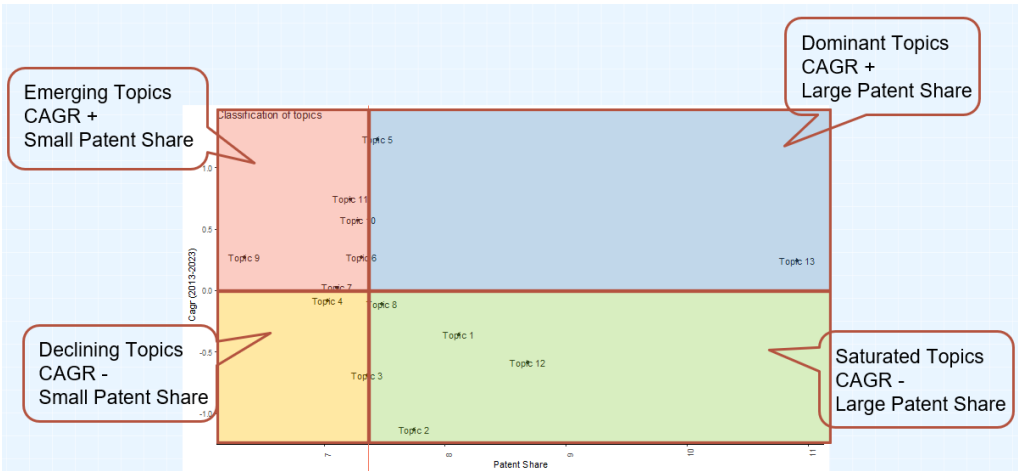
Table 1. *Topic information*

Topic No	Topic Title	Share (2023)	$CAGR$ (2013 – 2023)
1	Wearables, monitoring systems and genetic anomalies/cancer detection	8,109826	-0,36195
2	Blood pressure and blood vessels diagnostics	7,741197	-1,1312
3	Diagnostic ultrasound systems and risk measuring methods	7,353502	-0,68917
4	Artificial intelligence methods and neural processing	7,029363	-0,08043
5	Models and systems for therapy treatment plans in specific conditions (e.g. radiation, diabetes)	7,442481	1,229627
6	Monitoring systems for patient care, condition and status	7,309012	0,26861
7	Bio-enhanced management environments and systems	7,105631	0,027456
8	Anatomical anomalies and drug delivery	7,47426	-0,1095
9	Scanners, optical devices and data mining procedures	6,342952	0,269275
10	Image processing models and methods	7,28359	0,571198
11	Tissue analysis and cardiac monitoring	7,220033	0,744442
12	Deep Learning and neural networks for advanced reporting and evaluations	8,681836	-0,58146
13	Devices and information systems for data, records and case classification (e.g. pneumonia)	10,90632	0,243823

An early inspection of the above table reveals that the evolution of topics is evenly distributed, with some topics having a positive growth (for example 5, 6, 7) and some topics having a slight decline (topic 1,2,12). In addition, can also be observed that the

extracted topics cover a wide range of areas, with wearables, AI, treatment and monitoring plans, blood pressure diagnostics as well as anatomical issues, image processing, tissue analysis and deep learning. This is an indicator that medical diagnosis patents cover a wide range of fields. Moreover, according to the two thresholds, discussed in Section 3, the topic classifications are presented in Figure 5.

Figure 5. *Classification of topics*



The first area, where positive $CAGR$ and small patent share evaluations are apparent, depicts emerging topics that capture the interest of investors. Here the notable topics are the Scanners and Data Mining (9), Tissue Analysis (11) and Bioenhanced Management Systems (7) among others. The second area, where negative $CAGR$ and small patent share evaluations are apparent, depicts declining topics that have lost the interest of investors over time. Paradoxically, here the topic that is relevant to AI appears, but this may be attributed to the fact that patents with AI objectives may belong to other topics as well.

The third area, where positive $CAGR$ and large patent share evaluations are apparent, depicts dominant topics that drive innovation and retain their position of importance in a period of 10 years. The two dominant topics are associated with therapy treatment plans and specific conditions along with data records and case classification. These topics retain their position possibly due to their combination with innovative technologies.

Finally, the fourth area, where negative $CAGR$ but large patent share evaluations are apparent, depicts saturated topics that, while important, may have an overabundance of ideas, contributions and patens leading to confusion and mixed solutions. Multiple topics belong here such as wearables, blood pressure diagnostics, Anatomical anomalies and Deep learning. These topics are far from trivial but may need careful

inspection of the patent grants to discover the truly innovative ones. Finally, two topics Patient Care and Ultrasound and Risk Measuring balance between Emerging/Dominant and Declining/Saturated.

4.2 RQ_2: Which are the leading organizations in medical diagnosis patenting?

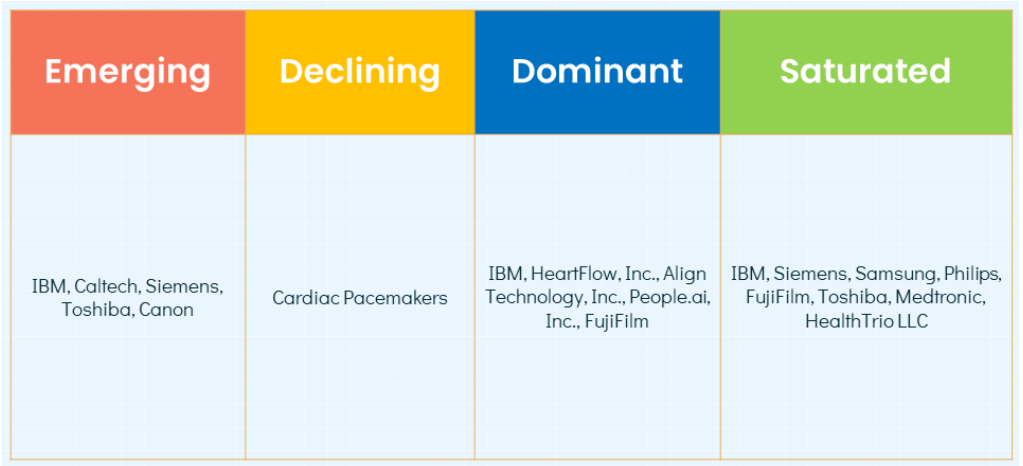
In this second Section, the top competitors in each topic are traced, where the following table (Table 2) reveals that multinational corporations such as IBM, Philips, Toshiba, Canon, and Siemens lead the way in patenting. Eventually, the main activities of the leading assignees reveal that a variety of technology domains are combined in developing medical diagnosis ICT. Overall, apart from the multinational corporations, other assignees, such as HeartFlow, Inc., Cardiac Pacemakers, Inc. and Medtronic, which are dedicated to developing products/innovations that are more relevant to the general field of medical technologies and devices, encompass medical diagnosis ICT as well.

Table 2. *Top assignees*

Topic No	Top assignees
1	IBM, Philips, FujiFilm, Toshiba
2	IBM, Siemens, Medtronic
3	IBM, ZOLL Medical Corporation
	Cardiac Pacemakers, Inc., IBM
5	HeartFlow, Inc., Align Technology, Inc., People.ai, Inc.
6	IBM, uBiome, Inc., General Electric Company
7	CERNER INNOVATION, INC., Hill-Rom Services, Inc.
8	IBM, Siemens, Samsung
9	IBM, Caltech, Siemens
10	IBM, Toshiba
11	IBM, Magic Leap, Canon
12	IBM, HealthTrio LLC
13	IBM, DexCom, Inc., FujiFilm

However, the next figure (Figure 6) is more important as the companies are distributed with respect to the different topic categories. Here it is evident that IBM participates in almost every topic category, indicating a top competitor in this patent dataset, while the Declining topic class involves only one competitor, Cardiac Peacemakers. In the Dominant category, an interesting finding is that AI companies, such as People.ai and Align Technology, have a significant involvement while companies with expertise in electronics mainly participate in the saturated topics (Siemens, Samsung, Philips, Toshiba).

Figure 6. *Top assignees per topic*



5. CONCLUSIONS

In this Section, the main conclusions of this study are presented, as indicated by the respective outcomes. Some potential ideas for future work that apply to the general field of patent analysis are also discussed. As shown by the analysis of this study, innovative technologies, and products are usually patented by large and small-medium companies that are relevant to the investigated domain, i.e., medical diagnosis, while also by technology companies that expand their interest in multiple areas of interest through AI and electronic products/technologies. Regarding the leading competitors, it was revealed that IBM is the main stakeholder which is also involved in the majority of the topics, while the following companies, in terms of overall involvement, include Samsung, Siemens and Toshiba.

As far as the topics are concerned, the primary extracted topics revolved around AI, medical procedures such as Tissue Analysis and Patient Care along with data mining. The drivers of innovation, i.e. dominant topic (topic 13), however seem to be the development of effective treatment plans and case classification. At the same time, the *therapy treatment plans*, i.e. topic 5, constitutes an interesting domain for future research as it obtains a relatively high patent share and carries the highest $CAGR$.

Thus, according to the findings of this study, it is suggested that researchers and companies should develop their experiments and future innovations around these main technology areas which seem to attract many activities. In addition, future investors and inventors should be encouraged to study the strategies and developed technologies of the leading assignees, i.e., assignees that are related to emerging and dominant

topics, in order to establish and obtain a significant strategic positioning in the medical diagnosis ICT.

6. FUTURE WORK

Regarding the future research, it should be mentioned that although the analysis of this study addressed some important research questions and provided valuable insights, there are still some unexplored research gaps and patent properties. Prior research showed that the various citation indicators can be used as knowledge basis in assessing the technological value and influence of patents, topics and assignees. In addition, while the titles and abstracts of the patents are usually leveraged to identify the main concepts of a patent dataset, the patent classifications also contain valuable information that can be used in a similar manner. Thus, future research can leverage this information to discover some additional or more detailed areas of interest, that were possibly not detected by the employed approach, complementing in this way the findings of this study.

Furthermore, despite the inclusiveness of $USPTO$ in many technological fields, it is reasonable that a single patent office does not cover all possible aspects of the relevant technologies, as many innovations that are either patented in different continents or not patented at all are overseen. Hence, future research may focus to including multiple data sources that can provide valuable information on medical diagnosis ICT. Also, since a single CPC subgroup was used to collect relevant data, it is proposed that additional subgroups or keywords should be evaluated to collect and analyze supplementary information as well as provide a more general overview.

ΠΕΡΙΛΗΨΗ

Η ιατρική έρευνα και οι τεχνολογία παίζουν ένα κρίσιμο ρόλο στην ευημερία της ανθρωπότητας, καθώς διασφαλίζουν ότι οι άνθρωποι έχουν πρόσβαση σε αποτελεσματική υγειονομική περίθαλψη. Η διάγνωση ασθενειών και η εφαρμογή προληπτικών μέτρων, σε συνδυασμό με τον έγκαιρο εντοπισμό και περιορισμό πανδημιών και επιδημιών, έχουν έρθει στο προσκήνιο τον τελευταίο καιρό, ιδιαίτερα κατά τη διάρκεια της πανδημίας COVID-19. Ιατρικές εταιρείες και τεχνολογικοί γίγαντες προσπαθούν να παρέχουν λύσεις Πληροφορικής για την ιατρική διάγνωση, που μπορούν να αξιοποιηθούν σε πληθώρα περιπτώσεων. Καθώς ο ανταγωνισμός είναι υψηλός και κάθε εταιρεία επωφελείται και εξαγάγει κέρδος από αυτές τις τεχνολογίες, πολλοί οργανισμοί προστατεύουν τις εφευρέσεις τους μέσω πατεντών. Οι πατέντες αποτελούν υψηλά αποτελεσματικό μέσο προστασίας των εμπορικών δικαιωμάτων μιας εφεύρεσης, ενώ έχουν αποδειχθεί επίσης εξαιρετικοί δείκτες της τεχνολογικής καινοτομίας και προόδου. Οι πληροφορίες που κρύβονται στις αιτήσεις πατεντών, οι

οποίες είναι συχνά σε κειμενική μορφή, αποκαλύπτουν πολύτιμες γνώσεις σχετικά με τις θεματικές τάσεις στις χορηγούμενες πατέντες, καθώς και στον εντοπισμό οργανισμούς και τις εταιρείες που έχουν το υψηλότερο μερίδιο αγοράς στις πατέντες ιατρικής διάγνωσης. Σε αυτήν τη μελέτη, συλλέγουμε ένα σύνολο δεδομένων ευρεσιτεχνιών που σχετίζονται με τις τεχνολογίες Πληροφορικής για ιατρική διάγνωση και παρακολούθηση επιδημιών από το Γραφείο Ευρεσιτεχνιών και Εμπορικών Σημάτων των Ηνωμένων Πολιτειών (USPTO) και εφαρμόζουμε αλγορίθμους pseudo-document based Topic Modeling (PTM), Correlated Topic Models (CTM), Hierarchical Pachinko Allocation (HPAM) και Latent Dirichlet Allocation (LDA). Ο βέλτιστος αλγόριθμος επιλέγεται στη συνέχεια για την ανάλυση των θεμάτων, ενώ σε συνδυασμό με έναν ειδικό δείκτη για την ανάπτυξη θεμάτων κατά τη διάρκεια των ετών. Τα αποτελέσματα από την ανάλυση των θεμάτων δείχνουν ότι οι πατέντες ιατρικής διάγνωσης καλύπτουν πολλούς θεματικούς άξονες, που κυμαίνονται από την πρόληψη ασθενειών (καρκίνος, διαβήτης, νευρολογική διέγερση) και την ανθρώπινη ευημερία (κιρκαδιανό ρυθμός, οδοντιατρικές πρακτικές, ορθοπεδικά) έως πιο ειδικούς ιατρικούς τομείς (διαταραχές ύπνου και διατροφής, εκφοβισμός) και τις καινοτόμες τεχνολογίες (έξυπνες συσκευές, βιομετρική επαυξημένη πραγματικότητα, διάγνωση μικροχλωρίδας).

