ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS
ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ & ΤΕΧΝΟΛΟΓΙΑΣ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ
ΜΕΤΑΠΤΥΧΙΑΚΟ ΣΤΑΤΙΣΤΙΚΗ
MSc IN STATISTICS

ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ & ΤΕΧΝΟΛΟΓΙΑΣ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ

AN ANALYSIS OF THE GREEK 1991-2016 CRIME DATA

Σπυριδούλα Ευαγγέλου Μαγγίνα

ΕΡΓΑΣΙΑ

Που υποβλήθηκε στο Τμήμα Στατιστικής
του Οικονομικού Πανεπιστημίου Αθηνών
ως μέρος των απαιτήσεων για την απόκτηση
Μεταπτυχιακού Διπλώματος
Συμπληρωματικής Ειδίκευσης στη Στατιστική
Μερικής Παρακολούθησης (Part-time)

Αθήνα
Νοέμβριος 2018


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΑΦΙΕΡΩΣΗ

Στους γονείς και την αδελφή μου Βιβή


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου,επίκουρο καθηγητή κ. Νικόλαο Δεμίρη για την πολύτιμη βοήθεια και ενθάρυνσή του στη διάρκεια της πολύ ουσιαστικής και όμορφης συνεργασίας μας.

Ευχαριστώ τους συμφοιτητές μου, με τους οποίους μοιραστήκαμε ευχάριστες στιγμές κατά τη διάρκεια της φοίτησής μας στο μεταπτυχιακό.

Ιδιαίτερα,ευχαριστώ τους γονείς και την αδελφή μου Βιβή,για την υποστήριξη και την αγάπη τους.


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΒΙΟΓΡΑΦΙΚΟ ΣΗΜΕΙΩΜΑ

Είμαι απόφοιτος του Τμήματος Μαθηματικών του Εθνικού και Καποδιστριακού Πανεπιστημίου Αθηνών.Η διδασκαλία στην εκπαίδευση έχει έως τώρα αποτελέσει τη βασική μου ενασχόληση.Έχω εργαστεί ως εξωτερικός συνεργάτης στο ερευνητικό εργαστήριο της Μονάδας Αλλεργιολογίας και Κλινικής Ανοσολογίας της Β' Παιδιατρικής Κλινικής του Πανεπιστημίου Αθηνών.Είμαι κάτοχος του Proficiency πτυχίου της Αγγλικής γλώσσας και γνωρίζω Γαλλικά.


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ABSTRACT

Spiridoula Maggina

AN ANALYSIS OF THE GREEK 1991-2016 CRIME DATA.

November 2018

The study of crimes has great scientific interest with practical implications regarding government future strategies aiming at society safety as well as and other social and macroeconomic impact.In this study, we have utilized annual historical crime data for the period 1991-2016, concerning seven common crimes in Greece's territory general,as well as focusing on the main 14 districts that form the country.The stationary time series is one of the tools for making predictions and autoregressive integrated moving average (ARIMA) models have been already successfully used in forecasting econometrics and other social science problems.The main goal of this study was the prediction of the best fitted models for depicting the crime trend and forecasting future values,regarding the whole country and Greece's 14 individual regions.

In this thesis, we have used Box Jenkins ARIMA methodology for the univariate time series analysis,after we removed the trend from the series.We suggest, for the time series analysis of count data,the implementation of a method based on GLM models,performed from the R package "tscount".We examined also, a multivariate analysis with VAR models.

The applied methods explain the time series adequately. According to the forecasts we do not expect significant changes in the crime pattern in the future. Regarding the variable “fraud” however, a substancial increase is been expected for the next 30 years in total Greece and some specific regions namely An.Makedonia-Thraki, Thessalia, Ipeiros, Peloponnisos, Attiki, Ionia Nisia, Voreio and Notio Aigaio according to the predictive models. ARIMA models could perform better in shorter run forecasts and with no doubt longer time series could provide better results.


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΠΕΡΙΛΗΨΗ

Σπυριδούλα Μαγγίνα

ΜΙΑ ΑΝΑΛΥΣΗ ΤΩΝ ΕΛΛΗΝΙΚΩΝ ΕΓΚΛΗΜΑΤΟΛΟΓΙΚΩΝ ΔΕΔΟΜΕΝΩΝ ΓΙΑ ΤΑ ΕΤΗ 1991-2016.

Νοέμβριος 2018

Η μελέτη των εγκλημάτων έχει μεγάλο επιστημονικό ενδιαφέρον με πρακτικές συνέπειες, αφορώντας μελλοντικές κυβερνητικές στρατηγικές που στοχεύουν στην κοινωνική ασφάλεια,καθώς και άλλες προεκτάσεις με κοινωνικό και μακροοικονομικό αντίκτυπο.Στην μελέτη αυτή,έχουμε χρησιμοποιήσει ετήσια ιστορικά εγκληματολογικά δεδομένα για τη περίοδο 1991-2016, που αφορούν σε επτά σύνηθη εγκλήματα στην επικράτεια της Ελλάδας,καθώς και στις 14 περιφέρειές της.Οι στάσιμες χρονολογικές σειρές είναι ένα από τα εργαλεία για να κάνουμε προβλέψεις, και τα αυτοπαλίνδρομα ολοκληρωμένα υποδείγματα κινητών μέσων (ARIMA) έχουν ήδη με επιτυχία εφαρμοστεί σε προβλέψεις προβλημάτων οικονομετρικών και άλλων κοινωνικών επιστημών.Ο βασικός στόχος αυτής της μελέτης ήταν η πρόβλεψη των καλύτερα προσαρμοσμένων μοντέλων που απεικονίζουν τις τάσεις της εγκληματικότητας και προβλέπουν μελλοντικές τιμές για την Ελλάδα στο σύνολο καθώς και για τις 14 περιφέρειές της.

Σε αυτήν την μελέτη,έχουμε χρησιμοποιήσει την μεθοδολογία των Box-Jenkins ARIMA για χρονοσειρές μιας μεταβλητής,άφού έχουμε αφαιρέσει την τάση από αυτές.Προτείνουμε για την ανάλυση δεδομένων μετρήσιμων θετικών ακέραιων τιμών,την εφαρμογή μεθόδου βασισμένη στα μοντέλα GLM εφαρμοσμένο από το πακέτο της R, 'tscount".Μελετήσαμε επίσης την μέθοδο της πολυμεταβλητής ανάλυσης με τα μοντέλα VAR.

Οι μέθοδοι που εφαρμόστηκαν εξηγούν τις χρονοσειρές μας σε ικανοποιητικό βαθμό. Σύμφωνα με τις προβλέψεις δεν περιμένουμε σημαντική διαφορά στο πεδίο της εγκληματικότητας για το μέλλον. Παρ' όλα αυτά, αναφορικά με την μεταβλητή “fraud”, αναμένεται σημαντική αύξηση μέσα στα επόμενα 30 χρόνια σε όλη την επικράτεια της Ελλάδας, καθώς και σε συγκεκριμμένες περιφέρειες, όπως η Αν. Μακεδονία-Θράκη, Θεσσαλία, Ήπειρος, Πελοπόννησος, Αττική, Ιόνια Νησιά, Βόρειο και Νότιο Αιγαίο, σύμφωνα με τα μοντέλα πρόβλεψης. Τα μοντέλα ARIMA θα μπορούσαν να συμπεριφερθούν καλύτερα σε μικρότερα διαστήματα πρόβλεψης και χωρίς αμφιβολία μεγαλύτερες χρονοσειρές θα μπορούσαν να παρέχουν καλύτερα αποτελέσματα.


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

TABLE OF CONTENTS

CHAPTER 1 : INTRODUCTION 1
CHAPTER 2: DATA 3
2.1 Crime data 3
2.2 Type of data 4
2.2.1 Administrative type of data 4
2.2.2 Victimization surveys and self-reporting of data 6
2.3 Data sources 6
2.4 Data collection 10
CHAPTER 3: MODELS AND METHODS 13
3.1 Time Series 13
3.2 Stohastic process -- Random Walk 14
3.3 Stationarity 15
3.4 The Box Jenkins Methodology 17
3.4.1 Trend Identification 17
3.4.2 Trend Elimination by Differencing 18
3.4.3.1 The Autocorrelation Function (ACF) 19
3.4.3.2 Unit Root Tests -- Augmented Dickey-Fuller Test 19
3.5 ARMA Models 20
3.6 ARIMA Models 21
3.6.1 The auto.arima function 22
3.7 Estimation of the model 22
3.8 Multivariate Analysis 23
3.8.1 The VAR Model (Vector Autoregressive Models) 24
3.8.2 VAR function in R 25
3.9 Count Data 26
R package 'tscount' 26
3.10 Missing Data 30
3.10.1 Imputation 30
3.10.2 Missing data estimation in time series 31
CHAPTER 4: RESULTS 33
4.1 Descriptives 33
4.1.1 Graphic visualization 33
4.2 Results 35
4.2.1 Univariate Time Series Analysis 35
4.2.2 Multivariate Analysis 43
4.2.3 Regional Analysis 45
4.2.3.1 Inference for count data 45
4.2.3.2 Inference with ARIMA Models Theory 55
1. Burglaries 55
2. Robberies 61
3. Motor and vehicle thefts 64
4. Fraud 67
5. Law about drugs 70
CHAPTER 5: DISCUSSION 73
APPENDIX 74
REFERENCES 81

ΕΙΚΟΝΕΣ

Figure 2.1 12

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

Table 1: ADF test p-values for different transformation of the time series. 38

Table 2 : Forecasts of the time series based on the smallest AICC. 39

Table 3. Estimations about count data in Dytiki Ellada. 51

Table 4. Predicted values for count data in Dytiki Ellada 53

Table 5. ADF tests for burglaries in Dytiki Makedonia 58

Table.6 Estimated coefficients for the time series "burglary"in Dytiki Makedonia. 58

