University of the Aegean

Doctoral Thesis

New Developments on Modelling Techniques for Dynamical Systems with Applications

Author
Emmanouil Nektarios
Kalligeris

Supervisor
Prof. Alex
Karagrigoriou

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

in the

Department of Statistics and Actuarial-Financial Mathematics

Samos, 5 September 2022

University of the Aegean

Doctoral Thesis

New Developments on Modelling Techniques for Dynamical Systems with Applications

Author
Emmanouil Nektarios
Kalligeris

Supervisor
Prof. Alex
Karagrigoriou

Seven-member Evaluation Committee

BARBU Vlad Stefan
Associate Professor, University of Rouen Normandy

CHALIDIAS Nikolaos (Three-member Evaluation Committee)
Professor, University of the Aegean

PARPOULA Christina
Assistant Professor, Panteion University of Social and Political Sciences

PLATIS Agapios
Professor, University of the Aegean

KARAGRIGORIOU Alexandros (Supervisor)
Professor, University of the Aegean

VASDEKIS Vasileios (Three-member Evaluation Committee)
Professor, Athens University of Economics and Business

ZIMERAS Stelios
Associate Professor, University of the Aegean

UNIVERSITY OF THE AEGEAN

Abstract

School of Sciences
Department of Statistics and Actuarial-Financial Mathematics

Doctor of Philosophy

New Developments on Modelling Techniques for Dynamical Systems with Applications

by Emmanouil Nektarios KALLIGERIS

This PhD thesis is conducted at the Department of Statistics and Actuarial-Financial Mathematics of the University of the Aegean. Its goal is to fill in the gap in the literature regarding the modelling of dynamical systems from the incidence data point of view. To that end, at first the state-of-the-art alongside the materials and methods required are presented and fully discussed. Then, several novel models and innovative methodologies for capturing the behavior of incidence data are proposed and fully investigated, accompanied by useful comparative studies, for practical purposes. Finally, we conclude with a short discussion on the findings of this PhD thesis.

ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ

Περίληψη

Σχολή Θετικών Επιστημών
Τμήμα Στατιστικής και Αναλογιστικών-Χρηματοοικονομικών Μαθηματικών

Διδακτορικό

Νέες Εξελίξεις στη Μοντελοποίηση Δυναμικών Συστημάτων με Εφαρμογές

του Εμμανουήλ Νεκτάριου ΚΑΛΛΙΓΕΡΗ

Η παρούσα διδακτορική διατριβή εκπονήθηκε στο Τμήμα Στατιστικής και Αναλο γιστικών-Χρηματοοικονομικών Μαθηματικών του Πανεπιστημίου Αιγαίου. Στόχος της, είναι η κάλυψη του κενού στη βιβλιογραφία όσον αφορά τη μοντελοποίηση δυναμικών συστημάτων που αφορούν δεδομένα επίπτωσης.

Παράγοντες όπως το Περιβάλλον, η Επιστήμη και η Τεχνολογία, αποτελούν συστήματα με ένα κοινό βασικό στοιχείο, τον χρόνο. Πρώτος, ο Άλμπερτ Αϊνστάιν μελέτησε τον χρόνο και τον χαρακτήρισε ως «ψευδαίσθηση». Σχεδόν κάθε σύστημα τείνει να εξελίσσεται, με διαφορετικούς ρυθμούς και συμπεριφορές, με την πάροδο του χρόνου. Το γεγονός αυτό κάνει τη φύση τέτοιων συστημάτων, δυναμική. Τα δυναμικά συστήματα αποτελούν ζωτικό εργαλείο για τη μοντελοποίηση φαινομένων που εξελίσσονται στον χρόνο. Η δημιουργία ενός τέτοιου συστήματος είναι σχετικά «απλή» καθώς χρειάζεται (1) ο προσδιορισμός της ποσότητας που εξελίσσεται στον χρόνο και (2) να τεθεί ο κανόνας που διέπει την εξέλιξη αυτή.

Υπάρχουν τρεις τύποι δυναμικών συστημάτων: (1) Διακριτά ή Συνεχή, (2) Πεπερασμένα ή Άπειρα, και (3) Ντετερμινιστικά ή Στοχαστικά. Αν και είναι διαθέσιμη μία πληθώρα θεωρητικών αποτελεσμάτων και εφαρμογών, εξακολουθούν να υπάρχουν αρκετά αναπάντητα ερωτήματα-κλειδί που αφορούν (κυρίως) τις γενικότερες πτυχές της δυναμικής και την έλλειψη επαρκούς συμφωνίας μεταξύ ποιοτικών και ποσοτικών αποτελεσμάτων. Στις μέρες μας, η μελέτη δυναμικών συστημάτων έχει να επιδείξει μεγάλα άλματα λόγω της ραγδαίας τεχνολογικής εξέλιξης η οποία έχει φέρει στο προσκήνιο νέους παράγοντες που πρέπει να ληφθούν υπόψη. Ως φυσική συνέπεια όλων των παραπάνω ο βαθμός πολυπλοκότητας, της ήδη πολύπλοκης έννοιας των δυναμικών συστημάτων, έχει αυξηθεί απότομα.

Η μοντελοποίηση των μηχανισμών αλλά και των συμπεριφορών που διέπουν τα δυναμικά συστήματα, βρίσκεται στο επίκεντρο διάφορων επιστημονικών πεδίων. Ένα από αυτά τα πεδία είναι αυτό της επιδημιολογίας το οποίο έχει επηρεαστεί σε μεγάλο βαθμό από τη δυναμική μοντελοποίηση. Τα επιδημιολογικά συστήματα μελετώνται εδώ και καιρό μέσω δυναμικών μοντέλων λόγω της άμεσης συσχέτισής τους με τη δημόσια υγεία και της πολυπλοκότητας που τα συνοδεύουν. Πολλοί

ερευνητές έχουν στρέψει το ενδιαφέρον τους στη δημιουργία επαρκών μοντέλων που είναι σε θέση να περιγράψουν και να προβλέψουν την εξέλιξη τέτοιων συστημάτων. Τα μοντέλα αυτά συνήθως προέρχονται από μηχανισμούς υποθέσεων, διαμορφώνοντας έτσι δυναμικά μαθηματικά μοντέλα που αντιπροσωπεύουν τους προαναφερθέντες μηχανισμούς τα οποία και εφαρμόζονται/δοκιμάζονται πάνω σε δεδομένα.

Τα δυναμικά συστήματα έχουν τη μοναδική ικανότητα να συνδυάζουν πολλά (αλληλεπιδρώντα) κομμάτια, οδηγώντας στη δημιουργία μιας νέας «όλα σε ένα» συμπεριφοράς. Αυτή η συνειδητοποίηση, δηλαδή ότι η συνολική συμπεριφορά ενός συστήματος δεν μπορεί να γίνει κατανοητή μελετώντας μόνο τη συμπεριφορά των επιμέρους συστατικών του, δημιούργησε μια πληθώρα νέων εννοιών και μαθηματικών εργαλείων. Ως αποτέλεσμα, η ανάπτυξη ενός δυναμικού μοντέλου δεν είναι εύκολη υπόθεση, καθώς το τελευταίο θα πρέπει να μπορεί να περιγράψει με επαρκή τρόπο τα χαρακτηριστικά της υπό μελέτη διαδικασίας.

Η παρούσα διατριβή στοχεύει στο να προσφερθούν νέες προοπτικές στον τομέα της δυναμικής μοντελοποίησης προτείνοντας νέα, ακριβή, ευέλικτα και εύκολα εφαρμόσιμα μοντέλα και τεχνικές μοντελοποίησης. Έτσι, στο πρώτο κεφάλαιο γίνεται μια εκτενής εισαγωγή στο αντικείμενο μελέτης της διατριβής παρουσιάζοντας όλες τις απαιτούμενες ορολογίες. Στο δεύτερο κεφάλαιο γίνεται λεπτομερής παρουσίαση όλων των απαραίτητων εργαλείων και μεθόδων που χρησιμοποιήθηκαν στα μετέπειτα κεφάλαια για την ανάπτυξη νέων καινοτόμων τεχνικών μοντελοποίησης δεδομένων επίπτωσης.

Η παρακολούθηση σε πραγματικό χρόνο της επιδημικής δραστηριότητας στα συστήματα επιδημιολογικής επιτήρησης, είναι συχνά δύσκολο να επιτευχθεί πλήρως λόγω της εποχικότητας που εμπλέκεται στη σειρά. Με αφορμή το γεγονός αυτό, στο κεφάλαιο τρία εξετάζεται η γενική μορφή καθώς και ειδικές περιπτώσεις περιοδικών αυτοπαλίνδρομων μοντέλων με σκοπό τη μοντελοποίηση δεδομένων επίπτωσης που αφορούν τα εβδομαδιαία εκτιμώμενα ποσοστά ασθενειών τύπου γρίπης στην Ελλάδα για την περίοδο 2014-2016.

Στο προηγούμενο κεφάλαιο επικεντρωθήκαμε αποκλειστικά στην αποτύπωση της συμπεριφοράς των δεδομένων επίπτωσης κατά τη διάρκεια τυπικών (μη ακραίων) περιόδων. Ωστόσο, ο προσδιορισμός της πλήρους πορείας τέτοιων δεδομένων, όπως τα εκτιμώμενα ποσοστά ασθενειών τύπου γρίπης στην Ελλάδα, είναι χρήσιμος για διάφορους λόγους. Για παράδειγμα ο εντοπισμός του τέλους μιας επιδημίας βοηθά τους υπεύθυνους δημόσιας υγείας στο να καθορίσουν εάν τα νέα κρούσματα αποτελούν μέρος μίας ήδη γνωστής (ή νέας) έξαρσης. Ως εκ τούτου, στο τέταρτο κεφάλαιο επιχειρούμε να αποτυπώσουμε τη συμπεριφορά τόσο των μη ακραίων όσο και των ακραίων περιόδων που εμφανίζονται σε δεδομένα επίπτωσης. Ο προσδιορισμός των ακραίων περιόδων καθίσταται δυνατός μέσω της ανάλυσης ανίχνευσης σημείων αλλαγής και αναπτύσσονται τεχνικές επιλογής μοντέλων προκειμένου να προσδιοριστεί το βέλτιστο περιοδικό αυτοπαλίνδρομο μοντέλο με συμμεταβλητές που περιγράφει καλύτερα το μοτίβο της υπό εξέταση χρονοσειράς. Επιπροσθέτως, αναπτύχθηκε ένας προηγμένος αλγόριθμος προκειμένου να βελτιωθεί η ακρίβεια του επιλεγμένου μοντέλου.

Το κεφάλαιο πέντε απαρτίζεται από δύο συγκριτικές μελέτες. Η πρώτη έχει ως στόχο την αξιολόγηση της προβλεπτικής ικανότητας του μοντέλου που προτείνεται στο κεφάλαιο τρία σε αντιπαραβολή με εναλλακτικά μοντέλα που σχετίζονται με τη μοντελοποίηση της νοσηρότητας της γρίπης. Στην δεύτερη, εφαρμόσαμε και αξιολογήσαμε πρωτοπόρες μεθόδους που βασίζονται στην ανάλυση σημείων αλλαγής για τον εντοπισμό αλλαγών σε δεδομένα τύπου γρίπης. Η εμπειρική συγκριτική μελέτη παρείχε στοιχεία από τα οποία διαφαίνεται δείχνουν ότι οι στατιστικές μέθοδοι που βασίζονται στην ανάλυση σημείων αλλαγής έχουν αρκετές ελκυστικές ιδιότητες σε σύγκριση με την τρέχουσα πρακτική για την ανίχνευση επιδημιών.

Συνεχίζοντας την προσπάθεια ανάπτυξης μιας αποτελεσματικής μεθοδολογίας για τη μοντελοποίηση της πλήρους συμπεριφοράς των δεδομένων επίπτωσης χρονοσειρών, στο έκτο κεφάλαιο προτείνεται μια επέκταση του Μαρκοβιανού εναλλασσόμενου μοντέλου. Τα συστατικά του επιλέγονται με τεχνικές ποινικοποιημένης πιθανότητας με στόχο την επίτευξη υψηλού βαθμού ευρωστίας όσον αφορά τη μοντελοποίηση των δυναμικών συμπεριφορών των επιδημιολογικών δεδομένων. Εκτός από τα στατιστικά συμπεράσματα, εφαρμόζεται ανάλυση ανίχνευσης σημείου αλλαγής για την επιλογή του αριθμού των εναλλαγών, η οποία μειώνει την πολυπλοκότητα της προαναφερθείσας διαδικασίας. Στο πλαίσιο αυτό προτείνεται μια διαδικασία τριών φάσεων για τη μοντελοποίηση δεδομένων επίπτωσης και ελέγχεται μέσω πραγματικών και προσομοιωμένων δεδομένων.

Βασιζόμενοι (κυρίως) στα αποτελέσματα του προηγούμενου κεφαλαίου, στο κεφάλαιο εφτά επιθυμούμε να ξεκινήσουμε τη διερεύνηση μιας «φυσικής» επέκτασης της μεθοδολογίας που παρουσιάστηκε στο προηγούμενο κεφάλαιο, δηλαδή τη χρήση ημι-Μαρκοβιανών εναλλασσόμενων μοντέλων για τη μοντελοποίηση δεδομένων επίπτωσης. Ως αποτέλεσμα, σε αυτό το κεφάλαιο ορίζουμε το διακριτού χρόνου ημι-Μαρκοβιανό εναλλασσόμενο μοντέλο δεσμευμένου μέσου με συμμεταβλητές. Θα πρέπει να σημειωθεί ωστόσο, ότι η μελέτη που παρουσιάζεται σε αυτό το κεφάλαιο βρίσκεται υπό εξέλιξη.

Η διδακτορική διατριβή ολοκληρώνεται με σύντομη συζήτηση σχετικά με τα ευρήματα καθώς και τις πιθανές επεκτάσεις της.

Acknowledgements

So this is it... the time has come. The most difficult part of writing this PhD thesis is to acknowledge each and everyone contributed, in each own way, to this challenging yet amazing journey.

First of all, I would like to acknowledge my supervisor Prof. Alexandros Karagrigoriou of the University of the Aegean. Mister Karagrigoriou words cannot describe the impact you had on me both academically and in real life. Please do not ever stop to be the great person you are.

I would like to also thank my friends Andreas, Christina, Effrosyni, Kimon and Thanasis for their continuous support all these years. Guys thank you for everything from the bottom of my heart. I am gratefully indebted to you.

Of course, I could not forget to acknowledge my colleagues: Assoc. Prof. Vlad Stefan Barbu (Univ. of Rouen Normandy), Dr. Andreas Makrides (Univ. of Cyprus), Kimon Ntotsis (Univ. of the Aegean) and Assist. Prof. Christina Parpoula (Panteion Univ. of Social and Political Sciences), with whom I had an amazing collaboration in all levels.

I would like to acknowledge both the three-member and seven-member evaluation Committees for their useful comments that improved significantly the overall quality of the thesis.

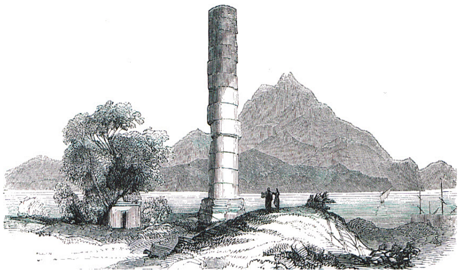
As this section comes to an end, I would like to thanks the woman who raised me to be a good person and stands by me from day zero. My mother Aristea. Mother, thank you for dedicating your life to me.

Contents

List of Figures

List of Tables

Dedicated to everyone that truly cares about me. . .



Temple of Hera, Samos, Greece

Chapter 1

Introduction

Environment, Humanity, Science, Technology. All these factors are systems with one essential element in common; time. First, Albert Einstein studied time and characterized it as an "illusion". Abhijit Naskar[1], influenced by Einstein, stated that "*time is basically an illusion created by the mind to aid in our sense of temporal presence in the vast ocean of space*". Almost every system tends to evolve, at different rates and behaviors, over time. This fact makes the nature of such systems, dynamical. Dynamical systems constitute a vital tool for modelling phenomena that evolve over time. Creating such a system, is rather simple; First, the quantity that evolves over time needs to be specified, and second, the rule that governs the latter evolution needs to be set.

Dynamical systems can be divided in three categories: (1) discrete or continuous; (2) finite or infinite, and; (3) deterministic or stochastic. Although many theoretical results and applications have been obtained, there are still several key-questions, regarding (mainly) global aspects of the dynamics and the lack of a sufficient agreement between qualitative and quantitative results, which remain unanswered. Nowadays, the study of dynamical systems has faced major changes due to the rapid technological advancement that has brought various new factors in the foreground, that need to be considered. As an example, in the early 80's differential equation models and techniques of reconstructing phase spaces from time series data, made their appearance.[2; 3] Furthermore, increased data availability along with the huge steps forward regarding the computer hardware/power, revived the interest for developing models that determine a system of governing dynamical equations from a given dataset.[4] As a natural consequence of all the above, the complexity, of an already complex concept such that of dynamical systems, has been sharply increased.

Modelling the governing mechanisms and behaviors of dynamical systems, lies at the heart of several scientific fields. The field of epidemiology for example, has been highly influenced by dynamical modelling.[5] Epidemiological systems, have long been studied via dynamic models due to their direct association to public health and the complexity that are accompanied by.[6]–[12] For almost four centuries now, many researchers have turned their focus in creating adequate models that are able to describe and predict the evolution of such systems.[13] These models usually derive through hypothesizing mechanisms, formulating dynamic mathematical models that represent the latter mechanisms, and finally testing those models against data.

The majority of applications of dynamical systems can be grouped into three main categories:

- **Predictive**

  The goal is the prediction of future states of the underlying system based on both past and present states of it. Prediction is usually forward reasoned from causes to effects.

- **Diagnostic**

  Here, the objective is to identify the possible path of the past states, that may have led to the present state of the system. Diagnosis is reasoned backwards from effects (e.g., symptoms) to causes (e.g., diseases).

- **Theoretical**

  There are applications which focus more on providing a theory for the physical phenomena rather than explaining the past or predicting the future. For example, a researcher might provide a theory for a particular problem in the form of a differential equations set, which could be possibly used for explaining or predicting an outcome. From the researcher's perspective though, this set of equations is his primary interest since it describes nodal aspects of the phenomenon under investigation.

The aforementioned three broad categories, represent the human need for explaining, understanding and predicting physical phenomena. Regarding the first two categories, note that some phenomena appear to be highly stochastic and some others, although deterministic, their governing mechanisms are either too complicated, or dependent (almost) solely, on accurate observations of the present state. As a result, the prediction or diagnosis of physical phenomena is not always feasible.

Dynamical systems have the unique ability of combining many (interacting) parts, leading to the generation of a new all-in-one behavior. This realization, i.e., that the total behavior of a system cannot be understood by just inferring on the behavior of its individual components, created a plethora of new concepts and mathematical tools. As a result, developing a dynamical model is anything but an easy task, as the latter needs to describe in a sufficient way the characteristics of the process under study. This doctoral thesis, intends in providing new insights into the field of dynamical modelling by proposing novel, accurate, flexible and easily applied models and modelling techniques.

The structure of the thesis is as follows. In Chapter 1 a thorough introduction is conducted on the topic(s) that will concern us throughout the thesis. In Chapter 2 the materials and methods used are presented in a detailed fashion. In Chapter 3, we develop an alternative approach in order to model seasonality of influenza, based on a periodic regression modelling. In Chapter 5, a comparative study is conducted for evaluating the forecasting performance of the model selected on Chapter 3 with other models associated with the modelling of influenza morbidity. In Chapter 4, we propose an algorithmic procedure based on change-point detection analysis and periodic-type ARMA modeling with covariates for capturing the behavior of time-series data that exhibit typical and non-typical periods. In Chapter 5.2, we implement and evaluate cutting-edge changepoint analysis-based methods, for detecting changes in location of univariate influenza like illness rate data. In Chapter 6, we propose an advanced regime switching modeling approach for incidence data. Finally, in Chapter 7 we discuss future aspects of our work related to the concept of semi-Markov switching modelling under the discrete time framework.

1.1 Aspects and Methods for Biosurveillance

Infections have plagued humans for millennia, mainly because of human pathogens which constantly evolve leading to new agents and mechanisms of transmission. The level of interaction between humans and pathogens, has been significantly increased over the years as a consequence of human intrusion into habits that where previously unknown (e.g., global traveling, international food trading, etc.). The aforementioned, put the health of the population, which constitutes a valuable asset for both society and health services,(14) at high risk. Continuous rapid changes in the environment and the socio-economic conditions, as well as the observed changes in the epidemiology of diseases and the burden they cause on humanity, are the main axes that impose the necessity for public health surveillance.(15)–(17) Public health surveillance can be defined as the "ongoing, systematic collection, analysis, interpretation, and dissemination of data regarding a health-related event that enables public health authorities to reduce morbidity and mortality".(18)

Epidemiological surveillance is a dynamic activity which continuously progresses and requires systematic monitoring in the field of health sciences and biostatistics. According to the World Health Organization(19) it is defined as "*the continuous, systematic collection, analysis and interpretation of health-related data needed for the planning, implementation, and evaluation of public health practice*". The aforementioned definition includes various aspects such as controlling the validity of data, analyzing data via advanced statistical methods, as well as extracting safe conclusions accompanied by scientific and methodological adequacy.(20) Although epidemiological surveillance is of high importance, it is related solely to the human population, fact that does not go hand in hand with todays "standards". During the past decades, along with the rapid advances in the scientific areas of computing, engineering, mathematics, statistics and public health, a potentially powerful surveillance-type science has been emerged, known as Biosurveillance.(21) It is defined as "the process of active data-gathering with appropriate analysis and interpretation of biosphere data that might relate to disease activity and threats to human or animal health whether infectious, toxic, metabolic, or otherwise, and regardless of intentional or natural origin in order to achieve early warning of health threats, early detection of health events, and overall situational awareness of disease activity" (Homeland Security Presidential Directive 21, 2014).

The key difference between traditional epidemiological surveillance and the emerging science of Biosurveillance, is the fact that while epidemiological surveillance tends to target only on human identified cases, Biosurveillance on the other hand requires the integration of data regarding the health of humans, animals and plants. As it is reasonable, in order to make such an integration work properly (i.e. counter possible biological threats) a massive coordinated operation is required among various stakeholders.

1.1.1 Biosurveillance systems and processes

The history of Biosurveillance, is relatively young in terms of origin since it constitutes a mixture between the well known processes of disease and public health surveillances. It is a continuous process (Figure 1.1) which, as opposed to epidemiological surveillance, monitors disease activity not only in people but also in plants or animals.(22; 23) By disease activity, it is meant to encompass not only the emergence and/or manifestation of the disease, but also the preliminary processes involved in the development and/or evolution of the disease. It mainly focuses in detecting and characterizing outbreaks of disease as well as monitoring the environment for several biological agents that are able to cause a disease such as bacteria, viruses, etc. The early outbreak detection of disease, constitutes the major challenge of Biosurveillance, since detecting outbreaks the moment they arise, in a sufficiently timely fashion, could prevent the affected individuals from getting sick or even more, killed.

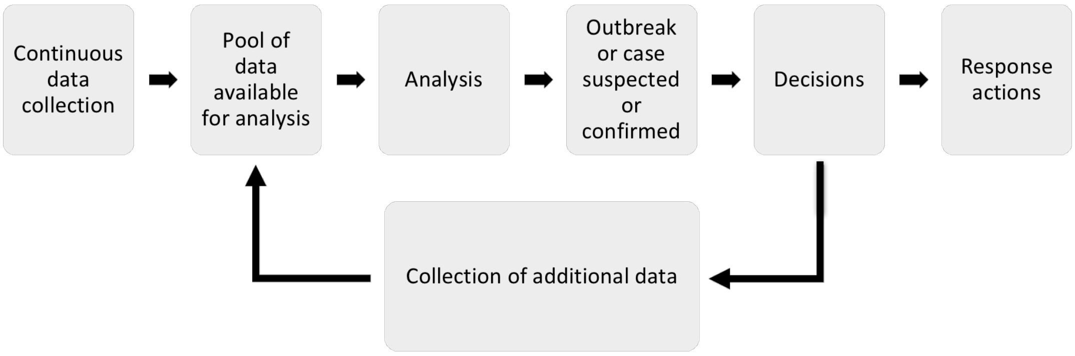

Continuous data collection
Pool of data available for analysis
Analysis
Outbreak or case suspected or confirmed
Decisions
Response actions
Collection of additional data

Figure 1.1: Biosurveillance process.

Biosurveillance is a multidisciplinary science which traditionally involves expertise from the scientific fields of epidemiology, medicine, microbiology, veterinary, public health, and health care. Nowadays, as part of the field's evolution, the increased possibility of more powerful biological threats and activities, has led this new scientific area to diversify its pool of expertise into more computer-oriented scientific fields such those of mathematics, (bio)statistics, computer science, and systems and quality engineering. The importance of the latter fields is reflected in the need of conducting biosurveillance at real-time and sometimes in forms of big data; hence, a necessity for timely and efficient automation is emerged as pointed out by Wagner et al.(22)

When it comes to biosurveillance, the leading agency nationally is the Center for Disease Control and Prevention (CDC). The CDC is responsible for collecting, analyzing, and disseminating national disease occurrence and mortality data to the public as well as promulgating goals and standards for biosurveillance systems and encouraging their adoption. However, a Biosurveillance process operates not only between functions of specialized local governments and health organizations such as CDC, but also accomplishes to expand these functions in cooperation with a variety of stakeholders such as laboratories, healthcare providers, etc. Thus, a "healthy" Biosurveillance process is based upon an effective communication between the aforementioned stakeholders.

The central element to each Biosurveillance process is the Biosurveillance system (Figure 1.2). The 21st century, has brought a plethora of advancements in the field of computing, which led to newer developments of automated Biosurveillance systems. However, such systems whether automated, manual, or both, must still be systematic in terms of their functionality. As with any engineering system, a Biosurveillance one should be able to meet its functional requirements (e.g., specifications of the diseases that must be detected and the time frame within which detection must occur) in order to be considered operational.

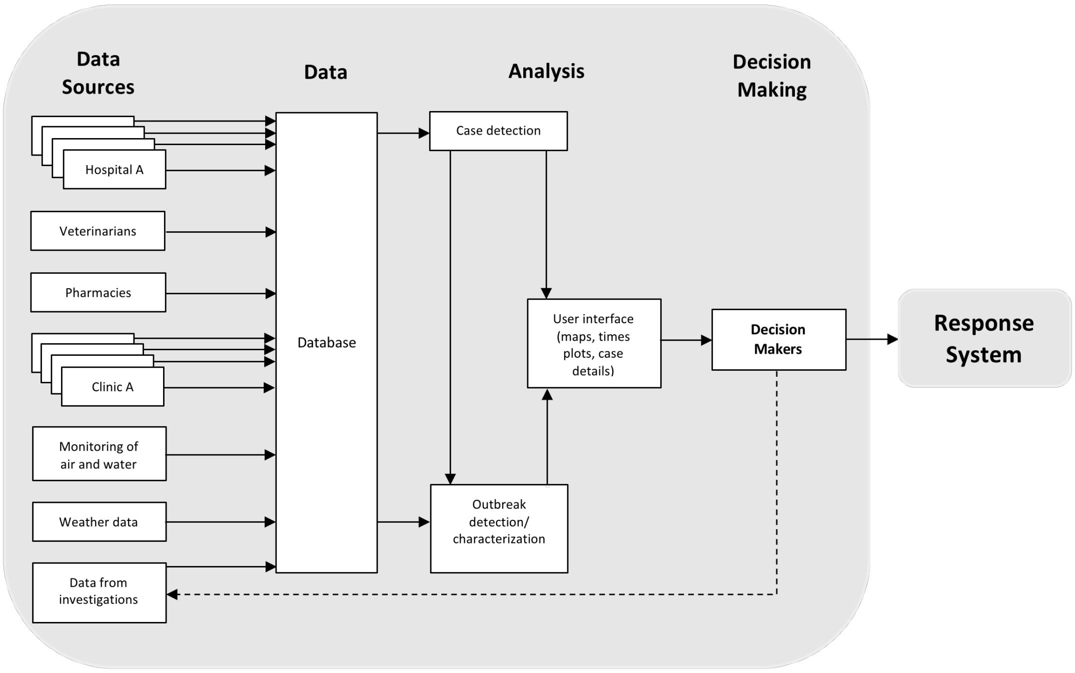


FIGURE 1.2: Generic biosurveillance system.

The main requirement of such systems is to be able to recognize threat patterns. After a thorough study on some of the major disease outbreaks in recent history, a set of nine fundamental threat patterns that a functional Biosurveillance system must be able to recognize, has been identified:

1. large aerosol release;

2. building or vessel contamination;
3. small premonitory release or contamination;
4. continuous or intermittent release of and agent;
5. person-to-person contagion;
6. transmission trough commercially distributed products;
7. water-borne transmission;
8. vector- or host- borne transmission; and
9. sexual or parental transmission.

Despite all the advances and advanced equipment, todays Biosurveillance systems can detect five out of the above nine patterns,(24) while their designer usually focuses on two or three patterns and partially covers some of the rest.

Finally, one of the most crucial aspects of biosurveillance is outbreak detection. The term "outbreak detection" refers to the methods that biosurveillance organizations recruit to detect the occurrence of an outbreak. In order to characterized as efficient, a biosurveillance system must be able to detect an outbreak as quickly as possible so that treatment could be enabled to those already sick and at the same time further illness is prevented. The aforementioned depends on two significant factors: (1) the cost, since early detection is usually expensive and; (2) the required timeliness, which varies by biological agent and route of transmission.

1.2 Surveillance Types and Systems

The human race has been threatened several times over its course from a plethora of diseases. The first references of disease outbreaks, come from Hippocrates (460-377 bc) who described an outbreak of mumps on the Greek island of Thasos. Moreover, he described several other outbreaks such those of malaria and influenza. The Black Plague (1346-1353), killed more than 60% of European population. Smallpox (18th century) killed annually, over 40000 Europeans. Severe Acute Respiratory Syndrome (SARS), generated widespread panic in 2003 and was caused by a previously unknown coronavirus (SARS-CoV-1) i.e., the same family of viruses that caused Middle East Respiratory Syndrome (MERS) and SARS-CoV-2. The above diseases, and many more, generated the need for the development of surveillance systems. Surveillance is a very challenging yet intriguing process that serves multiple public health functions such as estimating the burden of a disease or injury, determining the distribution and spread of illness, generating hypotheses and stimulating research, supporting disease control interventions, evaluating prevention and control measures, facilitating planning and guiding vaccine development.

Nowadays, several European countries[25] and Centers for Disease Control and Prevention (CDC) have developed epidemiological (or sentinel) surveillance systems in order to record, monitor and analyze the activity of both known and unknown infectious diseases. The fundamental objective of such systems, is the early and accurate identification of a single individual with a disease. There exist two primary types of disease surveillance, passive and active. In passive surveillance, case reports are supplied voluntarily from various sources, e.g., clinicians, laboratories, etc. The reliability of this kind of surveillance, in terms of completeness and accuracy, is based upon several factors, namely, whether reporting is legally mandated or a definitive diagnosis can be established, illness severity as well as the awareness of the medical condition among the public and the medical community. Although, passive surveillance is consider useful when it comes to routine surveillance activities, it has a significant flaw. Due to the fact that more (severe) illness is more likely to be reported, the severity of passively reported cases have a high chance to differ from those of all cases of an illness. In active surveillance, case finding can be retrospective, prospective, or both. Through retrospective case finding, the health status of individuals is identified from existing data, such as clinical records and death certificates, while through prospective the identification and collection of information regarding cases is done the moment they occur. Active surveillance, in which all cases are identified and reported in a specific geographic area, provides the most complete and unbiased ascertainment of disease and is optimal for describing the rate of a disease. In general, data collected through active surveillance, are considered superior in terms of information than those of passive surveillance.