REFERENCES

Abbas, A., Zhang, L., & Khan, S. U. (2014). A literature review on the state-of-the-art in patent analysis. World Patent Information, 37, 3-13.

Albino, V., Ardito, L., Dangelico, R. M., & Petruzzelli, A. M. (2014). Understanding the development trends of low-carbon energy technologies: A patent analysis. Applied energy, 135, 836-854.

Ardito, L., D'Adda, D., & Petruzzelli, A. M. (2018). Mapping innovation dynamics in the Internet of Things domain: Evidence from patent analysis. Technological Forecasting and Social Change, 136, 317-330.

Bakator, M., & Radosav, D. (2018). Deep learning and medical diagnosis: A review of literature. Multimodal Technologies and Interaction, 2(3), 47.

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

Bardi M., Mortagy A. and Alsayed A. (1998). A multi-objective model for locating fire stations. *European Journal of Operations Research,* **110**, 243-260.

Blei, D., & Lafferty, J. (2006). Correlated topic models. Advances in neural information processing systems, 18, 147.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.Choi, D., & Song, B. (2018). Exploring technological trends in logistics: Topic modeling-based patent analysis. Sustainability, 10(8), 2810.

Choi, D., & Song, B. (2018). Exploring technological trends in logistics: Topic modeling-based patent analysis. Sustainability, 10(8), 2810.
Chui, K. T., Alhalabi, W., Pang, S. S. H., Pablos, P. O. D., Liu, R. W., & Zhao, M. (2017). Disease diagnosis in smart healthcare: Innovation, technologies and applications. Sustainability, 9(12), 2309.
Dieng, A. B., Ruiz, F. J., & Blei, D. M. (2020). Topic modeling in embedding spaces. Transactions of the Association for Computational Linguistics, 8, 439-453.
Evangelista, A., Ardito, L., Boccaccio, A., Fiorentino, M., Petruzzelli, A. M., & Uva, A. E. (2020). Unveiling the technological trends of augmented reality: A patent analysis. Computers in Industry, 118, 103221.
Flores, N., Reyna, M. A., Avitia, R. L., Cardenas-Haro, J. A., & Garcia-Gonzalez, C. (2022). Non-Invasive Systems and Methods Patents Review Based on Electrocardiogram for Diagnosis of Cardiovascular Diseases. Algorithms, 15(3), 82.
He, J., Baxter, S. L., Xu, J., Xu, J., Zhou, X., & Zhang, K. (2019). The practical implementation of artificial intelligence technologies in medicine. Nature medicine, 25(1), 30-36.
Kwon, K., Jun, S., Lee, Y. J., Choi, S., & Lee, C. (2022). Logistics technology forecasting framework using patent analysis for technology roadmap. Sustainability, 14(9), 5430.
Lee, D., & Yoon, S. N. (2021). Application of artificial intelligence-based technologies in the healthcare industry: Opportunities and challenges. International Journal of Environmental Research and Public Health, 18(1), 271.
Liu, N., Shapira, P., Yue, X., & Guan, J. (2021). Mapping technological innovation dynamics in artificial intelligence domains: Evidence from a global patent analysis. Plos one, 16(12), e0262050.
Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., ... & Adam, S. (2021). Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. In Computational methods for communication science (pp. 13-38). Routledge.
Mimno, D., Li, W., & McCallum, A. (2007, June). Mixtures of hierarchical topics with pachinko allocation. In Proceedings of the 24th international conference on Machine learning (pp. 633-640). ACM.
Röder, M., Both, A., & Hinneburg, A. (2015, February). Exploring the space of topic coherence measures. In Proceedings of the eighth ACM international conference on Web search and data mining (pp. 399-408).
Sampaio, P. G. V., González, M. O. A., de Vasconcelos, R. M., dos Santos, M. A. T., de Toledo, J. C., & Pereira, J. P. P. (2018). Photovoltaic technologies: Mapping from patent analysis. Renewable and Sustainable Energy Reviews, 93, 215-224.
Soni, P., Sharma, R., & Dubey, A. (2022). Patent Landscape of COVID-19 Innovations: A Comprehensive Review.

Tian, S., Yang, W., Le Grange, J. M., Wang, P., Huang, W., & Ye, Z. (2019). Smart healthcare: making medical care more intelligent. Global Health Journal, 3(3), 62-65.

Wang, J., & Hsu, T. Y. (2023). Early discovery of emerging multi-technology convergence for analyzing technology opportunities from patent data: the case of smart health. Scientometrics, 1-30.

Zuo, Y., Wu, J., Zhang, H., Lin, H., Wang, F., Xu, K., & Xiong, H. (2016, August). Topic modeling of short texts: A pseudo-document view. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 2105-2114).

Greek Statistical Institute
Proc. of the 35th Panhellenic & 1st International Statistics Conf. (2023), pp.181-197

COMPARISON OF DIMENSIONALITY REDUCTION AND CLUSTERING METHODS ON THE MULTIDIMENSIONAL DATASET "FOREST COVER TYPE" WITH MIXED-TYPE DATA

Zacharenia Kyrana[1*], Emmanouil Pratsinakis[1], Nikolaos Papafilippou[1], Christos Dordas[1], Angelos Markos[2], George Menexes[1]
[1]Laboratory of Agronomy, School of Agriculture, Aristotle University of Thessaloniki, 54124 Thessaloniki
{kyranaze, epratsina, npapafil, chdordas, gmenexes}@agro.auth.gr
[2]Department of Primary Education, Democritus University of Thrace, 68100 Alexandroupoli
amarkos@eled.duth.gr

ABSTRACT

Multivariate datasets that contain mixed-type data offer researchers various opportunities to employ diverse statistical methods for dimensionality reduction and clustering. This study aims to compare several dimensionality reduction methods, including Principal Components Analysis, Factor Analysis, Multiple Correspondence Analysis, Categorical Principal Components Analysis with optimal scaling and Factor Analysis for Mixed Data. Additionally, the study examines clustering methods, including *k*-Means and Hierarchical Clustering. These methods are applied to the "Forest Cover Type" dataset, which pertains to the forest cover type within four wilderness areas. The dataset comprises 581,012 objects, obtained from forest sections measuring 30m × 30m, and it includes 55 mixed-type variables. Throughout the study, various strategies are implemented concerning the selection of the measurement scales for input variables and the coding of variable values. The aim of this study is to compare the aforementioned methods and strategies in terms of results and execution times, using Python and SPSS. The disadvantages of dimensionality reduction methods include the "curse of dimensionality", which pertains to determining the number of significant dimensions, the increased computational demands, the lack of software code for certain methods, variation in result calculations between software packages, and the limitations of some software in handling numerous binary coded variables or running Parallel Analysis. Clustering methods have their own disadvantages including software limitations in extracting Hierarchical Clustering results (and dendrograms), and the dependence of *k*-Means on the chosen analysis strategy.

Keywords: Multivariate data; Big data; Principal Components Analysis; Factor Analysis; Multiple Correspondence Analysis; Categorical Principal Components Analysis with optimal scaling; Factor Analysis for Mixed Data; Partitioning Clustering; Hierarchical Clustering

1. INTRODUCTION

Scientific research often relies on complex multidimensional and multivariate datasets, commonly known as big data, which contain a mix of data types including categorical (nominal, ordinal), binary, circular and scale variables [Hendrickson (2014); Menexes (2006)]. The challenge with using mixed-type data lies in effectively applying them to multivariate data analysis. To address this, two prominent schools of data analysis have proposed contrasting approaches. The Dutch School suggests converting categorical data into scale data [Van Rijckevorsel and De Leeuw (1988); Young (1981)], while the French School recommends transforming scale data into categorical data [Menexes (2006); Papadimitriou (1994)]. However, these conversions may lead to a loss of information compared to the original data [Menexes (2006)].

The development of multivariate and multidimensional analysis, including dimensionality reduction and clustering methods, traces its roots back to the French and Dutch Schools of data analysis and it continues to evolve in the fields of Statistics and Machine Learning. As a result, numerous methods and their extensions exist, demonstrating the intricate complexity of dimensionality reduction and clustering techniques [Cunningham and Ghahramani (2015)].

Dimensionality reduction methods aim to produce a compact representation of the original dataset that retains critical and useful information while reducing its overall "volume" [Sharma (1996)]. However, these methods also come with inherent disadvantages, such as the "curse of dimensionality" [Bellman (1961); Nguyen and Holmes (2019)], challenges in determining the number of important dimensions, high computing time requirements, handling missing values [Hair et. al. (2010)], outliers and extreme values [Menexes (2006); Hair et. al. (2010)] and difficulties in visualizing the results [Menexes (2006)].

On the other hand, clustering methods aim is to partition a set of objects into clusters based on pre-defined criteria-variables with the goal of grouping similar objects together and distinguishing them from objects in other clusters [Ahmad and Khan (2019); Jain and Dubes (1988)]. However, a significant limitation of clustering methods is their inability to handle datasets that consist of mixed-type data [Ahmad and Khan (2019)].

This study seeks to investigate the application of five dimensionality reduction and two clustering methods to a multidimensional mixed-type dataset. The goal is to compare these methods, in terms of their results and execution times, using two different statistical software packages. Through this comparison we aim to highlight significant computational and interpretive disadvantages of these methods.

2. MATERIALS AND METHODS

2.1 Methods and Computer system specifications

The dimensionality reduction methods employed in this study include Principal Components Analysis (PCA) [Anderson (1984); Hair et. al. (2010); Jolliffe (2002)], Factor Analysis (FA) [Anderson (1984); Hair et. al. (2010); Jolliffe (2002)], Multiple Correspondence Analysis (MCA) [Menexes (2006); Messaoud et. al. (2007)], Categorical Principal Components Analysis with optimal scaling (CATPCA) [Gifi (1990); Markos et. al. (2020); Menexes (2006)] and Factor Analysis for Mixed Data (FAMD) [Kassambara (2017); Markos et. al. (2020); Pagès (2014)]. The clustering methods employed in this study are Partitioning Clustering (*k*-Means) [Izenman (2008); Jobson (1992)] and Hierarchical Clustering (HC) [Hair et. al. (2010); Jobson (1992)]. To implement these methods, various strategies are employed in terms of selecting the main factorial axes, determining the measurement scale of the variables and coding the variable values.