Table 7. ARIMA models and AICC about burglaries time series. 60

Table 8. ARIMA models and AICC for the robberies time series. 62

Table 9.ARIMA models and AICC for the motor.vehicle theft time series. 65

Table 10. ARIMA models and AICC for the fraud time series. 68

Table 11. ARIMA models and AICC for the law about drugs time series. 71


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΚΑΤΑΛΟΓΟΣ ΓΡΑΦΗΜΑΤΩΝ

Diagram 4.1 crimes in Greece: robberies and frauds in the years 1991-2016 . 33

Diagram 4.2 crimes in Greece: homicides and rapes in the years 1991-2016. 34

Diagram 4.3 crimes in Greece: burglaries, motor and vehicle thefts, and crimes applied to the law about drugs. 34

Diagram 4.4 ACF plot of the time series fraud. 37

Diagram 4.5 ACF plot of log-values of time series fraud. 37

Diagram 4.6 ACF plot of the first differences of the log-values of time series fraud. 38

Diagram 4.7 Forecasts of homicides in Greece for the next 30 time units. 39

Diagram 4.8 Forecasts of burglaries in Greece for the next 30 time units. 40

Diagram 4.9 Forecasts of robberies in Greece for the next 30 time units. 40

Diagram 4.10 Forecasts of motor and vehicle thefts in Greece for the next 30 time units. 40

Diagram 4.11 Forecasts of rapes in Greece for the next 30 time units. 41

Diagram 4.12 Forecasts of frauds in Greece for the next 30 time units. 41

Diagram 4.13 Kalman imputation on 'law about drugs' missing data. 42

Diagram 4.14 Forecasts of the “law about drugs’ variable with the imputed values, for the next 30 time units. 43

Diagram 4.15 Cross - Correlations of the variables in multivariate analysis. 44

Diagram 4.16 Homicides in the 14 regions of Greece. 46

Diagram 4.17 Homicides in Dytiki Ellada and ACF plot of the time series. 49

Diagram 4.18 Diagnostic plots after model fitting to the Dytiki Ellada homicides data. 50

Diagram 4.19 Rapes in the regions of Greece,time series 1991-2016. 54

Diagram 4.20 Burglaries in Greece 1991-2016. 56

Diagram 4.21. ACF plots of the burglary time series, log-values of the series and the first differences of the log-values of the series and forecasts, in Dytiki Makedonia 57

Diagram 4.22 Forecasts about burglaries in the districts of Greece. 59

Diagram 4.23 Robberies time series in the region of Greece ,1991-2016. 61

Diagram 4.24 Robberies forecast plots . 63

Diagram 4.25 Motor and vehicle theft time series in the 14 districts of Greece. 64

Diagram 4.26 Forecasts in Motor and vehicle thefts in the districts of Greece. 66

Diagram4.27 Fraud in the regions of Greece. 67

Diagram 4.28 Forecasts in fraud in the districts of Greece. 69

Diagram 4.29 Crimes regarding the law about drugs in the regions of Greece. 70

Diagram 4.30 Forecasts for the law about drugs in the regions of Greece. 72

CHAPTER 1.

INTRODUCTION.

Criminal activity has always been to the interest of research in several scientific fields such as social sciences and economics. Moreover, governments of the countries process criminal data for the purpose of investigation, coordination and future strategies. With no doubt, a thorough study of the subject from many aspects, could provide useful and en-lightening information. Societies nowadays facing quite challenging circumstances and appear to have difficulties adjusting to them. Changes in the structure of a society, an economic crisis and violent population movements for instance, can transform the population's behavior in a city and a country as well.

At all times, there has been a wise strategy to gain knowledge from the past in order to infer better for the future.The ability to predict can serve as a valuable source of knowledge for law enforcement agencies, for example, both from tactical as well strategic perspectives. The purpose of this thesis is to provide a better understanding of a number of specific crimes in the country and the separate regions of Greece, as well as the relative predictive models.

In chapter 2 is presented an analytical structure of the data types and possible sources, based on Eurostat's regulations. Additionally, there is a thorough presentation of the stages of the data until they reach the justice authorities and the process that leads to the official data collection for this thesis analysis. Annually country level data and regional data as well have been obtained, concerning seven specific crimes.

Subsequently, in chapter 3 we introduce the time series and the methods for analyzing and forecasting them. Trend identification and elimination methods are provided, so as to gain stationarity. ARIMA (p,d,q) models are defined properly as a general form of ARMA (p,q) models. We discuss about the models estimation with the powerful AICC criterion. The approach of the multivariate analysis with VAR models is described and later on we elaborate on the count data anal-

ysis, that cannot be implemented with methods for continuous distributions. Finally, imputation and missing data estimation methods are given.

The purpose of the study is to draw inferences from the time series, hence in chapter 4 we depict the plots of the country's crime time series and following we proceed with the predicted models and the results. Firstly, we discuss the univariate forecast analysis for the seven selected crimes in Greece. Afterwards, we attempt a multivariate analysis on the time series 1991-2016, continuing with the regional Greece's time series, implemented with two different predictive methods, depending on the number of the observation in the series.

In chapter 5, a brief discussion on the analysis and some comments are mentioned.

CHAPTER 2.

DATA

2.1 Crime data.

Ackowledging the changes in Greece's population over the years and facing the economic crisis nowadays combined with various challenging circumstances,it appears to be quite important to analyze the crime pattern in all and in the specific country regions.Aiming to gain a thorough view of Greece's possible changes in crime and related trends through the last decades,the objective has been to find and collect official data with the most extensive time series sequence possible.

Generally,in Greece, according to EUROSTAT,there has been a decreasing in crime during 2007-2012, observating an increase in burglaries,a doubling in robberies during the years 2007-2012 and an increase of 38% of motor and vehicle theft for the same period of time.Crimes against property in all have shown a boost and homicides have a constant behavior,in contrast to the decrease in the other european countries.In Greece,the total of the recorded crimes has shown a reduction of 54%.However,there has been a discontinuance in the time series, that partly causes this particularly large decrease.

2.2 Type of data.

2.2.1 Administrative type of data

All the types of data featured in EU Statistics on crime and criminal justice are administrative data and they are produced in various agencies at each stage of the criminal justice system in each legislation.They constitute national domain and some of those international data collections existing from the national level are:

A. Eurostat data collection

Eurostat and UNODC,in 2014 began on a joined course collecting annual data on crime and criminal justice,using the UN and ad-hoc EUROSTAT questionnaires respectively.The data and metadata collected from national statistical institutes or/and the Police and the Justice Department of each country.These data collection allows to gather information on offense,victims,suspects,persons prosecuted and persons convicted,the number of police,judges and other judicial staff and the number of people detained in prison and prison capacity.

B. UN data collection (https://data.unodc.org/)

The Economic and Social Council,in its resolution 1984/48 of May 1984,requested "the Secretary -General to maintain and develop the United Nations crime-related data base by continuing to conduct quinquennial surveys of crime trends,operations of criminal justice systems and crime prevention strategies,and to report periodically to the Committee on Crime Prevention and Control on the progress made.The United Nations Surveys on Crime Trends and the Operations of Criminal Justice Systems(UN-CTS) has since involved into a biennial and since 2009 into an annual survey of criminal justice data.Since 2010 UNODC also includes a module on crime victimization survey data in UN-CTS.UNODCcollaborates with regional organizations to implement the data collection.Every year the UN-CTS is sent to 194 UN Member States plus 2 observer and 1 territory.About 50% of the questionnaires are received back,not all of them with complete data.

C. European Sourcebook Group (http:/wp.unil.ch/europeansourcrbook/)

The European Sourcebook Group is a group of mostly academic experts that produces on regular basis the European Sourcebook of Crime and Criminal Justice Statistics since 1996.The fifth and latest edition of the Sourcebook covers the years 2007-2011 and has been published in September 2014.It is similar to EU Statistics and covers police,prosecution,conviction,correctional statistics and victimization surveys.

D. Council of Europe SPACE I statistics (http:/wp.unil.ch/space)

SPACE is the Council of Europe Annual Penal Statistics of the populations held in custody and/or in other types of penal institutions across Europe ,compiled by researchers at a the University of Lausanne in Switzerland.

E. Sub-national data

Data available for the European Union member States,EFTA countries,EU candidate countries and EU potential candidate countries,on regional level for 2008,2009,2010 only,have been collected and are concerning domestic burglary,homicide,robbery and theft of a motor vehicle.These data have been collected with previous EU definitions,which are no longer valid accordingly with the collaboration of EU Statistics and UNODC.So,any attempt of comparing sub-national with national data,should be avoided.

F. City level data

Data on city level about crime,are asked from each jurisdiction for the largest[1] city and are in general provided for two different aggregations:

- the city proper
- the wider urban agglomeration

1 http://ec.europa.eu/regional policy/sources/docgener/focus/2012 01 city.pdf

2.2.2 Victimization surveys and self-reporting of data

Victimization surveys capture both criminal incidents reported to the police and criminal incidents not reported.As a result,they have the potential of uncovering crimes that are less well reported or recorded by the police.These surveys are randomly selected samples of population and can deepen the understanding of criminal actions and victim characteristics that might not be captured in police recorded data by asking directly persons about their victimization experiences.They give a better estimation of the prevalence of crimes.The advantage, though, is less important for crimes which rarely occur, as either sample sizes would have to increase or other methodologies and techniques would have to be applied in order to come with reliable estimates.
This kind of survey is conducted in several EU Member States but the different methodologies and definitions used, limit their use for international comparisons. Moreover,EUROSTAT does not publish data based on this kind of surveys.