In Greece, since 1999, a sentinel surveillance system is in operation which is based on voluntary participation of physicians, general practitioners and pediatricians of Primary Health Care (PHC) throughout Greece. The sentinel systems in PHC through registration, processing, analysis and results/conclusions export procedures, provide not only general guidelines for optimal decision making in health services but also the most important source of primary care epidemiological diseases data. Through such systems, the evolution of the frequency of certain diseases is recorded by carefully selected reporting sites and health workers who report cases of the disease or syndrome under surveillance, based on clinical diagnoses. In particular, the sentinel medical doctors send weekly epidemiological data regarding the number of consultations for all causes and the number of consultations for each syndrome under surveillance according to a specified clinical definition. These reporting forms, enable the National Public Health Organization of Greece to estimate the weekly number of syndrome cases per 1000 visits, i.e., the proportional morbidity, which reflects the activity of the syndrome under study.[26]

It is worth to be mentioned that, during the period 2014-2015 the sentinel system of Greece was reorganized from the ground up under the Operational Programme “Human Resources Development” of the National Strategic Reference Framework (NSRF) 2007-2013, action “Primary Health Care services (private and public) networking for epidemiological surveillance and control of communicable diseases". This reorganization redefined the national priorities regarding the syndromes monitored through the sentinel system, bringing influenza-like illness (ILI) and gastroenteritis to the center of interest. The study of the evolution of these two syndromes is a major public health concern, since despite the fact that they belong to the sentinel epidemiological surveillance priorities of the country, are also monitored traditionally by sentinel systems in the European region, while they are high in terms of international interest, due to their potential for widespread transmission (with ILI also representing a potential pandemic risk). Moreover, their surveillance through the sentinel system, enables studying the existence of seasonality, the determination of the signaled start and end weeks and the intensity of epidemic waves for ILI, as well as the determination of epidemic outbreaks for gastroenteritis nationwide.

1.3 Case and Outbreak Detection Sources

In various scientific fields, e.g., Medicine, Meteorology, etc., it is of high importance to monitor for abnormal activities. In the field of Public Health, case and outbreak detection, via proper surveillance systems, enables authorities for rapid investigation, pathogen identification, and response. The collected data can potentially bring to the attention of authorities diseases that were previously unrecognized and/or underappreciated. Several entities contribute to that cause such as people (e.g., physicians, veterinarians, nurse practitioners, etc.), laboratories, biosurveillance organizations and computers. Of course, in most cases, the final verdict of whether an individual has indeed a disease or just a syndrome, is left to the distinctive ability of the clinician in charge.

Some common sources of both case and outbreak detection, are the following:

i. **Sentinel clinicians**, due to the nature of the occupation they serve, constitute perhaps the most common case detection entity. At first, a sick individual seeks medical attention or is brought to a clinician, who examines the patient and then establishes a diagnosis. After, if the diagnosis is considered a notifiable disease, the clinician reports it to a local health department (for persons) or to a local department of agriculture (for animals).

ii. **Laboratories** constitute a valuable source of case reporting, since are able to detect, through rapid and reliable diagnostic tests, cases of notifiable diseases. The role of laboratories in case detection is nodal, since the latter are "process oriented" and therefore, may report cases in a more reliable way as opposed to clinicians. Note though that a laboratory, is not capable of detecting a case unless a clinician refers an individual to it.

iii. **Screening** has been widely used over the years, and even more during the SARS-CoV-2 pandemic, as a case detection technique. Among others, it involves interviewing and testing people during an outbreak so that additional cases (or carriers) of the disease can be identified with the goal of preventing further infections. A biosurveillance organization may use screening in a focused manner (e.g., screening of all students and teachers in a school), or deploy screening on a wide-scale basis. For example, the SARS-CoV-2 disease warranted screening all over the world. The majority of countries, screened citizens in multiple ways such as applying several testing methods, quarantining infected individuals for a certain period of time and imposing vertical or horizontal lockdowns.The intensity of the screening effort, depends mainly on the nature of each outbreak.

iv. **Computers** give a thunderous present to both case and outbreak detection processes. It is an indisputable fact, that during the past two decades computers have been through a tremendous development. As a result, they are considered as a powerful tool for detecting, storing and analyzing case data. Nowadays, surveillance organizations are using electronic laboratory reporting systems, which import case detection data from laboratories.(27)–(30) Because (almost) all infectious diseases initially make their appearance with a small number of syndromes, computer-based case detection systems focus on monitoring the most common of them, e.g., diarrhea, respiratory issues, influenza-like symptoms, rashes, hemorrhage and paralysis.

Despite all the advances towards prevention and treatment of infectious diseases, improved living conditions, and development of effective vaccines and antimicrobials, infectious diseases remain among the top causes of death worldwide. This, along with the fact that some outbreaks are never detected, implies that there is plenty of room for improving current methods of outbreak detection as well as developing new ones.

1.4 Incidence Data

Incidence data collected over time appear frequently in several scientific fields such as medicine,(31) meteorology(32) and public health.(33) Such data, for various reasons, do not follow a standard or typical model but often experience outbreaks. Consider, for example, recent work in genomics, looking at detecting changes in gene copy numbers or in the compositional structure of the genome(34)–(36) and in finance where interest lies in detecting changes in the volatility of time–series.(37)–(39)

The term "incidence", refers to the occurrence of new cases of disease or injury in a population over a specified period of time and it is usually expressed as a proportion or rate. As incidence proportion is defined the fraction of new cases of diseased or injured subject during a specified period over the initial size of the under study population. Incidence rate is defined in a similar way as incidence proportion, with the only difference lying in the denominator part. More specifically, while in incidence proportion the denominator is the initial size of the population, in the case of incidence rate the latter is being substituted by the sum of the period each subject was observed totaled.

1.5 State-of-the Art

During the past decade, a widespread implementation of various statistical modeling techniques is observed in the field of epidemiological surveillance. This fact, has led several authors to conduct detailed reviews on statistical surveillance in public health. Sonesson and Bock,(40) presented special aspects of prospective statistical surveillance as well as techniques for evaluating such methods. Farrington and Andrews,(41) described the methodological issues involved in outbreak detection, focusing mainly on infectious diseases, along with examples from a range of statistical techniques. Buckeridge et al.,(33) synthesized a research for algorithms dedicated to rapid outbreak detection. To that end, they examined how to use spatial and other covariate information from disparate sources to improve the timeliness of outbreak detection. Shmueli and Burkom,(21) discussed the statistical challenges in monitoring modern biosurveillance data through describing the current state of monitoring in the field and surveying the most recent biosurveillance literature. Unkel et al.,(42) reviewed several statistical methods for the prospective detection of infectious disease outbreaks.

Despite all the existing well-established reviews, detailed recommendations as to which statistical method is the "best" to use for outbreak detection is not possible. This is due to the fact that the latter, depends critically on the specific details of the application and implementation as well as its purpose and context.

Regression is listed among the top outbreak detection techniques. Perhaps the simplest regression model for outbreak detection is the one described by Stroup et al.(43) which, although does not incorporate time trends, it ensures that seasonal effects are automatically adjusted for by design rather than explicit modeling, thus providing some element of robustness. Another commonly used, fully parametric, regression model for outbreak detection, is that proposed by Serfling.(44) Serfling made use of a trigonometric function with linear trend, assuming Gaussian white noise errors, in order to model historical baselines. Costagliola et al.,(45; 46) based on Serfling's model and achieved the detection of the onset of influenza epidemics. Additionally, Pelat et al.,(47) developed an automated version of Serfling's model by considering cubic trend and trigonometric terms for prospective and retrospective surveillance purposes. The model selection method they used, was based on both ANOVA comparisons and Akaike Information Criterion (AIC).(48; 49) Parpoula et al.,(26) developed Serfling-type periodic regressions models and compared their performance to typical forecasting models, concluding that a periodic regression model with quadratic trend, annual, semi-annual and quarterly periodicity, as well as a moving average model of three terms had almost similar performance to the one of Pelat et al.

Changepoint analysis(50) has been proven a reliable "ally" for identifying outbreaks in scientific fields such as Medical,(51) Climate,(52; 53) Public Health,(54) Bioinformatics(55; 56) and Finance.(57) Changepoint is an instance in time where statistical properties before and after this time point differ. Such dynamical changes are often investigated through non-linear modeling. Erla et al.,(58) introduced a method to assess the complexity of the electro cortical activity, based on performing nonlinear prediction of single electroencephalogram (EEG) signals during photic stimulation (PS) and concluded that non-linear models can be a useful tool to characterize the dynamics of EEG during PS protocols. Faes et al.,(59) developed a method to perform time-varying nonlinear prediction of biomedical signals in the presence of non-stationarity and found the proposed prediction method to be suitable for quantifying the complexity of biomedical signals exhibiting nonlinear and/or nonstationary behaviors.(60) Kass-Hout et al.,(61) applied various changepoint detection methods to the active syndromic surveillance data in order to detect changes in the incidence of emergency department visits due to daily influenza-like illness. Monitoring in U.S.A. has tended to rely on detection algorithms such as Early Aberration Reporting (EARS) despite their limitation on detecting subtle changes and identifying disease trends. Hence, Kass-Hout et al. compared a combination of CUSUM method and EARS and concluded that EARS method in conjunction with change-point analysis is more effective in terms of determining the moving direction in influenza-like illness (ILI) trends between change-points. Texier et al.,(62) made use of change-point analysis for evaluating the ability of the method to locate the whole outbreak signal. Specifically, by using a kernel change-point model they were led in satisfactory results for the identification of the start and end of a disease outbreak in the absence of human resources. Christensen and Rudemo,(63) studied incidence data by using modifications of well-known hypothesis tests for retrospective detection of single and multiple change-points. By applying the multiple change-point methodology by means of modified forward selection, they concluded that the suggested method consists a useful tool for exploratory analysis of two datasets on disease incidence. Finally, Painter et al.(64) used both offline and online change-point algorithms for monitoring the quality of aggregate data. Painter and his colleagues, examined both offline and online detection using time series held at a constant lag and concluded that transient problems could be detected offline as neighboring changepoints with high posterior probability. By properly modifying their offline case conclusion, they achieved online monitoring for data quality problems.

Markov Switching Model,(65) also known as Regime Switching Model, hidden (semi-) Markov models, as well as multi-state systems, have been widely used with satisfactory results.(66)–(68). Shaby et al.,(69) approached meteorological/climatological phenomena, such as heat waves, via hidden Markov models for modeling and prediction purposes. Clements and Krolzig,(70) considered a three state MSM for modeling the Gross National Product of United States. Cao et al., (71) proposed a Markov Switching Susceptible Infected Recovered (SIR) epidemic ratio-dependent incidence rate and degenerated diffusion model. In addition, they obtained a threshold parameter useful for identifying the stochastic elimination and persistence of the disease under study. Shiferaw,(72) analyzed the dynamics of case fatality rate (CFR) of SARS-CoV-2, using Regime Switching Autoregressive (MSAR) models. He concluded that a two or three-regime MSAR approach could be more appropriate for capturing the non-linear behavior of CFR time series data for each one of the most infected countries around the world. Moreover, his results indicated that increases in CFRs are more volatile than decreases.

Control charts(73; 74) are a crucial tool for monitoring the characteristics of a process over time. The methods of Statistical Process Control (SPC) have a long history of application to problems in public health surveillance, and several approaches for detecting outbreaks of infectious diseases are directly inspired by, or related to, SPC methods.(75) Page(50) proposed control charts with memory, namely, the cumulative sum (CUSUM) and the exponentially weighted moving average (EWMA) control charts. Over the years, various extensions, modifications and variants of Page's control charts have made their appearance to serve the purposes of public health surveillance.(76)–(86)

There exist multiple reasons that may affect any assessment of the relative merits of different methods. Some of them are listed below:

1. The scope and the field of application of the public health surveillance system, e.g. how many parallel data series, which can range from one to several thousands, are to be monitored;

2. The quality of the data available, including the method of data collection, and the delay between event occurrence and reporting;

3. The spatio-temporal features of the data, such as count frequency, trend structure, seasonality, epidemicity, time step and spatial resolution;

4. The non-stationarity and the possible existence of correlations in the distribution of frequency of data;

5. The possible existence of overdispersion;

6. The features of the outbreaks that may occur, for example explosive or gradual onset, brief or long duration, low, moderate or high level of severity, or a mixture of all the above;

7. The use to which the system is to be put, including the post-signal processing protocols;

8. The availability of processing power and human resources to support the system and;

9. The choice of metrics to evaluate results.

As it may be understood, the assessment of the effectiveness of statistical and stochastic modeling techniques for outbreak detection as well as the validity of their results, which in their turn will result in safe conclusions, requires the use of evaluation criteria that are appropriately adjusted in order to serve public health surveillance purposes. However, in the scientific community, there are no widely accepted evaluation measures for this type of systems.(87; 88) Consequently, the issue which arises regarding the selection of the optimal statistical methodology for studying the changes of epidemic activity, and thus the early and accurate detection of outbreaks in public health surveillance, as well as the selection of the appropriate evaluation criteria of these methods, is a broad, complex and multifactorial research topic. This thematic area remains to some extent undeveloped, in spite of much discussion and the progress that have been made.(33; 89; 90)

1.5.1 Serfling model(44)

In 1963, Serfling analyzed pneumonia-influenza death cases with the goal of establishing an early quantitative measure of the severity of an influenza epidemic along with its geographic location. His attention focused on statistical techniques in order to construct standard curves of expected seasonal mortality against which reported deaths could be compared as they occur. The underlying idea was that an individual could use historical data to estimate seasonal trends in influenza and then, for a specific place and time, the interested researcher could be able to evaluate the amount of deaths occurred above this baseline rate. At that time, the Communicable Disease Center (i.e., todays Center for Disease Control and Prevention) was using two methods for estimating the expected weekly mortality, point-to-point linear and Fourier series with linear trend. According to Serfling these method came with some weak spots such as resulting seasonal curves which reflected distracting irregularities of the data. Hence, he proposed a model (1.1) that combined a linear term, for the modelling of trend, and sine and cosine terms, for the modelling of seasonality.

The model's structure is as follows:

$$y_t = \alpha_0 + \alpha_1 t + \sum_{i=1}^{n} \gamma_i cos\left(\theta(t)\right) + \sum_{i=1}^{n} \delta_i sin\left(\theta(t)\right) + \epsilon_t, \quad t = 1, ..., n, \tag{1.1}$$

where θ is a linear function of t.

Through a thorough investigation of (1.1), via Analysis of Variance (ANOVA) and least squares, he concluded that the optimal model for baseline influenza morbidity does not require many sine and cosine terms, rather than one of each:

$$y_t = \alpha_0 + \alpha_1 t + \gamma_1 cos\left(\frac{2\pi t}{n}\right) + \delta_1 sin\left(\frac{2\pi t}{n}\right) + \epsilon_t, \quad t = 1, ..., n,$$

where n the week cycles (usually $n = 52$) and $\epsilon_t \sim GWN\left(0, \sigma^2\right)$.

Nowadays, the paper of Serfling seems somewhat "outdated" considering the wide availability of advanced computer-based methods. Nevertheless, the concept he described keeps till now to be a reasonable and efficient approach to estimating influenza deaths.

1.5.2 Pelat et al. model(47)

In early 00's, time series data started to increase their availability in various scientific fields and especially on that of disease surveillance. The main problem with the latter field, was the lack of a dedicated tool to perform and guide analyses. To that end, Pelat et al. developed an online application, written in HTML, PHP, JavaScript and R, for the detection and quantification of epidemiological time series data (http://www.u707.jussieu.fr/periodic_regression/).

The core of the application is based on Serfling's model and a set of four basic principals:

i. **Determination of the Training Period**

   In many situations, a wide range of epidemiological time series data may be at the disposal of the researcher. This though, does not necessarily imply that all of them should be included in the training period. Over long periods of time, there is a high chance of occurring changes on both the way cases are being reported as well as various demographic characteristics. As a consequence, the goodness of fit of the baseline model to data will probably be affected. However, a minimum of one year historical data is required to fit the models of influenza morbidity. In case of more reliable predictions are of interest, then at least two or even three years of historical data should be considered.

ii. **Purge of the Training Period**

   In general, modelling long non-epidemic periods enables the identification of the so called "baseline level". the truly non-epidemic baseline level. When it comes to seasonal diseases such as influenza, this becomes hardly feasible due to the fact that epidemics occur at regular intervals (e.g., every half or one year). Two options exist for getting over such an issue. The first (less common) one requires explicit modeling of the epidemic periods. The second option is to exclude from the series data that lead to an epidemic. Regarding the second option, among various rules that are available in the literature for discarding data the most prevalent ones are: (1) exclusion of the top 15% (or 25%) values from the training period(91); (2) removal of data that fall above a given threshold(46) and; (3) exclusion of whole periods known to be epidemic prone.

iii. **Estimation of the Regression Equation**

   A wide gamma of regression equations exist. Linear(46), Linear on the log-transformed series,(92) Poisson,(93) and Poisson allowing for overdispersion,(94) are only a few. Linear regression is considered a better pick overall, when dealing with large datasets or incidences while, the use of log-transformed data or Poisson regression is advised when dealing with smaller sets of data. Although the regression equation usually incorporates a first(33; 45; 95) or a second degree polynomial(20; 44; 96) (with respect to time) for the modelling of trend, Pelat and his colleagues differentiated themselves. Specifically, they considered a third degree polynomial with respect to time into the model's structure, leading to the following formulation:

$$\begin{aligned} \mathrm{M33} = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \alpha_3 t^3 + \gamma_1 \cos\left(\frac{2\pi t}{n}\right) + \delta_1 \sin\left(\frac{2\pi t}{n}\right) \\ + \gamma_2 \cos\left(\frac{4\pi t}{n}\right) + \delta_2 \sin\left(\frac{4\pi t}{n}\right) + \gamma_3 \cos\left(\frac{8\pi t}{n}\right) + \delta_3 \sin\left(\frac{8\pi t}{n}\right) + \epsilon_t, \end{aligned} \tag{1.2}$$

where t denotes time, $\frac{i\pi t}{n}$ denotes periodicity of 1 year ($i = 2$), 6 months ($i = 4$) and 3 months ($i = 8$), respectively, and $\epsilon_t \overset{i.i.d.}{\sim} \mathcal{N}\left(0, \sigma^2\right)$. As for M33, it denotes the model that consists of a third degree polynomial with respect to time for the modelling of trend and three pairs of sine and cosine terms for the modelling of seasonality.

The model in (1.2), constitutes the pillar of their application since all the statistical analysis applies on that and eight more different variations of it (Table 1.1).

TABLE 1.1: Different variations of M33 model

	Trend			Periodicity		
Model	t	t^2	t^3	1 year	6 months	3 months
M11	*			*		
M12	*			*	*	
M13	*			*	*	*
M21	*	*		*		
M22	*	*		*	*	
M23	*	*		*	*	*
M31	*	*	*	*		
M32	*	*	*	*	*	

iv. Epidemic Alert Notification

The residuals standard error, could be utilized to asses the variation that accompanies model fit, as the baseline model is fitted to the observations. In this way, one is able to obtain forecast intervals for future observations, given the fact that the baseline model remains the same in the future. To obtain the epidemic thresholds which signal an unexpected change typically an upper percentile for the prediction distribution (assumed to be normal) is considered (mainly the upper 95th[46] or 90th[97] percentile). An increase on the percentile value, will lead inevitably to less observations outside the thresholds and more specific detection, whereas decreasing the threshold will increase sensitivity and timeliness of the alerts. Subsequently, a rule is used to define when epidemic alerts are produced, such as an observation(20) (or a series of observations) falls above the epidemic threshold (e.g., 2 weeks(98) or one month(97)). The latter step is of high importance since it prevents false alarms caused by isolated data points.

Nowadays, the application not only is still available but has also been upgraded to detect epidemics through hidden Markov models. The interested reader can access it through the following url https://periodic.sentiweb.fr.

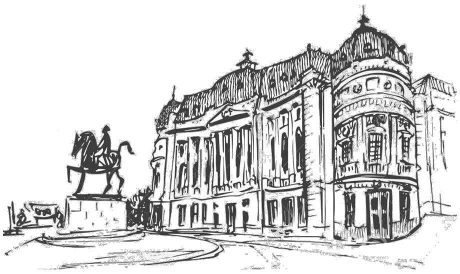
1.5.3 Parpoula et al. model(26)

Motivated by the work of Serfling and Pelat et al., Parpoula et al. applied a retrospective analysis on weekly ILI rate data concerning Greece between the period September 29, 2014 and October 2, 2016. After concluding that the best fitting model, in terms of AIC, BIC and R^2, was M23 (see Table 1.1), they conducted a comparative study between M23 and several well known forecasting models for the case of Greece, namely, linear trend (LT), simple moving average of 3 terms (MA3), a simple exponential smoothing with parameter equal to 0.1065 (SES), Holt's linear exponential smoothing with parameters equal to 0.0981 & 0.0246, Brown's quadratic exponential smoothing with parameter equal to 0.0296 (Brown's model), Winter's exponential smoothing with parameters equal to 0.1062, 0.1032 & 0.1036 (Winter's model) and that of Serfling's. Through an exhaustive comparison the best overall forecasting model, in terms of RMSE, AIC and BIC, turned out to be M23, with MA3 chasing closely (Table 1.2).

Table 1.2: Forecasting performance of each model considered

Model	RMSE	AIC	BIC
M23	12.02	547.89	572.77
LT	48.34	818.44	823.75
MA3	13.51	554.72	565.33
SES	16.06	585.05	587.00
Holt's	16.94	598.15	606.70
Brown's	18.87	618.87	620.65
Winter's	25.55	686.57	694.37
Serfling's	18.81	610.47	622.91

The models presented in Table 1.2, although adequate for modelling the baseline behavior of incidence data, they lack when it comes to the modelling of their non-typical period(s). In addition, the fact that such models do not incorporate covariates, at the time that incidence data are influenced by several factors, constitute their results somewhat “unrealistic”. Considering all the above, in this PhD thesis we will try to fill in the gap and propose robust models and techniques for improving the modelling of the whole behavior of time-series incidence data.



Bucharest, Romania

Chapter 2

Materials and Methods

In this Chapter we illustrate all the necessary materials and methods used throughout the thesis for establishing new innovative modelling techniques for incidence data.

2.1 Periodic Regression Modelling

Periodic regression models made their appearance in the early 60's(99; 100) and since then, they have been widely used in scientific areas such as Environment, (101)–(103) Hydrology(104) and Econometrics.(105)–(107) The main objective of such models is to capture the behavior of time-series data that exhibit both trend and seasonality. In this section, the distinctive characteristics of such models are presented in a detailed fashion.

In reality, periodic regression models are simply an extension of the well known seasonal ARMA models proposed by Box and Jenkins.(108) So, why "periodic" instead of "seasonal"? Mainly for two reasons; first and foremost, in periodic regression modelling the parameters constitute, as it will become evident throughout Subsections 2.1.1 to 2.1.3, periodic functions of time and second, to avoid any confusion with the seasonal ARMA models.

2.1.1 Stationary periodic autoregressive models $PAR_M(p_1, ..., p_M)$

Let us denote by $y_{m,t}$, $t = 1, ..., n$, $m = 1, ..., M$, the observations of a stationary univariate time-series of M seasons, over a period of n years. A $PAR_M(p_1, ..., p_M)$ model has the following form:

$$y_{m,t} = \sum_{i=1}^{p_m} \phi_i^{(m)} y_{m,t-i} + \epsilon_{m,t}, \tag{2.1}$$

where p_m the AR order per season m, $\phi_1^{(m)}, ..., \phi_{p_m}^{(m)}$ are autoregressive seasonally-varying parameters up to order p_m and $\epsilon_{m,t} \overset{i.i.d.}{\sim} GWN(0, \sigma^2)$.

In an essence, one could argue that a $PAR_M(p_1,...,p_M)$ model is a set of multiple $AR(p)$ models; one for each season of the year. This argument could be absolutely correct in base of $M=1$. Finally, note that some of the parameters $\phi_i^{(m)}$, $i=1,...,p_m$, may be equal to zero and as a consequence the order p_m in (2.1) is the maximum of all p_m.

Theoretical periodic autocorrelation and partial autocorrelation functions

In order to obtain the theoretical formulas for the Periodic Autocorrelation Function (PerACF) and the Periodic Partial Autocorrelation Function (PerPACF) of the model in (2.1), it is necessary to recall some useful quantities.

The first one, is the m^{th}-season theoretical periodic autocovariance function at lag k for $y_{m,t}$:

$$\begin{aligned}\gamma_k^{(m)} &= Cov\left(y_{m,t}, y_{m,t-k}\right)\\ &= E\left[\left(y_{m,t}-\mu_m\right)\left(y_{m,t-k}-\mu_{m-k}\right)\right], \quad k=0,1,2,..., \end{aligned} \tag{2.2}$$

where $\gamma_k^{(m)}$ does not depend on the year index t.

Notice that in case of $k=0$, (2.2) simply represents the variance of $y_{m,t}$.

The second quantity, is that of season's m theoretical PerACF of lag k for $y_{m,t}$:

$$\begin{aligned}\rho_k^{(m)} &= \frac{E\left[\left(y_{m,t}-\mu_m\right)\left(y_{m,t-k}-\mu_{m-k}\right)\right]}{\sqrt{E\left[\left(y_{m,t}-\mu_m\right)\left(y_{m,t}-\mu_m\right)\right]E\left[\left(y_{m-k,t}-\mu_{m-k}\right)\left(y_{m-k,t}-\mu_m\right)\right]}}\\ &= \frac{\gamma_k^{(m)}}{\sqrt{\gamma_0^{(m)}\gamma_0^{(m-k)}}}, \quad -1 \leq \rho_k^{(m)} \leq 1, \end{aligned} \tag{2.3}$$

and is independent of any scale of measurement.

Based on (2.2) and (2.3), the m^{th}-season PerACF of the model in (2.1) is defined as:

$$\rho_k^{(m)} = \frac{\sum_{i=1}^{p_m}\phi_i^{(m)}\gamma_{|k-i|}^{(m-i)}}{\sqrt{\sum_{i=1}^{p_m}\phi_i^{(m)}\gamma_i^{(m)}\sum_{i=1}^{p_m}\phi_i^{(m-k)}\gamma_i^{(m-k)}}}. \tag{2.4}$$

As for the m^{th}-season PerPACF of model (2.1), it is equal to its last AR parameter $\phi^{(p_m)}$.

Notice that k is defined in such a way so that it will be always positive. This is due to the fact that $\gamma_k^{(m)}$ is not defined for negative time lags since it is not symmetric with respect to lag k i.e., $\gamma_k^{(m)} \neq \gamma_k^{(-m)}$ which leads to $\rho_k^{(m)} \neq \rho_k^{(-m)}$.

Periodic Yule-Walker equations

The m^{th}-season periodic Yule-Walker equations, are obtained by setting $k = 1, 2, ..., p$ in the numerator of (2.4):

$$\begin{aligned}
\gamma_1^{(m)} &= \sum_{i=1}^{p_m} \phi_i^{(m)} \gamma_{|1-i|}^{(m-i)} \\
\gamma_2^{(m)} &= \sum_{i=1}^{p_m} \phi_i^{(m)} \gamma_{|2-i|}^{(m-i)} \\
&\vdots \\
\gamma_p^{(m)} &= \sum_{i=1}^{p_m} \phi_i^{(m)} \gamma_{|p-i|}^{(m-i)} .
\end{aligned}$$

2.1.2 Stationary periodic moving average models $PMA_M(q_1, ..., q_M)$

Periodic time-series models are governed by coefficients which change periodically in time. A class of periodic models suitable for the description of various seasonal time-series data, is that of periodic moving average processes.

A $PMA_M(q_1, ..., q_M)$ for $y_{m,t}$ can be written as

$$y_{m,t} = \epsilon_{m,t} + \sum_{j=1}^{q_m} \theta_j^{(m)} \epsilon_{m,t-j}, \tag{2.5}$$

where $\theta_1^{(m)}, ..., \theta_{q_m}^{(m)}$ are moving average parameters up to order q_m, which may vary with the season m and $\epsilon_{m,t} \overset{i.i.d.}{\sim} GWN(0, \sigma^2)$. For a theoretical analysis of $PMA_M(q_1, ..., q_M)$, see Cipra.(109)

2.1.3 Stationary periodic autoregressive moving average models $PARMA_M(p_1, q_1; ...; p_M, q_M)$

A wide variety of (seasonal) time-series cannot be filtered or standardized to achieve second-order stationarity.(110) This is due to the fact that the correlation structure of the series is season-depended. For example, consider a river where high runoff periods occur in the spring and low flows coupled with irrigation diversions occur in the summer. The stream-flow correlations between March, April and May, may differ from the correlations between June, July and August. In such cases, a useful class of models is that of *periodic autoregressive moving average* ($PARMA$) models,(101; 102; 111; 112) which constitute extensions of the widely used $ARMA$ models to allow parameters that depend on season.

A $PARMA_M(p_1, q_1; ...; p_M, q_M)$ for $y_{m,t}$ can be written as

$$y_{m,t} = \sum_{i=1}^{p_m} \phi_i^{(m)} y_{m,t-i} + \epsilon_{m,t} - \sum_{j=1}^{q_m} \theta_j^{(m)} \epsilon_{m,t-j}.$$

In all periodic models discussed in this Section, periodic second-order stationarity was assumed. This fact raises the necessity for imposing conditions upon the seasonal coefficients that are not easily expressed in terms of the PARMA. Luckily, one could circumvent this "problem" by using the results provided by Gladyshev.(100) As for the parameter estimation part, this is usually achieved through maximum likelihood estimation, with special cases where the moments technique is utilized. For more on this topic, and on PARMA modelling in general, the interested reader may refer to Vecchia(110) and Abu Jahel.(113)

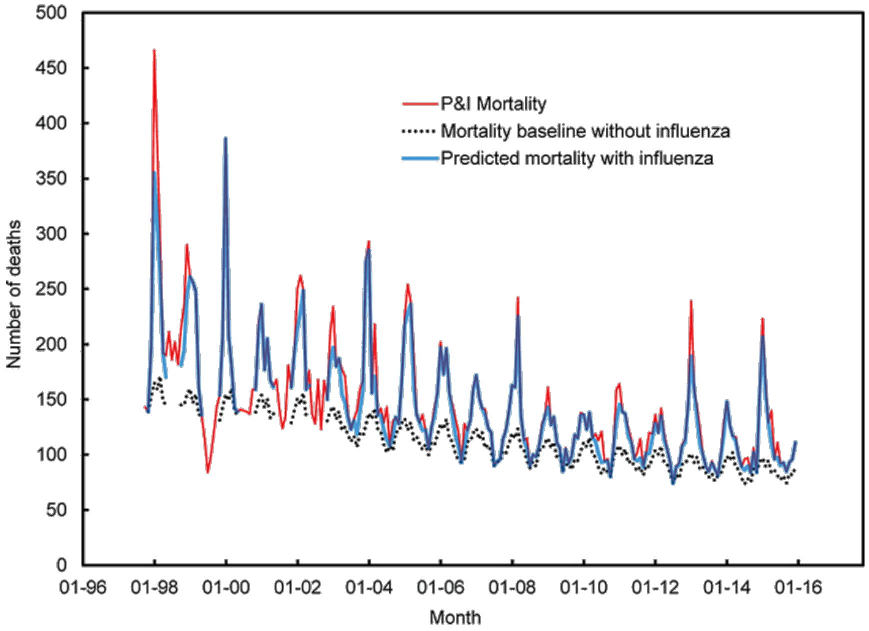


FIGURE 2.1: Monthly number of deaths due to influenza for a country (1/1996 - 1/2016) with heavy "notes" of periodicity

2.2 Changepoint Analysis

Abrupt changes (e.g., sudden jumps in level or volatility) is a common phenomenon occurring in the structure of time-series data. The points that the aforementioned changes occur, are known as changepoints and have the ability to split the data under analysis into distinct homogeneous segments. The latter is considered of high importance for various reasons such as validating an untested scientific hypothesis(114) as well as modelling assumptions(115) and monitoring and assessing safety critical processes.(116)

Let us denote by $\{y_1^n\} = \{y_1, ..., y_n\}$ a univariate time-series data set. The goal is to determine the number of changepoints, say m, along with their position τ, $\tau \in \tau_0^{m+1}$ where $\left\{\tau_0^{m+1}\right\} = \{\tau_0 = 0, \tau_1 = 1, ..., \tau_{m+1} = n\}$.