For the implementation and comparison of the above methods, a Dell Vostro 5490 laptop is used, with an Intel(R) Core(TM) i7-10510U CPU @ 1.80GHz (8 CPUs) processor, 2.3GHz and Windows 11 Pro 64-bit (10.0, Build 19043) (19041.vb_release.191206-1406) operating system. The RAM type is DDR4, with a speed of 2666MHz and 16GB memory. The hard drive type is M.2 PCIe NVMe, with a capacity of 512GB and with read and write speeds of 1483MB/sec and 786MB/sec, respectively. The statistical software packages used are Python 3.10, via the Anaconda platform and Jupyter Notebook 6.4.5, and IBM SPSS Statistics v26.0.

2.3 Dataset

The dataset used in this study is the "Forest Cover Type" dataset [UCI Machine Learning Repository]. This dataset concerns the forest cover types found in four wilderness areas of the Roosevelt National Forest in northern Colorado, USA. These four wilderness areas exhibit minimal human disturbance, resulting in the presence of seven forest cover types that are primarily influenced by ecological processes rather than forest management practices. The dataset comprises 581,012 objects, obtained from forest sections measuring 30m × 30m, and consists of 55 variables of mixed types. Specifically, the "Cover Type" variable is nominal variable and contains seven categories. The "Wilderness Area" variable is also nominal with four categories or can be represented by three independent binary dummy variables (coded as 0 or 1). Similarly, the "Soil Type" variable can be treated as a nominal variable with 40 categories or as 39 independent binary dummy variables (coded as 0 or 1). The "Elevation" (m) of the forests and the four variables related to the "Distances of forests from four nearby landmarks" (m) are scale variables. The three "Hillshade" variables at 9am, noon and 3pm are also scale variables measured on an index ranging from 0 to 255. Lastly, the "Aspect" and "Slope" variables, which represent the orientation and

steepness of the forests are scale variables that fall under the category of circular data. Their unit of measurement is in degrees (x^o), which are converted to radians (rad) using the following equation:

$$rad = x^o \left(\pi / 180^o\right), \text{ where } \pi \cong 3.14.$$

3. RESULTS

3.1 Dimensionality Reduction

Principal Components Analysis

The PCA method is applied to the dataset's 10 scale variables, as well as to the 40-1=39 binary dummy variables for "Soil Type" and the 4-1=3 binary dummy variables for "Wilderness Area". To account for the different units of measurement in these variables, Pearson's *r* correlation matrix is chosen as the input matrix. Additionally, the rotation of the factorial axes is performed using the Varimax method to enhance the interpretability of the extracted factors.

Consistent results are obtained in both software packages. The execution time for the results is 20 seconds in SPSS and 6.3 seconds in Python. Six different criteria (strategies) are employed to select the most significant principal components. The first criterion selects factorial axes with eigenvalues >1, resulting in the extraction of 38 principal components that explain 89% of the total variance (inertia). The second criterion selects factorial axes with eigenvalues >0.7, yielding 41 principal components that explain 94% of the total variance. The third criterion selects factorial axes with a percentage of explained variance ≥5%, resulting in the extraction of two factorial axes explaining 12% of the total variance. The fourth criterion selects factorial axes with a percentage of total variance explained ≥60%, leading to the extraction of 23 principal components that explain 60% of the total variance. The fifth criterion selects factorial axes based on the eigenvalues' diagram (Scree-plot, Figure 1), which is consistent across both software packages. This criterion extracts four principal components explaining 20% of the total variance. The sixth criterion selects factorial axes using Parallel Analysis, which compares the eigenvalues from the analysis with those resulted from a correlation matrix computed from a random sample with the same dimensions as the "Forest Cover Type" dataset. The application of Parallel Analysis in SPSS is applied using code in SPSS Syntax [O'Connor (2000)]. Factors and eigenvalues exceeding those of the random sample are selected. Using this criterion, Python extracts 38 principal components explaining 89% of the total variance (Table 1). However, SPSS fails to reproduce results.

Figure 1. *Scree-plot of PCA eigenvalues. The circle marks the point where the curve shows an elbow and indicates the number of the most important principal components (source: SPSS).*

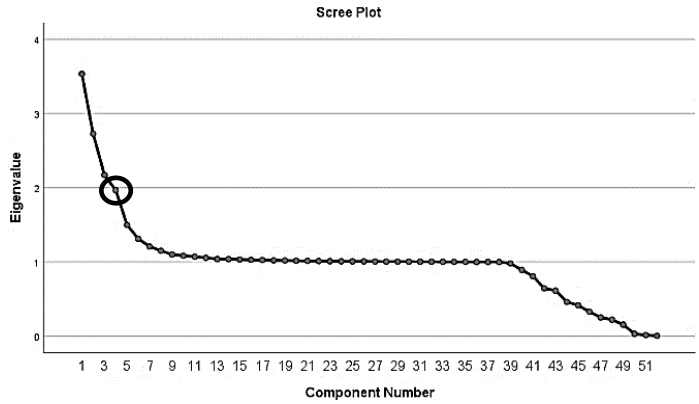
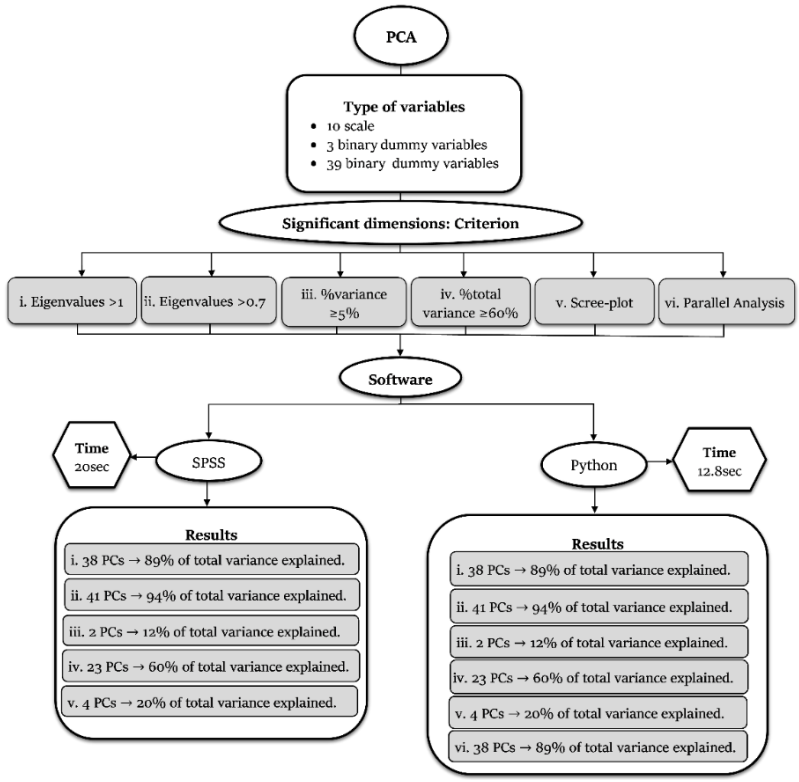


Table 1. *Table of PCA eigenvalues and eigenvalues of the random sample based on Parallel Analysis (source: Python)*

Factors	Eigenvalues of PCA	Eigenvalues of random sample
Factor 1	3.535	1.017
Factor 2	2.727	1.016
⋮	⋮	⋮
Factor 38	1.000	0.993
Factor 39	0.980	0.993
Factor 40	0.893	0.992
⋮	⋮	⋮

A comprehensive presentation of the PCA analysis and its output is provided in the flowchart of Figure 2.

Figure 2. *A comprehensive overview of PCA performed on a 581,012×52 data matrix. The analysis includes 10 scale variables and 42 binary dummy variables. The figure provides information on the types of variables involved in the analysis, the criteria used for selecting factorial axes, the statistical software packages employed for the analysis, the execution time of results in each software, and the number of extracted principal components (PCs) along with the corresponding percentages of total variance explained.*



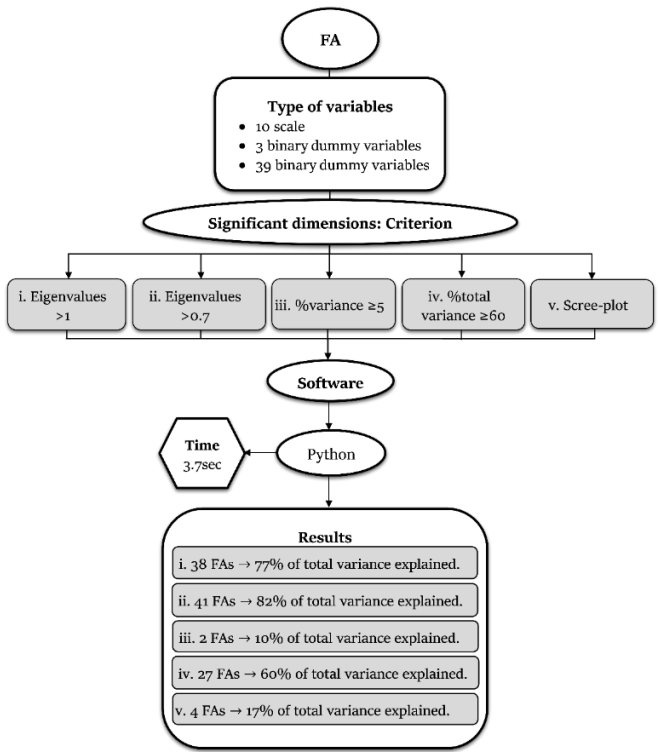
Factor Analysis

The FA method is applied to the dataset's 10 scale variables, as well as the 40-1=39 binary dummy variables for "Soil Type" and the 4-1=3 binary dummy variables for "Wilderness Area". Pearson's *r* correlation matrix is selected as the input matrix and the axes rotation is performed using the Varimax method. It is worth noting that both PCA and FA attempt to reduce the mathematical dimensions of a dataset, but whereas PCA assumes that the common variance is equal to the total variance, FA assumes that the total variance is split into common and unique variance [Fabrigar et. al. (1999)].

However, SPSS fails to extract results regardless of the number of factorial axes extracted and the number of iterations selected. This limitation is attributed to the extensive number of dummy variables involved in the analysis. Conversely, the execution time for obtaining the results in Python is 3.7 seconds. To select the most significant factors, five different criteria (strategies) are employed. The first criterion involves selecting factorial axes with eigenvalues >1, resulting in the extraction of 38 factorial axes that explain 77% of the total variance. The second criterion selects factorial axes with eigenvalues >0.7, leading to the extraction of 41 factorial axes that explain 82% of the total variance. The third criterion involves selecting factorial axes with a percentage of explained variance ≥5%, resulting in the extraction of two factorial

axes explaining 10% of the total variance. The fourth criterion selects factorial axes with a percentage of total variance explained ≥60%, leading to the extraction of 27 factorial axes explaining 60% of the total variance. The fifth criterion involves selecting factorial axes based on the scree-plot of eigenvalues, which is the same as in the PCA method (Figure 1). This criterion results in the extraction of four factorial axes explaining 17% of the total variance. Figure 3 presents a comprehensive overview of the FA analysis, along with the FA output.

Figure 3. *A comprehensive overview of FA performed on a 581,012×52 data matrix. The analysis includes 10 scale variables and 42 binary dummy variables. The figure provides information on the types of variables involved in the analysis, the criteria used for selecting factorial axes, the statistical software packages employed for the analysis, the execution time of results in each software, and the number of extracted factorial axes along with the corresponding percentages of total variance explained.*



Multiple Correspondence Analysis

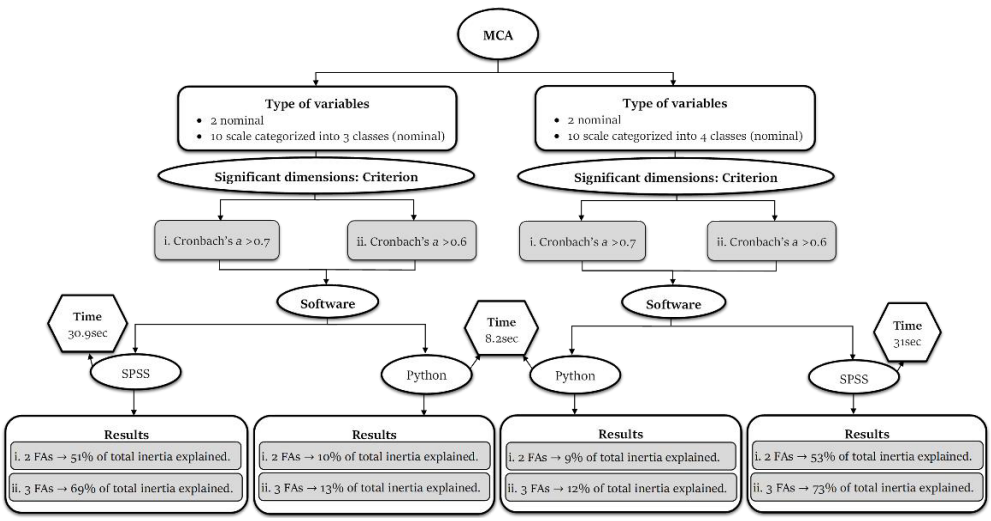
MCA is applied to the two nominal variables ("Wilderness Area" and "Soil Type" with four and 40 categories, respectively) and the 10 discretized scale variables of the dataset. "Cover Type" (seven categories) is included as a supplementary variable,

which does influence the analysis results. In SPSS, the selection of the most important factorial axes is based on the Cronbach's *alpha* criterion with thresholds of $a \geq 0.7$ and $a \geq 0.6$. However, no code for extracting Cronbach's *alpha* values is found in Python. As a result, two and three important factorial axes are extracted based on the SPSS results. Two strategies are employed to categorize the scale variables.