2.3 Data sources.

According to EUROSTAT 's overview, EUROSTAT collects data on crime and the operation of criminal justice systems in order to make polisy-relevant information and analysis available in a timely manner to the European community. Since 2014, the figures on crime and criminal justice are collected through a joint Eurostat-UNODC data collection across the EU countries, by using the UN crime questionnaire and an ad-hoc Eurostat questionnaire. Official sources such as the Police, the Ministry of the Interior, the Ministry of Justice and the National Prison Administration, report the data which are compiled by the National Statistics Office. The survey provides data on intentional homicide offense and victims recorded by Police,assault,sexual violence,rape,robbery and kidnapping offense recorded by Police,theft,motor vehicle theft and burglary,persons brought into formal contact with the police and/or the criminal justice system,persons prosecuted,count input/output statistics,official capacity of prisons and number of

persons held in prisons,number of police personnel,court personnel and prisons staff.Additional data on cases processed and drug trafficking are collected and published by EUROSTAT.

Data about criminal activities are difficult to collect and the impact in society is hard to measure and difficult to estimate.There are also various approaches in collecting data ,difficulties in process,validation and the lack of uniform definitions,standardized instruments and a common methodology between jurisdictions.For all the reasons mentioned above, a methodological guide has been formulated by EUROSTAT, which manages to give an insight into the process of annually collecting data.

The Crime and Criminal Justice Statistics -Methodological guide for users 2017 Version(updated May 2017)[1],presents how EU Statistics on crime and criminal justice are collected,which types of data are included and how they are classified in an international context.Additionally,it is been explained how data are processed and validated,how indicators are calculated, which limitations exist,which comparisons can be made,how different national definitions,legal systems and coverage impact comparability and why counting units and counting rules matter for comparisons.Many methods can be applied to achieve better results,and the most accurate one is recording administrative data on criminal acts brought to the attention of law enforcement and criminal justice procedures.A first collection with the reference year 2005 ,was organised in 2007 [2]

EU Statistics on crime and criminal justice include administrative data at four different stages of the criminal justice system according to the methodological guide.The first stage of administrative statistics is data recorded by law enforcement authorities.

Stages of data in EU Statistics on crime and criminal justice:

Police statistics usually provide the number of crimes recorded and the number of suspects and offences brought into formal contact by the police.They are usually a count of all criminal offense reported to or detected by the police.Nevertheless there is an undercoverage of crime the so called dark figure of crime.Not all criminal acts are reported to the law enforcement

1 http://ec.europa.eu/eurostat/documents/64346/2989606/Methodological+guide+for+users/bfd3bb4a-67b7-44de-860e-cb911df9e17a

2 http://ec.europa.eu/eurostat/statistics explained/index.phptitle=Crime_and_criminal_justice_statistics

authorities and there are also some decisions to be made in order a crime to be recorded,such as recognition by the victim,the knowledge that a crime has indeed occurred,the notification of the police,the possible failure of the police recording it,the lack of interest pursuing minor offenses etc. Eurostat also publishes data on the number of offenses of 13 criminal acts as defined by the ICCS[1].With an offense count,each contravention of an article of criminal law may be recorded separately.When counting offenses,data cannot be further disaggregated.On the level of police recorded data,a second counting unit is the person and more specifically the victim and the suspect too.

After a crime is reported by the police,an investigation is opened and the decision is to pass the case for prosecution.Between police and court level,prosecution is the intermediate stage,in which a large number of cases are dropped by the police so the prosecution level has to deal with more severe cases.Apart from this,if the police is obliged to hand over all the offenses to the prosecution service,then the criminal justice system has to allow the prosecution level to decide which cases will go to court.Also,the year in which a person is recorded as a suspect by the police and the the one that this person is recorded as prosecuted might not be the same.Finally,the collected data it is not possible to follow individual cases through the system.All data collected in EU Statistics on crime and criminal justice,are aggregated and no individual cases that can be tracked at any stage of the criminal justice system.

Furthermore,there is court statistics that provide important information on the number of legal cases processed in first instant courts,accordingly with the methodological guide of crime and criminal justice Statistics.When dealing with cases,the performance of the courts can be distinguished by cases brought to court,which is the number of proceedings newly initiated in court during a year ,cases resolved,which is the number of proceedings finalised or disposed of by a court decision during a year,and pending cases,which is the number of proceedings that are not finalized or disposed of as of 31 December of a given year.While cases resolved and cases pending are mutually exclusive and total the workload of courts,cases brought to court can only be a fraction of the sum of cases resolved and cases pending.Comparisons of levels in single year might not deliver reliable results,due to the fact that the time a case takes to pass through the court system can vary depending on the jurisdiction and the charges.Courts also generate data on

1 http://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/-/KS-GQ-17-010?inheritRedirect=true

persons convicted or acquitted of criminal offense.These data are suitable for analyzing the adequacy of the criminal justice system as a continuation of attrition of cases even though the time it takes for a person to pass through the court system,can vary greatly depending on the country and the charges-even if data are only recorded in first instance courts.This variation in length to come to a court decision has to be taken into account.At conclusion,court statistics are considered to be a more robust indicator of criminal justice especially if conviction figures can be broken down by different offense categories,than any administrative data on the previous levels of the criminal justice system.Conviction statistics are seen as the better indicator for levels of crime than police data,as convictions stand at the end of the decision chain in the criminal justice system.

Prison statistics also,provide important information on the number of prisoners and prison capacity.Prison data are stock data that give the number of prisoners incarcerated at a certain day of the year(usually 31st of December).Depending on the jurisdictions,some use differing reference dates or report an average prison population. Additionally amnesties can have a sudden effect and change counts of prison held.Data reported to EUROSTAT should exclude non-criminal prisoners,for example people held into their immigration status or foreign citizens without a legal right to stay.Some jurisdictions exclude persons in psychiatric facilities or in institutions for disciplinary detention for young offenders ,while others include persons in supervised probationary freedom.Apart from these,prison data consider to be a robust indicator

Finally,there has been an inclusion of data in EU Statistics on crime and criminal justice,on personnel according to the International Standard Classification of Occupations 2008(ISCO-08). By that,most jurisdictions exclude civilian stuff,customs officers,tax,military,court and secret service police,while others include other law enforcement personnel or not able to exclude support staff. EU Statistics on crime and criminal justice also include historical data on the police recorded offenses,police officers and prison population collected up to 2007.These data are based on different definitions and cannot be directly compared with the data published at a national level from 2008.

2.4 Data collection.

As stated so far,EUROSTAT and the other accurate sources that could possibly provide the data for this thesis,are been contributed mainly from administrative sources namely, the Police,Prosecution Courts and Prison statistics of each country member.Data from those organizations are annually listed by the type of crime and only for the main district of each country. During the research,there have also been clear severe breaks in the series in general,and also in the specific types of crimes we intended to describe and analyze.Our aim has been to obtain data,in order to gain the most extended time series sequence possible for the crimes we decided to focus on.

Considering all that,the decision that has been made ,was to proceed to the primarily and most direct source of that kind of observations.The statistical data that are been published by the Hellenic Police (www.astynomia.gr),regard criminality,car accidents,immigration, asylum etc.Additionally,annually reports are been published by the Coordination body of Drugs Prosecution(Σ.Ο.Δ.Ν) and EUROSTAT.

Therefore,administrative data have been collected (November 2017) concerning the district of all Greece for the years 1991 to 2016 and the main 14 regional territories that constitute the country,which officially are refered to the related general police directorate,for the same period of years.For the main district of Greece we have been provided with the 1991-2016 data,concerning specific crimes,namely homicides,burglaries,robberies,motor and vehicle thefts,rapes,fraud and crimes that are related to the law about drugs in general.Official data concerning the law about drugs crimes,are missing for the 1991-1999 period of time.Consistency checks have also been implemented through the data provided from the Ministry of Internal Affairs (astynomia.gr) and the database from Eurostat in order to compare the source's compatibility.Eurostat's database though, is quite more general and does not provide information for all the crimes and dates in demand.

Consequently,intending to direct the research into a more detailed field,regional data were examined out of the same source(December 2017),due to the fact that police records solely could made those data available.Nevertheless,there has been a scarsity of data from 1991 until 2007,concerning all the fourteen main regions that constitute the country.Obtainable data exist from 2008-2016 regarding the full length of regions and crimes.All the regions maintioned,are ordered by each territory's general police directorate.

Additionally,every region is classified according to the NUTS classification[1] (Nomenclature of territorial units for statistics),which is an hierarchical system for dividing up the economic territory of the EU for the purpose of the collection, development and harmonisation of European regional statistics, socio-economic analyses of the regions and framing of EU regional policies.The NUTS classification subdivides each Member State into regions at three different levels, covering NUTS 1, 2 and 3 from larger to smaller areas.As national data give an insight to the general picture of what is happening in each member state of the EU,about all aspects of human communities, regional data can give much more information about the reality of what is happening in a detailed level.Subnational level data,supply statistical information, aiding to analyze changing patterns and to describe the impact that policy decisions have in people's lives.The regional data in this thesis have been classified according to the NUTS 2[2],which is aligned with the corresponding general police directorates.

1 http://ec.europa.eu/eurostat/web/regions/background
2 http://ec.europa.eu/eurostat/documents/345175/7451602/nuts-map-EL.pdf

NUTS 2 regions in Greece, 2010 and 2013

Figure 2.1 Greece's regional segmentation according to NUTS 2 classification.

CHAPTER 3.

MODELS AND METHODS

3.1 Time Series.

Definition: A *time series* is a set of observations *Yt* ,each one been recorded at a specific time t.A *discrete- time time series* is one in which the observations refer to predetermined time intervals,eg.sallaries,whereas *continuous-time time series* are obtained when the observations are recorded continuously over some time interval,eg.temperature.

Time series analysis is usually progressed in two stages.The first is mainly descriptive, aiming to gain some information about the behavior of the series over time.The second is about the formation of time series models,in the efford to reproduct the stohastic process that possibly created the set of the observations.