The m changepoints will result in splitting the data into $m+1$ segments where the i^{th} segment will consist of $y_{\tau_{i-1}+1}^{\tau_i}$ data. For each segment there may be a differentiation on the mean, or/and on the variance or even on the whole distribution. The parameters associated with the i^{th} segment will be denoted by θ_i and the corresponding likelihood function is given by the following formula:

$$\mathscr{L}\left(m, \tau_1^m, \theta_1^{m+1}\right) = p\left(y_1^n | m, \tau_1^m, \theta_1^{m+1}\right).$$

From now on, we will denote by $p\left(\cdot|\cdot\right)$ the general conditional density function. In addition, we will assume (1) conditional independence of data across segments, so that

$$p\left(y_1^n | m, \tau_1^m, \theta_1^{m+1}\right) = \prod_1^{m+1} p\left(y_{\tau_{i-1}+1}^{\tau_i} | \theta_i\right)$$

and (2) that for any segment we can calculate the maximum likelihood estimator for the segment parameter. The latter is denoted by ($\hat{\theta}$ or $\hat{\theta}_i$ depending on the context). Thus we have

$$\max_\theta \left\{p\left(y_{\tau_{i-1}+1}^{\tau_i} | \theta\right)\right\} = p\left(y_{\tau_{i-1}+1}^{\tau_i} | \hat{\theta}\right).$$

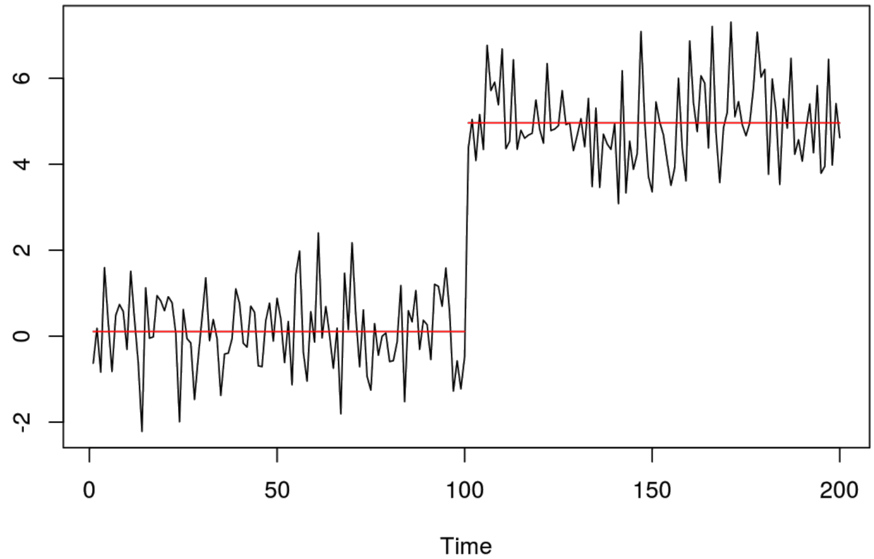


FIGURE 2.2: A times series with a shift (change) in mean around $t = 100$.

2.2.1 Single changepoint detection

In this Subsection we discuss a variety of methods for single changepoint detection.

Likelihood-ratio techniques

A common approach for detecting a single changepoint is via hypothesis testing. The hypothesis that needs to be investigated is as follows:

$$H_0 : \text{No changepoint},\ m = 0 \quad vs \quad H_1 : \text{A single changepoint},\ m = 1. \tag{2.6}$$

A general likelihood-ratio based approach for testing (2.6), was first introduced by Hinkley(117) in the late 70s, who derived the asymptotic distribution of the likelihood-ratio test statistic for a change in the mean within a sequence of normally distributed observations. Since then, Hinkley's approach has been extended to detect changes in both mean and variance.(118)–(122)

For the single changepoint detection problem, we need to specify a test statistic for concluding whether a change has occurred. The likelihood-ratio method requires calculation of the maximum log-likelihood value under both H_0 and H_1.

Under H_0, the maximum log-likelihood value is equal to

$$\log p\left(y_1^n|\hat{\theta}\right) \tag{2.7}$$

whereas under H_1, the maximum log-likelihood of a model with a changepoint at τ, $\tau \in \tau_1^m$, $\tau_1 = 1, ..., \tau_m = n - 1$, is given by the following relationship:

$$\max_{\tau}\left\{\log p\left(y_1^{\tau}|\hat{\theta}_1\right) + \log p\left(y_{\tau+1}^n|\hat{\theta}_2\right)\right\}. \tag{2.8}$$

Based on (2.7) and (2.8), the resulting test statistic is of the following form:

$$\lambda = -2\left[\log p\left(y_1^n|\hat{\theta}\right) - \max_{\tau}\left\{\log p\left(y_1^{\tau}|\hat{\theta}_1\right) + \log p\left(y_{\tau+1}^n|\hat{\theta}_2\right)\right\}\right]. \tag{2.9}$$

Of course, λ-test statistic requires selection of a threshold, say $\mathscr{C}$, so that when $\lambda > \mathscr{C}$ the null hypothesis is rejected. Rejection of H_0, corresponds to detection of a changepoint. In such a case, the estimation of its position is the value of τ, $\hat{\tau}$, that maximizes (2.8). It has to be noted that, changepoint detection constitutes a complex problem and as a result the usual asymptotic results of the likelihood-ratio statistic are no longer valid. There exist though some derivations on the asymptotic distribution of (2.9),(123) which can be used to give an approximate threshold for any desired significance level.

Penalized likelihood techniques

Penalized likelihood techniques have been quite appealing across the changepoint literature,(124; 125) mainly due to their parsimony property. As compared to likelihood-ratio based techniques, penalized likelihood ones can be extended more easily to detect multiple changepoints.

The general penalized likelihood ($\mathscr{PL}$) of a p_k-parameter model, say M_k, with unknown parameters Θ_k and a likelihood $L\left(\Theta_k\right)$, is defined as:

$$\mathscr{PL}\left(M_k\right) = -2\max_{\Theta_k}\left\{\log \mathscr{L}\left(\Theta_k\right)\right\} + p_k \phi\left(n\right),$$

where $\phi\left(n\right)$ is a user-determined penalty function (e.g., AIC ($\phi\left(n\right) = 2$), BIC ($\phi\left(n\right) = \log n$), Hannan-Quinn ($\phi\left(n\right) = 2\log\log n$)) to guard against overfitting.

It is clear that the resulted value of $\mathscr{PL}\left(M_k\right)$, is highly influenced by the choice of $\phi\left(n\right)$. For example, if one select as $\phi\left(n\right)$ the one of AIC, asymptotically he will face an overestimation problem related to the correct number of parameters.

For the detection of a single changepoint problem, M_k represents the model with k changepoints and $\Theta_k = \left(\tau_1^k, \theta_1^{k+1}\right)$ the associated parameter vector with dimension $p_k = k + (k+1)\dim\left(\theta\right)$.

Concluding, for estimating a single changepoint there is a close relation between the two techniques discussed in this Subsection, since both involve comparison of the maximum log-likelihood under H_0 and H_1, respectively. Their only difference lies upon the way of calculating the rejection region.

2.2.2 Multiple changepoint detection

Analyzing multiple changepoint models comes with a considerable computational challenge as the number of possible positions of m changepoints increases along with m. Consider for example a set of 1000 data points. Then, for a single changepoint there exist 999 possible positions while for 10 changepoints, there exist 2×10^{23} ones.

In the following Subsections we discuss various multiple changepoint techniques that are likelihood-ratio based and can be used for efficiently performing the maximization required in applying penalized likelihood methods.

Binary segmentation (BinSeg)

Perhaps the most robust search algorithm used across the changepoint literature, is that of BinSeg.(126)–(129) The main advantage of BinSeg is its ability to transform (almost) any single changepoint method to a multiple one.

The algorithm begins by testing if a single changepoint τ exists so that

$$\mathscr{R}\left(y_1^{\tau}\right) + \mathscr{R}\left(y_{\tau+1}^{n}\right) + \beta < \mathscr{R}\left(y_1^{n}\right). \tag{2.10}$$

If (2.10) stands true, then the dataset is dichotomized into two segments, namely $y_1^{\tau_\alpha}$ and $y_{\tau_\alpha+1}^{n}$ with τ_α representing the time-point where the change occurred.

Subsequently, (2.10) is applied to each segment, i.e.

$$\mathscr{R}\left(y_1^{\tau}\right)+\mathscr{R}\left(y_{\tau+1}^{\tau_\alpha}\right)+\beta<\mathscr{R}\left(y_1^{\tau_\alpha}\right) \tag{2.11}$$

and

$$\mathscr{R}\left(y_{\tau_\alpha}^{\tau}\right)+\mathscr{R}\left(y_{\tau+1}^{n}\right)+\beta<\mathscr{R}\left(y_{\tau_\alpha}^{n}\right) \tag{2.12}$$

If at least one of (2.11) and (2.12) holds, then the corresponding data are resplitted at the newly identified changepoint(s), and a hypothesis testing similar to the ones applied in (2.10), (2.11) and (2.12), is performed. The aforedescribed procedure is repeated until there are no other changepoints detected.

BinSeg is a computationally fast algorithm with a cost of $\mathscr{O}(n)$ (n the sample size). Despite its speed, the algorithm has a big flaw when it comes on the appropriate selection of $\mathscr{C}$ since different choices of it could lead to (substantially) different estimations of the number of changepoints, m.

For a more thorough discussion on the algorithmic aspect on BinSeg the interested reader may refer to Eckley et al. (130)

Segment neighborhood (SegNeigh)

An alternative search algorithm to that of BinSeg for changepoint detection is that of SegNeigh.(34; 131) In this algorithm, a criterion of data fit is defined, which is no other than the loss function ($\mathscr{R}(\cdot)$), for a segment. In the case of penalized likelihood we set

$$\mathscr{R}\left(y_s^t\right)=-\log p\left(y_s^t|\hat{\theta}\right). \tag{2.13}$$

Then a maximum number of segments, say M, is set, which corresponds to (at most) $M-1$ changepoints. The SegNeigh algorithm identifies the optimal partition of y_1^n across $m+1$ segments, $m=0,...,M-1$, through the minimization of the following loss function

$$\sum_{i=0}^{m}\mathscr{R}\left(y_{\tau_i}^{\tau_{i+1}}\right) \tag{2.14}$$

for a partition with changepoints at positions $\tau_1,\tau_2,...,\tau_m$. Finally, the algorithm outputs the best partition for $m=0,...,M-1$ along with the corresponding minimum value of (2.14).

The SegNeigh search has an $\mathscr{O}(n^2)$ computational cost, as compared to $\mathscr{O}(n)$ of the BinSeg algorithm. Hence, it is inevitable that this cost makes SegNeigh a little bit slow (in terms of execution speed). Through this disadvantage though, the algorithm achieves gaining a significantly improved predictive performance in simulation studies.(34)

Optimal partitioning (OP)

Another also quite popular changepoint detection algorithm, is that of OP. Proposed first by Yao[132] and later by Jackson et al.,[133] the method aims to the minimization of the following quantity:

$$\sum_{i=1}^{m+1} \left(\mathscr{R} \left(y_{\tau_i - 1}^{\tau_i} \right) + \beta \right). \tag{2.15}$$

The algorithmic procedure starts from setting a condition on the last changepoint, say τ_m, and calculates optimal segmentation of the data up to τ_m. Thereupon, τ_m is moved through from the beginning to the end of the data, leading to selection of an overally optimal segmentation as the final set of changepoints. Moving to more technical aspects of the algorithm, let us denote by $Q(n)$ the minimization from (2.15):

$$Q(n) = \min_{\tau} \left\{ \sum_{i=1}^{m+1} \left(\mathscr{R} \left(y_{\tau_i - 1}^{\tau_i} \right) + \beta \right) \right\}. \tag{2.16}$$

If we set $\tau_m = \tau^*$ and condition on its location, then:

$$Q(n) = \min_{\tau^*} \left\{ \min_{\tau | \tau^*} \left\{ \sum_{i=1}^{m+1} \left(\mathscr{R} \left(y_{\tau_i - 1}^{\tau_i} \right) + \beta \right) \right\} \right\}. \tag{2.17}$$

The above procedure can be repeated for the second to last, third to last,...,m^{th} to last, changepoints. This recursive "nature" of the above conditioning, becomes more evident if we set the inner minimization in (2.17) equal to $Q\left(\tau^*\right)$. In such a way, the latter equation can be re-written as:

$$Q(n) = \min_{\tau^*} \left\{ Q\left(\tau^*\right) + \mathscr{R}\left(y_{\tau^*+1}^{n}\right) \right\}. \tag{2.18}$$

Through (2.18), the global optimal segmentation can be identified by making use of (optimal) segmentations on subsets of the data. Specifically, (2.18) gives a recursive "subsistence" to the OP method since the optimal segmentation for $y_1^{\tau^*}$ is identified which is then utilized for informing the optimal segmentation for $y_1^{\tau^*+1}$. When $Q(n)$ is reached the optimal segmentation for y_1^n has been identified and thus the number as well the location of changepoints, have been recorded.

Finally, the computational cost of the OP procedure is $\mathscr{O}\left(n^2\right)$.

Pruned exact linear time (PELT)

Killick et al.,[134] on their attempt for improving the existing multiple changepoint detection search methods, proposed the PELT algorithm. Based mainly on the concept of the OP method, they proposed pruning of the values τ that can never be minima from the minimization performed at each iteration of $Q\left(\tau^*\right)$ (see (2.18) and Killick et al.[134]). Let us think a time s during the recursions of the OP algorithm so that

$$Q(s) = \min_{0 \leq \tau < s} \left\{ Q\left(\tau\right) + \mathscr{R}\left(y_{\tau+1}^{s}\right) + \beta \right\}.$$

Let also t, be a time such that $0 \leq \tau < s$ and

$$Q\left(\tau\right) + \mathscr{R}\left(y_{\tau+1}^{s}\right) + \beta > Q(s). \tag{2.19}$$

The way in which t defined, implies that the latter does not constitute the location of the last changepoint prior to time s.

If we consider a future time T, $T > s$, we can use the difference

$$Q\left(\tau\right) + \mathscr{R}\left(y_{\tau+1}^{s}\right) - Q(s)$$

to identify whether t is the location of the last changepoint prior to T. This is achieved by placing a condition on t that prevents it from being the future location of a last changepoint. In such a way, t can be removed from the minimization at each step of the PELT algorithm. According to the authors, removal of changepoint sequences that cannot be part of the final changepoints set, assures the exactness consistency of the PELT method.

Killick and her associates, summarized both the aforementioned condition and result under the following Theorem.

Theorem 1. *Suppose that there exists a constant K such that for all $t < s < T$,*

$$\mathscr{R}\left(y_{t+1}^{s}\right) + \mathscr{R}\left(y_{s+1}^{T}\right) + K < \mathscr{R}\left(y_{t+1}^{T}\right) \tag{2.20}$$

Then if

$$Q\left(\tau\right) + \mathscr{R}\left(y_{\tau+1}^{s}\right) + K > Q(s) \tag{2.21}$$

holds, at any future time $T > s$, t can never be the optimal last changepoint prior to T.

For the proof of Theorem 1, the reader is referred to Killick et al.(134)

As it follows, if (2.21) is true, then for any $T > s$ the optimal segmentation with the most recent changepoint before T being at s, will be better than any which has this most recent changepoint at t. The assumptions concerning (2.20) can be satisfied by considering a plethora of loss functions such as the log-likelihood or the penalized log-likelihood. Furthermore, it is necessary to highlight the condition set in Theorem 1 regarding the exclusion of a candidate changepoint t from future consideration. This is crucial since it reduces the amount of computations needed, and hence the computational cost, for obtaining the final set of changepoints. Finally, it has to be noted that the PELT algorithm is accompanied by a linear computational cost, i.e., $\mathscr{O}\left(n\right)$.

The pros and cons of the changepoint methods discussed in this Section, are summarized in Table 2.1.

Table 2.1: Characteristics of the changepoint methods discussed

Method	Computational Cost	Pros	Cons
BinSeg	$\mathscr{O}(n)$	Quick	Approximate
SegNeigh	$\mathscr{O}(n^2)$	Exact	Slow
OP	$\mathscr{O}(n^2)$	Exact	Slow
PELT	$\mathscr{O}(n)$	Quick & Exact	Not for all distributions

2.3 Regime Switching Modelling

2.3.1 Markov switching model of conditional mean

Various empirical studies have proven that the behavior of time-series variables may change overtime. Therefore, instead of using one model for the conditional mean of a variable, we usually use different models for capturing the aforementioned behavior. If we combine two or more dynamic models through a Markovian Switching mechanism, then the resulted model is Markov Switching Model (MSM), i.e., a time-series model of which the parameters change values based on the regime (state) that fall at time $t,\ t = 1, ..., T$.

Let us consider a random system with finite state space, say $\mathscr{D} = \{1, ..., K\}$, and an unobservable state variable $S_t,\ t,\ t = 1, ..., T$, which satisfies the first order Markov property assumption, i.e.,

$$p_{ij}(t) = P(S_t = j | S_{t-1} = i) = p_{ij}, \quad \sum_{j=1}^{K} p_{ij} = 1,$$

where the transition probabilities p_{ij} are time invariant i.e., $p_{ij}(t) = p_{ij},\ \forall t$.

Based on the above, and for a K-state model, the corresponding $K \times K$ constant transition probability matrix is as follows:

$$p_{K\times K} = \begin{bmatrix} P(S_t = 1|S_{t-1} = 1) & \cdots & P(S_t = K|S_{t-1} = 1) \\ P(S_t = 1|S_{t-1} = 2) & \cdots & P(S_t = K|S_{t-1} = 2) \\ \vdots & \ddots & \vdots \\ P(S_t = 1|S_{t-1} = K) & \cdots & P(S_t = K|S_{t-1} = K) \end{bmatrix}$$

$$= \begin{bmatrix} p_{11} & \cdots & p_{1K} \\ p_{12} & \cdots & p_{2K} \\ \vdots & \ddots & \vdots \\ p_{1K} & \cdots & p_{KK} \end{bmatrix}.$$

Let us consider a series of observable random variables $Y_t = y_t$ with corresponding hidden states $S_t = s_t,\ s_t \in \mathscr{D},\ t = 1, ...T$. A Markov Switching Model of

Conditional Mean (MSMCM) is defined as:

$$y_t = c_{s_t} + \sum_{i=1}^{p} \phi_{is_t} y_{t-i} + \epsilon_{s_t}, \;\; i \leq t \leq T, \tag{2.22}$$

where c_{s_t} is a switching intercept, ϕ_{is_t}, $i = 1, ..., p$, are autoregressive (AR) switching coefficients and ϵ_{s_t} are $i.i.d$ normally distributed random variables with zero mean and variance $\sigma^2_{s_t}$.

A distinctive characteristic of the model in (2.22), is that a set of variables can be significant in state j, but not necessarily in state i, $i \neq j$, $i, j \in \mathscr{D}$, and vice versa. Moreover, the formulation of the model, implies that (i) the dependent variable y_t is governed by K-different processes; and (ii) the stochastic transition through states, is controlled by the $p_{K\text{x}K}$ transition probability matrix.

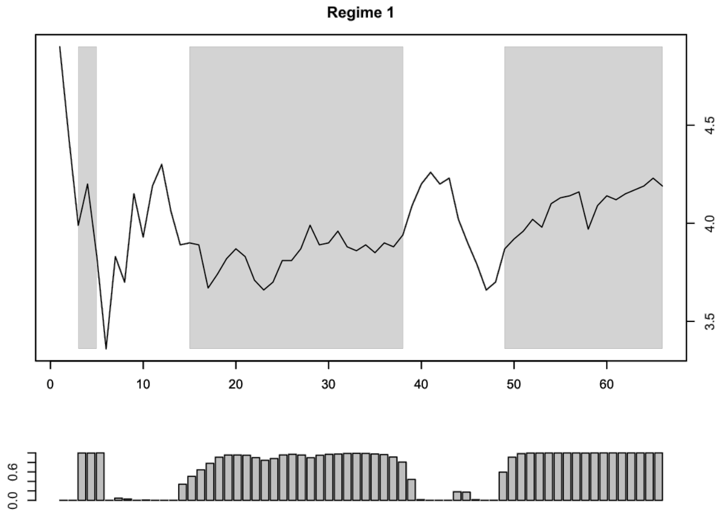


Figure 2.3: A time-series with two regimes (in white and gray, respectively).

Commonly, Markov Switching Models are also referred to as hidden Markov models, mainly due to their inclusion to the broader family of state-space models i.e., statistical models with hidden variables that control observable random variables. In general though, hidden Markov models are formulated so that the observable random variables at period t only depend on the hidden state variables at the same period, while in the Markov Switching context the observable random variables depend on their historical values as well as the hidden state variables (Figure 2.4). The latter setting makes the Markov Switching Models more "suitable" when it comes to time-series problems.

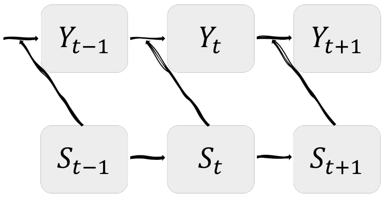


FIGURE 2.4: Mechanism of the Markov switching model.

2.3.2 Parameter inference on MSMCM – The expectation-maximization algorithm

Let $\boldsymbol{\theta}^*$ denote the vector of the unknown parameters of the model in (2.22), where:

$$\boldsymbol{\theta}^* = (c_{s_t}, \phi_{1s_t}, ..., \phi_{ps_t}, \sigma^2_{s_t}, p_{11}, ..., p_{KK}).$$

There exist several techniques for estimating $\boldsymbol{\theta}^*$, such as Maximum-Likelihood (ML), Gibbs Sampling, Monte-Carlo, etc. Among them, the widely used and well-known Expectation-Maximization (EM) algorithm(135) which constitutes an ML based technique and has been extensively used in the Regime Switching Modelling context.(136; 137)

Consider $\left\{y_1^T\right\} = \{y_1, ..., y_T\}$ to be the sequence of output observations and $\left\{s_1^T\right\} = \{s_1, ..., s_T\} \in \mathscr{D}$, the sequence of the corresponding states. The log-likelihood function of the complete data y_1^T and the unknown states s_1^T has the following form:

$$\mathscr{L}\left(y_1^T, s_1^T|\boldsymbol{\theta}^*\right) = \prod_{t=1}^{T}\sum_{s_t=1}^{K} f\left(y_t|S_t = s_t, y_1^t; \boldsymbol{\theta}^*\right) P(S_t = s_t|y_1^t)$$

$$\Leftrightarrow \log\mathscr{L}\left(y_1^T, s_1^T|\boldsymbol{\theta}^*\right) = \sum_{t=1}^{T}\log\sum_{s_t=1}^{K} f\left(y_t|S_t = s_t, y_1^t; \boldsymbol{\theta}^*\right) P\left(S_t = s_t|y_1^t\right). \tag{2.23}$$

As already mentioned, by construction, the EM algorithm is associated with ML which arises a serious issue since the necessity of MSMCM to have the state S_t depended on the previous one S_{t-1}, makes the latter insufficient for ML-type estimation. The solution to this problem, was provided by Hamilton(138) and is known as filtering. Specifically, he demonstrated that the likelihood function of the data y_t can be recursively calculated via y_t. Hamilton's filtering process though, considers that the K-states s_t are inextricably linked to the past values of y_t and thus, the resulted filtered probabilities may indicate false alarm regime changes. Kim(139) proposed a backward procedure, known as smoothing, that is applied on the filtered probabilities of Hamilton's so that probabilities conditioned on both past and future observations can be obtained. In such a way, the resulted smoothed probabilities have significantly reduced chance of misinterpreting an outlier as an actual regime change. For more on Kim's smoothing algorithm see Chen,(140) Kim and Nelson(141) and Di Persio and Vettori.(142)

Under the assumptions of model (2.22):

$$f\left(y_t|S_t = s_t, y_1^t; \boldsymbol{\theta}^*\right) = \frac{1}{\sqrt{2\pi\sigma_{s_t}^2}} exp\left(-\frac{(y_t - \mu_{s_t})^2}{2\sigma_{s_t}^2}\right),\ t = 1, ..., T,$$

where $\mu_{s_t} = c_{s_t} + \sum_{i=1}^{p} \phi_{is_t} y_{t-i}$.

The EM-algorithm constitutes an iterative process that consists of two steps: (i) the E-Step in which (smoothed) inferences on S_t are derived; and (ii) the M-Step in which the "new" ML estimates of $\boldsymbol{\theta}^*$ are calculated. The E & M Steps are repeated n-times, where n is the number of iterations needed to reach the (local) maximum of the likelihood function.

EM Algorithm

For the first iteration set $n = 0$, and set also some arbitrary initial values to $\boldsymbol{\theta}^*$, i.e.:

$$\left(\boldsymbol{\theta}^*\right)^{[n]} = \left(c_{s_t}^{[n]}, \phi_{1s_t}^{[n]}, ..., \phi_{ps_t}^{[n]}, \left(\sigma_{s_t}^2\right)^{[n]}, p_{11}^{[n]}, ..., p_{KK}^{[n]}\right).$$

E-Step

Hamilton's Filter

1. Set a (naive) guess for the initial probabilities ($t = 1$):

$$P\left(S_1 = s_1|y_1; (\boldsymbol{\theta}^*)^{[n]}\right),\ s_1 = 1, ..., K.$$

2. Set $t = 2$ and calculate:

$$P\left(S_t = s_t|y_1^t; (\boldsymbol{\theta}^*)^{[n]}\right) = \sum_{i=1}^{K} p_{is_t}^{[n]} P\left(S_t = i|y_1^t; (\boldsymbol{\theta}^*)^{[n]}\right),\ s_t \in \mathscr{D}.$$

3. Based on Bayes's rule and the total probability theorem, update the joint probability:

$$P\left(S_t = s_t|y_1^t; (\boldsymbol{\theta}^*)^{[n]}\right) = \frac{f\left(y_t|S_t = s_t, y_1^t; (\boldsymbol{\theta}^*)^{[n]}\right) P\left(S_t = s_t|y_1^t; (\boldsymbol{\theta}^*)^{[n]}\right)}{\sum_{s_t=1}^{K} f\left(y_t|S_t = s_t, y_1^t; (\boldsymbol{\theta}^*)^{[n]}\right) P\left(S_t = s_t|y_1^t; (\boldsymbol{\theta}^*)^{[n]}\right)}.$$

4. Set $t = t + 1$ and repeat Steps 2 & 3 until the sample size T is reached.

Kim's Smoother

1. Set $t = T - 1$ and calculate:

$$\begin{aligned}&P\left(S_t = s_t | y_1^T; (\boldsymbol{\theta}^*)^{[n]}\right)\\ &= \sum_{i=2}^{K} \frac{P\left(S_t = s_t | y_1^t; (\boldsymbol{\theta}^*)^{[n]}\right) P\left(S_{t+1} = i | y_1^T; (\boldsymbol{\theta}^*)^{[n]}\right) p_{s_t i}^{[n]}}{P\left(S_{t+1} = i | y_1^t; (\boldsymbol{\theta}^*)^{[n]}\right)}.\end{aligned}$$

2. Set $t = t - 1$ and repeat Step 1 until $t = 1$.

Notice that Kim's smoothing algorithm, requires the values of the filtered probabilities resulted through the filtering process of Hamilton.

M-Step

1. Set $n = n + 1$.
2. Calculate more accurate ML estimates for

$$\left(c_{s_t}^{[n]}, \phi_{1s_t}^{[n]}, ..., \phi_{ps_t}^{[n]}, \left(\sigma_{s_t}^2\right)^{[n]}\right),$$

   through explicit formulas which can be obtained by setting the partial derivatives of (7.2) equal to zero. Each formula will be a function of $P(S_t = s_t | y_1^T; (\boldsymbol{\theta}^*)^{[n-1]})$.
3. Calculate

$$p_{s_t i}^{[n]} = \sum_{t=1}^{T} \frac{P\left(S_t = i | y_1^T; (\boldsymbol{\theta}^*)^{[n-1]}\right) \frac{p_{s_t i}^{[n-1]} P\left(S_t = s_t | y_1^t; (\boldsymbol{\theta}^*)^{[n-1]}\right)}{P\left(S_t = i | y_1^t; (\boldsymbol{\theta}^*)^{[n-1]}\right)}}{P\left(S_t = s_t | y_1^T; (\boldsymbol{\theta}^*)^{[n-1]}\right)}.$$

Repeat the E & M Steps until convergence is achieved and consequently the final ML estimations for $\boldsymbol{\theta}^*$ are derived.

2.4 Discrete Time semi-Markov Chains

Let us assume a random system that has a finite state space $E = \{1, ..., N\}$, $N < \infty$, for which the time evolution is governed by a stochastic process $Z = (Z_t)_{t \in \mathbb{N}}$. In addition, let us denote by $S = (S_k)_{k \in \mathbb{N}}$ the successive time-points when state changes in $(Z_t)_{t \in \mathbb{N}}$ and by $J = (J_k)_{k \in \mathbb{N}}$ the associated visited states at these time-points.

Definition 1 (Markov renewal & semi-Markov chain). *If $(J, S) = (J_k, S_k)_{k \in \mathbb{N}}$ satisfies the relation*

$$\begin{aligned}&P(J_{k+1} = j, S_{k+1} - S_k = t | J_0, J_1, ..., J_k; S_1, ..., S_k)\\ &= P(J_{k+1} = j, S_{k+1} - S_k = t | J_k),\ j \in E,\ t \in \mathbb{N},\end{aligned} \tag{2.24}$$

then $Z = (Z_t)_{t \in \mathbb{N}}$ is a semi-Markov chain associated to the Markov renewal chain $(J, S) = (J_k, S_k)_{k \in \mathbb{N}}$, where

$$Z_t = J_{N(t)} \Leftrightarrow J_k = Z_{S_k},$$

with $N(t) = \max\{k \in \mathbb{N} | S_{k+1} \leq t\}, t \in \mathbb{N}$, being the counting process of the number of jumps in the time interval $(0, t]$. As a result, Z_t represents the state of the system at time t.(143) The fact that $(J_k, S_k)_{k \in \mathbb{N}}$ is a Markov renewal chain, implies that $(J_k)_{k \in \mathbb{N}}$ is an embedded Markov chain.

Note that throughout this thesis, we consider (J, S) to be homogeneous that is equation (2.24) is independent of k.

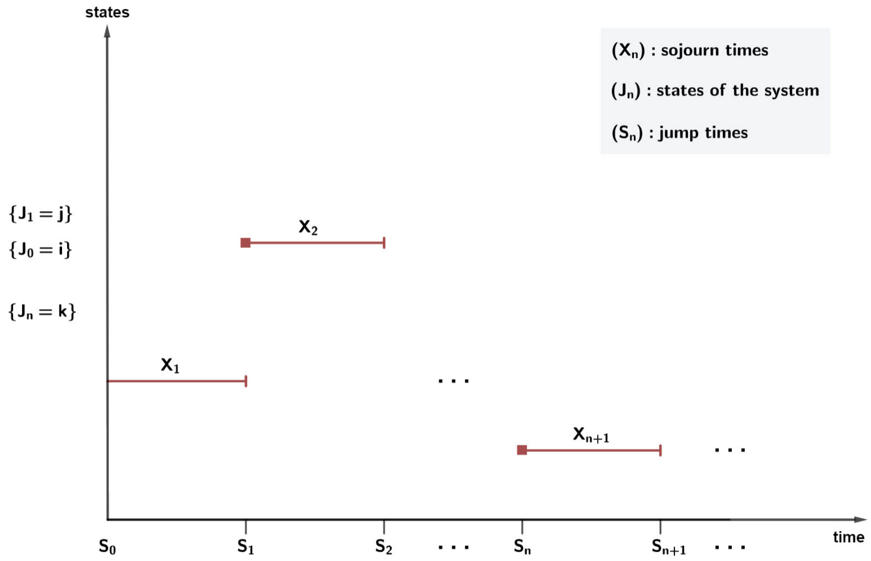


FIGURE 2.5: Sample path of a semi-Markov chain.