In the first strategy, the values of each scale variable are divided into three intervals or classes, ensuring that 33% of the values fall into each class. The execution time for obtaining the results is 30.9 seconds in SPSS and 8.2 seconds in Python. For $a \geq 0.7$, two factorial axes are extracted explaining 51% of the total inertia in SPSS and 10% in Python. For $a \geq 0.6$, three factorial axes are extracted explaining 69% of the total inertia in SPSS and 13% in Python.

In the second strategy, the values of each scale variable are divided into four intervals or classes, with 25% of the values falling into each class. The execution time of results is 31 seconds in SPSS and 8.2 seconds in Python. For $a \geq 0.7$, two factorial axes are extracted, explaining 53% of the total inertia in SPSS and 9% in Python. For $a \geq 0.6$, three factorial axes are extracted, explaining 73% of the total inertia in SPSS and 12% in Python. A comprehensive presentation of the MCA analysis and its output is provided in the flowchart of Figure 4.

Figure 4. *A comprehensive overview of MCA performed on a 581,012×12 data matrix. The analysis includes 12 nominal variables. The figure provides information on the types of variables involved in the analysis, the criteria used for selecting factorial axes, the statistical software packages employed for the analysis, the execution time of results in each software, and the number of extracted factorial axes along with the corresponding percentages of total variance explained.*



Categorical Principal Components Analysis with optimal scaling

The CATPCA method is employed using three different strategies, considering the measurement scale of the variables. Additionally, the variable “Cover Type” (with seven categories) is included as a supplementary variable. In SPSS, the selection of the most significant factorial axes is based on the Cronbach's *alpha* criterion, with a threshold of $a \geq 0.7$ and $a \geq 0.6$, respectively. However, there is no available Python code to perform CATPCA.

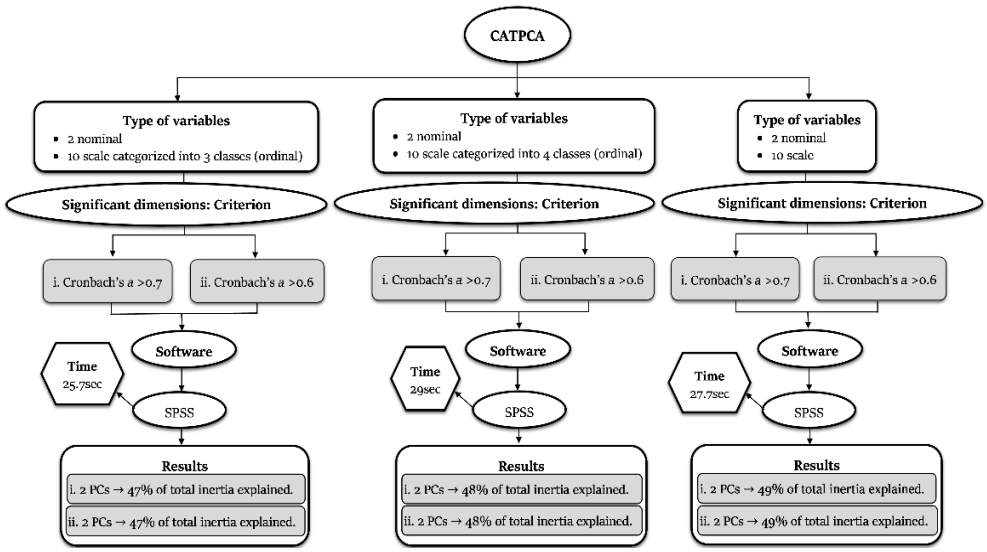
In the first strategy, the analysis includes two nominal variables (“Wilderness Area” and “Soil Type”) with four and 40 categories, respectively and 10 scale variables from the dataset. The execution time for obtaining the results in SPSS is 27.7 seconds. For $a \geq 0.7$ and $a \geq 0.6$, two principal components are extracted, explaining 49% of the total variance.

In the second strategy, the analysis includes the same two nominal variables (“Wilderness Area” and “Soil Type”) along with the 10 categorized scale variables from the dataset, which are treated as ordinal variables. Two different strategies are used to discretize the scale variables:

The first strategy divides the values of each scale variable into three intervals with 33% of the values in each class. The execution time for obtaining the results in SPSS is 25.7 seconds. For $a \geq 0.7$ and $a \geq 0.6$, two principal components are extracted, explaining 47% of the total variance in SPSS.

The second strategy divides the values of each scale variable into four intervals, with 25% of the values falling into each class. The execution time for obtaining the results in SPSS is 29 seconds. For $a \geq 0.7$ and $a \geq 0.6$, two principal components are extracted, explaining 48% of the total variance in SPSS. A comprehensive overview of the CATPCA analysis can be found in Figure 5.

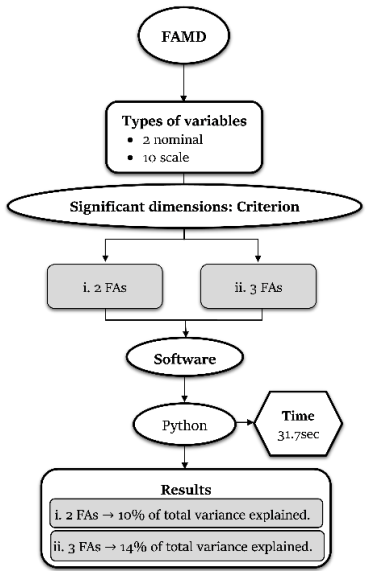
Figure 5. *A comprehensive overview of CATPCA performed on a 581,012×12 data matrix. The analysis includes two nominal and 10 ordinal or scale variables. The figure provides information on the types of variables involved in the analysis, the criteria used for selecting factorial axes, the statistical software packages employed for the analysis, the execution time of results in each software, and the number of extracted factorial axes along with the corresponding percentages of total variance explained.*



Factor Analysis for Mixed Data

The FAMD method is applied to the two nominal variables ("Wilderness Area" and "Soil Type" with four and 40 categories, respectively) and to the 10 scale variables of the dataset. In SPSS no option or code in the Syntax is found to run FAMD, while in Python no code is found to extract Cronbach's *alpha* values. The execution time of the results is 31.7 seconds in Python. Two different criteria (strategies) are used to select the most important factorial axes, based on the results of the MCA and CATPCA methods. In the first strategy two factorial axes are extracted that explain 10% of the total variance in Python. In the second strategy three factorial axes are extracted that explain 14% of the total variance in Python. A comprehensive overview of the FAMD analysis can be found in Figure 6.

Figure 6. *A comprehensive overview of Factor Analysis for Mixed Data (FAMD) performed on a 581,012×52 data matrix. The analysis includes two nominal and 10 scale variables. The figure provides information on the types of variables involved in the analysis, the criteria used for selecting factorial axes, the statistical software packages employed for the analysis, the execution time of results in each software, and the number of extracted factorial axes along with the corresponding percentages of total variance explained.*



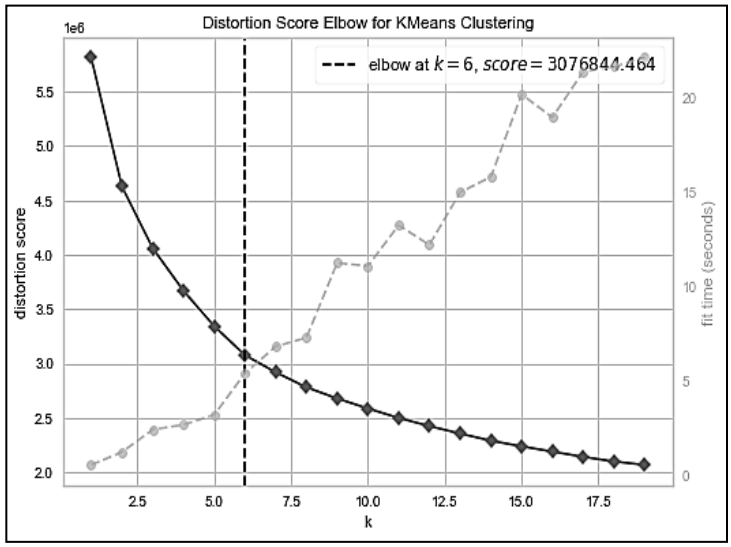
3.2 Clustering

Partitioning Clustering

The *k*-Means method is applied with five different strategies. More specifically:

In the first strategy, the method is applied to the 10 scale variables of the dataset. Using the Elbow Score criterion [Tripathi et. al. (2018)], as shown in Figure 7, six clusters are selected that include 24.8%, 3.7%, 17.1%, 12, 5%, 34.3% and 7.5% of the objects, respectively. Based on the coefficients η^2 [Michos et. al. (2012)], the horizontal distance of forests from roads has the greatest contribution to the formation of the clusters (η^2=0.831), followed by the horizontal distance of forests from fire points (η^2= 0.776). The execution time of the results is 5.7 seconds in SPSS and 1.2 seconds in Python.

Figure 7. *Score elbow plot for determining the number of clusters (source: Python).*



In the second strategy, the method is applied to the two factorial axes resulting from MCA (after categorizing the 10 scale variables into three classes, as nominal variables), based on the Cronbach's *alpha* criterion for $a \geq 0.6$ [Menexes (2006)], which explain 51% of the total inertia in SPSS. Using the Score Elbow criterion, four clusters (C1-C4) are selected that include 35.2%, 17.7%, 17.3%, and 29.8% of the objects, respectively. Based on the coefficients η^2, the first factor has the greatest contribution to the formation of the clusters (η^2=0.889), followed by the second factor (η^2=0.732). The execution time of the results is 2.8 seconds in SPSS and 0.8 seconds in Python.

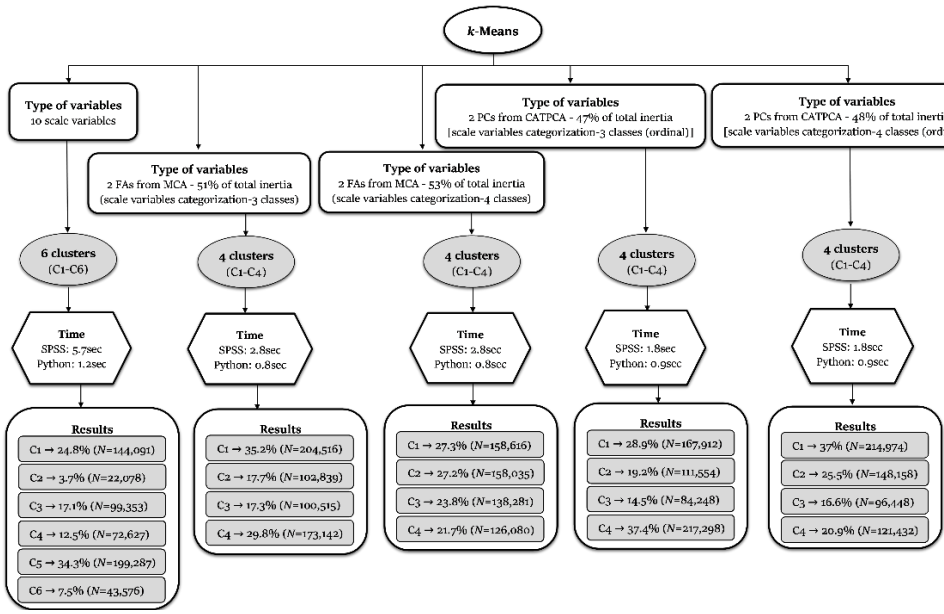
In the third strategy, *k*-Means is applied to the two factorial axes resulting from MCA (after categorizing the 10 scale variables into four classes, as nominal variables), based on the Cronbach's *alpha* criterion for $a \geq 0.6$, which explain 53% of the total inertia in SPSS. Using the Score Elbow criterion, four clusters (C1-C4) are selected that include 27.3%, 27.2%, 23.8%, and 21.7% of the objects, respectively. Based on the coefficients η^2, the first factor has the greatest contribution to the formation of the clusters (η^2=0.851), followed by the second factor (η^2=0.738). The execution time of the results is 2.8 seconds in SPSS and 0.8 seconds in Python.