The purpose of studying time series has always been in the use of making predictions.A basic feature of time series is the correlation between their successive values.Non causal time series models predictions are based on the previous values of the same series we intend to forecast.These time series models are distinguished in deterministic models and stohastic models,by the way the random factor affects their structure.In the latter,the randomness is what forms the structure of the time series.

3.2 Stohastic process

Stohastic models are based on the idea that each time series, is formed through a data producing mechanism,which is called a stohastic process.A *stohastic process* $\{Yt\}t \in N$ is a collection of random variables each one with a separate probability distribution, forming a family of joint distributed random variables combined altogether.

Analyze time series has to do with the production of a model which possesses similar qualites, alligned with the probability theory,as to the mechanism which produces the relevant stohastic process.

Assuming a time series $Y_1,Y_2,...,Y_N$,where each observation Yt of them is the result of the realization of the specific random variable Yt,from the set of the random variables $\{Y_1,Y_2,....,Y_N\}$ that form the stohastic process {Yt} $t \in N$,one method to discribe the stohastic process is to define the joint probability density function $f(Y1, Y2, ... YN)$ of the sequence of the variables $\{Y_1,Y_2,...Y_N\}$.In practice it is rather difficult to identify those multivariate distributions due to the large number of the parameters they own.It is quite practical to describe some properties of the joint distribution,such as first and second order moments.

A simple example of a stohastic process is given below:

Random Walk

Let a sequence Yt of independently and identically distributed random variables with zero mean and constant variance,e.g the *iid stohastic process*.An example of an iid process is the time series ε_t ,t = 1,2,3,..which is produced by consecutive drops of a coin (+1:head ,-1:tales)

$$\varepsilon t = \begin{cases} +1 & \text{with probability } 1/2 \\ -1 & \text{with probability } 1/2 \end{cases}$$

Adding cumulatively all the random variables of an iid process,is produced a time series known as *random walk process.*As a result,a random walk with zero mean and $Y_0 = 0$ is defined as follows

$$Y_t = \varepsilon_1 + \varepsilon_2 + ... + \varepsilon_t \quad \text{or} \quad Yt = Y_{t-1} + \varepsilon_t \quad , t=1,2,3,....\ ,$$

where ε_t is iid noise.(eg. random values)

This process assumes that Y_t indicates the pedestrian's position at time t and that it is relative to the previous steps that he has made.(A Random Walk is the route of a pedestrian,whereas he has to take a step on the right if the toss is heads or a step on the left, if it is tales,for the succesive tosses of the coin.)

3.3 Stationarity.

Time series are characterized from some statistical properties such as stationarity, autocovariance, autocorellation, partial autocorellation that are been explained in this paragraph.

A fundamental attribute concerning time series analysis is stationarity.We distinguish them in stationary and no stationary.When the properties of a time series seem to change overtime,then the stohastic process is no stationary, making difficult to adjust a model which interpret the series.On the other hand,if the stohastic process stays in balance close to a constant mean level,then we are able to analyze the procedure through a model with constant coefficients that can be processed based on historical data and use them to make predictions.We have the following definitions:

Let {Yt} be a time series with $E(Y^2_t)<\infty$.

The mean function of {Yt} is $\mu_Y(t) = E(Y_t)$.

The covariance function of {Yt} is $\gamma_Y(r,s) = Cov(Yr,Ys) = E[(Yr - \mu_Y(r))(Ys - \mu_Y(s))]$

for all integers r and s.

Definition 1: {Yt} is **(weakly) stationary** if

(i) $\mu_Y(t)$ is independent of t,

and

(ii) $\gamma_Y(t+h,t)$ is independent of t for each h.

Definition 2: **Strict stationarity** of a time series {Yt,t= 0,±1,...} is defined by the condition that $(Y_1,...,Y_n)$ and $(Y_{1+h},...,Y_{n+h})$ have the same joint distributions for all integers h and n>0.

In time series,we have one variable statistically processed with its own lag values,therefore whenever we use the term covariance function with reference to a stationary time series {Yt} we shall mean the function γ_Y of one variable, defined by $\gamma_Y(h) := \gamma_Y(h,0) = \gamma_Y(t+h,t)$. The function $\gamma_Y(\cdot)$ will be referred to as the autocovariance function and $\gamma_Y(h)$ as its value at lag h. As a result we conclude with the following definition:

Definition 3: Let {Yt} be a stationary time series. The **autocovariance function (ACVF)** of {Yt} at lag h is $\gamma_Y(h) = \text{Cov}(Y_{t+h}, Y_t)$.

The **autocorrelation function (ACF)** of {Yt} at lag h is $\rho_Y(h) \equiv \frac{\gamma_Y(h)}{\gamma_Y(0)} = \text{Cor}(Y_{t+h}, Y_t)$.

Definition 4: The **partial autocorrelation function (PACF)** of {Yt} is correlation coefficient between Y_t and Y_{t+k},when all the correlations of the in between values $Y_{t+1}, Y_{t+2}, ..., Y_{t+k-1}$ have been already taken into account.

The k-th φ_{kk} is the coefficient of the term Y_{t-k} in the regression:

$Y_t = \varphi_{k1} Y_{t-1} + \varphi_{k2} Y_{t-2} + ... + \varphi_{kk} Y_{t-k} + \varepsilon_t$.

The structure,both of ACF and PACF are extremelly usefull due to the fact that they can define the form of the stohastic process that created a specific time series.

Since in practical problems we start with a set of observed data $\{y_1, y_2, ... y_n\}$ and not a model, to assess the degree of dependence in the data and to select a model for the data that reflects this, one of the important tools we use is the sample autocorrelation function (ACF) of the data. If we believe that the data are realized values of a stationary time series {Yt}, then the sample ACF will provide us with an estimate of the ACF of{Yt}. This estimate may suggest which of the many possible stationary time series models is a suitable candidate for representing the dependence in the data.Following there are the analogous sample definitions.

Let $y_1, ..., y_n$ be observations of a time series.

The sample mean of $y_1, ..., y_n$ is $\bar{y} = \frac{1}{n} \sum_{t=1}^{n} y_t$.

The sample autocovariance function is $\hat{\gamma} := \frac{1}{n} \sum_{t=1}^{n-|h|} (y_{t+|h|} - \bar{y})(y_t - \bar{y}), -n < h < n$.

The sample autocorrelation function is $\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}, -n < h < n$.

The sample autocovariance and autocorrelation functions can be computed for any data set $\{y_1, ..., y_n\}$ and are not restricted to observations from a stationary time series. For data containing a trend for instance, $|\hat{\rho}(h)|$ will exhibit slow decay as h increases.

(see Peter J. Brockwell Richard A. Davis Introduction to Time Series and Forecasting, Second Edition).

3.4 The Box Jenkins Methodology.

3.4.1 Trend Identification.

Even though stationarity signifies a major concept in time series study and forecasting,it is a quite rare event in real data.Many of the time series being developed through time,often include mostly a trend, seasoning or/and cyclic component,features that make them non stationary.Most of the times,series that present trend are non stationary.

Therefore,the first step is to plot the series and inspect whether there is a trend,a seasonal component,any apparent sharp changes in behavior,or any outlying observations.Demonstrate data by graphical means can communicate the information in a clear and efficient way.Comparisons can also be made or determining causality and in time series,a single variable can be captured over a period of time and a line chart can denote the trend.

Following on,the goal is to remove the trend or/and the seasonal components to get stationary series.Sometimes,we need to transform the original data,so as to get to stationary series.For instance, if the magnitude of the fluctuations appears to grow roughly linearly with the level of the series, then the transformed series$\{\ln Y_1,...,\ln Y_n\}$will have fluctuations of more constant magnitude.There are several methods to remove trend and seasonal features,some of them could be trend estimation by exponential smoothing for instance,and another could be trend elimination by differencing,a method that developed extensively by Box and Jenkins(1976).

The classical decomposition model is,

$$Y_t = m_t + s_t + \varepsilon_t ,$$

where,

m_t is a slowly changing function known as a trend component,

s_t is a function with known period d referred to as a seasonal component,and

ε_t is a random noise component that is stationary.

In this thesis,the series are annually data of recorded crimes and for that reason we do not consider any seasonal component.Hence, in the absence of a seasonal component the model becomes the following

Nonseasonal Model with Trend:

$$Y_t = m_t + \varepsilon_t \text{ , } t = 1,...,n, \quad \text{where } E\varepsilon_t = 0.$$

3.4.2 Trend Elimination by Differencing.

This method,refers to the procedure of elliminating trend by differencing,and then finding the right stationary model for the defferenced series.It also has the advantage of fewer parameters to estimate.If $\{Y_t\}$ is a time series,then $Z_t = \Delta Y_t = Y_t - Y_{t-1}$,is defined as the lag-1 difference operator.In case where the new series $\{Z_t\}$ is not being able to modeled as stationary,the same procedure by differencing $W_t = \Delta Z_t = Z_t - Z_{t-1}$ is been applied repeatedly.

If the operator Δ is applied for example, to a linear trend function $m_t = c_0 + c_{1t}$, then we obtain the constant function $\Delta m_t = m_t - m_{t-1} = c_0 + c_{1t} - (c_0 + c_1(t-1)) = c_1$. In the same way any polynomial trend of degree k can be reduced to a constant by application of the operator k-times. In practice,the order k of differencing isoften times quite small,one or two.

3.4.3 Methods Achieving Stationarity.

Following the data transformation that provides series without trend,the intention lies in to modeling the estimated noise sequence eg.the residuals produced from the differencing method. The goal is to test the hypothesis whether there is no dependence among them,so we can see them as indepented and identically distributed random variables or there isdependence among the residuals,therefore the past observations of the noise sequence help to predict future values, implementing the appropriate methods for stationary series.Hence,stationary identification can be achieved either descriptively or by the application of various statistical tests.