In order to provide some basic definitions, we introduce now the proper notation. Consider $l, k \in \mathbb{N}$, $l \leq k$, two nonnegative integers and let $y_l, ..., y_k \in A = \{1, ..., s\}, s < \infty$. We will denote by Y_l^k the vector $Y_l^k = (Y_l, ..., Y_k)$ and we will write $\{Y_l^k = y_l^k\}$ for the event $\{Y_l = y_l, ..., Y_k = y_k\}$. In the case of a single state within the state space, i.e., $y_l, ..., y_k \equiv y \in A$, we denote by $\{Y_l^k = y\}$ the event $\{Y_l = y, ..., Y_k = y\}$. Finally, the notation $\{Y_l^k = \cdot\}$ refers to the event $\{Y_l = \cdot, ..., Y_k = \cdot\}$. It is obvious that all the above notations in terms of the chain Y can be easily expressed in terms of the chain Z.

Definition 2 (Hidden semi-Markov chain of order k). *Let $Y = (Y_t)_{t \in \mathbb{N}}$ be a homogeneous Markov chain of order k, $k \geq 1$, conditioned on the semi-Markov chain Z which means that $\forall y_0, ..., y_k \in A$, $i \in E, t \in \mathbb{N}^*$:*

$$\begin{aligned} & P(Y_{t+1} = y_k | Y_{t-k+1}^t = y_0^{k-1}, Y_0^{t-k} = \cdot, Z_{t+1} = i, Z_0^t = \cdot) \\ & = P(Y_{t+1} = y_k | Y_{t-k+1}^t = y_0^{k-1}, Z_{t+1} = i). \end{aligned} \tag{2.25}$$

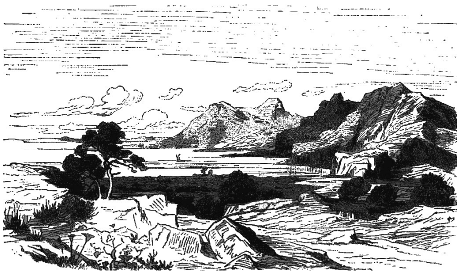
The chain $(Z, Y) = (Z_t, Y_t)_{t \in \mathbb{N}}$ is called a hidden semi-Markov chain of order k and the probability in (2.25) is known as the emission probability matrix of the conditional Markov chain Y.

If in (2.25) the observation process is characterized by the conditional independence property, then $\forall y \in A,\ i \in E, t \in \mathbb{N}^*$:

$$\begin{aligned}&P(Y_t = y|Y_0^{t-1} = \cdot, Z_t = i, Z_0^{t-1} = \cdot)\\&= P(Y_t = y|Z_t = i),\end{aligned}$$

where $\sum_{y_k} P(Y_{t+1} = y_k|Z_{t+1} = i) = 1$.

For more information on the topic of hidden semi-Markov chains the interested reader may refer to Barbu and Limnios.(144)



Pass Of Thermopylae, Lamia, Greece

Chapter 3

On Mixed PARMA Modeling of Epidemiological Time Series Data

Cite: **Kalligeris EN**, Karagrigoriou A and Parpoula C. On Mixed PARMA Modeling of Epidemiological Time Series Data. *Communications in Statistics-Case Studies, Data Analysis and Applications* 2020; **6(1)**: 36-49.

Chapter's Goal

Real time surveillance of epidemic activity in epidemiological surveillance systems, is often difficult to be fully achieved due to the seasonality involved in the series. In this Chapter, for the modeling of incidence data (see Section 1.4) concerning weekly estimated influenza like illness (ILI) rates[a], the general form as well as **special cases** of PARMA models (see Section 2.1) are considered. In the upcoming sections, the methodology followed is the one described in Subsection 1.5.2.

[a] The weekly estimated ILI rate, is a time series with specific characteristic properties, such as trend and seasonality.

3.1 Estimation of the Regression Equation

In this Section, an exhaustive search process (based on periodic mixed regression models) is performed in order to identify the optimal fit of the baseline model. Thus, linear, quadratic, cubic and quartic (for comparison purposes) trends are considered, and regarding the seasonal component, the most widely used periodicities are implemented, i.e., 12, 6, and 3 months. For a review of a general class of such models see Vasdekis et al. (145)

Note that other terms could also be included in the regression equation. For example, some authors incorporated variables, such as the day of the week, holiday, and post-holiday effects,(146) sex and age,(147) or climatological factors e.g., temperature and humidity.(148) Inclusion of all these terms may offers more flexibility, but it will be more prone to result in unidentifiable models or other problems related to model fitting. In the following sections, we focus on the study of the possible impact of several climatological factors (*temperature*, *humidity*, *wind force*, and *wind direction*) on influenza morbidity.

Several authors have studied the climate changes and how these affect public health.(149; 150) As a matter of fact, it has already been observed that higher temperatures are likely to increase heat-related mortality worldwide. In addition, there is strong evidence that high temperatures are associated with mortality.(151) The connection between temperature and mortality may be confounded by a range of measured (or unmeasured) confounders. Confounding factors are present when a covariate is strongly associated with both the outcome and exposure of interest, but it is not a result of the exposure and may distort the association being studied between two other variables. As pointed out by Touloumi et al.,(152) a fundamental consideration in epidemiological modeling is to properly control for all potential confounders. Such confounders may include meteorological indicators, such as relative humidity, seasonality and long-term trends. In addition, Tsangari et al.(153) concluded that high temperatures during warm months can result in increased mortality rates. Since influenza causes an estimated 290000-650000 deaths worldwide,(154) it is considered reasonable to study the possible effects of climatological factors on ILI.(148) Motivated by the aforementioned, additional climatological factors were incorporated into the model's structure namely, *minimum-maximum-median-mean temperature (temp)*, *minimum-maximum-median-mean wind direction (wd)*, and *minimum-maximum-median-mean wind force (wf)*. Moreover, first-second order auto-regressive terms and first-second order moving average terms, were also considered.

Combining all the above components, the regression equation is defined by the following *mixed PARMA(2,2)* model:

$$\begin{aligned} y_t &= \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \alpha_3 t^3 + \alpha_4 t^4 + \gamma_1 cos\left(\frac{2\pi t}{n}\right) + \delta_1 sin\left(\frac{2\pi t}{n}\right) \\ &+ \gamma_2 cos\left(\frac{4\pi t}{n}\right) + \delta_2 sin\left(\frac{4\pi t}{n}\right) + \gamma_3 cos\left(\frac{8\pi t}{n}\right) + \delta_3 sin\left(\frac{8\pi t}{n}\right) \\ &+ \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t + \lambda_1 \epsilon_{t-1} + \lambda_2 \epsilon_{t-2} \\ &+ \zeta_1 \text{minwd} + \zeta_2 \text{maxwd} + \zeta_3 \text{medianwd} + \zeta_4 \text{meanwd} \\ &+ \theta_1 \text{minwf} + \theta_2 \text{maxwf} + \theta_3 \text{medianwf} + \theta_4 \text{meanwf} \\ &+ \omega_1 \text{mintemp} + \omega_2 \text{maxtemp} + \omega_3 \text{mediantemp} + \omega_4 \text{meantemp}, \end{aligned} \tag{3.1}$$

where α_i, γ_i, δ_i, ϕ_i, λ_i, ζ_i, θ_i, ω_i, are the appropriate coefficients of the relevant terms and $\epsilon_t \sim WN(0, \sigma^2)$.

Next, a thorough comparison takes place among all candidate periodic mixed models, i.e.:

- *mixed* $PAR(1)$
- *mixed* $PAR(2)$
- *mixed* $PMA(1)$

- *mixed* $PMA(2)$
- *mixed* $PARMA(1,1)$
- *mixed* $PARMA(2,1)$
- *mixed* $PARMA(1,2)$
- *mixed* $PARMA(2,1)$

with respect to the significance of the climatological factors. Note that all regression equations for the observed value y_t are special cases of Equation (3.1).

The **initial** model selection process is described step by step as follows: We start from the simplest model, labeled as $PAR(1)$, by examining the significance of each of the climatological explanatory variables of the mixed model. If there is at least one significant, then we keep the model and go on to the next one (e.g., $PAR(2)$). The procedure continuous until the significance of the climatological factors for each model is examined. Finally, the models kept by the process, are being compared with respect to Modified Divergence Information Criterion ($MDIC$) criterion[155] and the model with the lowest $MDIC$ value is selected. In our case, and based on the above procedure, a $PARMA(2,1)$ mixed model with *minimum temperature* as a significant covariate ($p-value < .001$) was selected.[156]

As a result, the time period and the minimum temperature are the explanatory variables, the observed time series values, weekly ILI rate (number of ILI cases per 1000 visits), is the dependent variable and all regression equations for the observed value y_t are special cases of the following $PARMA(2,1)$ mixed model:

$$\begin{aligned} y_t = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \alpha_3 t^3 + \alpha_4 t^4 + \gamma_1 cos\left(\frac{2\pi t}{n}\right) + \delta_1 sin\left(\frac{2\pi t}{n}\right) \\ + \gamma_2 cos\left(\frac{4\pi t}{n}\right) + \delta_2 sin\left(\frac{4\pi t}{n}\right) + \gamma_3 cos\left(\frac{8\pi t}{n}\right) + \delta_3 sin\left(\frac{8\pi t}{n}\right) \\ + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t + \lambda_1 \epsilon_{t-1} + \omega_1 \text{mintemp}, \end{aligned}$$

where $\epsilon_t \sim WN(0, \sigma^2)$, n denotes the sample size, and parameter coefficients are estimated by least squares regression.

Selection of the best fitting model, in terms of trend and seasonality, is made possible by an exhaustive search among twelve candidate models (combining the four trends, namely, linear, quadratic, cubic and quartic, and the three seasonal periodicities, namely, 12, 6, and 3 months), and the selection process is relied on Analysis of Variance ($ANOVA$) comparison (significance level α is chosen to be 5%) to select between nested models, and on Akaike Information Criterion (AIC), or Modified Divergence Information Criterion ($MDIC$), or Bayesian Information Criterion (BIC) to select between non-nested models.

The latter process is described step by step as follows:

Step 1: The process starts by comparing (via ANOVA) the simplest model labeled as $M11$ (linear trend and $12-$month seasonal periodicity), and defined as

$$\begin{aligned} M11: \quad y_t = {} & \alpha_0 + \alpha_1 t + \gamma_1 cos\left(\frac{2\pi t}{n}\right) + \delta_1 sin\left(\frac{2\pi t}{n}\right) + \phi_1 y_{t-1} \\ & + \phi_2 y_{t-2} + \epsilon_t + \lambda_1 \epsilon_{t-1} + \omega_1 \text{mintemp}, \end{aligned} \tag{3.2}$$

with the two models within it is nested, labeled as $M12$ (linear trend and $12-$ and $6-$month seasonal periodicities) and $M21$ (quadratic trend and $12-$month seasonal periodicity), which are defined as

$$\begin{aligned} M12: \quad y_t = {} & \alpha_0 + \alpha_1 t + \gamma_1 cos\left(\frac{2\pi t}{n}\right) + \delta_1 sin\left(\frac{2\pi t}{n}\right) + \gamma_2 cos\left(\frac{4\pi t}{n}\right) \\ & + \delta_2 sin\left(\frac{4\pi t}{n}\right) + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t + \lambda_1 \epsilon_{t-1} + \omega_1 \text{mintemp}, \end{aligned}$$

and

$$\begin{aligned} M21: \quad y_t = {} & \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \gamma_1 cos\left(\frac{2\pi t}{n}\right) + \delta_1 sin\left(\frac{2\pi t}{n}\right) + \gamma_2 cos\left(\frac{4\pi t}{n}\right) \\ & + \delta_2 sin\left(\frac{4\pi t}{n}\right) + \gamma_3 cos\left(\frac{8\pi t}{n}\right) + \delta_3 sin\left(\frac{8\pi t}{n}\right) + \phi_1 y_{t-1} + \phi_2 y_{t-2} \\ & + \epsilon_t + \lambda_1 \epsilon_{t-1} + \omega_1 \text{mintemp}. \end{aligned}$$

Step 2: In the case that none of the alternative models ($M12$ and $M21$) is significantly better than the initial one ($M11$), the process retains $M11$ and stops.

Step 3: If one of the two alternative models is better than the initial one ($p-value < 0.05$), the algorithmic process keeps it and goes on.

Step 4: If both alternative models are better than the initial one, the algorithmic process keeps the one with the lowest AIC, $MDIC$ or BIC respectively, and goes on.

The procedure is repeated until finding the "best overall" model over the twelve candidate models.

3.1.1 Epidemic alert notification

The epidemic thresholds which signal an unexpected change are typically obtained by taking an upper percentile for the prediction distribution (assumed to be normal), usually the upper 95th percentile,(46) or upper 90th percentile.(97) In addition, a minimum period above the epidemic threshold is also required. The latter step is important since in this way we avoid making alerts for isolated data points. In this paper, the rule was set to be "*a series of observations fall above the epidemic threshold during 2 weeks*"(47; 91)). This way, the beginning of the epidemic is signaled the first time the series exceeds the threshold, and the end, the first time the series returns below the threshold.

3.1.2 Model identification

We conducted a retrospective analysis; the whole time series with 105 observations was therefore included in the training period (as done for example in Parpoula et al.(26)). Then, we chose to exclude the top 15% observations from the training period (89 values kept out of 105). Based on ANOVA comparisons, and AIC, BIC and $MDIC$ criteria values, the model selected was $\boldsymbol{M11}$ with a linear trend, an annual periodic term (one year harmonics), first and second order auto-regressive terms, a first order moving average term, and the minimum temperature. The forecast interval was set to be 95%, that is the upper limit of the prediction interval which is used as a threshold to detect epidemics. The alert rule, was chosen to be "*an epidemic is declared when 2 weekly successive observations are above the estimated threshold*".

The mathematical form of $\boldsymbol{M11}$ is described as follows:

$$y_t = \alpha_0 + \alpha_1 t + \gamma_1 cos\left(\frac{2\pi t}{n}\right) + \delta_1 sin\left(\frac{2\pi t}{n}\right) + \phi_1 y_{t-1} \\ + \phi_2 y_{t-2} + \epsilon_t + \lambda_1 \epsilon_{t-1} + \omega_1 \text{mintemp}.$$

Table 3.1 presents the estimated parameters, the standard errors (sd), the test statistic values (t-value) and the associated p-values of the selected model. The twelve periodic regression mixed models are described in Table 3.2, in which the components included in each model are indicated by "*", along with AIC, $MDIC$, BIC and $R^2_{GLLM(m)}$(157) values of each model. The model finally kept $M11$ is in bold italics.

Table 3.1: Selected Model M11

Parameter	Estimate	sd	t-value	p-value
α_0	10.468	2.128	4.920	$< .001$
α_1	-0.075	0.027	-2.789	0.007
γ_1	-10.315	1.444	-7.144	$< .001$
δ_1	12.726	2.144	5.937	$< .001$
ϕ_1	0.811	0.114	7.098	$< .001$
ϕ_2	-0.234	0.090	-2.613	0.011
λ_1	0.261	0.135	-2.613	0.058
ω_1	0.729	0.172	4.247	$< .001$

Figure 3.1 illustrates the model selection pathway using MDIC criterion. The model selection pathway for *ANOVA & AIC* and *ANOVA & BIC* respectively, is exactly the same as in Figure 3.1 and therefore is chosen not to be presented. In addition, Figure **??** illustrates the plots of the time series, the predicted baseline as well as the threshold. In addition, the epidemics detected by the selected model (***M*11**) appear in light red. Table 3.3 presents the dates and the results of the retrospective evaluation of the excess influenza morbidity[1] in Greece for 2014-2016 along with excess percentages[2], using the $M11$ periodic regression mixed model.

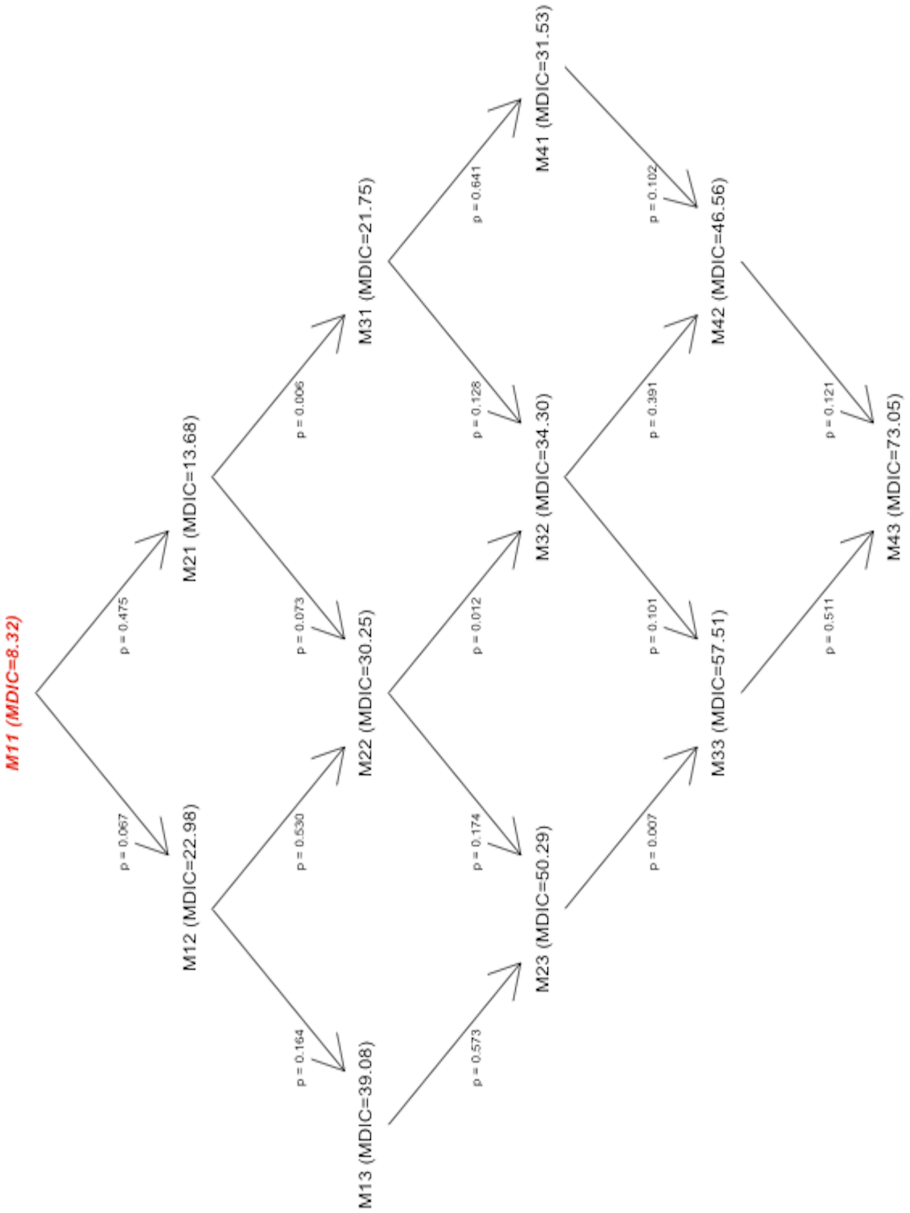


FIGURE 3.1: Model selection pathway (ANOVA & MDIC).

[1] The excess morbidity is defined as the cumulative difference between observations and baseline over the entire epidemic period.

[2] Excess percentages were calculated as the observed size divided by the sum of expected values throughout each epidemic.

It is worth to be noted that we chose $MDIC$ over AIC or BIC, because of the interesting characteristics that seems to appear and will be discussed in Section 3.1.3.

TABLE 3.2: Models selected through the algorithm pathway

	T[a]				P[b]			ARMA			LV[c]	IC[d]			R^2
M[e]	t	t^2	t^3	t^4	1 y[f]	6 m[g]	3 m	AR(1)	AR(2)	MA(1)	MT[h]	AIC	MDIC	BIC	$R^2_{GLMM(mar)}$
M11	*				*			*	*	*	*	409.31	8.32	429.29	0.927
M12	*				*	*		*	*	*	*	406.96	22.98	431.37	0.934
M13	*				*	*	*	*	*	*	*	406.56	39.08	435.42	0.928
M21	*	*			*			*	*	*	*	410.72	13.68	432.92	0.938
M22	*	*			*	*		*	*	*	*	408.49	30.25	435.11	0.934
M23	*	*			*	*	*	*	*	*	*	408.17	50.29	439.24	0.937
M31	*	*	*		*			*	*	*	*	403.82	21.75	428.23	0.938
M32	*	*	*		*	*		*	*	*	*	402.83	34.30	431.67	0.941
M33	*	*	*		*	*	*	*	*	*	*	401.04	57.51	434.33	0.937
M41	*	*	*	*	*			*	*	*	*	405.56	31.53	432.19	0.946
M42	*	*	*	*	*	*		*	*	*	*	403.90	46.56	434.97	0.942
M43	*	*	*	*	*	*	*	*	*	*	*	402.48	73.05	437.98	0.946

[a] "T" denotes *trend*; [b] "P" denotes *periodicity*; [c] "LV" denotes *latent variable*; [d] "IC" denotes *information criterion*; [e] "M" denotes *model*; [f] "y" denotes *year*; [g] "m" denotes *months*; [H] "MT" denotes *minimum temperature.*

Table 3.3: Retrospective evaluation of the excess influenza morbidity, Greece 2014-2016

SW [a]	EW [a]	Cases	Expected cases	Excess cases	Excess percentage
201501	201512	1151	891	260	29%
201605	201608	316	225	91	40%

[a] SW and EW denote the signaled start and end weeks for epidemics, respectively.

3.1.3 Model performance evaluation

In this section we evaluate the predictive performance of the selected model (M11) in comparison with other significance models considered in this analysis. The selected periodic regression mixed model, identified as the optimal one, was $M11$ defined in Equation (3.2) (see also Table 3.2). However, one can easily observe from Figure 3.1, that the algorithm could have proceed and therefore stopped at $M12$, since the p-value is close to $\alpha = 5\%$ (0.067). It is worth to be noted that if we exclusively take into account the AIC, BIC and $R^2_{GLMM(mar)}$ values, then the “best overall” models would be $M21$, $M31$ and $M33$, respectively. Thus, one could make a comparison between the aforementioned models ($M11$, $M12$, $M21$, $M31$ and $M33$) in order to ensure the selection of the “best overall” model with respect to several measures of prediction accuracy of a forecasting model, such as:

- *Mean Error (ME)*,
- *Root Mean Squared Error (RMSE)*,
- *Mean Absolute Error (MAE)*,
- *Mean Percentage Error (MPE)*,
- *Mean Absolute Percentage Error (MAPE)* and
- *Mean Absolute Scaled Error (MASE).*(158)

The results of the comparative study are presented in Table 3.4 where the full model (M43) has also been included for the shake of completeness.

TABLE 3.4: Common accuracy measures

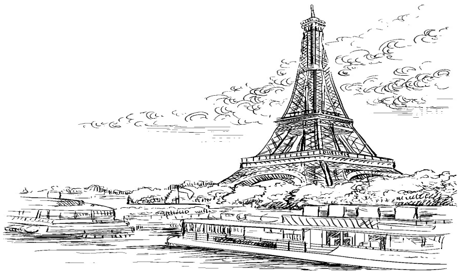
Model	ME	RMSE	MAE	MPE	MAPE	MASE
M11	$-2.359224e^{-16}$	4.298974	3.040053	0.2782516	48.19166	0.230213
M12	$2.179757e^{-16}$	4.10272	2.857186	−23.29906	75.60382	0.2163651
M21	$4.937659e^{-17}$	4.280225	2.9915	7.396555	50.17105	0.2265362
M31	$-3.469447e^{-17}$	4.009004	2.853861	22.26226	63.17711	0.2161133
M33	$7.6736e^{-17}$	3.703522	2.757	−69.34156	127.622	0.2087783
M43	$9.953614e^{-17}$	3.68828	2.783278	−69.73654	132.0232	0.2107683

We observe from that models $M21$ and $M31$ are not satisfactory since their results are never the best for none of the accuracy measures examined. Although, the results for models $M11$, $M12$, and $M33$, are similar, $M11$ clearly outperforms the others in terms of MPE and $MAPE$. Even if model $M33$ was better, we would choose again $M11$, since it has the advantage of less explanatory variables and a $R^2_{GLMM(mar)}$ value very close to the one of $M43$. Thus, $M11$ is the "best overall" model.

It is worth to be noted that $MDIC$, AIC and BIC values of the models are in full support of the above results. Indeed, $MDIC$ is clearly in favor of model $M11$ (smallest $MDIC$ value). In respect to AIC and BIC values, we observe that $M11$ is not the best; however, by choosing alternative models such as $M12$, $M33$ or $M43$, the gain is not significant enough to balance the complexity associated with these models. Moreover, the $MDIC$ values tend to get bigger as more explanatory variables are included in the model. In fact, the penalty given by $MDIC$ to the models is much bigger compared to the penalty given by AIC. Thus, the addition of explanatory variables makes $MDIC$ a somewhat "stricter metric" in comparison with AIC, and nearly the exact opposite of $R^2_{GLMM(mar)}$.

3.2 Concluding Remarks

Conclusively, in this study, we conducted a retrospective analysis of epidemiological time series data (week40/2014 to week39/2016) for Greece. We developed an alternative approach in order to model seasonality of influenza, based on a periodic regression model which incorporates an additional auto-regressive and moving average component into Serfling's model including additionally climatological and meteorological covariates associated with ILI, with the ultimate aim of the early and accurate outbreak detection. The model selected (via an exhaustive search process) as the optimal one succeeded in estimating accurately the influenza-like syndrome morbidity burden in Greece for the period 2014-2016 as well as the duration of the epidemic waves. Within this framework, the present work provided general recommendations to serve critical needs of Public Health for the very early and accurate detection of epidemic activity.



Eiffel Tower, Paris, France

Chapter 4

Periodic-type ARMA Modeling with Covariates for Time-Series Incidence Data via Changepoint Detection

Cite: **Kalligeris EN**, Karagrigoriou A and Parpoula C. Periodic-Type Auto-Regressive Moving Average Modeling with Covariates for Time-Series Incidence Data via Changepoint Detection. *Statistical Methods in Medical Research* 2020; **29(6)**: 1639-1649.

Chapter's Goal

In Chapter 3, we focused **solely** on capturing the behavior of non-extreme periods of incidence data. However, identifying the **full** time course of data such as the ILI rates for Greece (2014-2016) is useful for several reasons. Indeed, identifying the end of an epidemic, helps public health officials in both knowing when response activities can cease and determining whether new cases are part of an already known (or a new) outbreak.

Hence, in this Chapter we attempt to capture the behavior of both non-extreme and extreme periods that occur in time-series incidence data. The identification of extreme periods is made possible via changepoint detection analysis (see Section 2.2) and model selection techniques are developed in order to identify the optimal PARMA model with covariates that best describes the pattern of the time-series. Finally, in the context of incidence data modeling, an advanced algorithm was developed in order to improve the accuracy of the selected model.

4.1 Changepoint Detection, Periodic ARMA Modelling, and Estimation Performance

In this Chapter, we make use of the Segment Neighbourhood method which is arguably the most widely used changepoint search method and like *PELT*, is exact. It is worth to be noted that the *PELT* method was also examined with derived results almost identical to the ones by *SegNeigh*. The *PELT* method gave satisfactory results but the *SegNeigh* was preferred since it seems to give more reliable results as compared with previous analyses based on Serfling-type periodic regression modeling of the same data.(26) For a comprehensive survey on changepoint see Fryzlewicz(159) and Aminikhanghahi and Cook.(160)

4.1.1 Modeling of time-series incidence data

In Chapter 3 we extended the work of Pelat et al.(47) and Parpoula et al.(26) using mixed PARMA models for estimating the non-epidemic period of ILI rate data. Specifically, they made use of 105 weekly ILI rate observations $(\tau_1, ..., \tau_{105})$ for Greece, between September 29, 2014 and October 2, 2016 that were provided by the National Public Health Organization (NPHO) of Greece.

Identification of extreme periods

In this work, we provide a general, user friendly and computationally fast algorithmic procedure for estimating not only the baseline level but also the extreme periods of the time-series via changepoint analysis. Firstly, we applied the changepoint method to the time-series data, using the *SegNeigh* algorithm,(161) and changes in mean were found at time-points

- τ_{13},
- τ_{25},
- τ_{66}, and
- τ_{74},

as also seen in Figure 4.1.

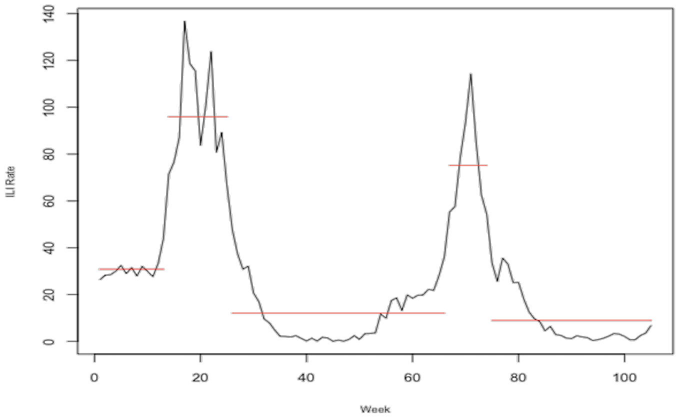


Figure 4.1: Changepoints in mean

The method detected five periods described below in terms of time-points τ. Two of them were found to be epidemic and are highlighted in red as follows.

- $\tau_1 - \tau_{13}$ (week40/2014-week52/2014),
- $\tau_{14} - \tau_{25}$ (week1/2015-week12/2015),
- $\tau_{26} - \tau_{66}$ (week13/2015-week53/2015),
- $\tau_{67} - \tau_{74}$ (week1/2016-week8/2016),
- $\tau_{75} - \tau_{105}$ (week9/2016-week39/2016).

Remark 1 The resulted epidemic periods could be explained by both an epidemiological and a statistical perspective. From an epidemiological perspective there are usually certain weeks that are expected to be epidemic prone. However, the burden of disease varies from year to year according to the specific epidemiological features of ILI syndrome, meteorological factors or even socio-economic conditions e.g. vaccinations, health policies etc. From a statistical perspective, in a disease (as well as an ILI rate) surveillance problem, there is little to no control over disease incidence, the distribution of disease incidence is non-stationary, and disease incidence (usually) returns to its original state once an outbreak has run its course. To aid interpretation, the detected changepoints enable experimenters to divide the whole time series into two types of segments based on the direction and magnitude of the incidence of disease: severe outbreaks (non-typical activity) and non-epidemic activity (typical activity), from highest to lowest public health interest in terms of alarm signals. Therefore, a beginning of an epidemic trend is a change point whose timely detection will predict occurrence of a new epidemic. Similar explanations as the ones given above, could be used in several scientific fields that incidence data occur such as seismology(66)–(68) and meteorology.(69)

Remark 2 The changepoint algorithm identifies all possible time points τ_κ where the mean changes significantly. Then the time series is divided according to the identified time points τ_κ and then two sample t-tests take place. The *Segment Neighbourhood* (*SegNeigh*) approach by default enables the experimenter to guarantee a prescribed alarm false detection probability (usually $\alpha = 5\%$). Irrespectively of the theoretical false detection probability set, the main characteristic of *SegNeigh* though, is its accuracy which is accomplished through an exhaustive iteration process (with a computational cost of $\mathscr{O}\left(n^2\right)$(162) where $n = \textit{sample size}$). Through the above iteration process the final changepoints selected have a Pvalue almost equal to 0. As a result, the method almost always achieves to select the correct time points τ_κ where the mean changes significantly. This feature is fully verified in the simulation study (see Section 4.3) where in all cases the true changepoints have been identified correctly.