In the fourth strategy, the method is applied to the two factorial axes resulting from CATPCA (after categorizing the 10 scale variables into three classes, as ordinal variables), based on the Cronbach's *alpha* criterion for $a \geq 0.6$, which explain 47% of the total inertia in SPSS. Using the Score Elbow criterion, four clusters (C1-C4) are selected that include 28.9%, 19.2%, 14.5%, and 37.4% of the objects, respectively. Based on the coefficients η^2, the first principal component has the greatest contribution to the formation of the clusters (η^2=0.892), followed by the second principal component

(η^2=0.702). The execution time of the results is 1.8 seconds in SPSS and 0.9 seconds in Python.

In the fifth strategy, *k*-Means is applied to the two factorial axes resulting from CATPCA (after categorizing the 10 scale variables into four classes, as ordinal variables), based on the Cronbach's *alpha* criterion for $a \geq 0.6$, which explain 48% of the total inertia in SPSS. Using the Score Elbow criterion, four clusters (C1-C4) are selected that include 37.0%, 25.5%, 16.6%, and 20.9% of the objects, respectively. Based on the coefficients η^2, the first principal component has the greatest contribution to the formation of the clusters (η^2=0.839), followed by the second principal component (η^2=0.687). The execution time of the results is 1.8 seconds in SPSS and 0.9 seconds in Python. A comprehensive overview of the *k*-Means analysis can be found in Figure 8.

Figure 8. *Flowchart of k-Means (types of variables involved in the analysis, number of selected clusters based on Score Elbow criterion, execution time of results in each software, percentages and number of objects in each cluster).*



Hierarchical Clustering (HC)

The HC method with the respective dendrograms cannot be performed as both software packages "crashed", due to the large sample size.

4. CONCLUSIONS

This study shows that at least five dimensionality reduction methods and one clustering method can be used for the same dataset, combined with different analysis strategies in

terms of how the data are coded and how the measurement scale of the variables is determined. It also shows that it is necessary to use several analysis strategies, until a satisfactory dimensionality reduction and clustering solution is found to strengthen the validity of the results. Furthermore, Python has the fastest execution times.

Regarding the dimensionality reduction methods, the above analyses show that for the “Forest Cover Type” dataset, the most satisfactory reduction of the mathematical dimensions is achieved with MCA and CATPCA. When the Parallel Analysis is applied, SPSS is unable to extract results, possibly due to the limited memory space of the system. Also, when FA is applied, SPSS “crashes” due to the large number of dummy variables involved in the analysis. Specifically, the Varimax rotation of the axes is not performed, and the factorial loadings and scores are not extracted. Furthermore, it is observed that when applying PCA and FA, removing a different dummy variable each time produce different results. There are also slight differences between the results of the PCA and FA methods. These small differences are considered insignificant in the context of an empirical study. In addition, regarding MCA, it is observed that the calculation of the percentage of the explained total variance is differed between SPSS and Python due to different method of inertia correction used or/and the type of input matrix (Burt matrix or logical matrix with binary 0-1 coding). Additionally, for MCA and FAMD, no code is found in Python for the values of the Cronbach's *alpha* criterion. Also, no code is found in Python to run the CATPCA method, while the FAMD method is not supported by the SPSS statistical package. For future research, it is proposed to develop code in Python or SPSS Syntax for the CATPCA and FAMD methods, respectively.

Disadvantages of dimensionality reduction methods are identified for this particular dataset, are the "curse of dimensionality", in the sense of determining the number of important dimensions-axes, the increased computing power required in some cases, the lack of software code for certain methods and criteria, the differentiation between the software packages in terms of calculations and by extension the extraction of different results, the collapse of SPSS in terms of managing too many dummy variables and running the Parallel Analysis and finally the difficulty of highlighting the most appropriate method for the reduction of mathematical dimensions.

Regarding the clustering methods, the above analyses show that for the “Forest Cover Type” dataset there are many clustering strategies, as more than one strategy can be used. However, the composition and number of clusters are different depending on the analysis strategy. Regarding the limitations of HC, it seems that the method cannot be performed on this dataset or on datasets of similar size (mainly in terms of the number of cases), since both software packages "crashed". This fact is rather a disadvantage, because HC may not have "meaning" in "Tall" datasets (dendrogram construction), however it gives the possibility to estimate the number of clusters, combining various distances and joining methods. Finally, the main limitation of *k*-Means is that the user cannot choose a distance and clustering method.

ΠΕΡΙΛΗΨΗ

Τα πολυμεταβλητά σύνολα δεδομένων μεικτού τύπου ωθούν τους ερευνητές στην εφαρμογή πολλών στατιστικών μεθόδων μείωσης των διαστάσεων και ταξινόμησης. Στην παρούσα μελέτη οι μέθοδοι μείωσης των διαστάσεων που συγκρίνονται είναι η Ανάλυση σε Κύριες Συνιστώσες, η Παραγοντική Ανάλυση, η Παραγοντική Ανάλυση των Πολλαπλών Αντιστοιχιών, η Κατηγορική Ανάλυση σε Κύριες Συνιστώσες με βέλτιστη κλιμακοποίηση και η Παραγοντική Ανάλυση για Μεικτού τύπου Δεδομένα. Οι μέθοδοι ταξινόμησης που συγκρίνονται είναι η *k*-Means και η Ιεραρχική Ταξινόμηση. Αυτές οι μέθοδοι εφαρμόζονται στο σύνολο δεδομένων "Forest Cover Type", το οποίο αφορά τον τύπο δασικής κάλυψης τεσσάρων περιοχών άγριας φύσης. Το σύνολο δεδομένων αποτελείται από 581.012 αντικείμενα, τα οποία ελήφθησαν από τμήματα δάσους 30m × 30m και από 55 μεταβλητές μεικτού τύπου. Η μελέτη εφαρμόζει διάφορες στρατηγικές ως προς την επιλογή της κλίμακας μέτρησης των μεταβλητών και την κωδικοποίηση των τιμών των μεταβλητών. Σκοπός της παρούσας μελέτης είναι η σύγκριση αυτών των μεθόδων και στρατηγικών όσον αφορά τα αποτελέσματα και τους χρόνους εξαγωγής των αποτελεσμάτων, με τη χρήση της Python και του SPSS. Τα μειονεκτήματα των μεθόδων μείωσης των διαστάσεων είναι η “κατάρα των διαστάσεων”, με την έννοια του καθορισμού του αριθμού των σημαντικών διαστάσεων, η αυξημένη υπολογιστική ισχύς που απαιτείται, η έλλειψη κώδικα των λογισμικών για ορισμένες μεθόδους, η διαφοροποίηση ως προς τους υπολογισμούς μεταξύ των λογισμικών, η κατάρρευση ορισμένων λογισμικών ως προς τη διαχείριση πολλών ψευδομεταβλητών και την εκτέλεση της Parallel Analysis. Τα μειονεκτήματα των μεθόδων ταξινόμησης είναι η αδυναμία εξαγωγής των αποτελεσμάτων της Ιεραρχικής Ταξινόμησης (και των δενδρογραμμάτων) και το γεγονός ότι τα αποτελέσματα της *k*-Means εξαρτώνται από την εκάστοτε στρατηγική ανάλυσης.

REFERENCES

Ahmad A. and Dey L. (2007). A k-mean clustering algorithm for mixed numeric and categorical data. Data and Knowledge Engineering, 63(2), 503-527.

Anderson, T. W. (1984). An Introduction to Multivariate Statistical Analysis, 2nd ed.; John Wiley and Sons Inc.: New York, USA.

Bellman, R. E. (1961). Adaptive Control Processes: A Guided Tour. Princeton University Press: New Jersey, USA.

Cunningham J. P. and Ghahramani Z. (2015). Linear Dimensionality Reduction: Survey, Insights, and Generalizations. Journal of Machine Learning Research, 16(89), 2859−2900. https://doi.org/10.48550/arXiv.1406.0873

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C. and Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. Psychological Methods, 4(3), 272–299. https://doi:10.1037/1082-989x.4.3.272

Gifi, A. (1990). Nonlinear Multivariate Analysis. Wiley: New York, USA.

Izenman, A. J. (2008). Modern multivariate statistical techniques. Regression, Classification and Manifold Learning. XXV. New York: Springer.
Hair, J. F., Black, W. C., Babin, B. J. and Anderson, R. E. (2010). Multivariate Data Analysis: A Global Perspective, 7th ed.; Pearson Education Inc.: New Jersey, USA.
Hendrickson, J. L. (2014). Methods for Clustering Mixed Data. Doctoral dissertation, University of South Carolina, Columbia.
Jain, A. K. and Dubes, R. C. (1988). Algorithms for clustering data. Englewood Cliffs: Prentice Hall.
Jobson, J. D. (1992). Applied multivariate data analysis. Volume II: Categorical and Multivariate methods. New York: Springer-Verlag, Inc.
Jolliffe, I. T. (2002). Principal component analysis, 2nd ed.; Springer: New York, USA.
Kassambara, A. (2017). Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning, 1st ed.; STHDA, Volume 1.
Markos A., Moschidis O. and Chadjipantelis T. (2020). Sequential dimension reduction and clustering of mixed-type data. Int. J. Data Analysis Techniques and Strategies, 12(3), 228-246. https://doi.org/10.1504/IJDATS.2020.108043
Menexes, G. (2006). Experimental Designs in Data Analysis. Doctoral dissertation, University of Macedonia, Thessaloniki.
Messaoud R. B., Boussaïd O., Loudcher-Rabaseda S. (2007). A Multiple Correspondence Analysis to Organize Data Cubes. Frontiers in Artificial Intelligence and Applications, 155(1), 133-146. https://halshs.archives-ouvertes.fr/halshs-00476483
Michos, M. C., Mamolos, A. P., Menexes, G. C., Tsatsarelis, C. A., Tsirakoglou, V. M., & Kalburtji, K. L. (2012). Energy inputs, outputs and greenhouse gas emissions in organic, integrated and conventional peach orchards. *Ecological indicators*, *13*(1), 22-28.
Nguyen L. H. and Holmes S. (2019). Ten quick tips for effective dimensionality reduction. PLoS Computational Biology, 15(6), e1006907. https://doi.org/10.1371/journal.pcbi.1006907
O'connor, B. P. (2000). SPSS and SAS Programs for Determining the Number of Components Using Parallel Analysis and Velicer's MAP Test. Behavior Research Methods, Instruments, & Computers, 32, 396-402. http://dx.doi.org/10.3758/BF03200807
Pagès, J. (2014). Multiple Factor Analysis by Example Using R, 1st ed.; Chapman and Hall/CRC: USA.
Papadimitriou, G. (1994). Data Analysis Methods: University Lectures, University of Macedonia Economics and Social Sciences, Thessaloniki, Greece.
Sharma, S. (1996). Applied Multivariate Techniques. John Willey and Sons Inc.: New York, USA.
Tripathi, S., Bhardwaj, A., & Poovammal, E. (2018). Approaches to clustering in customer segmentation. *International Journal of Engineering & Technology*, *7*(3.12), 802-807.