3.4.3.1 The Autocorrelation Function (ACF).

The sample autocorrelations of an iid sequence $Y_1,...,Y_n$ with finite variance are approximately iid with distribution N(0,1/n),when n is large.Therefore, if $y_1,...,y_n$ is a realization of such an iid sequence, about 95% of the sample autocorrelations should fall between the bounds $\pm 1.96/\sqrt{n}$. If we compute the sample autocorrelations up to lag 40 and find that more than two or three values fall outside the bounds, or that one value falls far outside the bounds, we therefore reject the iid hypothesis.

3.4.3.2 Unit Root Tests.

Since the autocorellation coefficient's statistical significance check is more descriptive and quite subjective,better and more acurate stationarity criteria have been developed in bibliography,including unit root checks by Dickey and Fuller (1981).The unit root problem in time series arises when either the autoregressive or moving average polynomial of an ARMA model has a root on or near the unit circle. A unit root in either of these polynomials has important implications for modeling.Unit tests refer to stationarity control in non stationary time series which turn into stationary after taken some differences.These statistical tests, define the number of the differencing needed in order to get a stationary series,e.g the integration degree d of the I(d) integrated time series.Apart from this,unit root tests can solve the problem of autocorrelations between the residuals.

- **Augmented Dickey-Fuller Test.**

In case where a time series has an autoregressive model of grater order than one,then the use of an AR(1) model for unit root testing will result the autocorrelation of the residuals.This is in contradiction with the Dickey-Fuller tests,where the residuals are white noise(random component).

Let the real model is AR(p) : $Y_t = \delta + a_1 Y_{t-1} + a_2 Y_{t-2} + \ldots.. + a_p Y_{t-p} + \varepsilon_t$

and we adjust an AR(1) :$Y_t = \delta + a_1 Y_{t-1} + \varepsilon_t$,$\varepsilon_t = a_2 Y_{t-2} + \ldots. + a_p Y_{t-p} + \varepsilon_t$,

then the missing lags of the Y_t are incorporated in the residuals ε_t,where they become autocorrelated.

To overcome this issue,a higher order AR model is assumed ,where apart from the Yt-1 term,the lags ΔY_{t-1} and ΔY_{t-2} of the ΔY_t variable are also included.In that way the residuals autocorrelation is been corrected.

$\Delta Y_t = \delta + (a_1-1)Y_{t-1} + a_2 Y_{t-2} + a_3 Y_{t-3} + \varepsilon_t$ or

$\Delta Y_t = \delta + \beta Y_{t-1} + \delta_1 \Delta Y_{t-1} + \delta_2 \Delta Y_{t-2} + \varepsilon_t$, where $\beta=(a_1+a_2+a_3)-1$, $\delta_1=-(\alpha_2+\alpha_3)$, $\delta_2=-a_3$

$\Delta Y_{t-j} = Y_{t-j} - Y_{t-j-1}$,j=1,2.

In general,the regression equation becomes $\Delta Y_t = \delta_0 + \gamma_t + \beta Y_{t-1} + \sum_{j=1}^{p-1} \delta_j \Delta Y_{t-j} + \varepsilon_t$,with constant and trend.

Then,the augmented Dickey-Fuller test is : H_0: $\beta=0$ vs Ha : $\beta<0$.

The test statistic is $t_{\hat{\beta}} = \frac{\hat{\beta}}{s_{\hat{\beta}}}$,with the appropriate DF critical values.

Rejecting the null hypothesis,means that the series Y_t is stationary process.

3.5 ARMA Models.

As stated above,our aim is to have stationary transformed series,that resemble white noise,which is a stationary series with random values and the lack of dependence among them.In practice though,time series appear to have strong dependence among their values.Consequently,the development of models that are based on the idea of strong correlation among the series,can explain efficiently concepts that take place in real economy and other scientific fields.

The general form of an ARMA model is:

$$Y_t = \delta + a_1 Y_{t-1} + a_2 Y_{t-2} + ... + a_p Y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - .... - \theta_q \varepsilon_{t-q}.$$

where δ, is the constant

$a_1, a_2, \ldots a_p$, are the autoregressive coefficients (AR)

$\theta_1, \theta_2, \ldots, \theta_q$, are the moving average coefficients (MA)

ε_t , constitute white noise.

They are the *autoregressive moving average models of order (p,q)* ,represented as ARMA(p,q). They are regression models with depended variable the series Y_t and explanatory variables the lag-p previous values of the same series and a number of(lag-q) previous errors of the ranom variable ε_t .

3.6 ARIMA Models.

When time series are expressed in absolute levels such as GPD(gross domestic product), employnment level and other financial measures,they are not stationary.Such series are usually display trend or seasonal flunctuations,and their mean changes level over time,but still behave uniformly in specific time intervals.This behavior is common in series that gain stationarity after some differencing of specific order ,as mentioned above, by the Box-Jenkins methodology.

An ARMA (p,q) model with d order differencing, forms a new type of models,the *Autoregressive Integrated Moving Averages of order (p,d,q),* the so called ARIMA (p,d,q).

A form of the general ARIMA (p,d,q) is:

$$Y_t = \delta + \varphi_1 Y_t{}_{-1} + \ldots + \varphi_{p+d} Y_t{}_{-p-d} + \varepsilon_t - \theta_1 \varepsilon_t{}_{-1} - \ldots - \theta_q \varepsilon_t{}_{-q}.$$

which depicts an ARMA (p+d,q) model,with (p+d) previous values of Yt, ε_t random error of current time,q number of previous errors.

On the contrary to the general ARMA model,this model is not stationary.ARIMA model has three components,the AR autoregressive term, I differencing term and MA moving average term.

- The *AR* term is about the past values that are used for forecasting the next value.The PACF plot is used to determine the AR component (parameter p of the model).
- The *MA* term defines the past errors that are used to predict future the values.The ACF plot is used to determine the MA part of the model (parameter q).

- The *I* term refers to the order of differencing of the series to make it stationary (parameter p).The ADF test (Augmented Dickey-Fuller test) is used to determine whether the series is stationary or not.

3.6.1 The auto.arima function.

When dealing with a very large number of variables,although ARIMA is a powerfull model for forecasting time series data,it can be extremelly time consuming.Hence,the auto.arima function implemented in the software R,makes the series stationary,creates the ACF anf PACF plots and determines the values of p and q by using these plots.

Finally,auto.arima takes into account the AIC,BIC and AICC values generated to determine the best combination of parameters. AIC (Akaike Information Criterion),BIC (Bayesian Information Criterion) and AICC values are estimators to compare models. The lower these values, the better is the model.We prefer the AICC criterion as it has a more extreme penalty for large-order models, which counteracts the overfitting tendency of the AIC. (see chapter 5, paragraph5.5 Order Selection Peter J. Brockwell Richard A. Davis Introduction to Time Series and Forecasting, Second Edition).

3.7 Estimation of the model.

Continuing with the Box-Jenkins methodology,since the desirable stationarity has been achieved with the necessary differences,following is to estimate the autoregressive parameters $a_1, a_2, \dots, a_p$ and the $\theta_1, \theta_2, \dots, \theta_q$ moving averages of the ARIMA(p,d,q) which have been identified in the previous step.The examination of ACF and PACF plots help to get an idea of the order of p and q of ARMA(p,q) that adjusts better in the differencied data.

There are numerous methods to estimate an ARIMA(p,d,q) model,and since it is equivalent to an ARMA(p,q) in the d -differences of Yt,it is efficient to analyze the stationary ARMA model parameters,assuming that Yt is a stationary series.The maximum likelihood method (ML) is considered to be a better method,since it provides estimators that can operate asymptotically.

Proceeding,the problem is to find the suitable values of p and q, avoiding the higher values for them because trying to estimate a large number of parameters, introduces estimation errors that

affect in a negative manner the use of the fitted model for prediction.Hence,the selection of p and q ,occurs using several criteria.Each of these criteria includes a penalty term to discourage the fitting of too many parameters. We base our choice of p and q on the minimization of the AICC statistic as it is considered to be a bias-corrected version of the AIC (Akaike criterion).

The **AIC Akaike statistic** is: $AIC(\beta) := -2\ln L_X(\beta, \frac{S_X(\beta)}{n}) + 2(p+q+1)$

The definition of **AICC criterion** is:

$$AICC(\varphi, \theta) = -2\ln L(\varphi, \theta, S\frac{(\varphi, \theta)}{n}) + 2(p+q+1)\frac{n}{(n-p-1-2)},$$

where $L(\varphi, \theta, \sigma^2)$ is the likelihood of the data under the Gaussian ARMA model with parameters φ,θ,σ^2 and S(φ,θ) is the residual sum of squares.

For any fixed values of p and q, the maximum likelihood estimates of φ and θ are the values that minimize the AICC. The minimum AICC model can be found by computing the maximum likelihood estimators for each fixed p and q and choosing from these the maximum likelihood model with the smallest value of AICC.

When the smallest AICC value provides the model,then it is required to check the goodness of fit of the model,by checking whether the residuals resemble white noise.

3.8 Multivariaye Analysis.

Quite often,several phenomena apear to have some complexibility in order to understand and analyze.Speaking for the crime pattern of a country for instance,it may be unsuficient for a single variable analyzed seperately,to describe it in all.The interaction of the response variables has a significant role in order to gain insight in that direction.Multivariate analysis uses simultaneously multiple response variables,by that meaning that a multivariate time series has more than one time-dependent variable.

Each variable depends not only on its past values but also has some dependency on other variables. This dependency is used for forecasting future values.

A multivariate approach is preferred to univariate methods as they form a powerfull technique since they reserve statistical power and they save time.Additionally,taking multiple variables into account simultaneously may reveal patterns that would not be detectable by univariate methods.

3.8.1 The VAR Model.(Vector Autoregressive Models).

Vector autoregression (VAR) is a stohastic process used to capture the linear dependencies among multiple time series and has been introduced by Sims in 1980.