Note that the purpose of the changepoint analysis is the identification of epidemic periods which justifies the splitting of the original dataset into non-extreme and extreme periods. Several rules have been suggested in the literature in this respect, such as excluding the 15% or 25% higher values from the training period,(91) removing all data above a given threshold,(46) or excluding whole periods known to be epidemic prone. One of the contributions of the present work lies on the fact that it does not rely on arbitrary pruning and in that sense is filling up the gap in the relevant literature.

The modeling of non-extreme periods

For modeling of non-epidemic time-series data, we pruned the observations that were characterized as epidemic through the changepoint analysis in the previous subsection. Then, the algorithm proposed in Chapter 3 was executed, with respect to *AIC* and *ANOVA* comparisons, for several models with trend, periodicity, AR, and MA terms as well as the average minimum weekly temperature which was the only covariate identified as significant among a number of possibly significant meteorological variables considered. The aforementioned Periodic-type Auto-Regressive Moving Average (ARMA) models with covariates are denoted by Mij, were $i = 1, 2, 3, 4$ corresponds to linear (1), quadratic (2), cubic (3) and quartic (4) trend, respectively and $j = 1, 2, 3$ corresponds to annual (1), 6-month (2) and 3-month (3) periodicity, respectively. Furthermore, an ARMA(1,1) model was chosen by a preliminary model selection process for all models considered. The process starts comparing by *ANOVA* the simplest model labeled as M11 with the two models in which it is nested, labeled as M12 and M21. In the case that none of the alternative models (M12 and M21) is significantly better than the initial one (M11), the process retains M11 and stops. If one of the two alternative models is better than the initial one, the algorithmic process keeps it and goes on. In the case that both of the alternative models are better than the initial one, in terms of the *ANOVA* comparison, the algorithmic process keeps the one with the lowest *AIC* and then goes on. The procedure is repeated until finding the “best overall” model over the twelve candidate models Mij, $i = 1, 2, 3, 4$, $j = 1, 2, 3$. Figure 4.2 illustrates the model selection pathway using *ANOVA* and *AIC* criterion. The model selected as the optimal one for baseline influenza morbidity was the M23 with quadratic trend, 12-month, 3-month and 6-month seasonal periodicity, ARMA(1,1) terms and minimum temperature as meteorological covariate, which is described as follows:

$$\begin{aligned} y_t = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \gamma_1 cos\left(\frac{2\pi t}{n}\right) + \delta_1 sin\left(\frac{2\pi t}{n}\right) \\ + \gamma_2 cos\left(\frac{4\pi t}{n}\right) + \delta_2 sin\left(\frac{4\pi t}{n}\right) + \gamma_3 cos\left(\frac{8\pi t}{n}\right) \\ + \delta_3 sin\left(\frac{8\pi t}{n}\right) + \phi_1 y_{t-1} + \epsilon_t + \lambda_1 \epsilon_{t-1} + \omega_1 \text{mintemp}. \end{aligned} \tag{4.1}$$

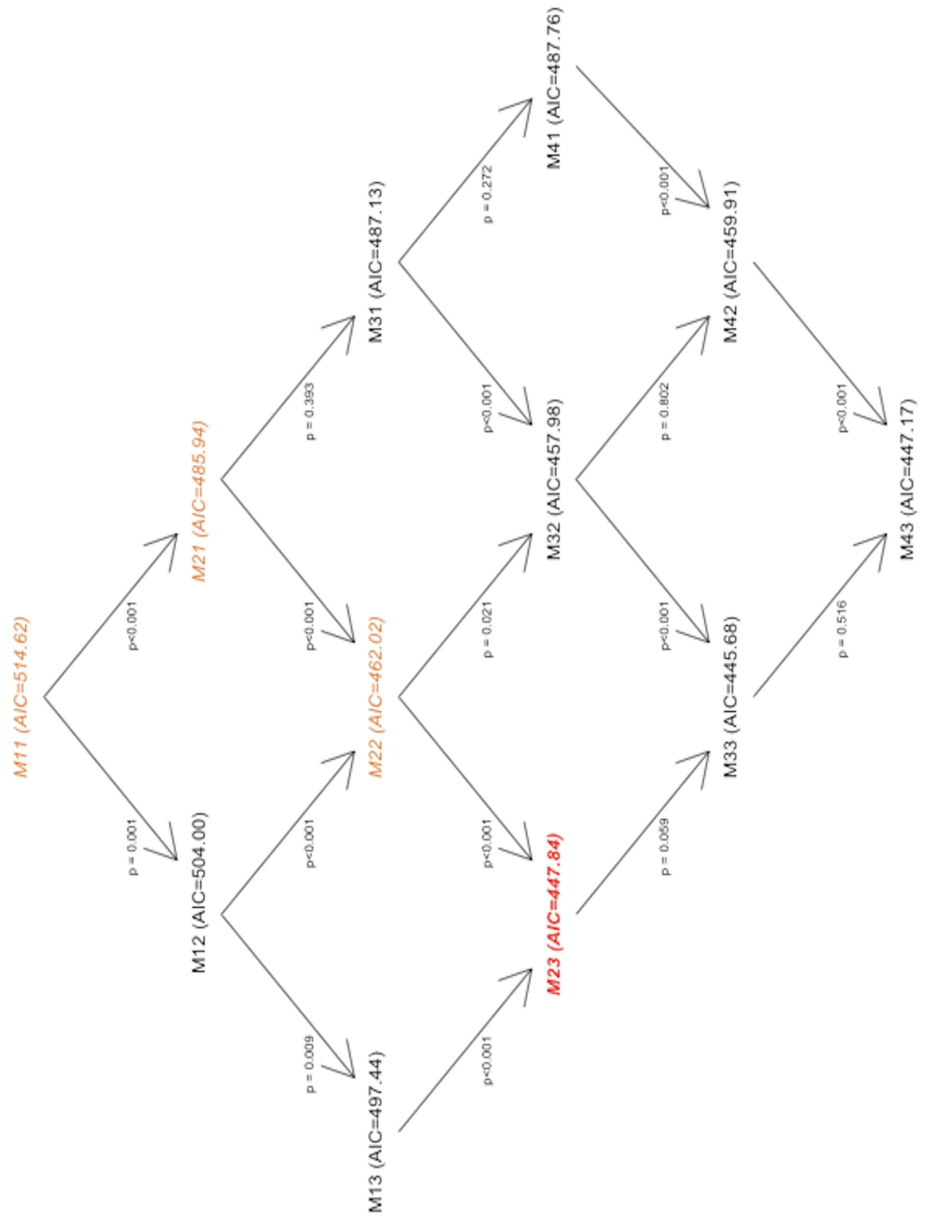


Figure 4.2: Model selection pathway (*ANOVA* & *AIC*)

Note that BIC has also been considered for all 12 candidate models with results similar to the ones of AIC. The use of BIC is a guard against overfitting since it is a consistent criterion that downweighs the penalty as compared to the sample size, and thus does not suffer from the phenomenon of overfitting that also affects AIC.

In addition, in order to fully verify the results given by AIC and BIC, we have calculated AICc which is used in situations of small sample sizes in order to guard against overfitting. The results, based on the real data, found to be similar to the ones given by the aforementioned criteria.

The modeling of extreme periods and estimation accuracy

The selected model M23 described in the previous section resulted a Mean Squared Error (*MSE*) of 507.77 which is quite high, indicating a relatively poor model fit. This is due to the fact that the baseline model is adequate for capturing non-epidemic periods but not epidemic ones. For improving our technique we introduce a polynomial approximation of the behavior of the time-series for each epidemic period (non-typical) identified in the previous subsection by changepoint analysis and evaluate the estimated rate value $\hat{y}^*_{\tau_{\kappa;l}}$ of the τ_k^{th} non-typical time-point by the polynomial of the l^{th} non-typical period. The chosen polynomials, with respect to the significance of their coefficients, provide a satisfactory approximation of the epidemic periods. Note that higher degree polynomials could be used but such a choice was found to increase the complexity of the procedure without a significant gain.

Let us denote by

- L the number of non-typical periods
- K_l, $l = 1, ..., L$, the number of time points within the l^{th} non-typical period
- $\tau_{\kappa;l}$ the κ - time point within the l^{th} non-typical period.
- $\hat{y}^b_{\tau_{\kappa;l}}$, $\kappa = 1, ..., K_l$, $l = 1, ..., L$, the estimated **baseline** rate values by baseline model of each non-typical time-point
- $\hat{y}^*_{\tau_{\kappa;l}}$, $\kappa = 1, ..., K_l$, $l = 1, ..., L$, the **excess** from the baseline model $\hat{y}^b_{\tau_{\kappa;l}}$, $\kappa = 1, ..., K_l$, $l = 1, ..., L$, to the actual response $y_{\tau_{\kappa;l}}$ and
- $\hat{y}^m_{\tau_{\kappa;l}}$, $\kappa = 1, ..., K_l$, $l = 1, ..., L$, the **modified** estimated rate values of each non-typical time-point.

The steps below describe the algorithmic procedure:

Step 1 The proper baseline model is identified, via *ANOVA* comparisons and *AIC*, and is fitted to *the whole* dataset.

Step 2 The excess y^* from the baseline model $\hat{y}^b$ to the actual response y is calculated for the time points of the non-typical period(s):

$$y_{\tau_{\kappa;l}} - \hat{y}^b_{\tau_{\kappa;l}} = y^*_{\tau_{\kappa;l}}, \quad \kappa = 1, ..., K_l, \quad l = 1, ..., L.$$

Step 3 (**Optional**) The non-typical period(s) could be split into subperiods if the researcher feels that more than one patterns are present (a graphical representation may help in deciding).

Step 4 A n^{th}-degree polynomial with respect to time is fitted to the excess values $y^*_{\tau_{\kappa;l}}$, $\kappa = 1, ..., K_l$, for each non-typical period separately, $l = 1, ..., L$. If necessary, for better fitting purposes, use Step 3 (optional).

$$\hat{y}^b_{t_{\kappa;l}} = \hat{\beta}_0 + \hat{\beta}_1 \tau_{\kappa;l} + \hat{\beta}_2 \tau^2_{\kappa;l} + ... + \hat{\beta}_n \tau^n_{\kappa;l},$$

$$\kappa = 1, ..., K_l, \qquad l = 1, ..., L.$$

Step 5 The final estimated rates corresponding to the time-points τ_κ, $\kappa = 1, ..., K_l$, $l = 1, ..., L$ are calculated according to the following expression:

$$\hat{y}^m_{\tau_{\kappa;l}} = \begin{cases} \hat{y}^b_{\tau_{\kappa;l}} + \hat{y}^*_{\tau_{\kappa;l}}, \text{ if } \tau_{\kappa;l} \text{ is a time point of the } l^{th} \\ \text{non-typical period} \\ \\ \hat{y}^b_{\tau_{\kappa;l}}, \text{ otherwise} \end{cases}, \tag{4.2}$$

where the upper and lower legs correspond to the estimated rate of typical and non–typical time–points, respectively.

Note that in this work we made use of Step 3 and thus the new epidemics (highlighted in red) are

$$\tau_1 - \tau_{13}, \; \tau_{14} - \tau_{20}, \; \tau_{21} - \tau_{25}, \; \tau_{26} - \tau_{66}, \; \tau_{67} - \tau_{74}, \; \tau_{75} - \tau_{105}.$$

and thus

$$\tau_{\kappa;l} = \begin{cases} \tau_{14;1}, ..., \tau_{20;1}, \text{ for the } l = 1^{st} \text{ non-typical period} \\ \\ \tau_{21;2}, ..., \tau_{25;2}, \text{ for the } l = 2^{nd} \text{ non-typical period} \\ \\ \tau_{67;3}, ..., \tau_{74;3}, \text{ for the } l = 3^{rd} \text{ non-typical period} \end{cases}.$$

The methodology proposed in this paper, is quite general and it can be applied to any type of incidence data as long as extreme/non-typical periods occur. Indeed, to any set of incidence data we propose the implementation of the changepoint analysis for identifying the periods of interest (both typical and non-typical). Then, for the standard/typical observations the baseline model is identified while for each period of non-typical observations an appropriate polynomial approximation is considered with respect to the significance of its coefficients. The combination of the proper baseline and polynomial approximation provides the complete model describing the entire set of incidence data. In case of a complex behavior in a certain non-typical period, it is recommended to cut-off the non-typical period optional Step 3 of the algorithm and consider a different polynomial for each piece of it in order to secure a more accurate fitting.

Remark 3 Traditionally, epidemic thresholds which signal an unexpected change are obtained by taking the upper 95th percentile of the baseline prediction distribution. A rule(47),(26) that is "a series of observations fall above the epidemic threshold during 2 weeks", is then used to define when epidemic alerts are produced, in order to avoid making alerts for isolated data points. If such a rule was applied in the present work the resulted epidemic period obtained almost coincides with the one obtained via the changepoint methodology. The threshold technique although simple lacks justification since it does not rely on a formal statistical procedure for the identification of the epidemic periods as opposed to the changepoint methodology. Changepoint analysis is based on hypothesis testing and Likelihood Ratio Tests in order to test if the mean and/or variance of a segment (which is differently defined in the algorithmic procedure between each changepoint method) differs from the mean and/or variance of another segment. Thus, no arbitrary thresholds, as the aforementioned, are needed in order to define the start and end of a non-typical period. Note that neither SegNeigh nor any other changepoint method in general, requires a threshold. This is due to the fact that there is almost no false detection in changepoint identification (see Remark 2). As a result, the proposed method achieves the accurate identification of points of change without entering a waiting period to confirm the identification.

Remark 4 (a) Following the aforementioned step-by-step methodology, we managed to reduce the *MSE* to 33.34 which is considerably smaller than the previous one. Alternative prediction performance measures could also be used providing similar results.
(b) Note that the residuals are not normally distributed which is expected since the sample size is too small (despite the fact of being sufficient for retrospective analysis). As for the correlation, the results given from the real data indicate that the residuals are uncorrelated for all usual significance levels with a Pvalue of 0.2811.

In conclusion, according to the proposed methodology, the estimation of the epidemic time-points for the ILI rate for Greece 2014-2016 ($\tau_1 - \tau_{13}$, $\tau_{26} - \tau_{66}$, $\tau_{75} - \tau_{105}$) is obtained by combining the polynomial approximation of the epidemic periods with the baseline model based exclusively on non-epidemic time-points ($\tau_{14} - \tau_{20}$, $\tau_{21} - \tau_{25}$, $\tau_{67} - \tau_{74}$).

4.2 The Overall Model - Final Results

Table 4.1 presents the dates and results of the retrospective evaluation of the excess influenza morbidity in Greece for 2014-2016 along with excess percentages, using the model M23. The excess morbidity is defined as the cumulative difference between observations and baseline over the entire epidemic period. Excess percentages were calculated as the observed size divided by the sum of expected values throughout each epidemic.

TABLE 4.1: Retrospective evaluation of the excess influenza morbidity, Greece 2014-2016

SW [a]	EW [a]	Cases	Expected cases	Excess cases	Excess percentage
201501	201512	1151	446	704	158%
201601	201608	601	264	337	128%

[a] SW and EW denote the signaled start and end weeks for epidemics, respectively.

Figure 4.3 illustrates the plots of:

- the observed ILI rate (black line),
- the estimated values following expression (4.2) (red line),

- the predicted baseline level based on M23 (green line), and
- the estimated threshold (upper bound of 95% prediction interval - red dotted line) for the whole time-series under study.

Due to the reduction of the *MSE* value by the methodology discussed in Subsection 4.1.1, the estimated rate values describe satisfactorily the underlying mechanism of the time-series under investigation. It is worth to be noted that through the proposed methodology we managed to capture the two peaks that occurred during the first epidemic period as well as the single peak occurred during the second epidemic period.

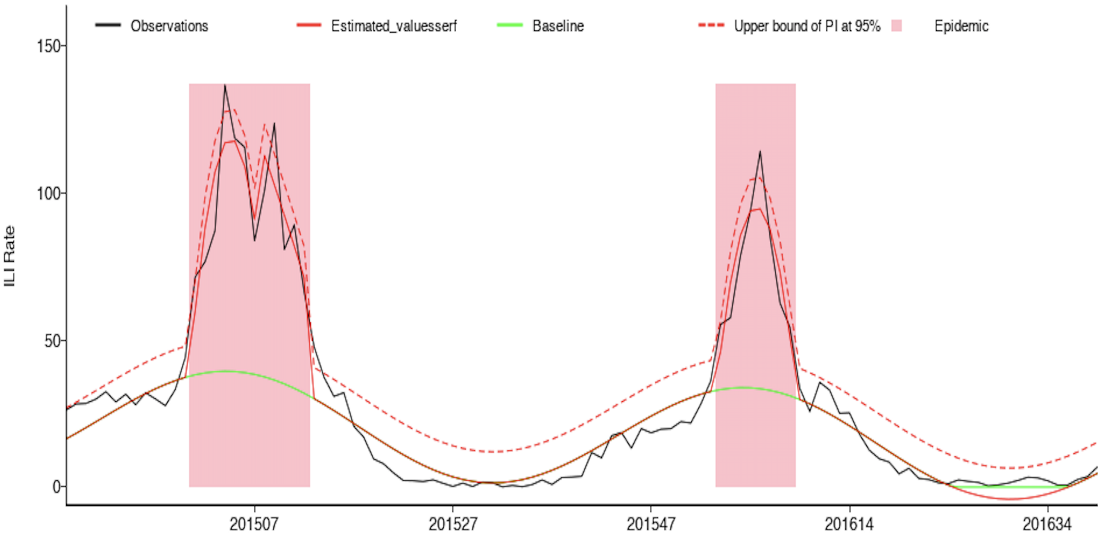


FIGURE 4.3: Estimated influenza morbidity and detected epidemics in Greece 2014-2016

Forecasting in the sense of predicting accurately the next time point(s) is a feature of prospective and not of retrospective analysis as done in this work. The novelty of the model though in terms of forecasting, lies in the fact that based on historical data (greater than one year)[47] the model developed can describe satisfactorily the behavior of the whole time series and thus provides to the public health officials and agents visual representation of what is expected for the upcoming year. Both public health agents and public in general could be prepared properly in order to treat/avoid the ILI syndrome.

4.3 Simulation Study

In this section a simulation study is conducted in order to verify the general applicability but also explore several capabilities of the proposed methodology.

Combining the following M23 model

$$\begin{aligned} z_t = 21.789133 &- 0.290946t + 0.001489t^2 \\ &- 9.567230cos\left(\frac{2\pi t}{n}\right) + 12.150802sin\left(\frac{2\pi t}{n}\right) \\ &+ 0.532738cos\left(\frac{4\pi t}{n}\right) - 1.411999sin\left(\frac{4\pi t}{n}\right) \\ &+ 1.380258cos\left(\frac{8\pi t}{n}\right) + 2.256325sin\left(\frac{8\pi t}{n}\right) \\ &+ 0.176566\text{mintemp} + 0.94z_{t-1} + u_t + 0.09u_{t-1}, \end{aligned}$$

where $u_t \sim \mathscr{N}(0, 0.896)$, $t = 1, ..., 105$, $n = 52.179$ (# of 2 calendar years) and the proposed methodology described in the previous section, an exhaustive iteration process (1000 iterations) performed and the mean value of three information criteria, namely *AIC*, *BIC* and *AICc*, for each of the 12 candidate models is presented in Table 4.2.

TABLE 4.2: Information criteria mean for each of the 12 candidate models

MODEL ID	AIC	BIC	AICc
M11	426.774	443.7892	428.2479
M12	366.0996	387.9763	368.5325
M13	365.3339	392.0721	369.0013
M21	401.3934	420.8393	403.3137
M22	315.5189	365.8263	325.5331
M23	294.475	323.644	298.8703
M31	402.1699	424.0466	404.6028
M32	305.8306	334.5688	307.498
M33	295.9189	327.5186	301.1199
M41	392.1575	416.465	395.1718
M42	299.8777	331.0466	304.2729
M43	296.9612	330.9916	303.0494

The results show that the top 3 model choices for all 3 criteria are the same, i.e. M23, M33, M43, with M23 and M33 being very close to each other and M43 coming relatively close in the 3^{rd} place with the corrected one selected in all cases. Similar results are obtained irrespectively of the model considered for simulations.

4.4 Concluding Remarks

Conclusively, in this study, we established that the changepoint detection analysis in conjunction with Periodic-type ARMA modeling with covariates is capable of modeling time-series data with typical and non-typical parts and identifying effectively the beginning and end of the extreme periods that occurred. It is reminded that the changepoint analysis was implemented as typical yet powerful technique for the purpose of identifying the extreme periods. In addition, it is worth to be noted that the proposed approach succeeds in:

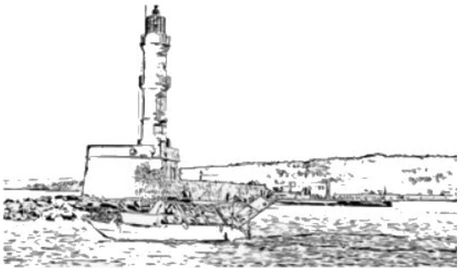
1. Modeling both the typical and non-typical activity of incidence data, and

2. Improving the estimation performance by increasing the accuracy of the selected model.

Moreover, there are crucial advantages of the proposed methodology over the existing ones. Firstly, the changepoint method identified directly the extreme periods, and thus there was no need for an epidemic threshold that was mostly chosen ad hoc and alert rule to be considered. Secondly, the pruning of the time-series data is no longer arbitrary, since the time-points identified as extreme through changepoint analysis, have been automatically pruned. Finally, the proposed approach captures the behavior of the whole time-series with no significant loss of accuracy, and hence the derived results could be used for forecasting purposes.

The proposed methodology was implemented for analyzing ILI rate data for Greece and also explored through a series of simulated data. Based on these results, we could safely recommend the use of the proposed methodology to any incidence data. Indeed, the almost no-false changepoint detection provides a powerful and accurate mechanism of identifying the beginning and end of extreme periods which in turn makes the proposed technique useful in any situation with such type of structure. Furthermore, the use of changepoint detection analysis along with time-series modeling techniques presented in this work, seem to provide a useful tool for modeling outbreaks that may occur in incidence data and at the same time beneficial to the society due to the consequences associated with the early detection and prevention of extreme, possibly harmful, events.

Chapter 5



Old Port, Chania, Greece

Comparative Studies

Chapter's Goal

In this Chapter, two comparative studies take place. The first one, aims to evaluate the predictive ability of the model proposed in Chapter 3 in comparison with alternative models related to the modeling of influenza morbidity. In the second, motivated by the work done in Chapter 4, we applied and evaluated pioneering methods based on changepoint analysis to detect changes in influenza type data. The empirical comparative study provided evidence that statistical methods based on change point analysis have several attractive properties compared to current practice for outbreak detection.

5.1 Part I: A Comparative Study for the Use of Periodic Regression Models for Detection of Influenza Outbreaks

Cite: **Kalligeris EN**, Karagrigoriou A, Lambrou A and Parpoula C. A Comparative Study for the Use of Periodic Regression Models for Epidemiological Surveillance. *Proceedings* 31st *Panhellenic Statistical Conference* 2018; 292-301.

In 2017, Parpoula et al.(26) modeled ILI rate data for Greece (2014-2016) by considering a polynomial of 3rd degree for the modelling of trend and *sine* and *cosine* terms (with period one year, six months and three months) for the modelling of seasonality.

Their analysis was based on the following model:

$$\begin{aligned} M33: \quad y_t = {} & \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \alpha_3 t^3 + \gamma_1 \cos\left(\frac{2\pi t}{n}\right) + \delta_1 \sin\left(\frac{2\pi t}{n}\right) \\ & + \gamma_2 \cos\left(\frac{4\pi t}{n}\right) + \delta_2 \sin\left(\frac{4\pi t}{n}\right) + \gamma_3 \cos\left(\frac{8\pi t}{n}\right) + \delta_3 \sin\left(\frac{8\pi t}{n}\right) + \epsilon_t. \end{aligned}$$

Through an exhaustive process, using ANOVA comparisons along with AIC, BIC information criteria, they concluded that the best fitting model is **M23**, that is:

$$\begin{aligned} M23: \quad y_t = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \gamma_1 \cos\left(\frac{2\pi t}{n}\right) + \delta_1 \sin\left(\frac{2\pi t}{n}\right) \\ + \gamma_2 \cos\left(\frac{4\pi t}{n}\right) + \delta_2 \sin\left(\frac{4\pi t}{n}\right) + \gamma_3 \cos\left(\frac{8\pi t}{n}\right) + \delta_3 \sin\left(\frac{8\pi t}{n}\right) + \epsilon_t. \end{aligned} \tag{5.1}$$

Additionally, using evaluation criteria such as RMSE, AIC, and BIC, the forecasting capability of the selected model (**M23**) was examined in contrast to typical trend or/and seasonality detection models, i.e.:

- a linear trend (LT),
- a moving average of 3 terms(MA3),
- an exponential smoothing (SES),
- a Holt's model,
- a Brown's model,
- a Winter's model, and
- the standard CDC algorithm for flu detection (Serfling's model)

According to the aforementioned comparisons, **M23** was found to be more accurate (in terms of RMSE values) than the rest of the models under comparison. Also, based on AIC and BIC criteria, **M23** had the smaller and second smaller value, respectively, and thus, it suited best to the dataset examined. It has to be noted that in the comparisons, the MA3 model was proven a strong competitor since it scored a smaller BIC value and satisfactory AIC and RMSE values).

Since the above results constituted the motivating force of the work presented in Chapter 3, we conducted a comparative study among the model in (3.2) (labeled here as MXM11), the M23 model in (5.1) and the strong competitor of the latter, MA3, which is expressed as:

$$MA3: \quad y_t = \epsilon_t + \sum_{j=1}^{3} \epsilon_{t-j}, \quad t = 1, ..., 105.$$

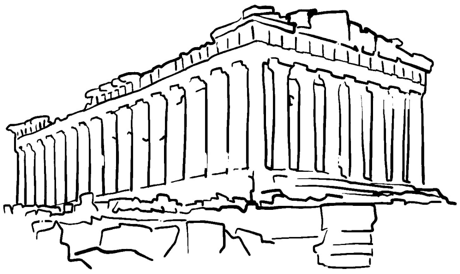
The goal of this study is to evaluate the forecasting performance of the model MXM11 proposed in Chapter 3 with all models considered by Parpoula et al.(26) Looking at the results of Table 5.1, we conclude that MXM11 clearly outperforms M23 and MA3 since it succeeds a significant improvement in the values of all three criteria considered i.e., RMSE, AIC and BIC.

Table 5.1: Comparative Performance of Forecasting Models

Model	RMSE	AIC	BIC
MXM11	4.30	409.31	429.29
M23	12.02	547.89	572.77
LT	48.34	818.44	823.75
MA3	13.51	554.72	565.33
SES	16.06	585.05	587.00
Holt's	16.94	598.15	606.70
Brown's	18.87	618.87	620.65
Winter's	25.55	686.57	694.37
Serfling's	18.81	610.47	622.91

5.1.1 Concluding Remarks

In this Chapter, a comparative study was conducted with the goal of evaluating the forecasting performance of the model proposed in Chapter 3 with alternative models associated with the modelling of influenza morbidity. Specifically, previously proposed models describing the behavior of influenza morbidity were considered, and the resulted most competitive models, namely, MXM11, M23 and MA3, were put under comparison in terms of RMSE, AIC and BIC. The results derived through this study, indicated that among the aforementioned models the one with the best forecasting performance is the one obtained via the proposed methodology in Chapter 3, namely MX11.



Parthenon, Athens, Greece

5.2 Part II: A Comparative Study of Changepoint Analysis Techniques for Outbreak Detection

Cite: Parpoula C, **Kalligeris EN** and Karagrigoriou A. A Comparative Study of Change-Point Analysis Techniques for Outbreak Detection. *Proceedings* 21st *European Young Statisticians Meeting, Milosevic B & Obradovic, M eds.*, Faculty of Mathematics, University of Belgrade Publications 2019; 85-89.

The basic tool of the methodology presented in Chapter 4, was the one of changepoint analysis. Since there exist several changepoint analysis techniques, in this Chapter a comparative study is conducted among three techniques, namely

- recursive segmentation and permutation (RS/P),
- MXM11 (see Chapter 3), and
- SegNeigh (see Subsection 2.2.2).

The comparison is performed in the basis of Receiver Operating Characteristic (ROC) curve analysis, and its related statistics/metrics (Accuracy-ACC, Sensitivity-SENS, Specificity-SPEC, Area Under the ROC curve-AUC) metrics.

5.2.1 Phase I Distribution-Free Changepoint Analysis

Several approaches for detecting outbreaks of infectious diseases in the literature are directly inspired by, or related to, methods of Statistical Process Control (SPC). In an epidemiological surveillance problem, the underlying process distribution is not normal and usually unknown. Hence statistical properties of commonly used SPC charts could be highly affected. In this paper, we implement important aspects of univariate distribution-free Phase I change-point analysis and apply some of the recent developments in this area, in order to develop a novel SPC charting method that works best for monitoring and outbreak detection processes.

Let y_i represent the ith observation, $i = 1, \ldots, m$, collected from the distribution of a quality characteristic, either continuous or discrete, Y. When the process is in-control (IC), these observations are assumed independent with an unknown but common cumulative distribution function (c.d.f.) $F_0(y)$, whereas the out-of-control (OC) state can be described by a multiple change-point model, that is

$$F_r(*) = \begin{cases} \mathrm{F}_0(y) & \text{if } 0 < i \leq \tau_1 \\ \mathrm{F}_1(y) & \text{if } \tau_1 < i \leq \tau_2 \\ & \vdots \\ \mathrm{F}_k(y) & \text{if } \tau_k < i \leq m \end{cases} \quad ,$$

where $0 < \tau_1 < \tau_2 < \ldots < \tau_k < m$ denote k change points and $F_r(*)$, $r = 0, \ldots, k$, are unknown c.d.f. which, at one or several times, may shift in position. Note here that the shift times τ_i are also assumed to be unknown. This Phase I analysis procedure provides a statistical test for verifying the hypothesis system

$$H_0 : \text{The process was IC } (k = 0) \quad vs. \quad H_1 : \text{The process was OC } (k > 0)$$

and identifying the time of the changes when the hypothesis of an IC process is rejected. This hypothesis testing system (performed in Phase I) requires the specification of a nominal false alarm probability (FAP). Following the recursive segmentation and permutation (RS/P) approach of Capizzi and Masarotto,(163) choosing an acceptable/reasonable FAP value, say α, we test the stability over time of the level parameter. The following steps need to be executed for level-changes detection, i.e., detection of single or multiple level shifts.

Let us consider the problem of testing the null hypothesis that the process was IC against the alternative hypothesis that the process mean experienced an unknown number of step shifts. In such a case, a set of test (control) statistics is needed for detecting $1, 2, \ldots, K$ step shifts with K denoting the maximum number of hypothetical change points. The mean values $\mu_0, \ldots, \mu_k$, and the change points are assumed unknown. Further, defining $\tau_0 = 0$ and $\tau_{k+1} = m$, it is also assumed that $\tau_r - \tau_{r-1} \geq l_{MIN}$, $r = 1, \ldots, k+1$, where l_{MIN} is a (user pre-specified) constant giving the minimum number of successive observations allowed between two change points. For a sequence of individual observations, the control statistic and the possible change points are computed using a simple forward recursive segmentation approach. The algorithm starts with $k = 0$ and then proceeds in K successive stages. At the beginning of stage k, the interval $[1, m]$ is partitioned into k subintervals, each having a length greater or equal to l_{MIN}. At stage k, one of these subintervals is split, adding a new potential change point. The new change point is selected maximizing

$$\sum_{i=1}^{k+1} (\hat{\tau}_i - \hat{\tau}_{i-1}) (\bar{y}\,(\hat{\tau}_{i-1}, \hat{\tau}_i) - \bar{y}_{om})^2 \tag{5.2}$$

conditionally on the results of the previous stages. Here $\bar{y}_{om}$ represents the overall mean (om) of observations, $\bar{y}(\alpha, b) = \frac{1}{b-\alpha} \sum\limits_{i=\alpha+1}^{b} y_i$, and $0 = \hat{\tau}_0 < \hat{\tau}_1 < \cdots < \hat{\tau}_k < \hat{\tau}_{k+1} = m$ are the boundaries of the new partition. The control statistic T_k, $k = 1, \dots, K$ is equal to the attained maximum value of Equation (5.2). Therefore, given a test statistic, its p-value can be calculated, as the proportion of permutations under which the statistic value exceeds or is equal to the statistic computed from the original sample of observations. Choosing an acceptable FAP, say α, then, for p-value$< \alpha$, the null hypothesis that the process was IC is rejected.