UCI Machine Learning Repository. Covertype Data Set. Available online: https://archive.ics.uci.edu/ml/datasets/covertype (accessed on 9 March 2023).
Van Rijckevorsel, J. and De Leeuw, J. (1988). Component and Correspondence Analysis. Dimension Reduction by Functional Approximation, John Willey and Sons Ltd.: Chichester, England, pp. 103-114.
Young F. (1981) Quantitative Analysis of Qualitative Data. Psychometrika, 46(4), 357-388. https://doi.org/10.1007/BF02293796

Greek Statistical Institute

Proc. of the 35th Panhellenic & 1st International Statistics Conf. (2023), pp.198-212

A MISSION RELIABILITY REDUNDANCY MODEL FOR AN AIRCRAFT FLEET WITH COLD STANDBY SPARE PARTS

Konstantinos A. Tasias, Panagiotis Mpistintzanos, Fotios Pekridis
University of Western Macedonia, Department of Mechanical Engineering
ktasias@uowm.gr, panagiotismpistis@gmail.com, pekridis.fotis@gmail.com

ABSTRACT

A fleet of aircraft is often deployed in a location to accomplish a specific mission. In such cases, the aircraft maintenance managers face the problem of defining the required number of spare parts that should be available during the mission to maximize its success, under a total weight and/or volume restriction. A reliability-based statistical analysis is a prerequisite for dictating the optimal solution to this problem. This study uses a cold standby redundancy strategy to examine the mission reliability problem for multiple, non-identical, parts. The formulated optimization problem is solved using dynamic programming and simulation-based optimization. To demonstrate the applicability of the model, an amphibious aircraft type for firefighting operations is employed as a case study.

Key Words: Reliability; Redundancy; Cold standby; Spare parts

1. INTRODUCTION

System reliability optimization is critical in the real world, and significant work has been undertaken over the last decades in this direction. The improvement of reliability may be achieved through the following main approaches: a. improving the reliability of the system's components; b. employing redundant components (Kuo & Prasad, 2000; Yeh and Hsieh, 2011).

The main drawback of the first approach is that the system's reliability can be improved only to some extent. Specifically, the overall reliability of a system is limited by its least reliable component, and improvements to other components cannot surpass this

constraint. To this end, the second approach is a complementary approach that helps overcome this limitation by providing alternatives in case of failures and aims to maximize reliability by utilizing backup components. This widely applied approach for improving system reliability (Reihaneh et al. 2022) is known as the Reliability Redundancy Allocation Problem (RRAP). RRAP deals with determining the optimal allocation of redundant components in a system, based on the importance and criticality of each element, to maximize the overall system reliability, while considering various constraints such as cost, weight, and space limitations (Yeh et al. 2022).

The RRAP typically considers two types of components: primary components, which are essential for the system to function, and redundant components, which serve as backups in case of primary component failures. The parts that are in standby mode can be further classified as hot, warm, or cold standby, depending on the failure probability during inactivity. Specifically, in the hot standby strategy, the failure probability is the same compared to the respective active component, whereas a very low and zero failure probability is considered in the warm and cold standby strategies, respectively (Amari and Dill, 2009). Nevertheless, most of the studies in the relevant literature implemented the active standby strategy, i.e., the redundant components are subject to operational stresses, and only a few papers have dealt with the cold-stand redundancy strategy (Ardakan et al. 2016). Specifically, Coit (2001, 2003) was the first to introduce the cold standby strategy in the redundancy allocation problem. Furthermore, Wang et al. (2015, 2016, 2018) used the cold standby strategy by considering degrading components, parameter uncertainty, and periodic inspection and maintenance, respectively. Ardakan and Hamadani (2014) studied the cold standby strategy in the RRAP. Finally, Garg and Sharma (2013) and Ardakan and Rezvan (2018) developed multi-objective RRAP with conflicting objectives, i.e., cost and reliability.

The RRAP can be formulated as a mathematical optimization model, often using: a. exact optimization methods (e.g., Kuo, 2001; Reihaneh et al. 2022); b. meta-heuristic algorithms, including, inter alia, genetic algorithm (e.g., Tavakkoli-Moghaddam et al., 2008; Lins and Droguett, 2011), tabu search (e.g., Kulturel-Konak et al., 2003; Ouzineb et al., 2008), particle swarm optimization (e.g., Dos Santos Coelho, 2009; Beji et al., 2010), memetic algorithm (e.g., Pourdarvish and Ramezani, 2013), ant colony optimization (e.g., Liang and Smith, 2004; Nahas et al., 2007), artificial immune algorithms (Chen, 2006), artificial bee colony (Garg et al., 2013), and imperialistic competitive algorithms (Afonso et al. 2013), stochastic fractal search (Mellal and Zio, 2016); c. hybridization techniques, which have been recently proposed, that aim to enhance the performance of the abovementioned methods (Thymianis et al. 2023).

In this study, the deployment of an aircraft fleet for a mission and the problem of defining the components that should be also deployed to support the fleet and minimize

the mission abortion probability is formulated as an RRAP problem. Multiple redundant components that can be shared among the aircraft are considered, and a cold standby strategy is applied during their inactivity. Moreover, the use of any distribution of parts' time-to-failure enhances the model's applicability. The objective function of the model represents the system's reliability, and the constraints capture the limitations and requirements of the system under a realistic framework. Dynamic programming and simulation-based optimization are used to define the optimal redundancy allocation of components to achieve a high level of system reliability and so, improve system performance, reduce downtimes, and enhance the overall system robustness.

2. PROBLEM SETTING

Let us assume a fleet of aircraft located permanently in a specific airbase, denoted hereafter as home base. Occasionally, there is an operational demand for several aircraft to deploy to another airbase to fulfill a mission. Specifically, part of the fleet must relocate from its home base to a different airfield to carry out a specific military operation. These missions could include, for example, participation in a military exercise, humanitarian assistance, or the deployment of firefighting aerial means across a region to maximize the proximity to potential fire incidents. Therefore, the necessity of aircraft deployment is very common, especially in military aviation applications.

Nevertheless, aircraft are very complex systems that are prone to stochastic failures. Consequently, there is a high possibility that after deployment, one or more aircraft may face a problem and either become unairworthy or unable to fulfill their mission. In such cases, there are usually the following options: a. All aircraft, or only the failed ones, come back to their home base because their mission has not been completed. However, this option is available only if the discrepancy does not affect their airworthiness, e.g., a failure in the firefighting system which does not prevent the aircraft from returning to its home base (ferry flight), whatsoever; b. Another aircraft, usually a cargo aircraft, must transport the required spare parts to restore the failure, which is a very time-consuming and costly option. It should be noted that in such applications, military transportation options are preferable because they offer greater security and direct control over the shipment process, whereas commercial transportation options require, apart from additional security measures, coordination with civilian partners, as well.

Hence, to avoid the abovementioned scenarios and ensure sustained operations during the deployment interval, efficient and extensive logistical planning is necessary. Maintenance engineers should define the quantity of redundant spare parts of each type of component that should be relocated to the new airbase to support the mission, or equivalently maximize the mission reliability. These backup parts are in a cold standby

mode because they are considered not to have any operating stresses before their installation on an aircraft. However, in practice, capacity and/or weight limitations constrain the number of redundant parts that can be carried out for the deployment. Moreover, in real-world applications, each type of part may follow a different failure probability density function, thereby arising an additional challenge.

In this study, an appropriate mathematical model is developed, which defines the optimal solution to the abovementioned realistic problem, characterized by complexity and uncertainty.

3. MODEL FORMULATION

3.1 Notations

For the model formulation, the following parameters are considered:

a. The number of aircraft to be deployed (N).

b. The different types of components (n).

c. The weight and volume of each type of component $i \in [1, n]$, denoted by w_i and v_i, respectively.

d. The total available weight and capacity for redundant components, denoted by W and V, respectively.

e The optimal number of spare parts of each type $i \in [1, n]$, which is the decision variable of the model, denoted by m_i.

3.2. Assumptions

The proposed model is developed on the basis of the following assumptions:

1. Each type of component of the deployed fleet is considered independent of another. This is a reasonable assumption since, in practical applications in aviation, the failure probability of one type of component is not related to the failure probability of another type of component.

2. If any type of component fails, then the entire system fails. The reasoning behind this assumption is that the successful completion of the mission prerequisites the operability of all deployed aircraft during the interval.

3. Perfect switching to standby parts is considered.

4. The redundant parts are assumed to perfectly restore the failure.

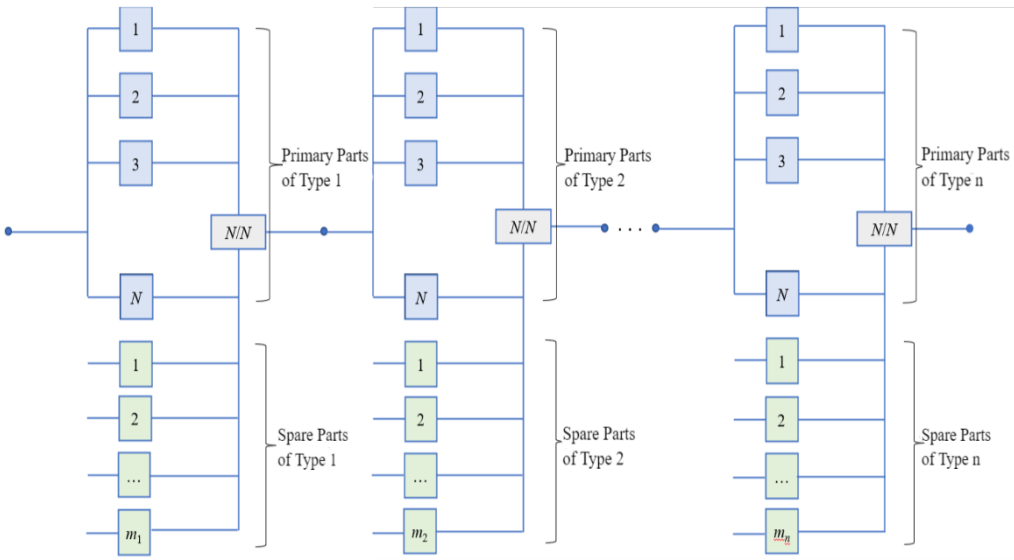
5. The failure probability for the standby parts is assumed to be zero (cold standby).

6. The system's reliability is not affected by the time required for part replacement. This assumption is justified by the routine post-flight inspections and servicing procedures that every aircraft undergoes, regardless of the reason for landing, i.e. failure or not. In practice, the duration of these post-flight maintenance actions is generally adequate for the replacement of the majority of aircraft parts.

3.3 Model Description

The system, i.e., the aircraft fleet, is represented as a series of blocks. To facilitate the quantitative assessment of the effect of multiple, non-identical, types of redundant parts on the system's reliability, the different types of parts, i.e., appurtenances, devices, and accessories, are used as the distinctive components of the system. Each part has two possible states, i.e., operating and failing to operate. The assumption that the failure of any type of part causes a system's failure (see Assumption No.2), entails an in-series connection among the different types. Furthermore, each type of part consists of N primary components, which are put on the aircraft fleet and are initially operating, and a number of spare parts in cold standby mode, in case one or more out of the primary components of the same type fails. The utility of carrying a specific number of backup parts for each type of component is related to its effect on the system's reliability. The reliability block diagram of the system is presented in Figure 1.

Figure 7 *Reliability Block Diagram of the system*



Consequently, the problem is formulated as an RRAP problem with multiple non-identical components with a cold stand-by strategy.

According to Yeh et al (2022), the most usual structure in RRAP is the series-parallel, i.e., subsystems in series and components are parallel to each subsystem. Nevertheless, most studies in the relevant literature focus on different subsystems, each supported by components dedicated to specific subsystems (see e.g., Tian et al. 2008; Khalili-Damghani et al. 2013; Mellal and Zio 2016; Reihaneh, et al. 2022). However, in our study, redundant components may be used in any of the subsystems, i.e., aircraft, that fail. This flexibility allows for a different configuration compared to studies that restrict the use of redundant components to specific subsystems. Hence, the model employs a connection of types of parts rather than aircraft. This choice facilitates an extension to a *k*-out-of-*N* redundancy model, a scenario often met in practice when deploying more aircraft than required to an airbase to fulfill a mission. Such situations arise either due to compliance with pilot crew rest regulations or the unique requirements of the mission.