A VAR model describes the evolution of a set of K endogenous variables $y_t=(y_{1t},...y_{kt},...,y_{Kt})$,for k = 1,2,...,K .

A p - order VAR model ,VAR(p) is :

$y_t=c+A_1 y_t-_1+A_2 y_t-_2+.....+Ap\, y_t-_p+\varepsilon_t$,where c is $k\times 1$ verctor of intercepts,

A_i is a $k\times k$ matrix and ε_t is a $k\times 1$ vector of error terms,that satisfy :

$E(\varepsilon_t)=0$, $E(\varepsilon_t \varepsilon'_t)=\Omega$, $E(\varepsilon_t \varepsilon'_{(t-k)})=0$ e.g every error term has zero mean,

and time invariant positive definite covariance matrix.

VAR models generalize the univariate AR model, by allowing for more than one evolving variable.In VAR models every variable is expressed with an equation which explains its changing pattern through its own lagged values,the other model variables lagged values and the error term.

As already been said,VAR is vector autoregression, that is, an autoregression on a vector-valued time series. A *K*-variate VAR(*p*) model consists of *K* time series. In other words, it is a generalization of an autoregressive (AR) model to a multivariate setting. However, there are no moving average (MA) components there.

Estimating a VAR model means finding the coefficient values. A K-variate VAR(p) model has a K-long vector of intercepts and p square matrices of dimension $K \times K$ of autoregressive coefficients. All of these need to be estimated.

VAR models can be estimated applying ordinary least squares (OLS) to each of the K model equations separately. For an unrestricted model, this is as good as feasible generalized least squares (feasible GLS, or FGLS). If the VAR model has some coefficient restrictions, FGLS will be asymptotically more efficient than equation-by-equation OLS. There are other estimation alternatives such as maximum likelihood estimation which may be preferable if the model errors are non-normally distributed.

One important characteristic of a VAR(p) process is its stability. This means that it generates stationary time series with time invariant means, variances and covariance structure, given sufficient starting values.

3.8.2 VAR function in R.

In practice,with the implementation of "MTS" package and the function VAR in R software environment,the analysis and estimation of the variables is achieved,performing least squares estimation of a VAR model.

For the multivariate linear time series analysis, the package performs model specification, estimation, model checking, and prediction for many widely used models, including vector AR models, vector MA models, vector ARMA models, seasonal vector ARMA models, VAR models with exogenous variables, multivariate regression models with time series errors, augmented VAR models, and error correction VAR models for co-integrated time series.

(Package 'MTS',All-Purpose Toolkit for Analyzing Multivariate Time Series (MTS) and Estimating Multivariate Volatility Models, Author:Ruey S. Tsay and David Wood).

3.9. Count Data.

Oftentimes, there are time series where the response variable of interest is discrete, not continuous as it is supposed to be when implementing ARIMA theory.When the response variable is the counted number of occurrences of an event,it is not feasible to produce normally distributed errors.The distribution of counts is discrete, not continuous, and is limited to non-negative values.There are two problems with applying an ordinary linear regression model to these data. First, many distributions of count data are positively skewed with many observations in the data set having a value of 0.

The high number of 0's in the data set prevents the transformation of a skewed distribution into a normal one. Second, it is quite likely that the regression model will produce negative predicted values, which are theoretically impossible.Examples such as daily addmissions in a bank, the prediction of the number of times a person perpetrated domestic violence against his or her partner in the last year,and others in finance and industrial business for detecting defected products,all form count data.Therefore,it has arised a great interest in regression models for time series of counts.

Models for count time series should take into account that the observations are nonnegative integers and they should capture suitably the dependence among observations as mentioned also previously. A convenient and flexible approach is to employ the generalized linear model (GLM) methodology (Nelder and Wedderburn 1972) for modeling the observations conditionally on the past information.

R package "tscount".

Accordingly,there has been bublished a study from Tobias Liboschik(TU Dortmund University), Konstantinos Fokianos (University of Cyprus) and Roland Fried (TU Dortmund University), "tscount: An R Package for Analysis of Count Time Series Following Generalized Linear Models " in the Journal Statistical Software (November 2017 Volume 82, Issue 5.),in order to give a further insight to the subject and provide some useful methods for statistical software package R application.

In the study it is stated that,the package "tscount" provides likelihood-based estimation methods for analysis and modeling of count time series following generalized linear models (GLMs).

The conditional mean of the process is linked to its past values, to past observations and to potential covariate effects. The package provides allowance for models with the identity and with the logarithmic link function. The conditional distribution can be Poisson or negative binomial. An important special case of this class is the so-called INGARCH model and its log-linear extension.

A more general dependence structure can be implemented by the package.General time series models whose conditional mean may depend on time-varying covariates, previous observations and similar to the conditional variance of a GARCH model, on its own previous values,are been considered.It is specified that the usage and output of the functions are in parts inspired by the R functions arima and glm in order to provide a familiar experience to the user.

Let, a count time series $\{Y_t, t \in N\}$ and $\{X_t, t \in N\}$ a time varying r-dimensional covariate vector, $X_t = (X_{t,1}, \ldots, X_{t,r})^T$.

The main goal is model the conditional mean $E(Y_t | F_{t-1})$ of the count time series by a process $\{\lambda_t : t \in N\}$, such that $E(Y_t | F_{t-1}) = \lambda_t$.

F_t indicates the history of the joint process $\{Y_t, \lambda_t, X_{t+1} : t \in N\}$ up to time t including the covariate information at time t +1.

The general form of the model is :

(1) $g(\lambda_t) = \beta_0 + \sum_{k=1}^{p} \beta_k g^{\sim}(Y_{t-ik}¿) + \sum_{l=1}^{q} a_l g(\lambda_{t-jl}¿) + \eta^T X_t$, where g: $R^+ \rightarrow R$ is a link function,

$g^{\sim} : N_0 \rightarrow R$ is a transformation function and $\eta = (\eta_1, \ldots, \eta_T)^T$,applies to the covariate effects.In GLM we call $v_t = g(\lambda_t)$ the linear predictor.

In practice,the central function for fitting a GLM for count time series is “tsglm”.The “tscount” package fits models of the form (1) by quasi conditional maximum likelihood (ML) estimation . With that,a comparison of the fit of a Poisson with that of a negative binomial conditional distribution is been is put into effect.The resulting fitted models have class ‘tsglm’, for which a number of methods is provided, including summary for a detailed model summary and plot for diagnostic plots.

Concerning the model evaluation,there are several types of response residuals,and the empirical autocorrelation function of these residuals is useful for diagnosing serial dependence which has not been explained by the fitted model. So,a plot of the residuals against time can reveal changes of the data generating process over time.Also a plot of squared residuals r_t^2 against the corresponding fitted values $\hat{\lambda}_t$ exhibits the relation of mean and variance and might point to the Poisson distribution if the points scatter around the identity function or to the negative binomial distribution if there exists a quadratic relation.

At a subsequent time, for assessing the probabilistic calibration of the predictive distribution,a method tool is the probability integral transform (PIT),which follow a uniform distribution if the predictive distribution is the proper one.(see Gneiting et al.2007).

An other evaluation step,concerns marginal calibration,defined as the difference of the average predictive c.d.f. and the empirical c.d.f. of the observations.A plot of the marginal calibration for values y in the range of the original observations is implemented,applying the function “marcal” in the R code(Christou and Fokianos 2015b) . If the predictions from a model are appropriate the marginal distribution of the predictions resembles the marginal distribution of the observations and the difference of the average predictive c.d.f. and the empirical c.d.f. of the observations should be close to zero. Major deviations from zero point to model deficiencies.

In the study it is clearly stated that,Gneiting et al. (2007) have shown that the calibration assessed by a PIT histogram or a marginal calibration plot is a necessary but not sufficient condition for a perfect forecast.A prefered model is the one with the maximal sharpness among all sufficiently calibrated models. Sharpness is the concentration of the predictive distribution and can be measured by the width of prediction intervals. A simultaneous assessment of calibration and sharpness summarized in a single numerical score can be accomplished by proper scoring rules (Gneiting et al. 2007).There are a variety of scoring rules,put into effect with the function "scoring" in the software R code,where every different scoring rule captures different characteristics of the predictive distribution and its distance to the observed data. Except for the normalized error score, the model with the lowest score is preferable. The mean squared error score is the only one which does not depend on the distribution and is also known as mean squared prediction error. The mean normalized squared error score measures the variance of the Pearson residuals and is close to one if the model is adequate. The Dawid-Sebastini score is a variant of this with an extra term to penalize overestimation of the standard deviation.There are also other criteria for model selection,such as the AIC ,BIC,or QIC.The model with the smallest value,is the preferable.

(For probing theory,see [1]Tobias Liboschik(TU Dortmund University), Konstantinos Fokianos (University of Cyprus) and Roland Fried (TU Dortmund University)"tscount: An R Package for Analysis of Count Time Series Following Generalized Linear Models " in the Journal Statistical Software (November 2017) Volume 82, Issue 5.)

1 https://www.jstatsoft.org/article/view/v082i05

3.10 Missing Data.

Missing data, is a common issue that appears in most statistical research projects, concerning all the aspects of human activity,economics,government reports etc., and can have a huge effect on the validity of the conclusions that can be made.Missing values can occur either at random (MAR),where the lack of data is not at random,but an amount of the sample population avoids to answer a specific question,for instance,and the results may be biased.There are also missing data completely at random (MCAR),where the reasons that lead to any particular value being missing are independent both of observable variables and of unobservable parameters of interest, and occur entirely at random.And finally,there are the (MNAR) type of missing data that are not MAR,or MCAR and the reason that they are missing is related to the incomplete or missing variable.

The sample fails to be representative due to missing values and the conclusions for the population can be wrong or distorted.Therefore,there exist several methods from the bibliography to cope with missing data.