5.2.2 Comparative Study

This paper focuses on the study of weekly ILI rate data (provided from the Hellenic Center for Disease Control and Prevention) for Greece, between September 29, 2014 (week40/2014) and October 2, 2016 (week39/2016), which were used for analysis purposes. Here, we perform the RS/P approach for both periods under study (1^{st} period: week40/2014-week39/2015, 2^{nd} period: week40/2015-week39/2016) executing $L = 100000$ permutations with $K = \max\left(3, \min\left(50, \left[\frac{m}{15}\right]\right)\right)$ and $l_{MIN} = 5$. Our procedure signals possible changes of the mean (p-value < 0.001 for a change in level). The extracted signaled start (sw) and end weeks (ew) of the epidemics were sw01-ew14/2015 and sw01-ew08/2016.

In our study, the ability of RS/P method to detect the true (and correct amount of) change-points is tested through benchmarking. Therefore, RS/P derived change-points are compared with those derived after executing:

1. the standard CDC and ECDC flu detection algorithm (Serfling's model)[44]

$$\textbf{M11:} \quad y_t = \alpha_0 + \alpha_1 t + \gamma_1 \cos\left(\frac{2\pi t}{m}\right) + \delta_1 \sin\left(\frac{2\pi t}{m}\right) + \epsilon_t,$$

where y_t are the observed time series values (weekly ILI rate), ϵ_t are centered zero-mean random variables with variance σ^2, m denotes the number of observations within one year, and model coefficients are estimated by least squares method,

2. an extended Serfling's model presented by Parpoula et al.[26]

$$\begin{aligned}\textbf{M23:} \quad y_t = {} & \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \gamma_1 \cos\left(\frac{2\pi t}{m}\right) + \delta_1 \sin\left(\frac{2\pi t}{m}\right) \\ & + \gamma_2 \cos\left(\frac{4\pi t}{m}\right) + \delta_2 \sin\left(\frac{4\pi t}{m}\right) + \gamma_3 \cos\left(\frac{8\pi t}{m}\right) + \delta_3 \sin\left(\frac{8\pi t}{m}\right) + \epsilon_t,\end{aligned}$$

3. a mixed model with a linear trend, 12-month seasonal periodicity, Auto-Regressive Moving Average (ARMA) terms, that is ARMA(2,1), and the minimum temperature (mintemp) as a random meteorological

covariate presented in Chapter 3:

$$\textbf{MXM11:} \quad y_t = \alpha_0 + \alpha_1 t + \gamma_1 \cos\left(\frac{2\pi t}{m}\right) + \delta_1 \sin\left(\frac{2\pi t}{m}\right) + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \epsilon_t + \lambda_1 \epsilon_{t-1} + \omega_1 \text{mintemp},$$

and

4. Segment Neighbourhood (SegNeigh) algorithm which uses an optimization step that searches over all previous change-point locations and picks the one that gives the optimal segmentation up to time t, presented in Subsection 2.2.2.

As aforementioned, the current approach to influenza surveillance is based on Serfling's cyclic regression model (**M11**). Parpoula et al.,(26) developed extended Serfling-type periodic regressions models, and through an exhaustive search process (using ANOVA comparisons and AIC, BIC information criteria) the best fitting model **M23** was selected. The aforementioned procedure allowed Parpoula et al.(26) to extract the signaled start and end weeks of the epidemics, i.e., sw01-ew13/2015, sw01-ew08/2016. It is worth to be noted that the signaled start and end weeks were found to be identical considering either Serfling's model (**M11**) or extended Serfling's model (**M23**). Then, the above results motivated us (see Chapter 3) to incorporate ARMA terms and random meteorological covariates in the model structure, for identifying the epidemics (sw01-ew12/2015, sw05-ew08/2016). In addition, in Chapter 4 we established that the changepoint detection analysis in conjunction with mixed effects periodic ARMA time series modeling, is capable of modelling time-series data with typical and non-typical parts and identifying the beginning and end of the extreme periods that occurred (sw01-ew12/2015, sw01-ew08/2016).

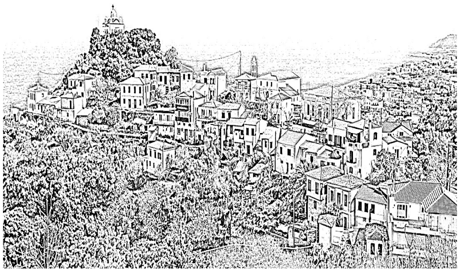
Therefore, we then examine the ability of the RS/P, **MXM11** and SegNeigh approaches to detect the true change-points compared to the standard and extended CDC and ECDC flu detection algorithm (models **M11** & **M23**). The diagnostic performance of a test to discriminate between two groups (here, epidemic from non-epidemic) is typically evaluated using Receiver Operating Characteristic (ROC) curve analysis, and its related statistics/metrics (Accuracy-ACC, Sensitivity-SENS, Specificity-SPEC, Area Under the ROC curve-AUC). Hence we estimated these metrics along with their 95% Confidence Interval (CI) (exact Clopper-Pearson CIs for ACC, SENS and SPEC, exact binomial CI for each derived AUC) for each method (as shown in Table 1). Table 1 indicates that RS/P and SegNeigh approaches (higher ACC, SENS and AUC values) outperform **MXM11**, and seem to detect successfully the true change-points compared to the standard approach to influenza surveillance.

TABLE 5.2: Metrics for RS/P, MXM11 and SegNeigh approaches

Metric	RS/P	MXM11	SegNeigh
ACC (95% CI)	99.05% (94.81% to 99.98%)	95.24% (89.24% to 98.44%)	99.05% (94.81% to 99.98%)
SENS (95% CI)	100.0% (83.89% to 100.0%)	76.19% (52.83% to 91.78%)	95.24% (76.18% to 99.88%)
SPEC (95% CI)	98.81% (93.54% to 99.97%)	100.0% (95.71% to 100.0%)	100.0% (95.71% to 100.0%)
AUC (95% CI)	0.988 (0.944 to 0.999)	0.881 (0.803 to 0.936)	0.976 (0.926 to 0.996)

5.2.3 Concluding Remarks

In this Chapter, we implemented and evaluated cutting-edge changepoint analysis-based methods for detecting changes in location of univariate ILI rate data. The empirical comparative study provides evidence that statistical methods based on change-point analysis have several appealing properties compared to the current practice for the detection of epidemics. In particular, RS/P and SegNeigh approaches, both succeeded in early and accurate outbreak detection. Both RS/P and SegNeigh approaches are advantageous since they can be applied to historical data without the need for distinguishing between epidemic and non-epidemic periods in the data, and single or multiple mean shifts can be detected. Further, RS/P Phase I distribution-free changepoint analysis method is able to guarantee a predefined false alarm probability without any knowledge about the (in-control) underlying distribution, whereas SegNeigh algorithm in conjunction with mixed effects periodic ARMA time series modeling is capable of modeling time series data with typical and non-typical parts.



View of Agia Triada in Karlovasi, Samos, Greece

Chapter 6

On Stochastic Dynamic Modeling of Incidence Data

Cite: **Kalligeris EN**, Karagrigoriou A and Parpoula C. On Stochastic Dynamic Modeling of Incidence Data (Under Revision, 2022).

Continuing our attempt(s) to develop an effective methodology for modelling the complete behavior of time-series incidence data, in this Chapter an extension of the model presented in (2.22), is proposed. Its components are selected by penalized likelihood techniques with the goal of achieving a high level of robustness regarding the modeling of dynamic behaviors of epidemiological data. In addition to statistical inference, Changepoint Detection Analysis is performed for the selection of the number of regimes, which reduces the complexity associated with Likelihood Ratio Tests. Within this framework, a three-phase procedure for modeling incidence data is proposed and tested via real and simulated data.

Let as recall the model of (2.22):

$$y_t = c_{s_t} + \sum_{i=1}^{p} \phi_{is_t} y_{t-i} + \epsilon_{s_t},\ \ i \leq t \leq T.$$

We now define the Markov Switching Model of Conditional Mean (MSMCM) with covariates which results by incorporating covariates $\mathbf{\Omega}_j$, $j = 1, ..., q$, into the above model's structure, given that $S_t = s_t$:

$$y_t = c_{s_t} + \sum_{i=1}^{p} \phi_{is_t} y_{t-i} + \sum_{j=1}^{q} \gamma_{js_t} \mathbf{\Omega}_{jt} + \epsilon_{s_t},\ \ i \leq t \leq T \tag{6.1}$$

where γ_{js_t} the coefficient associated with the $\mathbf{\Omega}_j$ covariate and $c_{s_t}, \phi_{is_t}, \epsilon_{s_t}$ as in (2.22).

The unknown parameters of (6.1), denoted by $\boldsymbol{\theta^*}$, are

$$\boldsymbol{\theta^*} = (c_{s_t}, \phi_{1s_t}, ..., \phi_{ps_t}, \gamma_{1s_t}, ..., \gamma_{qs_t}, \sigma^2_{s_t}, p_{11}, ..., p_{KK}),$$

and are estimated via the EM algorithm discussed in Subsection 2.3.2.

6.1 Nonnegative Garrote Variable Selection & Bias

Regularization has been intensely studied on the interface between statistics and computer science. From various regularization methods that exist, we will focus on those usually referred to as penalized likelihood techniques. The latter are based on the idea of nonnegative garrote(164) and aim on the reduction of model's complexity by considering a penalty along with the quadratic loss function which is defined as:

$$L(\hat{y}, y) = \sum_{t=1}^{T} (y_t - \hat{y}_t)^2.$$

Below, some of the most widely used penalized likelihood techniques are being presented:

1. L^1, also known as **Lasso** (Least absolute shrinkage and selection operator(165)) is defined by:

$$L(\hat{y}, y) + \lambda \left[\sum_{i=1}^{p} \sum_{j=1}^{q} (|\phi_{is_t}| + |\gamma_{js_t}|) \right]$$

2. L^2, also known as **Ridge**,(166) is defined by:

$$L(\hat{y}, y) + \lambda \left[\sum_{i=1}^{p} \sum_{j=1}^{q} (\phi_{is_t}^2 + \gamma_{js_t}^2) \right]$$

3. **Elastic-Net**(167) is defined by:

$$\frac{1}{2\tau} L(\hat{y}, y) + \lambda \left[\alpha \left(\sum_{i=1}^{p} \sum_{j=1}^{q} (|\phi_{is_t}| + |\gamma_{js_t}|) \right) + \left(\frac{1-\alpha}{2} \right) \left(\sum_{i=1}^{p} \sum_{j=1}^{q} (\phi_{is_t}^2 + \gamma_{js_t}^2) \right) \right], \tag{6.2}$$

where τ the sample size and $\alpha \in [0, 1]$ & λ the tuning parameters that result the penalty in the loss function. The tuning parameter α, balances the amount of emphasis given to minimize the loss function versus minimizing the sum of squared coefficients and/or the sum of absolute coefficients. Notice that if in (6.2) we set $\alpha = 0$, then Elastic-Net reduces to Ridge, while if $\alpha = 1$, then Elastic-Net reduces to Lasso. Generally, Elastic-Net is considered preferable over Lasso and Ridge as it neutralizes the limitations of the two techniques, while it includes them as special cases.(168)

There exists a quite rich literature on nonnegative garrote based techniques and their application on time-series models. Nardi,(169) studied the Lasso estimator for fitting autoregressive time-series models via adopting a double asymptotic framework where the maximal lag may increase with the sample size.

Chen,(170) proposed an adaptive Lasso regression of the time series on its lags and the lags of the residuals for identifying the optimal subset autoregressive moving average model. Furthermore, Medeiros and Eduardo,(171) studied the asymptotic properties of the adaptive Lasso in sparse, high-dimensional and linear time-series models. Such techniques though, suffer from two serious drawbacks: (1) lack of estimation accuracy and; (2) biasedness. In order to handle the aforementioned issues, k-fold Cross-Validation ($CV_{(k)}$) will be used which, as pointed out by Bergmeir et al.,(172) can be applied in cases of autoregressive models provided the models considered have uncorrelated errors. Mosteller and Tukey,(173) introduced the $CV_{(k)}$, which constitutes one of the most popular CV techniques and divides the data into k-groups of equal size (classical choice $k = 10$, (174)). Note that higher values of k, lead to less biased model. Subsequently, the first group is used for testing purposes, while the remaining $k - 1$ are used for training the model. The procedure is repeated k times with the testing group changing each time. In each repetition, the Mean Squared Error (MSE) is calculated in order to estimate an overall standard error of $CV_{(k)}$ based on the following formula:

$$CV_{(k)} = \frac{1}{k} \sum_{i=1}^{k} MSE_i.$$

6.2 The Proposed Methodology

In this Section, the MSMCM with covariates and the methods discussed in Section 6.1 will be jointly applied in a series of simulated data for identifying:

1. the required order of the autoregressive process;

2. the covariates that have a significant contribution to the problem under investigation;

3. the regimes occurred.

A key problem that arises in empirical studies is the determination of the number of states required for an MSM so that an adequate characterization of data could be achieved. Hamilton(138) offered suggestive evidence that a MSM of two-states outperforms linear models in terms of forecasts, but no statistical test can be applied. Likelihood Ratio Tests (LRTs) consist a more formal statistical procedure for determining the number of regimes but are considered too complicated, since the usual necessary regularity conditions required to apply the asymptotic theory are no longer met.(175) Another common problem, is the plethora of available variables. The Markov Switching mechanism, by its nature, tends to discard only a few of the variables considered due to the fact that most of them are significant in at least one regime. This may lead to a complex, and thus difficult to interpret model, and as a result the following question arises “Is such complexity worth it in terms of information gained?”

In order to handle the aforementioned issues, we propose the following three -

phase procedure:

Phase 1 Apply a penalized likelihood technique, e.g., Lasso, Ridge, Elastic-Net, etc., to the list of candidate explanatory variables (including the appropriate order of the AR process that are required for applying later the MSM method) with the goal of extracting the significant ones. The tuning parameter λ, is the one resulted through $CV_{(k)}$.

Phase 2 Perform Changepoint Detection Analysis(50; 176) to the response variable and identify the number of changepoints occurred. Based on the associated graphical representation, define the number of regimes required.

Phase 3 Apply the MS mechanism, considering as explanatory variables and number of regimes the ones resulted through Phases 1 & 2, respectively. The regimes are created based on K different models of which the coefficients are being estimated through the EM algorithm as presented in Subsection 2.3.2.

The advantage of the proposed methodology lies in the fact that it makes use of a more flexible model while retaining only the most significant variables for building the switching model, resulting, as it will become evident in the simulation section, in increased accuracy of the latter, in terms of Mean Squared Error (MSE) and Akaike's Information Criterion (AIC). In addition, by implementing Changepoint Detection Analysis into the procedure, the complexity of LRTs is avoided resulting into a robust selection of regimes.

6.3 Performance Evaluation of the Proposed Methodology

In this Section a real case as well as a simulation study are conducted for testing the proposed methodology. Both studies, are based on a dataset concerning 105 weekly ILI incidence data for Greece, between September 29, 2014 and October 2, 2016. The state space considered across the studies is $\mathscr{D} = \{0, 1\}$, where $s_t = 0$ is the non-epidemic state and $s_t = 1$ is the epidemic one, $t = 1, ..., 105$. Finally, note that the R-package used for fitting the MSMCM with covariates to the data is *MSwM.*(177)

6.3.1 Real case study

The vector $\underset{\sim}{X}$ contains the set of variables in which the proposed methodology will be tested on:

$$\underset{\sim}{X}^T = \left(t, t^2, sin\left(\frac{i\pi t}{\tau}\right), cos\left(\frac{i\pi t}{\tau}\right), \text{Tmin}, \text{Tmax}, \text{Tmean}, \text{Tmed}, \text{WDmed}, \right.$$
$$\left. \text{WFmax}, \text{WFmean}, \text{WFmed}\right), \quad i=\{2,4,8\}.$$

Observe that the exploratory variables include trend (linear and quadratic) and periodicity (of 12, 6 and 3 months for $i = \{2, 4, 8\}$) as well as 8 meteorological variables due to their possible effect on ILI(148) namely, the minimum, maximum, mean and median of the daily temperature (T), maximum, mean and median wind force (WF) and median wind direction (WD). The data were provided by the Hellenic National Meteorological Service (HNMS).

The process starts by applying the penalized likelihood technique of preference on $\underset{\sim}{X}$ in order to extract the variables which have a significant impact to the dependent variable (ILI incidence rate). In our case, all 3 techniques, namely, Lasso, Ridge and Elastic-Net were used but the preferred one is that of Elastic-Net since it resulted the lowest overall MSE. Furthermore, for the selection of the tuning parameter λ, a $CV_{(10)}$ is performed for achieving estimation accuracy and unbiassedness. According to Figure A.1, which represents the behavior of MSE for different values of $log\lambda$, the value of λ which results the lowest MSE is 0.0734.

Since Elastic-Net constitutes a trade-off between Ridge ($\alpha = 0$) and Lasso ($\alpha = 1$), the value of α ranges from 0 to 1. Based on Figure 6.1, which illustrates the behavior of various values of α as opposed to both different values of $log\lambda$ and their corresponding MSE, the selected value of α is that of 0.729.

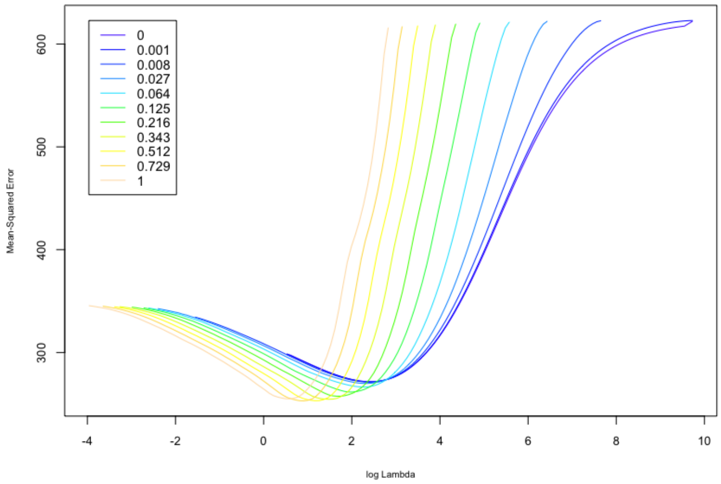


FIGURE 6.1: Behavior of various values of α as opposed to both different values of $log\lambda$ and their corresponding MSE.

Finally, after performing Changepoint Detection Analysis to the dataset for the identification of the number of regimes, we apply the Markov Switching mechanism to the variables selected as significant from the Elastic-net method. Figures 6.2 and 6.3 illustrate the regimes occurred, which are described by the models presented in (6.3) and (6.4), respectively. In particular, we are interested in identifying two regimes corresponding to the outbreak ($week52/2014 - week14/2015$

and $week01/2016 - week08/2016$) and non-outbreak periods, respectively. Observe that both models include, besides trend and periodicity, 4 meteorological variables, namely, minimum (Tmin), mean (Tmean), median (Tmed) temperature and mean (WFmean) wind force.

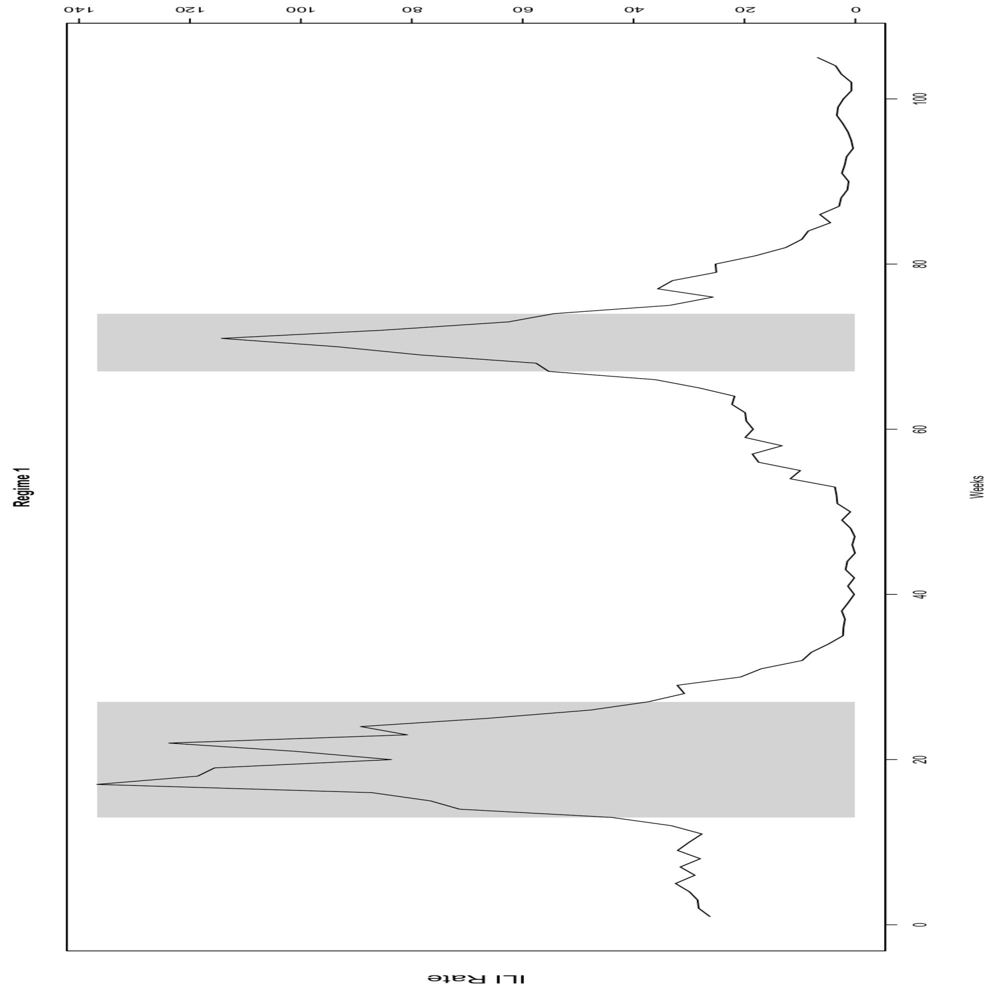


FIGURE 6.2: The first regime (outbreak) resulted through the proposed methodology.

1^{st} Regime

$$\hat{y}_t = 146.47 - 0.37t - 37.48sin\left(\frac{2\pi t}{\tau}\right) - 2.68cos\left(\frac{2\pi t}{\tau}\right)$$

$$- 44sin\left(\frac{4\pi t}{\tau}\right) - 23.94cos\left(\frac{4\pi t}{\tau}\right) + 8.32sin\left(\frac{8\pi t}{\tau}\right) - 14.24cos\left(\frac{8\pi t}{\tau}\right)$$

$$+ 0.03\text{Tmin} + 15.32\text{Tmean} - 11.08\text{Tmed} + 10.04\text{WFmean} - 0.59\hat{y}_{t-1}. \quad (6.3)$$

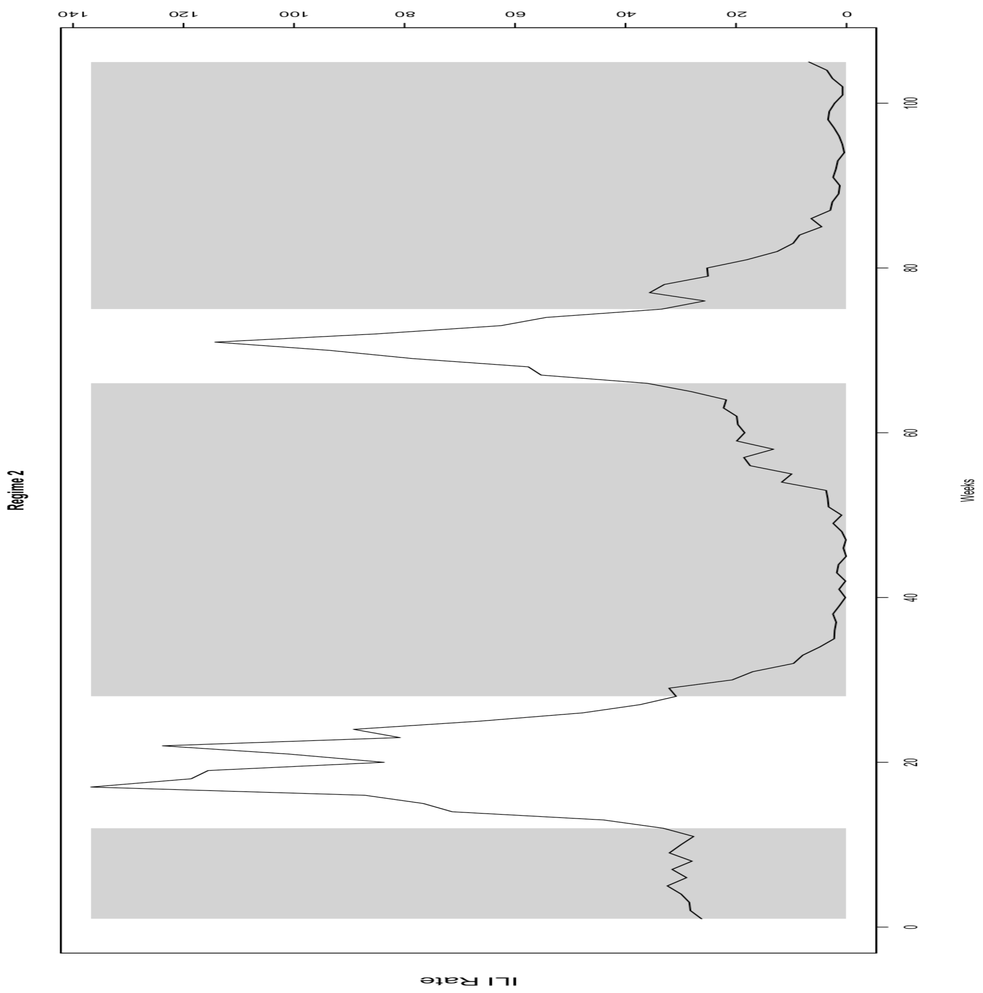


FIGURE 6.3: The second regime (non-outbreak) resulted through the proposed methodology.

2^{nd} Regime

$$\hat{y}_t = 19.55 - 0.005t + 18.93sin\left(\frac{2\pi t}{\tau}\right) - 10.65cos\left(\frac{2\pi t}{\tau}\right)$$

$$- 5.11sin\left(\frac{4\pi t}{\tau}\right) - 2.24cos\left(\frac{4\pi t}{\tau}\right) + 0.97sin\left(\frac{8\pi t}{\tau}\right) - 0.31cos\left(\frac{8\pi t}{\tau}\right)$$

$$- 0.23\text{Tmin} + 4\text{Tmean} - 4.13\text{Tmed} + 1.65\text{WFmean} + 0.69\hat{y}_{t-1}. \qquad (6.4)$$

The models presented in (6.3) and (6.4) succeed in fully recognizing the regimes of the dataset under investigation. Specifically, since the data concern weekly ILI rate observations, the first regime refers to an epidemic (outbreak) while the second one to a non-epidemic (non-outbreak).

The transition probability matrix given in Table 6.1, indicates that the probability of changing a regime from t to $t+1$ is very low. As a result, if y_t falls into a regime (outbreak or non-outbreak) then y_{t+1} has a high probability of falling into the same regime. This “consistency” designates how well-behaved is the proposed methodology considering that the occurrence of a false alarm is extremely rare.

TABLE 6.1: Transition probabilities resulted through the proposed methodology.

	Regime 1	Regime 2
Regime 1	0.971	0.036
Regime 2	0.029	0.964

For the resulted model, the values of MSE and AIC show that the model is preferable than the one obtained by the classical MSMCM *without* a penalized likelihood technique. This conclusion will be verified in the following section where a series of simulations will be conducted for 5 different type of models.

6.3.2 Simulation study

In Chapter 4 we proposed an advanced methodology for the modeling of incidence data concerning various scientific fields such as medicine, seismology, meteorology, etc. Based on (1) the aforementioned methodology and; (2) on the real data of the previous section, a series of simulations is performed for verifying the findings of the real case study as well as establishing the general applicability of the proposed methodology. Specifically, based on the following general model:

$$y_{1t} = 21.78 - 0.29t + 0.001t^2 + 12.15sin\left(\frac{2\pi t}{\tau}\right) - 9.56cos\left(\frac{2\pi t}{\tau}\right)$$

$$- 1.41sin\left(\frac{4\pi t}{\tau}\right) + 2.99cos\left(\frac{4\pi t}{\tau}\right) + 0.17\text{Tmin} + 0.95y_{t-1} + \epsilon_{1s_t}, \qquad (6.5)$$

where

$$\epsilon_{1s_t} \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2_{1s_t} = 15.3),$$

and its 4 variants y_{2t}-y_{5t} (Table A.1), 500 datasets, consisting of 105 observations each, were simulated. Note that although, for presentation purposes, the results discussed concern a single dataset from the 500 simulated, similar results apply for all datasets as it becomes evident in Table 6.3.

After selecting the proper value of α and λ, as discussed in the *Real Case Study* section, the Markov Switching modeling mechanism is applied to the variables extracted through Elastic-Net. Figures 6.4 and 6.5 illustrate the regimes occurred, which are described by the models presented in (6.6) and (6.7), respectively with the transition probability matrix given in Table 6.2.

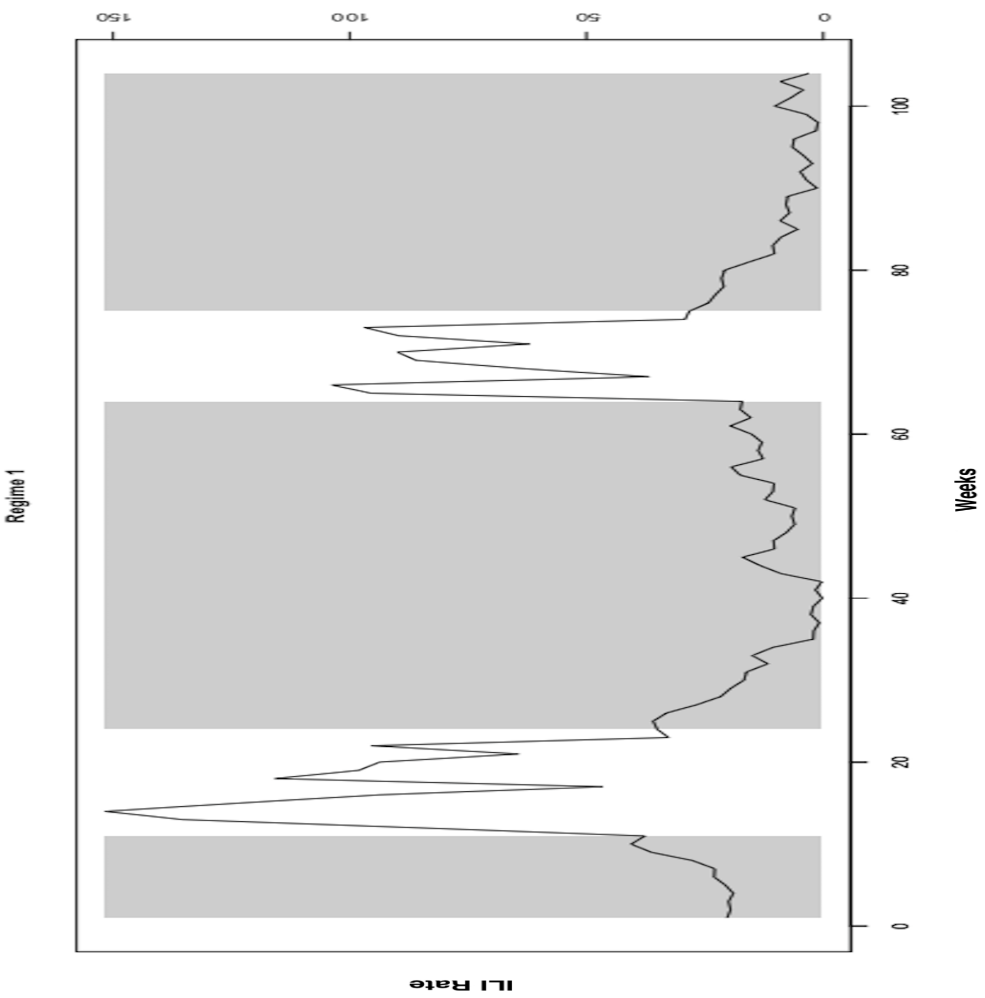


FIGURE 6.4: The first regime resulted through the proposed methodology.