4. THE OPTIMIZATION PROBLEM

The optimization purpose of the problem is to maximize the system's mission reliability, by considering the optimal number of redundant parts of each of the n different types, i.e., m_i, $\forall i \in [1, n]$. Based on the assumption that the system functions only if all mission-related aircraft are airworthy and capable of fulfilling the mission (see Assumption No. 2, Section 3.2), the system's reliability is determined by assessing the respective probability of success for each of the n types of parts given that m_i items are decided to be available in cold standby mode for each type i $\left(i \in [1, n]\right)$. Hence, the overall system reliability is derived as the product of the respective reliabilities of each type:

$$\max \; R_{System} = R_{1|m_1} \cdot R_{2|m_2} \cdot \ldots \cdot R_{n|m_n} \tag{1}$$

Furthermore, the reliability optimization includes total weight and capacity constraints, which are inherent in the aviation industry. These constraints arise from the safe operating limits set by the aircraft manufacturer and/or aviation regulations. In particular, the combined weight and volume of the backup parts cannot exceed the total available weight and capacity:

$$\sum_{i=1}^{n} m_i \cdot w_i \leq W \tag{2}$$

$$\sum_{i=1}^{n} m_i \cdot v_i \leq V \tag{3}$$

The following constraints set the boundaries of the decision and nominal variables of the problem. In particular, the number of deployed aircraft for the mission and the number of different types of parts, are, by default, positive integer numbers (Equations (4)-(5)), whereas the number of backup parts for each type is a non-negative integer, with zero indicating the absence of redundant part for a specific type (Equation (6)):

$$N \in \square^{+} \tag{4}$$

$$n \in \square^{+} \tag{5}$$

$$m_i \in \square,\ \forall i \in [1, n] \tag{6}$$

Moreover, the weight and volume of each type of part and the total available weight and volume, are inherently positive numbers (Equations (7)-(8)):

$$w_i, v_i > 0,\ \forall i \in [1, n] \tag{7}$$

$$W, V > 0 \tag{8}$$

By default, the reliability of each distinct type of component lies within the range of zero to unity (Equation (9)):

$$0 < R_{i|m_i} < 1,\ \forall i \in [1, n],\ m_i \in \square \tag{9}$$

Finally, the maximum weight and volume of each type, must not exceed the total available weight and capacity, respectively; otherwise, the type of component associated with the maximum value of weight and volume could not be considered (Equations (10)-(11)):

$$W > \max_i w_i,\ \forall i \in [1, n] \tag{10}$$

$$V > \max_i v_i,\ \forall i \in [1, n] \tag{11}$$

5. SOLUTION PROCEDURE

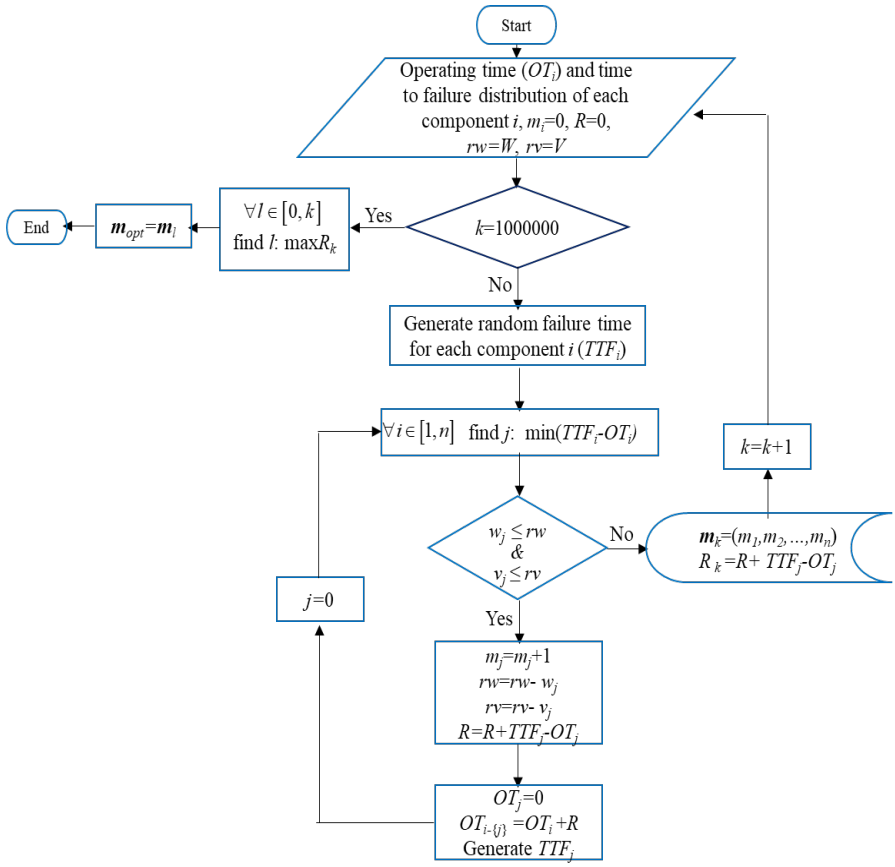
Dynamic programming, an exact algorithm that results quickly in the optimal solution, is used to solve the problem. In our study, the problem is divided into separate stages, where each stage corresponds to a different type of component i $\left(i \in [1, n]\right)$. The decision made at each stage involves selecting the number of items of the specific type to be available during the mission. The possible states, at each stage, represent the remaining weight and capacity, denoted by *rw* and *rv*, respectively. Then, the optimal solution is obtained through backward recursion, i.e., moving from the terminal stage to the initial one, while applying the following recursive formula:

$$R_i^*\left(rw,rv\right)=\max\left\{R_{i|m_i}\cdot R_{i+1}^*\left(rw-m_i\cdot w_i,rv-m_i\cdot v_i\right)\right\} \tag{12}$$

where $R_i^*\left(rw,rv\right)$ for $i\in\left[1,n-1\right]$ denotes the optimal system's reliability when considering stages i to n with available weight rw and capacity rv at the i^{th} stage $\left(R_n^*\left(rw,rv\right)=\max\left\{R_{n|m_n}\right\}\right)$. Subsequently, $R_1^*\left(W,V\right)$ represents the maximum reliability of the system, i.e., $\max R_{System}$, achieved when optimal decisions are made at each stage, considering the initial available weight and capacity.

Moreover, the problem was addressed through simulation, which concurrently validated the results obtained from the respective dynamic programming model. Figure 2 presents the flow diagram of the simulation process.

Figure 2 *Flow chart of the simulation process*



A computer program developed in MATLAB R2017a dictates the optimum solution to the optimization problem both for the dynamic programming and the simulation-based solution approach.

6. CASE STUDY

To demonstrate the applicability of the proposed model, a case study of an aerial firefighting fleet is employed. Aircraft failure data were collected from a maintenance department, covering a nominal period of a ten (10), fully operational, years span regarding a ten aircraft fleet. Thereby, trustworthy information about the system failure mechanism was available. The data set was in a two-dimensional matrix format, where each row is a single restoration action and the following variables were the columns in the data frame: Aircraft, affected subsystem, failure description, maintenance actions, date, and time of failure, i.e., when the failure happened, date and time of restoration, i.e., at what time the repair was completed. Therefore, a rectangular array of data consisting of six (6) columns and approximately seven thousand (7000) rows was available. An indicative example of the format of the data set is provided in Figure 3. Unfortunately, the real data set is confidential but can be made available upon request to the authors.

Figure 3 *Format of the available raw data set*

Aircraft	Subsystem	Failure Description	Restoration Actions	Date/Time of Failure	Date/Time of Restoration
S/N 1	Hydraulic	Failure 1	Restoration Action 1	01/01/2020 08:15	01/01/2020 12:15
			Restoration Action 2		
S/N 4	Avionics	Failure 2	Restoration Action 1	10/01/2020 13:45	10/01/2020 15:22
S/N 3	Engine	Failure 3	Restoration Action 1	02/02/2020 16:05	04/02/2020 10:44
…	…	…	…	…	…

6.2. Data Preprocessing

The raw data had to be cleaned and transformed to be used for our model. To this end, the data preprocessing phase consisted of the following steps:

a. Identification and handling of missing values. The missing data can be categorized as completely random, and the reason for missing is reasonably considered to be human errors. These missing values of the data set were handled in several different ways. In particular, missing values in the subsystem column could be easily completed

manually, since the failure description indicated the affected subsystem in all cases. However, given that the categorization of failures into affected subsystems added no value to our analysis, this column was deleted. On the other hand, missing values in the restoration actions column were either manually completed in case additional information could be provided from the aircraft's logbook, or the respective failure was deleted, otherwise. Furthermore, missing values in the date/time columns of restoration were filled by inserting the mean time to repair the same failure if applicable or deleted otherwise.

It should be mentioned that the affected aircraft and the failure description are mandatory fields to insert a new record. Moreover, each record's failure date is automatically filled as the insertion date. Therefore, there were no missing values in these columns.

b. Identification and handling of incorrect values. The problem of inaccurate values pertains solely to the date and time of restoration of a failure. At this phase, only data that are obviously entered incorrectly are characterized as outliers. Similarly to the former category of missing values, incorrect data were handled either through imputation, by substituting inaccurate values with the expected data and time of restoration, or by deleting the respective entries.

c. The restoration of a failure may include replacing a component and/or other maintenance actions. For example, a failure in controlling the aircraft's movable surfaces can be rectified through proper cable rigging, a maintenance action that does not require the use of a spare part. Furthermore, a failure's restoration may necessitate multiple actions. For instance, the recovery process might involve removing corrosion, applying lubrication, and subsequently replacing a specific part. To align with the focus of this study on replaceable items, and to enhance the utilization of data, the column initially designated for restoration actions has been converted into a column specifying the necessary backup parts.

d. A failure may be restored by multiple restoration actions, which implies that a specific action did not correct the issue, and so, another action was required. Therefore, the replacement of a part does not necessarily signify that the part failed.

On the other hand, apart from corrective maintenance actions, i.e., restoration actions performed in response to an unexpected failure, a proactive maintenance approach is implemented to prevent a failure before it happens. Age-based preventive maintenance policies are usual in the aviation industry; thus, specific aircraft parts may be proactively replaced before failure.

Consequently, in case a part was replaced erroneously or proactively, the exact time an aircraft part would have failed if continued its operation is unknown, resulting in right-

censored failure data. Therefore, for the reliability analysis, failure data had to be categorized into exact and right-censored failure data, and so, an additional column was added that divides the entries into these two categories: exact (E) and right-censored (RC).

Figure 4 *Format of the final data set*

Aircraft	Failure Description	Spare Part	Date/Time of Failure	Date/Time of Restoration	Exact (E)/Right-Censored (RC)
S/N 1	Failure 1	Spare Part 1	01/01/2020 08:15	01/01/2020 12:15	E
S/N 4	Failure 2	Spare Part 2	10/01/2020 13:45	10/01/2020 15:22	E
S/N 3	Failure 3	Spare Part 1	02/02/2020 16:05	04/02/2020 10:44	RC
…	…	…	…	…	

6.3 Reliability Analysis

A reliability analysis, which is a prerequisite for defining the optimal solution to this RRAP problem, was conducted. Since the study focuses on the successful completion of a mission, the reliability analysis was limited to parts that belong to at least one of the following categories: a. Minimum Equipment List (MEL) of the aircraft, i.e., the minimum equipment requirements for the specific type of aircraft to be considered airworthy. The parts that belong to the MEL allow an aircraft to continue operating even if other parts or systems are inoperative; b. Mission-Related (MR) parts, i.e., specific parts that are related to the mission. These parts do not affect the airworthiness of an aircraft, but their failure results in its inability to fulfill the mission, e.g., a failure in the fire system of a firefighting aircraft, or a failure in the reconnaissance pod of an aircraft during an aerial photography mission. Hence, although the MEL is the same for a specific type of aircraft, the MR parts should be considered ad-hoc, especially for multi-role aircraft.

To perform the reliability analysis of the above-mentioned parts, the final data set along with flight data and maintenance-related data, and the analysis was conducted using the Minitab software. First, the trend and correlation for the failure data of each part and each aircraft were tested to validate the assumption of independent and identically distributed data.