- Imputation : where estimated values are filled in the place of the missing data.
- Omission : where samples with invalid data are discarded from further analysis.
- Analysis : where methods are applied unaffected by the missing values.

3.10.1 Imputation.

Imputation is the process when replacing missing data with substituted values that are been estimated based on other available information.When one or more values are missing,any case that has a missing value is been discarded,because it may produce biased results or affect the representativeness of them.When all missing values have been imputed, the data set can then be analysed using standard techniques for complete data.

There exist many theories to account for missing data but the majority of them introduce large amounts of bias. Some of the most common efforts to deal with missing data are : hot deck and cold deck imputation, listwise and pairwise deletion, mean imputation ,regression imputation,last observation carried forward,stochastic imputation and multiple imputation.

Quite a few software based statistical packages are putting into effect imputation methods for substituting missing values on a data set.The R package “imputeTS” : “ Time Series Missing

Value Imputation”,performs replacement of missing values in univariate time series.It offers several imputation functions and missing data plots,as well as some imputation algorithms,namely Kalman Smoothing on ARIMA models among all other.The “na.kalman” function of the package,uses Kalman Smoothing on structural time series models (or on the state space representation of an arima model) for imputation.

3.10.2 Missing data estimation in time series.

It is often been observed severe discontinuances in time series data sets,so that makes difficult to the analysis and future estimated values prediction.Due to a variety of reasons,like short communication between agencies and/or government or private offices,or because of incomplete responting to questionaires, it is not uncommon values of some cases to be missing.

With noninformative reasons of why data are missing,there are many methods to apply in order to deal with the obstacle.It could rather be multiple imputation or fill-in the missing values,using the rest of the observations.

In this thesis,as it is mentioned in chapter 2,there are serious discontinuances in the series;

1. firstly in the main territory of Greece for the variable concerning the law about drugs(missing values for the period of years1991-1999);where it is been apllied the method of imputation.
2. secondly in the 14 regions that constitute the country of Greece.For the time series 1991-2016 for every response variable there are missing data for the years 1991 – 2007.

The missing values, were substituted using the knowledge of the rest of the observations,by assuming that every regional variable shares a specific rate of the total number (of crimes)in the district of Greece.

When the full time series(years 1991-2016) for all the response variables concerning the territory of Greece have been obtained, and for the regional data; the time series 2008-2016 is

disposable,it is feasible to calculate the mean of the percentages for each variable (crime) to the total (i.e number of crimes in whole Greece) ,applied to the years 2008-2016.

The final step to fill the missing regional data for the years 1991-2007 ,is to compute those mean values,multiplying them to the total number of a specific response variable(crime) each time and for each separate chronological date.

CHAPTER 4.

RESULTS.

4.1 Descriptives.

4.1.1 Graphic visualization.

We have obtained the raw data as mentioned in the previous paragraphs,specifically homicides, burglaries,robberies,rapes,frauds,motor and vehicle thefts and crimes concerning the law about drugs for the 1991-2016 period of time.For the law about drugs variable ,we have missing data from 1991 until 1999. **Diagrams 4.1 ,4.2 ,4.3** depict the original data,which have been created with the implementation of the statistical program R Studio (with R version 3.4.3 ,released in November 2017) and the r package 'plotly' using the appropriate commands in order to illustrate the form of the data in this time series and also the trend and possible correlations between variables.

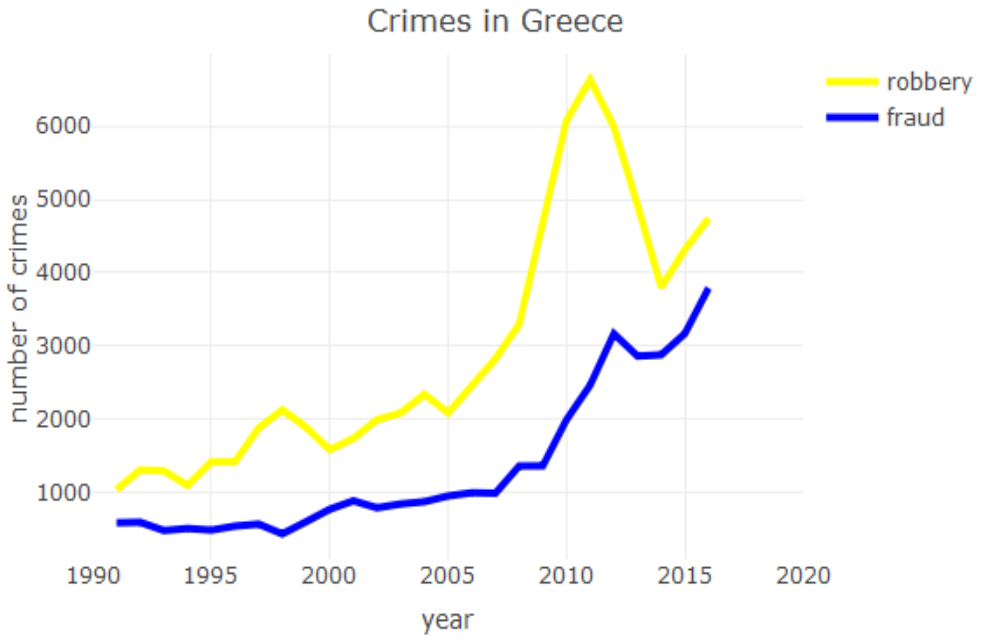

Diagram 4.1 crimes in Greece: robberies and frauds in the years 1991-2016 .

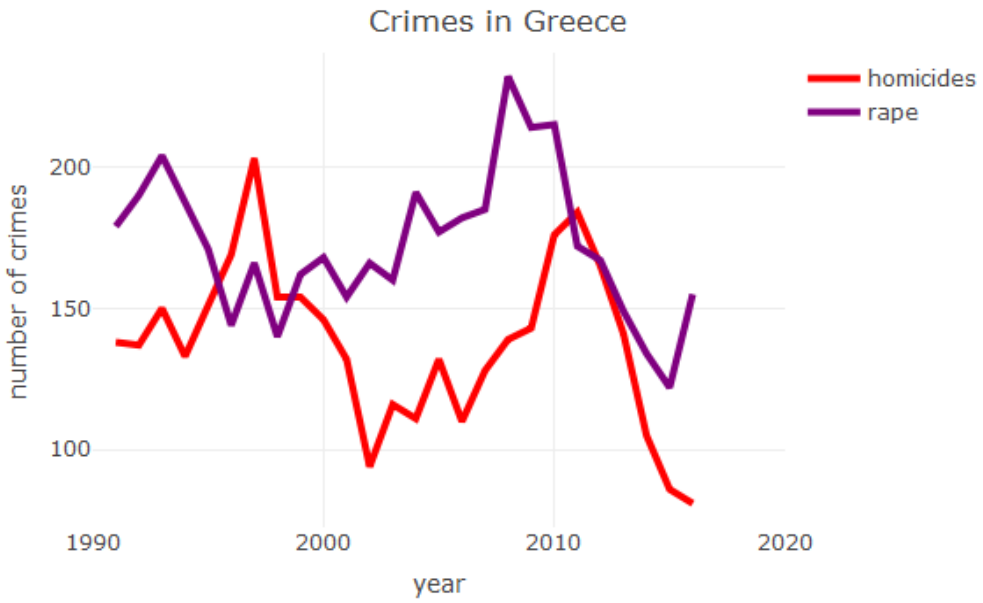

Diagram 4.2 crimes in Greece: homicides and rapes in the years 1991-2016.

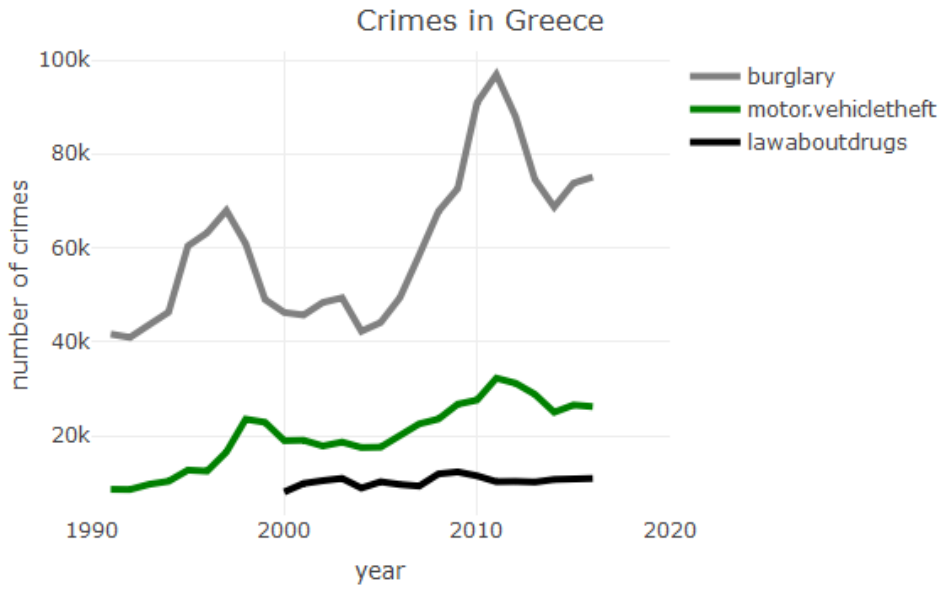

Diagram 4.3 crimes in Greece: burglaries, motor and vehicle thefts, and crimes applied to the law about drugs.

The diagram 4.1 depicts the time series for the period of time 1991-2016 of the response variables robbery and fraud in Greece.It is clearly demonstrated an upwards trend through the years in both graphs,and a tense growth after 2009.A pick concerning the two crimes is shown the year 2013 to follow a downward trend after that.The two plots,seem to be correlated.