1^{st} Regime

$$\hat{y}_t = 82.63 - 0.15t - 14.94sin\left(\frac{2\pi t}{\tau}\right) - 14.94cos\left(\frac{2\pi t}{\tau}\right) - 59.94sin\left(\frac{4\pi t}{\tau}\right)$$
$$+ 12.14cos\left(\frac{4\pi t}{\tau}\right) - 11.62sin\left(\frac{8\pi t}{\tau}\right) + 17.22cos\left(\frac{8\pi t}{\tau}\right) - 0.31\hat{y}_{t-1} - 0.36\hat{y}_{t-2}. \tag{6.6}$$

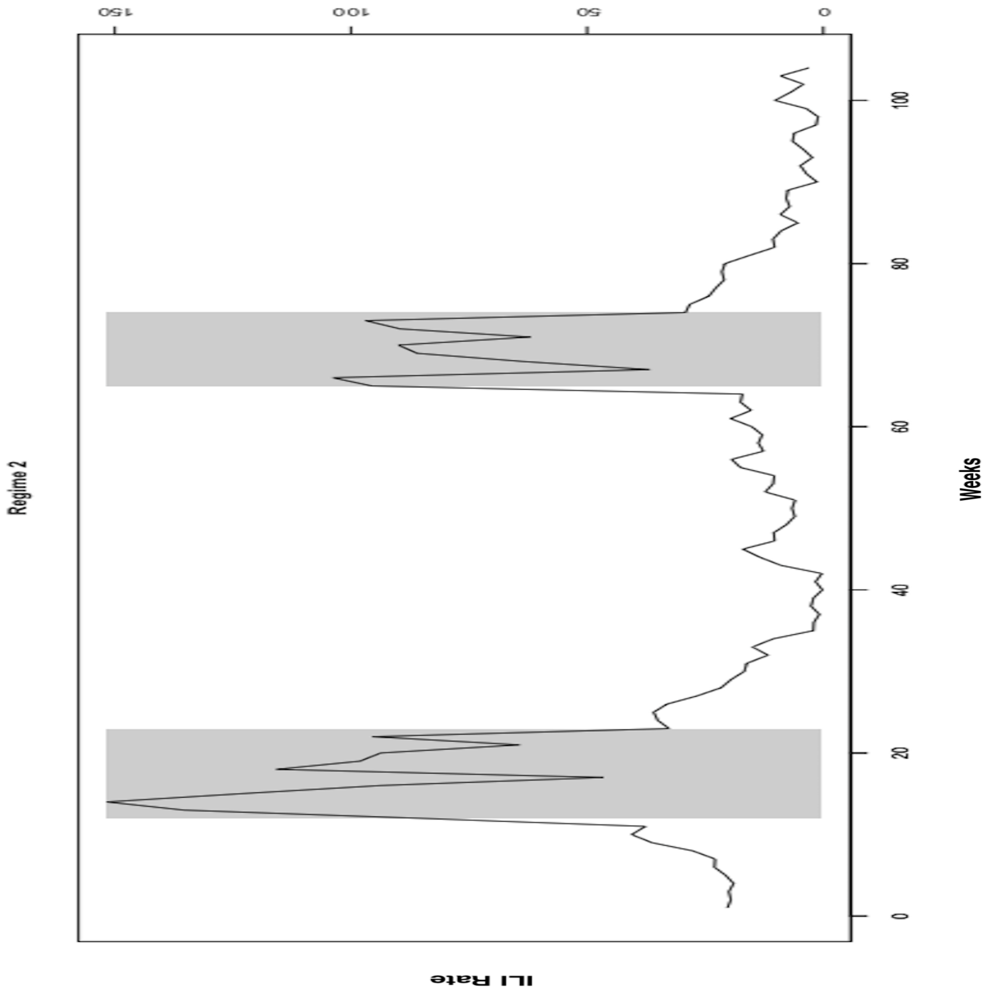


FIGURE 6.5: The second regime resulted through the proposed methodology.

2^{nd} Regime

$$\hat{y}_t = 12.1 - 0.06t - 1.44sin\left(\frac{2\pi t}{\tau}\right) - 2.64cos\left(\frac{2\pi t}{\tau}\right) + 0.05sin\left(\frac{4\pi t}{\tau}\right)$$

$$+ 0.86cos\left(\frac{4\pi t}{\tau}\right) - 0.08sin\left(\frac{8\pi t}{\tau}\right) + 0.75cos\left(\frac{8\pi t}{\tau}\right) + 0.63\hat{y}_{t-1} - 0.05\hat{y}_{t-2}. \tag{6.7}$$

The models presented in (6.6) and (6.7) once again achieve to fully recognize the regimes occurred to the simulated incidence dataset (first regime–non-epidemic, second regime–epidemic).

TABLE 6.2: Transition probabilities resulted through the proposed methodology.

	Regime 1	Regime 2
Regime 1	0.9	0.02
Regime 2	0.09	0.97

For comparative purposes, we present models (6.8) and (6.9) for the two regimes resulted by applying only the MSMCM with covariates (that is, without the proposed methodology) to $\underset{\sim}{X}$ (see Figures A.2 and A.2), meaning that no penalized likelihood technique has been preceded.

1^{st} Regime

$$\begin{aligned}\hat{y}_t = {} & 11.54 - 0.10t - 0.0001t^2 + 5sin\left(\frac{2\pi t}{\tau}\right) - 5.13cos\left(\frac{2\pi t}{\tau}\right) \\ & + 1.1sin\left(\frac{4\pi t}{\tau}\right) + 1.04cos\left(\frac{4\pi t}{\tau}\right) - 0.36sin\left(\frac{8\pi t}{\tau}\right) + 0.87cos\left(\frac{8\pi t}{\tau}\right) \\ & + 0.3854\text{Tmin} - 0.10\text{Tmax} + 0.31\text{Tmean} - 0.22\text{Tmed} - 0.02\text{WDmed} \\ & + 0.0266\text{Wmax} + 0.2401\text{Wmean} + 1.3301\text{Wmed} + 0.5851\hat{y}_{t-1} - 0.1286\hat{y}_{t-2}. \end{aligned} \tag{6.8}$$

2^{nd} Regime

$$\begin{aligned}\hat{y}_t = {} & -20.67 + 4.28t - 0.0572t^2 + 26.6sin\left(\frac{2\pi t}{\tau}\right) - 10.9cos\left(\frac{2\pi t}{\tau}\right) \\ & - 56.94sin\left(\frac{4\pi t}{\tau}\right) + 7.43cos\left(\frac{4\pi t}{\tau}\right) - 9.82sin\left(\frac{8\pi t}{\tau}\right) + 12.66cos\left(\frac{8\pi t}{\tau}\right) \\ & + 2.1\text{Tmin} + 2.49\text{Tmax} - 41.18\text{Tmean} + 37.7\text{Tmed} - 0.44\text{WDmed} \\ & + 2.9\text{Wmax} - 19.53 + 17.66\text{Wmed} - 0.61\hat{y}_{t-1} - 0.07\hat{y}_{t-2}. \end{aligned} \tag{6.9}$$

Results like the above are obtained for each of the 500 simulated datasets and for each of the 5 models considered ($y_{1t} - y_{5t}$).

Table 6.3 presents the resulted overall mean values of MSE (along with the corresponding standard deviations) and AIC of each model considered for the simulations. Notice that the proposed methodology achieves a considerable reduction of MSE and AIC in almost all cases.

TABLE 6.3: Overall mean values (standard deviations) of MSE and AIC of the five models considered for the 500 simulations.

Model	MSE[a]	MSE[b]	AIC[a]	AIC[b]
y_{1t}	241.00 (71.81)	241.25 (73.15)	596.68 (30.94)	605.31 (31.78)
y_{2t}	243.75 (77.10)	243.51 (76.74)	598.33 (31.81)	606.12 (31.97)
y_{3t}	619.14 (94.74)	628.65 (99.33)	698.25 (15.89)	709.26 (16.75)
y_{4t}	635.53 (105.36)	645.54 (111.62)	700.81 (17.47)	711.73 (18.73)
y_{5t}	638.52 (104.80)	646.94 (108.42)	700.60 (17.38)	712.17 (17.17)

The overall values of MSE and AIC of each model considered for the simulations with[a] and without[b] applying the proposed methodology;

Moreover, Table 6.4 indicates that over 90% of the simulated models resulted through the proposed methodology, performed better (in terms of AIC) than the ones where the typical methodology applied. The aforementioned, are also imprinted into Figures A.2 and A.3, where the occurring regimes fail to be accurately recognized. Hence, by comparing the models in (6.6) and (6.7) with the ones in (6.8) and (6.9), the latter are:

1. far to complex and thus difficult to interpret;
2. based solely on the EM algorithm with the risk of overfitting.

TABLE 6.4: Best performance (%) of AIC with and without applying the proposed methodology to the models considered.

Model	AIC[a]	AIC[b]
y_{1t}	92.76%	7.24%
y_{2t}	91.45%	8.55%
y_{3t}	94.74%	5.26%
y_{4t}	93.42%	6.58%
y_{5t}	94.08%	5.92%

a = with and b = without applying the proposed methodology;

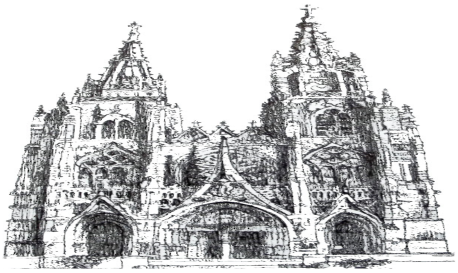
6.4 Concluding Remarks

Conclusively, in this study we proposed an advanced regime switching modeling approach for incidence data. The Elastic-Net method was used for variable pre-selection purposes, along with k-fold Cross-Validation for bias encountering purposes. After applying Changepoint Detection Analysis to the dataset for selecting the number of regimes, the Markov Switching mechanism was applied to the variables selected through the Elastic-Net process.

One of the main innovative features of the proposed methodology, is the use of the Elastic-Net method for variable pre-selection purposes (along with k-fold Cross-Validation for bias encountering purposes). Although the "screening before fitting" is an arguable issue, our findings clearly show that the variables discarded through a regularization process (i.e., Elastic-Net), were the ones with insignificant (or negligible significant) impact to the overall explanatory capability of the switching model. Hence, the results of the proposed methodology found to be superior to those associated with the "after fitting" screening, i.e., the selection process relied solely on the Markov Switching mechanism in terms of:

- MSE and AIC;
- robustness regarding the selection of regimes;
- simplicity/interpretability of the final selected model;
- estimation performance since there was an improvement (in terms of MSE and AIC) compared to the model resulted through the Markov Switching mechanism.

Consequently, the proposed three-phase procedure results in a robust, easily interpretable, user friendly and low-dimensional modeling scheme and as such, it is highly recommended to be applied to any type of data that experience typical and non-typical phases in fields such as epidemiology, medicine, seismology, meteorology and finance.



Notre Dame Cathedral, Rouen, France

Chapter 7

On Some Future Extensions & Generalizations

Cite: **Kalligeris EN**, Karagrigoriou A, Makrides A, Parpoula C and Barbu VS. On Some Aspects of Semi-Markov Switching Models. *Proceedings* 22nd *European Young Statisticians Meeting, Makrides, A. et al. eds*, Panteion University Publications 2021; 46-50.

Based (mainly) on the results of Chapter 6, in this Chapter we wish to initiate the exploration of a natural extension of the methodology presented in the previous Chapter, namely the use of semi-Markov switching models for the modelling of incidence data. As a result, in this Chapter we define the discrete time semi-Markov Switching model of conditional mean with covariates. Note though, that the work presented in this Chapter constitutes a work in progress on some aspects of discrete-time semi-Markov switching models.

Based on Definition 2, we now define the semi-Markov switching model of conditional mean with covariates.

Let us suppose a series of observations $\left\{y_0^{T-1}\right\}$ and $\left\{z_0^{T-1}\right\}$ a hidden state variable which follows a first order semi-Markov chain which is characterized by the following semi-Markov kernel $\boldsymbol{q}$:

$$q_{ij}(t) = P(J_k = j, S_{k+1} - S_k = t | J_{k-1} = i).$$

Definition 3 (Discrete Time semi-Markov Switching Model of Conditional Mean with Covariates)**.** *A semi-Markov switching model of conditional mean with covariates* $\Omega_1, ..., \Omega_q$ *for* $y_t,\ t \in \mathbb{N}^*$, *is defined by:*

$$y_t = c_{z_t} + \sum_{i=1}^{p} \phi_{iz_t} y_{t-i} + \sum_{d=1}^{q} \gamma_{dz_t} \Omega_d + \epsilon_t,\ t = 0, 1, ..., T-1, \tag{7.1}$$

where c_{z_t} *is a switching intercept,* ϕ_{iz_t}, $i = 1, ..., p$, *are autoregressive (AR) switching coefficients,* γ_{dz_t} *the coefficient associated with the* Ω_d *covariate,* $d =$

$1, ..., q$, *and* ϵ_t *are i.i.d zero-mean normally distributed random variables with variance* $\sigma^2_{z_t}$.

Under the model in (7.1) and for a N-state setting, one could consider various underlying (discrete) distributions for the waiting (sojourn) times between states.

7.1 Parameter Inference

Consider a series of observations $\left\{y_0^{T-1}\right\}$ and $\left\{z_0^{T-1}\right\}$ a hidden state variable as in Section 2.4. Moreover, suppose that the number of sojourn (waiting) times, denoted by $v_0, v_1, ..., v_R$, fulfills the obvious equality:

$$v_0 + v_1 + ... + v_R = T.$$

The relationship between the sojourn times and the state sequence can be simplified by reducing the entire sequence of states $z_0, z_1, ..., z_{T-1}$ to the sequence of states $j_0, j_1, ..., j_R$ which have been visited:

$$\begin{aligned}
j_0 &:= \{z_0, z_1, ..., z_{v_0-1}\}\\
j_1 &:= \{z_{v_0}, z_{v_0+1}, ..., z_{v_0+v_1-1}\}\\
&\vdots\\
j_R &:= \{z_{v_0+v_1...+v_{R-1}}, z_{v_0+v_1...+v_{R-1}+1}, ..., z_{T-1}\}.
\end{aligned}$$

Ferguson in (178) introduced the classical form of the complete (noncensored) data likelihood which allows only for sequences in which the last observation coincides with an exit from the hidden state. This form though, comes with some limitations since the summation includes all the possible paths considered in the complete-data likelihood and as a result the probability of obtaining an analytical solution is negligible. Furthermore, it assumes that the exit from a state coincides with the end of the sequence of observations Y_0^{T-1} since the sojourn times $v_r, r = 0, ..., R$ sum up to T. This results in the forbiddance of the consideration of semi-Markov chains with absorbing states which is unrealistic for most applications. Considering the aforementioned, Guédon[179] proposed the implementation of the survivor function into (7.1):

$$\begin{aligned}
&\underset{\sim}{L}'_{complete}\left(z_0^{T-1}, Y_0^{T-1}|\theta\right)\\
&= \sum_{t=0}^{T-1} log \sum_{r=1}^{R-1} f(y_t|Z_t, Y_0^t; \theta) P_{j_0} w_{j_0}(v_0) P_{j_r|j_{r-1}} w_{j_{r-1}}(v_{r-1}) P_{j_R|j_{R-1}} W_{j_R}(v_R),
\end{aligned} \tag{7.2}$$

where

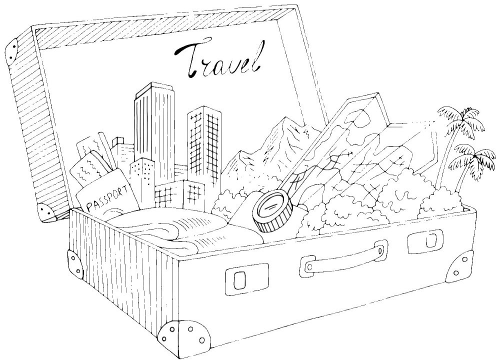
$$W_{j_r}(v_r) = \sum_{u_r \geq v_r} w_{j_r}(u_r),$$

is the survivor function for the sojourn time in state j_r and $\theta \in \mathbb{R}^m, m \in \mathbb{N}$, is the a parameter vector.

The estimator resulting through $\underset{\sim}{L}'_{complete}\left(z_0^{T-1}, Y_0^{T-1}|\theta\right)$ is known as *partial likelihood estimator*. Estimating the likelihood of semi-Markov switching model constitutes an incomplete (censored) data problem since the only accessible quantity is the observations. This fact makes the Expectation-Maximization (EM) algorithm the most suitable ML estimation technique for such models. For more on the estimation of censored semi-Markov switching models the interested reader may refer to Barbu(144) and Guédon.(179)

7.2 Concluding Remarks

In this Chapter, we discussed the concept of semi-Markov switching models under the discrete time framework. The fundamental aspects of such models were presented. To that end, the proper notations as well as the formulation of discrete time semi-Markov switching models of conditional mean with covariates were provided together with the associated parameter inference.



A journey full of new places and experiences

Chapter 8

Discussion and Conclusion

Before closing the thesis, we wish to briefly recall on the tools and findings provided.

In Chapter 3, we developed an alternative approach in order to model seasonality of influenza, based on a periodic regression model which incorporates additional auto-regressive and moving average components into Serfling's classical model including additionally climatological and meteorological covariates associated with ILI, with the ultimate aim of the early and accurate outbreak detection. The model selected (via an exhaustive search process) as the optimal one succeeded in estimating accurately the influenza-like syndrome morbidity burden in Greece for the period 2014-2016 as well as the duration of the epidemic waves.

In Chapter 4, we established that the changepoint detection analysis in conjunction with Periodic-type ARMA modeling with covariates, is capable of modeling time-series data with typical and non-typical parts and identifying effectively the beginning and end of the extreme periods that occurred. The proposed approach captures the behavior of the whole time-series with no significant loss of accuracy, and hence the derived results could be used for forecasting purposes.

In Chapter 5, a two comparative studies were conducted. The first one had the goal of evaluating the forecasting performance of the model MXM11 selected via the proposed methodology of Chapter 3 with other models associated with the modelling of influenza morbidity like the ones of M23 and MA3 (see Parpoula et al.[26] and Serfling,[44] respectively). The results derived indicated that among them the one with the best forecasting performance was the one of MX11. In the second comparative study, we implemented and evaluated cutting-edge changepoint analysis-based methods for detecting changes in location of univariate ILI rate data. The empirical comparative study provided evidence that statistical methods based on changepoint analysis have several appealing properties compared to the current practice for the detection of epidemics. RS/P and SegNeigh approaches, both succeeded in early and accurate outbreak detection and they can be applied to historical data without the need for distinguishing between epidemic and non-epidemic periods in the data, and single or multiple mean shifts can be detected. We concluded that, RS/P Phase I distribution-free changepoint analysis method is able to guarantee a prescribed false alarm probability without any knowledge about the (in-control) underlying distribution, whereas SegNeigh algorithm in conjunction with mixed effects periodic ARMA time series modeling is capable of modeling time series data with typical and non-typical parts.

In Chapter 6, we proposed an advanced regime switching modeling approach for incidence data. The Elastic-Net method was used for variable pre-selection purposes, along with k-fold Cross-Validation for bias encountering purposes. After applying Changepoint Detection Analysis to the dataset for selecting the number of regimes, the Markov Switching mechanism was applied to the variables selected through the Elastic-Net process. The proposed three-phase procedure resulted in a robust, easily interpretable, user friendly and low-dimensional modeling scheme.

Finally, in Chapter 7 we discussed future aspects of our work related to the concept of semi-Markov switching models under the discrete time framework and the proper notations as well as the formulation of discrete time semi-Markov switching models of conditional mean with covariates were provided together with the associated parameter inference.

This PhD thesis focused on the development of new modelling techniques for capturing the behavior of dynamical systems. To that end, and having as a reference point a dataset containing 105 ILI-rate incidence data for Greece (2 year window, 2014-2016), we achieved in:

1. Developing three novel methodologies; one for capturing the non-extreme (baseline) behavior of incidence data (Chapter 3), and two for capturing both the extreme and non-extreme behavior of incidence data (Chapters 4 and 6).

2. Conducting two useful comparative studies (Chapters 5 and 5.2) to evaluate the fitting and forecasting performance of the aforementioned methodologies, and;

3. Providing the appropriate formulation of a future plan on the semi-Markov switching model of conditional mean with covariates (Chapter 7).

Concluding, the work done throughout the thesis has been devoted solely on providing new insights and perspectives on the modelling of dynamical systems. Having studied the latter from the incidence data point of view, useful tools for capturing the behavior of such data were provided that could be proven beneficial to the society due to the consequences associated with the early detection and prevention of extreme, possibly harmful, events. We would like to close, with the hope that this PhD thesis will constitute a small useful handbook for the modelling of incidence data and dynamical systems in general.

Appendix A

Supplementary Material for Chapter 6

TABLE A.1: Models used for the simulation of the datasets.

	Trend		Periodicity		Autoregressive	Covariate
Model	t	t^2	1 year	6 months	AR(1)	Minimum Temperature
y_{2t}	*	*	*	*	*	
y_{3t}	*		*	*		*
y_{4t}	*		*	*		
y_{5t}	*		*			

TABLE A.2: Transition probabilities resulted through the typical methodology.

	Regime 1	Regime 2
Regime 1	0.94689649	0.1705281
Regime 2	0.05310351	0.8294719

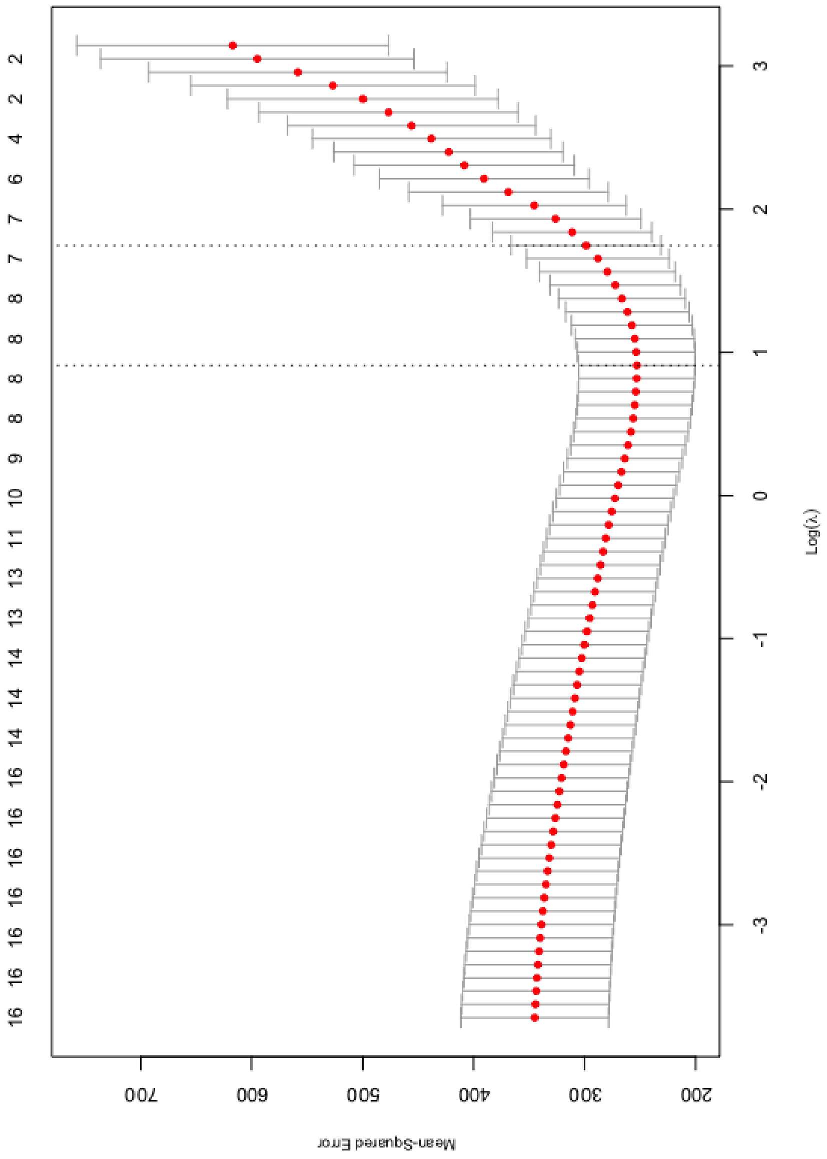


FIGURE A.1: Behavior of MSE for different values of $(log\lambda)$.

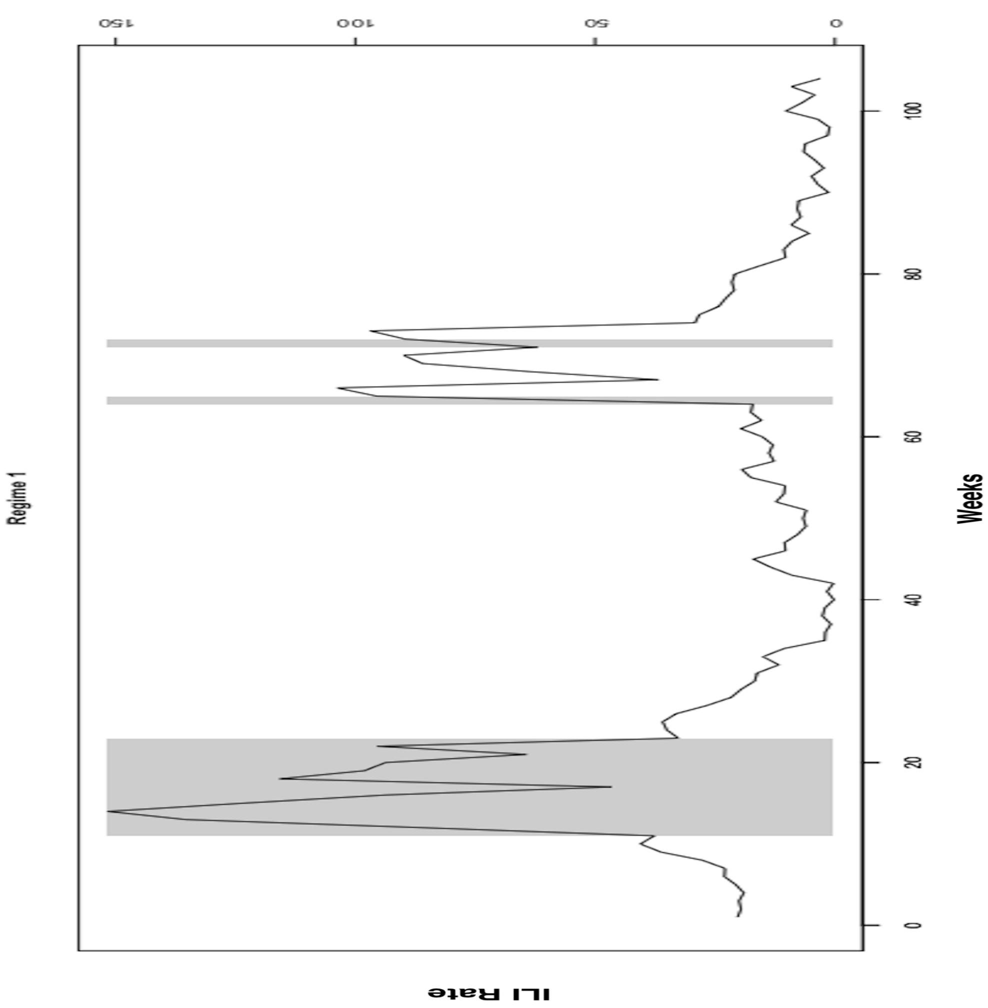


FIGURE A.2: First regime resulted through the typical methodology.

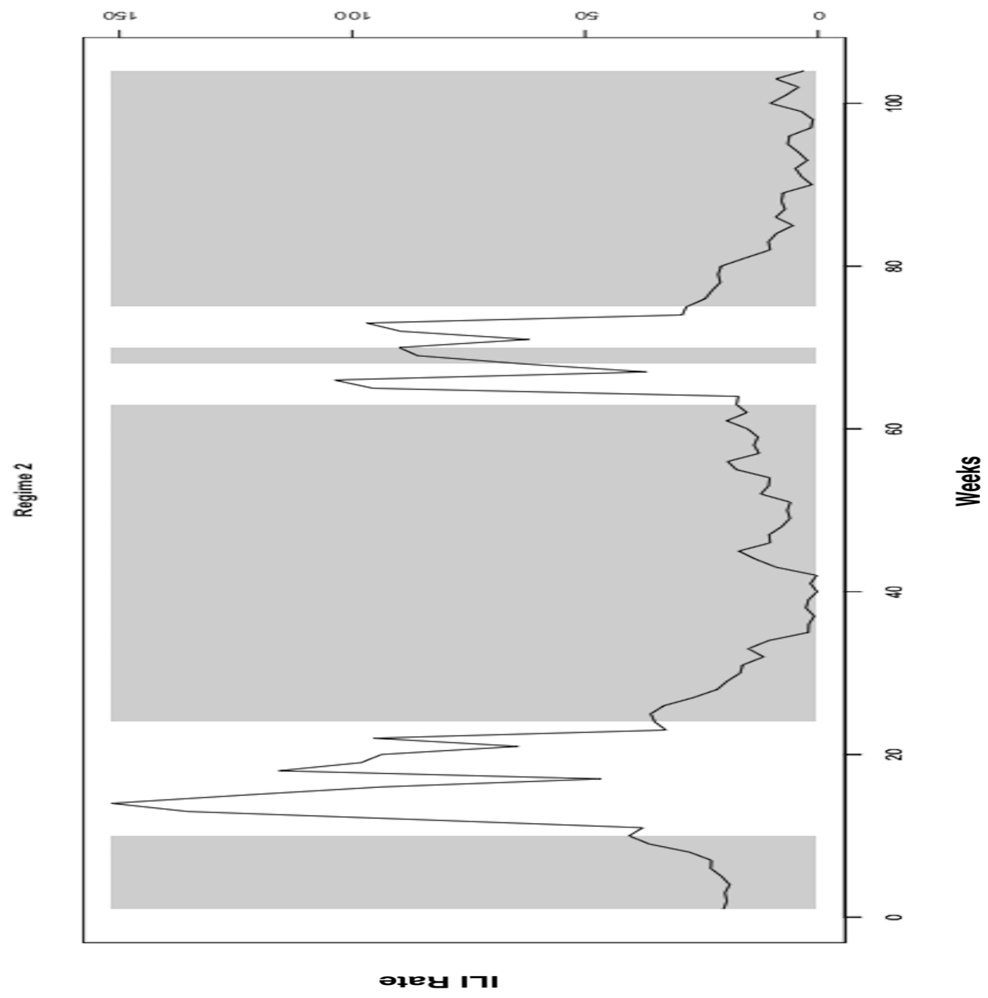


Figure A.3: Second regime resulted through the typical methodology.

Bibliography

[1] Naskar A. Love, God & Neurons: Memoir of a Scientist Who Found Himself by Getting Lost 2016.

[2] Packard NH, Crutchfield JP, Farmer JD and Shaw RS. Geometry from a Time Series. Phys Rev Lett 1980; **45**: 712-716.