Furthermore, several candidate distributions were studied for suitability. In particular, the dataset was fitted to Weibull, lognormal, exponential, loglogistic, normal, and logistic distributions. The choice of candidate distribution functions is justified based on the following considerations: a. Aircraft parts can be broadly categorized into electronic components and mechanical components (see e.g., Sadananda Upadhya and Srinivasan 2012; Guo et al. 2019); b. From a practical applications perspective, the exponential distribution is a good fit for modeling the failure of electronic components (Wessels 2007), whereas the Weibull distribution is the best fitting distribution for modeling problems associated with mechanical components' aging, wear, and deterioration (Billinton and Allan, 1983); c. Normal, lognormal, and loglogistic distributions are also quite effective for modeling positively skewed reliability data, which is often the case for mechanical parts during the wear-out phase (Raqab et al. 2018; Ebeling, 2019); d. The logistic distribution offers versatility in reliability modeling and describes well the degradation process of many systems in practice (Wang and Pham. 2006). The maximum likelihood estimation method was used for each distribution and the suitability of fitted parameters was evaluated through the Anderson-Darling (AD) goodness-of-fit test.

The output of the optimal solution contains the total reliability without available redundant parts, the optimal quantity of each type of part, the available free space, and the total reliability with the optimal solution.

7. SUMMARY AND FUTURE RESEARCH

In this study, a decision-making tool on reliability redundancy allocation problem in case of aircraft deployment for a mission, is developed. The model incorporates many important and realistic factors to dictate the optimal quantity of many different types of components that should be available during the mission to avoid the undesirable and costly scenario of an aircraft aborting the mission. The total number of spare parts that are in a cold standby mode is subject to weight and/or volume constraints, and each component type is considered to follow different failure probability density functions. The solution procedure entails both a dynamic programming and a simulation-based optimization approach. The model was evaluated in the case study of an aerial firefighting aircraft. Several possible directions of this study are suggested: a. The extension to a *k*-out-of-*N* redundancy model; b. The use of a metaheuristic algorithm to find a near-optimal solution but in less computational time; c. The incorporation of a cost function into the model. This extension would facilitate decision-making not only on an operational level but on a tactical level as well. For example, it would be possible to define whether the deployment of a larger number of backup parts by a cargo airplane to the mission airbase is a more cost-efficient option compared to the transportation of fewer spare parts by the deployed aircraft; d. The relaxation of the assumptions of a

perfect switch in case of a part replacement, and zero failure probability in standby, would enhance the model's applicability.

ΠΕΡΙΛΗΨΗ

Στην αεροπορική βιομηχανία, και κυρίως στην στρατιωτική αεροπορική βιομηχανία, εγείρεται συχνά η απαίτηση μεταστάθμευσης ενός συγκεκριμένου αριθμού αεροσκαφών σε μια διαφορετική από τη μητρική αεροπορική βάση προκειμένου να εκπληρώσουν μια συγκεκριμένη αποστολή. Ωστόσο, τα αεροσκάφη αποτελούν σύνθετα συστήματα τα οποία υπόκεινται σε στοχαστικές βλάβες. Ως εκ τούτου, τα στελέχη που είναι επιφορτισμένα με τη συντήρηση και υποστήριξη των αεροσκαφών αντιμετωπίζουν το πρόβλημα προσδιορισμού των απαιτούμενων ανταλλακτικών που θα πρέπει να είναι διαθέσιμα κατά τη διάρκεια της αποστολής ώστε να μεγιστοποιηθεί η πιθανότητα επιτυχούς ολοκλήρωσης αυτής. Η επιλογή των ανταλλακτικών θα πρέπει να λαμβάνει υπόψη περιορισμούς βάρους και όγκου οι οποίοι υφίστανται για τη μεταφορά αυτών σε συνδυασμό με κατάλληλη μελέτη αξιοπιστίας. Σε αυτή την εργασία εξετάζεται το συγκεκριμένο πρόβλημα για πολλαπλούς, μη-όμοιους τύπους ανταλλακτικών τα οποία βρίσκονται σε ψυχρή κατάσταση αναμονής. Το πρόβλημα επιλύεται με χρήση δυναμικού προγραμματισμού καθώς και μέσω προσομοίωσης. Ένας τύπος αεροσκαφών που χρησιμοποιείται για αποστολές αεροπυρόσβεσης έχει χρησιμοποιηθεί ως μελέτη περίπτωσης

REFERENCES

Afonso, L.D., Mariani, V.C., & dos Santos Coelho, L. (2013). Modified imperialist competitive algorithm based on attraction and repulsion concepts for reliability-redundancy optimization. *Expert Systems with Applications*, 40 (9), 3794-3802.

Amari, S.V. & Dill, G. (2009) A new method for reliability analysis of standby systems, *Proceedings of the IEEE Annual Reliability and Maintainability Symposium*, IEEE Press, Piscataway, NJ, 417–422.

Ardakan, M.A., & Hamadani, A. Z. (2014). Reliability–redundancy allocation problem with cold-standby redundancy strategy. *Simulation Modelling Practice and Theory*, 42, 107-118.

Ardakan, M. A., & Rezvan, M. T. (2018). Multi-objective optimization of reliability–redundancy allocation problem with cold-standby strategy using NSGA-II. *Reliability Engineering & System Safety*, 172, 225-238.

Ardakan, M.A., Sima, M., Hamadani, A.Z., & Coit, D. W. (2016). A novel strategy for redundant components in reliability-redundancy allocation problems. *IIE Transactions*, 48(11), 1043-1057.

Beji, N., Jarboui, B., Eddaly, M., & Chabchoub, H. (2010). A hybrid particle swarm optimization algorithm for the redundancy allocation problem. *Journal of Computational Science*, 1 (3), 159-167.

Billinton, R., & Allan, R. N. (1992). Reliability evaluation of engineering systems (Vol. 792). New York: Plenum press.

Chen, T.C. (2006). IAs based approach for reliability redundancy allocation problems. *Applied Mathematics and Computation*, 182 (2), 1556-1567.

Coit, D.W. (2001). Cold-standby redundancy optimization for nonrepairable systems. *IIE Transactions*, 33(6), 471-478.

Coit, D.W. (2003). Maximization of system reliability with a choice of redundancy strategies. *IIE transactions*, 35(6), 535-543.

Dos Santos Coelho, L. (2009) An efficient particle swarm approach for mixed-integer programming in reliability–redundancy optimization applications. *Reliability Engineering & System Safety*, 94 (4), 830–837.

Ebeling, C. E. (2019). An introduction to reliability and maintainability engineering. Waveland Press.

Garg, H., & Sharma, S.P. (2013). Multi-objective reliability-redundancy allocation problem using particle swarm optimization. *Computers & Industrial Engineering*, 64 (1), 247-255.

Garg, H., Rani, M., & Sharma, S.P. (2013). An efficient two phase approach for solving reliability–redundancy allocation problem using artificial bee colony technique. *Computers & operations research*, 40 (12), 2961-2969.

Guo, Y., Sun, Y., Li, L., & Tang, X. (2019). Reliability assessment for multi-source data of mechanical parts of civil aircraft based on the model. Journal of Mechanical Science and Technology, 33, 3205-3211.

Khalili-Damghani, K., Abtahi, A. R., & Tavana, M. (2013). A new multi-objective particle swarm optimization method for solving reliability redundancy allocation problems. *Reliability Engineering & System Safety*, 111, 58-75.

Kulturel-Konak, S., Smith, A.E. & Coit, D.W. (2003) Efficiently solving the redundancy allocation problem using tabu search. *IIE Transactions*, 35(6), 515–526.

Kuo, W. (2001) Optimal Reliability Design: Fundamentals and Applications. Cambridge University Press, New York.

Kuo, W., & Prasad, V. R. (2000). An annotated overview of system-reliability optimization. *IEEE Transactions on Reliability*, 49(2), 176-187.

Liang, Y.C., & Smith, A.E. (2004). An ant colony optimization algorithm for the redundancy allocation problem (RAP). *IEEE Transactions on reliability*, 53 (3), 417-423.

Lins, I.D., & Droguett, E.L. (2011). Redundancy allocation problems considering systems with imperfect repairs using multi-objective genetic algorithms and discrete event simulation. *Simulation Modelling Practice and Theory*, 19 (1), 362–381.

Mellal, M. A., & Zio, E. (2016). A penalty guided stochastic fractal search approach for system reliability optimization. *Reliability Engineering & System Safety*, 152, 213-227.

Nahas, N., Nourelfath, M. & Ait-Kadi, D. (2007) Coupling ant colony and the degraded ceiling algorithm for the redundancy allocation problem of series-parallel systems. *Reliability Engineering & System Safety*, 92 (2), 211–222.

Ouzineb, M., Nourelfath, M., & Gendreau, M. (2008). Tabu search for the redundancy allocation problem of homogenous series–parallel multi-state systems. *Reliability Engineering & System Safety*, 93 (8), 1257-1272.

Pourdarvish, A. & Ramezani, Z. (2013) Cold standby redundancy allocation in a multi-level series system by memetic algorithm. *International Journal of Reliability, Quality and Safety Engineering*, 20 (03), 1340007

Raqab, M. Z., Al-Awadhi, S. A., & Kundu, D. (2018). Discriminating among Weibull, log-normal, and log-logistic distributions. Communications in Statistics-Simulation and Computation, 47 (5), 1397-1419.

Reihaneh, M., Ardakan, M.A., & Eskandarpour, M. (2022). An exact algorithm for the redundancy allocation problem with heterogeneous components under the mixed redundancy strategy. *European Journal of Operational Research*, 297 (3), 1112-1125.

Sadananda Upadhya, K., & Srinivasan, N. K. (2012). Availability estimation using simulation for military systems. *International Journal of Quality & Reliability Management*, 29 (8), 937-952.

Tavakkoli-Moghaddam, R., Safari, J. & Sassani, F. (2008). Reliability optimization of series-parallel systems with a choice of redundancy strategies using a genetic algorithm. *Reliability Engineering & System Safety*, 93(4), 550-556.

Tian, Z., Zuo, M. J., & Huang, H. (2008). Reliability-redundancy allocation for multi-state series-parallel systems. *IEEE Transactions on Reliability*, 57(2), 303-310.

Thymianis, M., Tzanetos, A., Dounias, G., & Koutras, V. (2023). Hybridization in nature inspired algorithms as an approach for problems with multiple goals: An application on reliability–redundancy allocation problems. *Engineering Applications of Artificial Intelligence*, 121, 105980.

Wang, H., & Pham, H. (2006). Reliability and Optimal Inspection-maintenance Models of Multi-degraded Systems. *Reliability and Optimal Maintenance*, 171-202.

Wang, W., Xiong, J., & Xie, M. (2015). Cold-standby redundancy allocation problem with degrading components. *International Journal of General Systems*, 44(7-8), 876-888.

Wang, W., Xiong, J., & Xie, M. (2016). A study of interval analysis for cold-standby system reliability optimization under parameter uncertainty. *Computers & Industrial Engineering*, 97, 93-100.

Wang, W., Wu, Z., Xiong, J., & Xu, Y. (2018). Redundancy optimization of cold-standby systems under periodic inspection and maintenance. *Reliability Engineering & System Safety*, 180, 394-402.

Wessels, W. R. (2007). Use of the Weibull versus exponential to model part reliability. In 2007 *Annual Reliability and Maintainability Symposium* (pp. 131-135). IEEE.

Yeh, W. C., & Hsieh, T. J. (2011). Solving reliability redundancy allocation problems using an artificial bee colony algorithm. *Computers & Operations Research*, 38(11), 1465-1473.

Yeh, W. C., Zhu, W., Tan, S. Y., Wang, G. G., & Yeh, Y. H. (2022). Novel general active reliability redundancy allocation problems and algorithm. *Reliability Engineering & System Safety*, 218, 108167.

Χορηγίες

Η Οργανωτική Επιτροπή οφείλει να ευχαριστήσει τις πιο κάτω Εταιρείες και Φορείς που συνεισέφεραν στην επιτυχή διοργάνωση του Συνεδρίου.

Goldair tourism

Πανεπιστήμιο Δυτικής Αττικής
ΔΠΜΣ Επιστήμες της Αγωγής μέσω Καινοτόμων Τεχνολογιών & Βιοϊατρικών Προσεγγίσεων

ΚΡΙΤΙΚΗ

propobos PUBLICATIONS