[3] Crutchfield JP and McNamara BS. Equation of Motion from a Data Series. Complex Syst 1987; **1**: 417-452.

[4] Daniels BC and Nemenman I. Automated Adaptive Inference of Phenomenological Dynamical Models. Nat Commun 2015; **6**: 8133.

[5] Strogatz SH. Nonlinear Dynamics and Chaos: with Applications to Physics, Biology, Chemistry and Engineering 2018.

[6] Olsen LF and Schaffer WM. Chaos versus Noisy Periodicity: Alternative Hypotheses for Childhood Epidemics. Science 1990; **249(4968)**: 499-504.

[7] Ellner SP, Bailey BA and Bobashev GV. Noise and Nonlinearity in Measles Epidemics: Combining Mechanistic and Statistical Approaches to Population Modeling. Am Nat 1998; **151(5)**: 425-440.

[8] Allen LJS. An Introduction to Stochastic Epidemic Models. In Brauer F, van den Driessche P and Wu Jianhong (Eds.). Mathematical Epidemiology 2008; **1945**: 81-130.

[9] Bolker BM and Grenfell BT. Chaos and Biological Complexity in Measles Dynamics. Proc R Soc Lond B: Biological Sciences 1993; **251(1330)**: 75-81.

[10] Earn DJD, Rohani P, Bolker BM and Grenfell BT. A Simple Model for Complex Dynamical Transitions in Epidemics. Science 2000; **287(5453)**: 667-70.

[11] Rohani P, Earn DJD and Grenfell BT. Opposite Patterns of Synchrony in Sympatric Disease Metapopulations. Science 1999; **286(5441)**: 968-971.

[12] Bauch CT and Earn DJD. Transients and Attractors in Epidemics. Proc R Soc Lond B 2003; **270**: 1573-1578.

[13] Schichl H. Models and History of Modeling. Modeling Languages in Mathematical Optimization 2004; **Chap. 2**: 25-39.

[14] Starfield B, Shi L and Mackinko J. Contribution of Primary Care to Health Systems and Health. Milbank Q 2005; **83**: 457-502.

[15] Heath I and Smeeth L. Tackling Health Inequalities in Primary Care [Editorial]. Br Med J 1999; **318**: 1020-1021.

[16] Norbury M, Mercer SW, Gillies J, Furler J and Watt GCM. Time to Care: Tackling Health Inequalities Through Primary Care. Fam Pract 2011; **28**, 1-3.

[17] Teutsch SM and Thacker SB. Planning a Public Health Surveillance System. Epidemiol Bull 1995; **16**: 1-6.

[18] Sosin DM. Syndromic Surveillance: The Case for Skillful Investment. Biosecur Bioterror 2003; **1(4)**: 247-253.

[19] World Health Organization <http://www.who.int/topics/public_health_surveillance/en/> 2017.

[20] Fleming DM and Rotar-Pavlic D. Information from Primary Care: Its Importance and Value. A Comparison of Information from Slovenia and England and Wales, Viewed from the "Health 21" Perspective. Eur J Public Health 2002; **12**: 249-253.

[21] Shmueli G and Burkom H. Statistical Challenges Facing Early Outbreak Detection in Biosurveillance. Technometrics 2010; **52**: 39-51.

[22] Wagner MM, Gresham LS and Dato V. Case Detection, Outbreak Detection, and Outbreak Characterization. In Wagner MM, Moore AW and Aryel RM (Eds.). Handbook of Biosurveillance 2006; 27-50.

[23] Dato V, Shephard R and Wagner MM. Outbreaks and Investigation. In: Wagner MM, Moore AW and Aryel RM (Eds.). Handbook of Biosurveillance 2006; 13-26.

[24] Wagner MM. The Space Race and Biodefense: Lessons from NASA about Big Science and the Role of Medical Informatics. J Am Med Inform Assoc 2002; **9**: 120-122.

[25] Hulth A, Andrews N, Ethelberg S, Dreesman J, Faensen D, van Pelt W and Schnitzler J. Practical Usage of Computer-Supported Outbreak Detection in Five European Countries. Euro Surveill 2010; **15(36)**: 1-6.

[26] Parpoula C, Karagrigoriou A and Lambrou A. Epidemic Intelligence Statistical Modelling for Biosurveillance. Blömer J, Kotsireas I and Kutsia T (Eds.): MACIS 2017, LNCS 10693 2017; 1-15.

[27] Effler P, Ching-Lee M, Bogard A, Man-Cheng L, Nekomoto T and Jernigan D. Statewide System of Electronic Notifiable Disease Reporting from Clinical Laboratories: Comparing Automated Reporting with Conventional Methods. JAMA 1999; **282**: 1845-1850.

[28] Hoffman MA, Wilkinson TH, Bush A, Myers W and Griffin RG. Multijurisdictional Approach to Biosurveillance, Kansas City. Emerg Infect Dis 2003; **9**: 1281-1286.

[29] Overhage JM, Suico J and McDonald CJ. Electronic Laboratory Reporting: Barriers, Solutions and Findings. J Public Health Manag Pract 2001; **7**: 60-66.

[30] Panackal AA, M'ikanatha NM, Tsui FC, McMahon J, Wagner MM, Dixon BW, Zubieta J, Phelan M, Mirza S, Morgan J, Jernigan D, Pasculle AW, Rankin JT Jr, Hajjeh RA and Harrison LH. Automatic Electronic Laboratory - Based Reporting of Notifiable Infectious Diseases. Emerg Infect Dis 2001; **8**: 685-691.

[31] Greene SK, Huang J, Abrams AM, Gilliss D, Reed M, Platt R, Huang SS and Kulldorff M. Gastrointestinal Disease Outbreak Detection Using Multiple Data Streams from Electronic Medical Records. Foodborne Pathog Dis 2008; **9(5)**: 431-441.

[32] Ramadona AL, Lazuardi L, Hii YL, Holmner A, Kusnanto H and Rocklöv J. Prediction of Dengue Outbreaks Based on Disease Surveillance and Meteorological Data. PLoS Med 2016; **11(3)**: e0152688.

[33] Buckeridge DL, Burkom H, Campbell M, Hogan WR and Moore AW. Algorithms for Rapid Outbreak Detection: A Research Synthesis. J Biomed Inform 2005; **38(2)**: 99-113.

[34] Braun JV, Braun RK and Muller HG. Multiple Changepoint Fitting via Quasilikelihood, with Application to DNA Sequence Segmentation. Biometrika 2000; **87**: 301-314.

[35] Olshen AB, Venkatraman ES, Lucito R and Wigler M. Circular Binary Segmentation for the Analysis of Array-Based DNA Copy Number Data. Biostatistics 2004; **5**: 557-572.

[36] Picard F, Robin S, Lavielle M, Vaisse C and Daudin JJ. A Statistical Approach for Array CGH Data Analysis. Bioinformatics 2005; **6**: 27.

[37] Aggarwal R, Inclan C and Leal R. Volatility in Emerging Stock Markets. J Financ Quant Anal 1999; **34**: 33-55.

[38] Andreou E and Ghysels E. Detecting Multiple Breaks in Financial Market Volatility Dynamics. J Appl Econ 2002; **17**: 579-600.

[39] Fernandez V. Detection of Breakpoints in Volatility. IESA 2004; **11**: 1-38.

[40] Sonesson C and Bock D. A Review and Discussion of Prospective Statistical Surveillance in Public Health. J R Stat Soc Ser A Stat Soc 2003; **166(1)**: 5-21.

[41] Farrington P and Andrews N. Outbreak Detection: Application to Infectious Disease Surveillance. In: Brookmeyer R and Stroup DF (Eds.). Monitoring the Health of Populations: Statistical Principles and Methods for Public Health Surveillance 2003; 203-231.

[42] Unkel S, Farrington P, Garthwaite PH, Robertson C and Andrews N. Statistical Methods for the Prospective Detection of Infectious Disease Outbreaks: A Review. J R Stat Soc Ser A Stat Soc 2012; **175(1)**: 49-82.

[43] Stroup DF, Williamson GD, Herndon JL and Karon JM. Detection of Aberrations in the Occurrence of Notifiable Diseases Surveillance Data. Stat Med 1989; **8**: 323-329.

[44] Serfling R. Methods for Current Statistical Analysis of Excess Pneumonia-Influenza Deaths. Public Health Rep 1963; **78**: 494-506.

[45] Costagliola D, Flahault A, Galinec D, Garnerin P, Menares J and Valleron AJ. When is the Epidemic Warning Cut-Off Point Exceeded? Eur J Epidemiol 1994; **10**: 475-476.

[46] Costagliola D, Flahault A, Galinec D, Garnerin P, Menares J and Valleron AJ. A Routine Tool for Detection and Assessment of Epidemics of Influenza-Like Syndromes in France. Am J Public Health 1991; **81**: 97-99.

[47] Pelat C, Boëlle PY, Cowling BJ, Carrat F, Flahault A, Ansart S and Valleron AJ. Online Detection and Quantification of Epidemics. BMC Med Inform Decis Mak 2007; **5**: 29.

[48] Bengtsson T and Cavanaugh JE. An improved Akaike Information Criterion for State-Space Model Selection. Comput Stat Data Anal 2006; **50**: 2635-2654.

[49] Shang J and Cavanaugh JE. Bootstrap Variants of the Akaike Information Criterion for Mixed Model Selection. Comput Stat Data Anal 2008; **52**: 2004-2021.

[50] Page ES. Continuous Inspection Schemes. Biometrika 1954; **41**: 100-115.

[51] Liu S, Wright A and Hauskrecht M. Change-Point Detection Method for Clinical Decision Support System Rule Monitoring. Artif Intell Med Conf Artif Intell Med (2005-) 2017; **10259**: 126-135.

[52] Gallagher C, Lund R and Robbins M. Changepoint Detection in Climate Time Series with Long-Term Trends. J Clim 2013; **26(14)**: 4994-5006.

[53] Jaxk R, Chen J, Wang XL, Lund R and QiQi L. A Review and Comparison of Changepoint Detection Techniques for Climate Data. J Appl Meteorol Climatol 2007; **6**: 900-915.

[54] Parpoula C. A Distribution-Free Control Charting Technique Based on Change-point Analysis for Detection of Epidemics. Stat Methods Med Res 2022; https://doi.org/10.1177/09622802221079347.

[55] Lio P and Vannucci M. Wavelet Change-Point Prediction of Transmembrane Proteins. Bioinformatics 2000; **16(4)**: 376-382.

[56] Erdman C and Emerson JW. A Fast Bayesian Change Point Analysis for the Segmentation of Microarray Data. Bioinformatics 2008; **24(19)**: 2143-2148.

[57] Spokoiny V. Multiscale Local Change Point Detection with Applications to Value-at-Risk. Ann Stat 2009; **37**: 1405-1436.

[58] Erla S, Faes L, Tranquillini E, Orrico D and Nollo G. k-Nearest Neighbour Local Linear Prediction of Scalp EEG Activity During Intermittent Photic Stimulation. Med Eng Phys 2011; **33(4)**: 504-512.

[59] Faes L, Chon KH and Nollo G. A Method for the Time-Varying Nonlinear Prediction of Complex Nonstationary Biomedical Signals. IEEE Trans Biomed Eng 2009; **56(2)**: 205-209.

[60] Tahmasbi R and Rezaei S. Change Point Detection in GARCH Models for Voice Activity Detection. IEEE Transactions on Audio Speech and Language Processing 2008; **16(5)**: 1038-1046.

[61] Kass-Hout TA, Xu Z, McMurray P, Park S, Buckeridge DL, Brownstein JS, Finelli L and Groseclose SL. Application of Change Point Analysis to Daily Influenza-Like Illness Emergency Department Visits. J Am Med Inform Assoc 2012; **19(6)**: 1075-1081.

[62] Texier G, Farouh M, Pellegrin L, Jackson ML, Meynard JB, Deparis X and Chaudet H. Outbreak Definition by Change Point Analysis: A Tool for Public Health decision? BMC Med Inform Decis Mak 2016; **16**: 33.

[63] Christensen J and Rudemo M. Multiple Change-Point Analysis Applied to the Monitoring of Salmonella Prevalence in Danish Pigs and Pork. Prev Vet Med 1998; **36(2)**: 131-43.

[64] Painter I, Eaton J and Lober WB. Using Change Point Detection for Monitoring the Quality of Aggregate Data. Statistics in Defense and National Security. Annual Conference Proceedings 2012; **5**: 169.

[65] Lindgren G. Markov Regime Models for Mixed Distributions and Switching Regressions. Scand J Stat 1978; **5(2)**: 81-91.

[66] Barbu VS, Karagrigoriou A and Makrides A. Semi Markov Modelling for Multi State Systems, Methodol Comput Appl Prolab 2017; **19**: 1011-1028.

[67] Karagrigoriou A, Makrides A, Tsapanos T and Vougiouka G. Earthquake Forecasting Based on Multi State System Methodology. Methodol Comput Appl Probab 2016; **18**: 547-561.

[68] Votsi I, Limnios N, Tsaklidis G and Papadimitriou E. Hidden Markov Models Revealing the Stress Field Underlying the Earthquake Generation. Physica A 2013; **392**: 2868-2885.

[69] Shaby BA, Reich B, Cooley D and Kaufman CG. A Markov Switching Model for Heat Waves. Ann Appl Stat 2016; **10(1)**: 74-93.

[70] Clements MP and Krolzig HM. A Comparison of the Forecast Performance of Markov-Switching and Threshold Autoregressive Models of US GNP. Econom J 1998; **1(1)**: C47-C75.

[71] Cao Z, Liu X, Wen X, Liu L and Zu L. A Regime-Switching SIR Epidemic Model with a Ratio-Dependent Incidence Rate and Degenerate Diffusion. Sci Rep 2019; **9**: 10696.

[72] Shiferaw YA. Regime Shifts in the COVID-19 Case Fatality Rate Dynamics: A Markov-Switching Autoregressive Model Analysis. Chaos Solit Fractals: X 2021; **6**: 100059.

[73] Montgomery DC. Introduction to Statistical Quality Control 2013.

[74] Oakland JS. Statistical Process Control 2008.

[75] Woodall WH. Use of Control Charts in Health Care and Public Health Surveillance (with discussion). J Qual Technol 2006; **38**: 88-103.

[76] Borror CM, Champ CW and Rigdon SE. Poisson EWMA Control Charts. J Qual Technol 1998; **30**: 352-361.

[77] Gan FF. Monitoring Observations Generated from a Binomial Distribution Using Modified Exponentially Weighted Moving Average Control Charts. J Stat Comput Simul 1991; **37**: 45-60.

[78] Lucas JM. Counted Data CUSUM's. Technometrics 1985; **27**: 129-144.

[79] Burkom HS, Murphy SP and Shmueli G. Automated Time Series Forecasting for Biosurveillance. Stat Med 2007; **26**: 4202-4218.

[80] Dong Y, Hedayat AS and Sinha BK. Surveillance Strategies for Detecting Changepoint in Incidence Rate Based on Exponentially Weighted Moving Average Methods. J Am Stat Assoc 2008; **103**: 843-853.

[81] Elbert Y and Burkom HS. Development and Evaluation of a Data-Adaptive Alerting Algorithm for Univariate Temporal Biosurveillance Data. Stat Med 2009; **28**: 3226-3248.

[82] Höhle M and Paul M. Count Data Regression Charts for the Monitoring of Surveillance Time Series. Comput Stat Data Anal 2008; **52**: 4357-4368.

[83] Hutwagner L, Thompson WW, Seeman GM and Treadwell T. The Bioterrorism Preparedness and Response Early Aberration Reporting System (EARS). J Urban Health 2003; **80**: 89-96.

[84] Nobre FF and Stroup DF. A Monitoring System to Detect Changes in Public Health Surveillance Data. J Epidemiol 1994; **23**: 408-418.

[85] Rogerson PA and Yamada I. Monitoring Change in Spatial Patterns of Disease: Comparing Univariate and Multivariate Cumulative Sum Approaches. Stat Med 2004; **23**: 2195-2214.

[86] Rossi G, Lampugnani L and Marchi M. An Approximate CUSUM Procedure for Surveillance of Health Events. Stat Med 1999; **18**: 2111-2122.

[87] Fricker RD. Introduction to Statistical Methods for Biosurveillance, With an Emphasis on Syndromic Surveillance. Naval Postgraduate School, Monterey, California 2013.

[88] Fricker RD. Some Methodological Issues in Biosurveillance. Stat Med 2011; **30**: 403-415.

[89] Watkins RE, Eagleson S, Hall RG, Dailey L and Plant, AJ. Approaches to the Evaluation of Outbreak Detection Methods. BMC Public Health 2006; **6**: 263.

[90] Fraker SE, Woodall WH and Mousavi S. Performance Metrics for Surveillance Schemes. Qual Eng 2008; **20**: 451-464.

[91] Viboud C, Boëlle PY, Pakdaman K, Carrat F, Valleron AJ and Flahault A. Influenza Epidemics in the United States, France, and Australia, 1972-1997. Emerg Infect Dis 2004; **10**: 32-39.

[92] Brillman JC, Burr T, Forslund D, Joyce E, Picard R and Umland E. Modeling Emergency Department Visit Patterns for Infectious Disease Complaints: Results and Application to Disease Surveillance. BMC Med Inform Decis Mak 2005; **5(1)**: 4.

[93] Thompson WW, Shay DK, Weintraub E, Brammer L, Cox N, Anderson LJ and Fukuda K. Mortality Associated with Influenza and Respiratory Syncytial Virus in the United States. JAMA 2003; **289**: 179-186.

[94] Vergu E, Grais RF, Sarter H, Fagot JP, Lambert B, Valleron AJ and Flahault A. Medication Sales and Syndromic Surveillance, France. Emerg Infect Dis 2006; **12**: 416-421.

[95] Langmuir AD. The Surveillance of Communicable Diseases of National Importance. In: Buck C, Liopis A, Najera E and Terris M (Eds.). The Challenge of Epidemiology. Issues and Selected Readings, 3rd edn. 1995; 855-867.

[96] Chai T and Draxler RR. Root Mean Square Error (RMSE) or Mean Absolute Error (MAE)?-Arguments Against Avoiding RMSE in the Literature. Geosci Model Dev 2014; **7**: 1247-1250.

[97] Simonsen L, Reichert TA, Viboud C, Blackwelder WC, Taylor RJ and Miller MA. Impact of Influenza Vaccination on Seasonal Mortality in the US Elderly Population. Arch Intern Med 2005; **165**: 265-272.

[98] Lui KJ and Kendal AP. Impact of Influenza Epidemics on Mortality in the United States from October 1972 to May 1985. Am J Public Health 1987; **77**: 712-716.

[99] Gladyshev EG. Periodically and Almost-Periodically Correlated Random Processes with a Continuous Time Parameter. Theory Probab its Appl 1963; **8(2)**: 173-177.

[100] Gladyshev EG. Periodically Correlated Random Sequences. Sov math, Dokl 1961; **2**: 385-388.

[101] Jones RH and Brelsford WM. Time Series with Periodic Structure. Biometrika 1967; **54**: 403-407.

[102] Pagano M. On Periodic and Multiple Autoregressions. Ann Stat 1978; **6**: 1310-1317.

[103] Vecchia AV, Obeysekera JT, Salas JD and Boes DC. Aggregation and Estimation for low-order Periodic ARMA Models. Water Resour Res 1983; **9**: 1297-1306.

[104] Hipel KW and McLeod AI. Time Series Modelling of Water Resources and Environmental Systems. Elsevier sci. Ltd 1994.

[105] Birchenhall CR, Bladen-Hovell RC, Chui APL, Osborn DR and Smith JP. A Seasonal Model of Consumption. Econ J 1989; **99**: 837-43.

[106] Osborn DR and Smith JP. The Performance of Periodic Autoregressive Models in Forecasting Seasonal U.K. Consumption. J Bus Econ Stat 1989; **7**: 117-27.

[107] Osborn DR. Seasonality and Habit Persistence in a Life-Cycle Model of Consumption. J Appl Econom 1988; **3**: 255-66.

[108] Box GEP and Jenkins GM. Time Series Analysis: Forecasting and Control. Holden-Day, San Francisco, California 1976.

[109] Cipra T. Periodic Moving Average Processes. Apl mat 1985; **30**: 218-229.

[110] Vecchia AV. Maximum Likelihood Estimation for Periodic Autoregressive Moving Average Models. Technometrics 1985; **27**: 375-384.

[111] Troutman BM. Some Results in Periodic Autoregression. Biometrika 1979; **66**: 219-228.

[112] Tiao GC and Grupe MR. Hidden Periodic Autoregressive-Moving Average Models in Time Series Data. Biometrika 1980; **67**: 365-373.

[113] Abu Jahel AM. Generation of PAR Time Series Models Using Periodic Levinson-Durbin Algorithm. Master Thesis, The Islamic University of Gaza 2013.

[114] Henderson R and Matthews JNS. An Investigation of Changepoints in the Annual Number of Cases of Haemolytic Uraemic Syndrome. J R Stat Soc Ser C Appl Stat 1993; **42**: 461-471.

[115] Fryzlewicz P and Subba Rao S. Multiple-Change-Point Detection for Auto-Regressive Conditional Heteroscedastic Processes. J R Stat Soc Series B Stat Methodol 2014; **76(5)**: 903-924.

[116] Elsner JB, Xu FN and Jagger TH. Detecting Shifts in Hurricane Rates Using a Markov Chain Monte Carlo Approach. J Clim 2004; **17**: 2652-2666.

[117] Hinkley DV. Inference About the Change-Point in a Sequence of Random Variables. Biometrika 1970; **57**: 1-17.

[118] Chen J and Gupta AK. Testing and Locating Variance Changepoints with Application to Stock Prices. J Am Stat Assoc 1997; **92**: 739-747.

[119] Gupta AK and Tang J. On Testing Homogeneity of Variances for Gaussian Models. J Stat Comput Simul 1987; **27**: 155-173.

[120] Haccou P, Meelis E and Geer S. The Likelihood Ratio Test for the Change Point Problem for Exponentially Distributed Random Variables. Stoch Process Their Appl 1988; **27**: 121-139.

[121] Hinkley DV and Hinkley EA. Inference About the Change-Point in a Sequence of Binomial Random Variables. Biometrika 1970; **57**: 477-488.

[122] Hsu DA. Detecting Shifts of Parameter in Gamma Sequences with Applications to Stock Price and Air Traffic Flow Analysis. J Am Stat Assoc 1979; **74**: 31-40.

[123] Chen J and Gupta AK. Parametric Statistical Change Point Analysis. Birkhauser, 2000.

[124] Gupta AK and Chen J. Detecting Changes of Mean in Multidimensional Normal Sequences with Applications to Literature and Geology. Comput Stat 1996; **11**: 211-221.

[125] Yao Y. Estimating the Number of Change-Points via Schwarz's Criterion. Stat Probab Lett 1988; **6**: 181-189.

[126] Scott AJ and Knott M. A Cluster Analysis Method for Grouping Means in the Analysis of Variance. Biometrics 1974; **30(3)**: 507-512.

[127] Sen A and Srivastava MS. On Tests for Detecting Change in Mean. Ann Stat 1975; **3(1)**: 98-108.

[128] Venkatraman ES. Consistency Results in Multiple Change-Point Problems. PhD thesis, Stanford University 1993.

[129] Vostrikova LJ. Detecting Disorder in Multidimensional Random Processes. Sov math, Dokl 1981; **24**: 55-59.

[130] Eckley IA, Fearnhead P and Killick R. Analysis of Changepoint Models. In: Bayesian Time Series Models, Barber D, Cemgil AT, Chiappa S (Eds.). Cambridge University Press 2011.

[131] Braun JV and Muller HG. Statistical Methods for DNA Sequence Segmentation. Stat Sci 1998; **13(2)**: 142-162.

[132] Yao Y. Estimation of a Noisy Discrete-Time Step Function: Bayes and Empirical Bayes Approaches. Ann Stat 1984; **12**: 1434-1447.

[133] Jackson B, Sargle JD, Barnes D, Arabhi S, Alt A, Gioumousis P, Gwin E, Sangtrakulcharoen P, Tan L and Tsai TT. An Algorithm for Optimal Partitioning of Data on an Interval. IEEE Signal Process Lett 2005; **12**: 105-108.

[134] Killick R, Fearnhead P and Eckley IA. Optimal Detection of Changepoints With a Linear Computational Cost. J Am Stat Assoc 2012; **107(500)**: 1590-1598.

[135] Dempster A, Laird N and Rubin D. Maximum Likelihood from Incomplete Data via the EM Algorithm. J R Stat Soc Series B Stat Methodol 1977; **39(1)**: 1-38.

[136] Hamilton JD. Analysis of Time Series Subject to Changes in Regime. J Econometrics 1990; **45**: 9-70.

[137] Liporace LA. Maximum Likelihood Estimation for Multivariate Observations of Markov Sources. IEEE Trans Inf Theory 1982; **28**: 729-734.

[138] Hamilton JD. A New Approach of the Economic Analysis of Nonstationary Time Series and the Business Cycle. Econometrica 1989; **57(2)**: 357-384.

[139] Kim CJ. Dynamic Linear Models with Markov-Switching. J Econom 1994; **60**: 1-22.

[140] Chen MY. Markov Switching Models. Department of Finance, National Chung Hsing University 2013.

[141] Kim CJ and Nelson CR. State-Space Models with Regime Switching: Classical and Gibbs-Sampling Approaches with Applications. The MIT Press 1999.

[142] Di Persio L and Vettori S. Markov Switching Model Analysis of Implied Volatility for Market Indexes with Applications to S&P 500 and DAX. J Math 2014; **2014**.

[143] Limnios N and Oprişan G. Semi-Markov Processes and Reliability. Birkhäuser, Boston 2001.

[144] Barbu VS and Limnios N. Semi-Markov Chains and Hidden Semi-Markov Models Toward Applications, Their Use in Reliability and DNA Analysis. Lecture Notes in Statistics, Springer 2008.

[145] Vasdekis VGS, Rizopoulos D and Moustaki I. Weighted Composite Likelihood Estimation for a General Class of Random Effects Models. Biostatistics 2014; **15(4)**: 677-689.

[146] Mostashari F, Fine A, Das D, Adams J and Layton M. Use of Ambulance Dispatch Data as an Early Warning System for Community Wide Influenza-Like Illness, New York City. J Urban Health 2003; **80(2 Suppl 1)**: i43-49.

[147] Brinkhof MW, Spoerri A, Birrer A, Hagman R, Koch D and Zwahlen M. Influenza-Attributable Mortality Among the Elderly in Switzerland. Swiss Med Wkly 2006; **136(19-20)**: 302-309.

[148] Wong CM, Yang L, Chan KP, Leung GM, Chan KH, Guan Y, Lam TH, Hedley AJ and Peiris JS. Influenza-Associated Hospitalization in a Subtropical City. PLoS Med 2006; **3(4)**: e121.

[149] Gosling SN, Lowe JA, McGregor GR, Pelling M and Malamud BD. Associations Between Elevated Atmospheric Temperature and Human mortality. A Critical Review of the Literature. Clim Change 2009; **92**: 299-341.

[150] Kovats RS and Hajat S. Heat Stress and Public health. A Critical Review. Ann Rev Public Health 2008; **29**: 41-55.

[151] Armstrong B. Models for the Relationship Between Ambient Temperature and Daily Mortality. Epidemiology 2006; **17**: 624-631.

[152] Touloumi G, Atkinson R, Tertre L, Samoli AE, Schwartz J, Schindler C, Vonk JM, Rossi G, Saez M, Rabszenko D and Katsouyanni K. Analysis of Health Outcome Time Series Data in Epidemiological Studies Environmetrics 2004; **15**: 101-117.

[153] Tsangari H, Paschalidou A, Vardoulakis S, Heaviside C, Konsoula Z, Christou S, Georgiou KE, Ioannou K, Mesimeris T, Kleanthous S, Pashiardis S, Pavlou P, Kassomenos P and Yamasaki EN. Human Mortality in Cyprus. The Role of Temperature and Particulate Air Pollution. Reg Environ Change 2016; **16**: 1905-1913.

[154] World Health Organization <http.//wwwwhoint/mediacentre/factsheets/fs211/en/> 2017.

[155] Mantalos P, Mattheou K and Karagrigoriou A. An Improved Divergence Information Criterion for the Determination of the Order of an AR Process. Commun Stat Simul Comput 2010; **39**: 865-879.

[156] Toma A. Model Selection Criteria Using Divergences. Entropy 2014; **16(5)**: 2686-2698.

[157] Nakagawa S and Schielzeth H. A General and Simple Method for Obtaining R^2 from Generalized Linear Mixed-Effects Models. Methods Ecol Evol 2013; **4**: 133-142.

[158] Forecasting <https.//enwikipediaorg/wiki/Forecasting> as published in 25 June 2022.

[159] Fryzlewicz P. Wild Binary Segmentation for Multiple Change-Point Detection. Ann Stat 2014; **42(6)**: 2243-2281.

[160] Aminikhanghahi S and Cook DJ. A Survey of Methods for Time Series Change Point Detection. Knowl Inf Syst 2017; **51(2)**: 339-367.

[161] Killick R and Eckley IA. changepoint: An R Package for Changepoint Analysis. J Am Stat Assoc 2014; **58(3)**: 1-19.

[162] Black PE. “big-O notation”. Dictionary of Algorithms and Data Structures. U.S. National Institute of Standards and Technology 2005.

[163] Capizzi G and Masarotto G. Phase I Distribution-Free Analysis of Univariate Data. J Qual Technol 2013; **45**: 273-284.

[164] Breiman L. Better Subset Regression Using the Nonnegative Garrote. Technometrics 1995; **37(4)**: 373-384.

[165] Tibshirani R. Regression Shrinkage and Selection via the Lasso. J R Stat Soc Series B Stat Methodol 1996; **58(1)**: 267-288.

[166] Hoerl AE and Kennard RW. Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics 1970; **12(1)**: 55-67.

[167] Zou H and Hastie T. Regularization and Variable Selection via the Elastic Net. J R Stat Soc Series B Stat Methodol 2005; **67(2)**: 301-320.

[168] Hastie T, Tibshirani R and Friedman J. The Elements of Statistical Learning: Data Mining, Inference and Prediction, Second Edition. Springer 2016.

[169] Nardi Y and Rinaldo A. Autoregressive Process Modeling via the Lasso Procedure. J Multivariate Anal 2011; **102(3)**: 528-549.

[170] Chen K and Chan KS. Subset ARMA Selection via the Adaptive Lasso. Stat Interface 2011; **4(2)**: 197-205.

[171] Medeiros CM and Eduardo M. L1-Regularization of High-Dimensional Time-Series Models with Non-Gaussian and Heteroskedastic Errors. J Econometrics 2016; **191(1)**: 255-271.

[172] Bergmeir C, Hyndman RJ and Koo BA. Note on the Validity of Cross-Validation for Evaluating Time Series Prediction, Comput Stat Data Anal 2018; **120(C)**: 70-83.

[173] Mosteller F and Tukey JW. Data Analysis, Including Statistics, In Handbook of Social Psychology. Addison-Wesley, Reading, MA 1968.

[174] McLachlan GJ, Do KA and Ambroise C. Analyzing Microarray Gene Expression Data. Wiley 2004.

[175] Di Sanzo S. Testing for Linearity in Markov Switching Models: A Bootstrap Approach. Stat Method Appl 2009; **18**: 153-168.

[176] Lee S and Lee S. Change Point Test for the Conditional Mean of Time Series of Counts Based on Support Vector Regression. Entropy 2021; **23(4)**: 433.

[177] Sanchez-Espigares JA and Lopez-Moreno A. MSwM: Fitting Markov Switching Models. CRAN 2018. R package version 14, https://CRANR-projectorg/package=MSwM.

[178] Ferguson JD. Variable Duration Models for Speech. Proceedings of the Symposium on the Applications of Hidden Markov Models to Text and Speech. Princeton, New Jersey 1980; 143-179.

[179] Guédon Y. Estimating Hidden semi-Markov Chains from Discrete Sequences. J Comput Graph Stat 2003; **12(3)**: 604-639.