In the next diagram 4.2,it is displayed the crime pattern in Greece of homicides and rapes,showing an oposite trend between the two graphs,the first years from 1991 until 1999,with rapes to decrease and homicides to increase.In 1995 it takes place a growth in both crimes,to continue with the same pattern over the years and show decrease from 2011 onwards.Both plots appear to have a trend,showing either times an upward or an downward behavior.

Burglaries also depict an upward trend shift,in diagram 4.3,presenting a changepoint in 2011,turning to a downwards trend.Thefts of vehicles in general also have a trend with an upward notion,but more close to a mean.As for the crimes that refered to the law about drugs,due to the fact that there are missing values and possibly, because very few crimes of such kind are been dislocated ,we gain an image constant to a mean.

Apparently,as stated above,the series own a trend,therefore they are not stationary,and for that reason,several techniques and methods have been applied in order to become stationary and possible to analyze.

4.2 Results.

4.2.1 Univariate Time Series Analysis.

A first atempt for analysing the time series 1991-2016 of specific crimes in Greece has been a univariate anlysis of each variable implemented with ARIMA theory for non stationary series with trend.

For each time series,namely homicides,robberies,rapes,burglaries,frauds,motor and vehicle thefts and crimes regarding the law about drugs for the years 1991-2016,there has been a trend identification from the plots with additional autocorrelation plots to the time series values,the log-values and the first differences of the log-values of the time series,targeting to make them

stationary.Following,ADF tests (Augmented Dickey-Fuller test) have been applied ,that examine the hypothesis of stationarity.

By the time stationarity has been achieved,a fit of the model is the next step with the auto.arima function.Finally,predictions for the next 30 time units have put into effect,applying the arima forecast function,joined with the correspondingly forecast plots and a summary of the predicted best model,evaluated by the smallest AICC criterion.

The corresponding commands for the code in statistical software R,presented with the example of one crime(e.g. fraud),would be :

```
# for the autocorrelation plots :
acf(ts(fraud))
acf(log(ts(fraud )))
acf(diff(log(ts(fraud ))))
#ADF test.(Augmented Dickey-Fuller test):
adf.test(ts(fraud), alternative="stationary", k=0)
adf.test(log(ts(fraud)), alternative="stationary", k=0)
adf.test(diff(log(ts((fraud)))), alternative="stationary", k=0)
# fit of the model:
arima_fit<-auto.arima(fraud)
arima_forecast<-forecast(arima_fit,h=30)
autoplot(arima_forecast)
summary(arima_forecast)
```

For the variables “homicides”,”burglary”,”robbery”,”motor.vehicletheft”,”rape”,and”fraud”, stationarity has been achieved taking the first differences on the univariate analysis of every variable.With the first differences stationarity on mean has been achieved and with the log-values of the series,on variance.Combined the first differences in the log-values of the data,we accoblished to make both mean and variance stationary.In the ACF plots there has been a slow reduction up to lag-1 after we took the first differences in the log-values of the series.

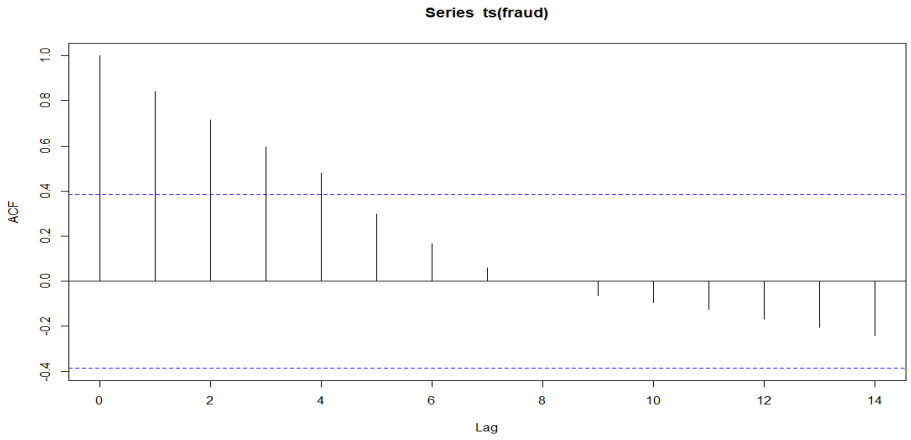

Diagram 4.4 ACF plot of the time series fraud.

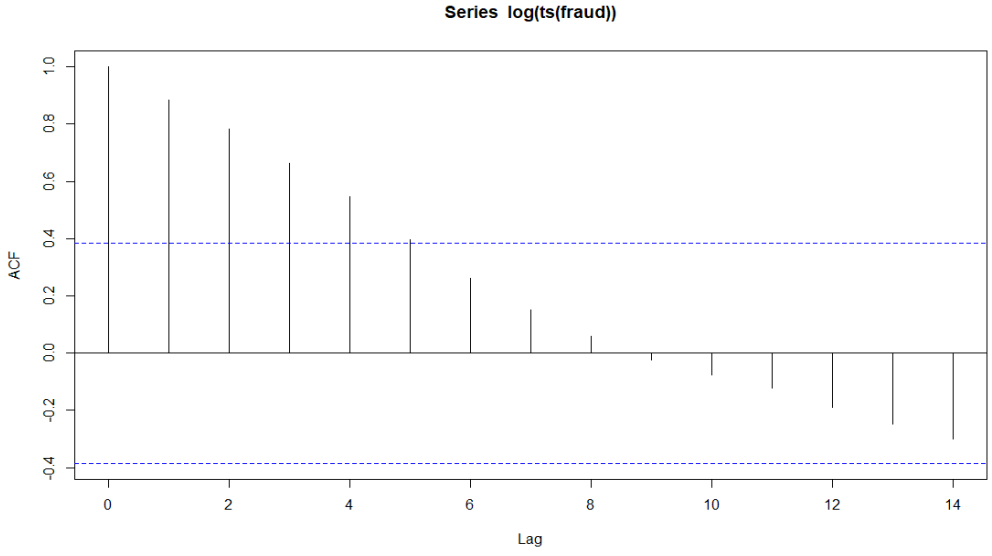

Diagram 4.5 ACF plot of log-values of time series fraud.

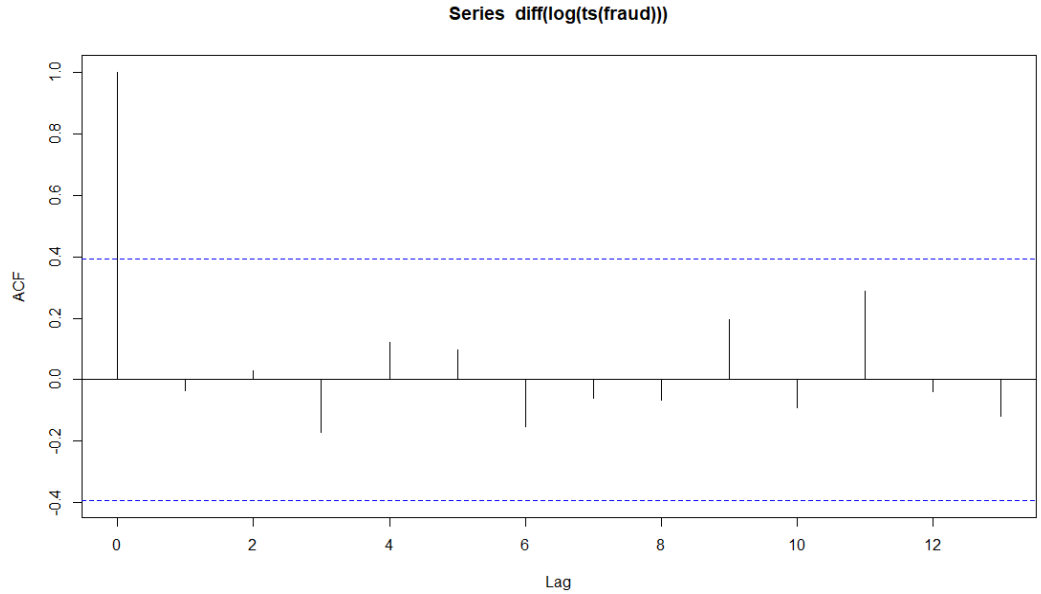

Diagram 4.6 ACF plot of the first differences of the log-values of time series fraud.

Diagrams 4.4 - 4.6 depict the ACF plots of the transformations of the time series 1991-2016 concerning fraud in Greece,and the steps to achieve stationarity.

For testing the stationarity hypothesis,ADF tests have been implemented.The **p-values** of the tests are given below,to each variable (crime) respectively .

crimes	ADF ts	ADF log (ts)	ADF diff (log (ts))
homicides	0.6722	0.7613	0.01
burglary	0.7535	0.7595	0.2333
robbery	0.7218	0.6458	0.04239
Motor.vehicletheft	0.6784	0.7386	0.02278
rape	0 .538	0.5201	0.01
fraud	0.96	0.4211	0.01

Table 1 : ADF test p-values for different transformation of the time series.

With the accomplisment of stationarity in the series,we have been able to fit a model to the series and make forecasts with 80% and 95% prediction intervals respectively, with the ARIMA theory.With the auto.arima function we have been provided the following models that describe the series and forecast its future values,based on the smallest AICC criterion value.

crimes	**ARIMA model**	**AICC**	**AIC**
homicides	ARIMA(1,0,0) with non-zero mean	238.72	237.63
burglary	ARIMA(0,1,1)	515.32	514.77
robbery	ARIMA(0,1,1)	384.99	384.45
Motor.vehicle theft	ARIMA(0,1,0)	465.19	465.01
rape	ARIMA(1,0,0) with non-zero mean	234.3	233.2
fraud	ARIMA(0,2,1)	338.65	338.08

Table 2 : Forecasts of the time series based on the smallest AICC.

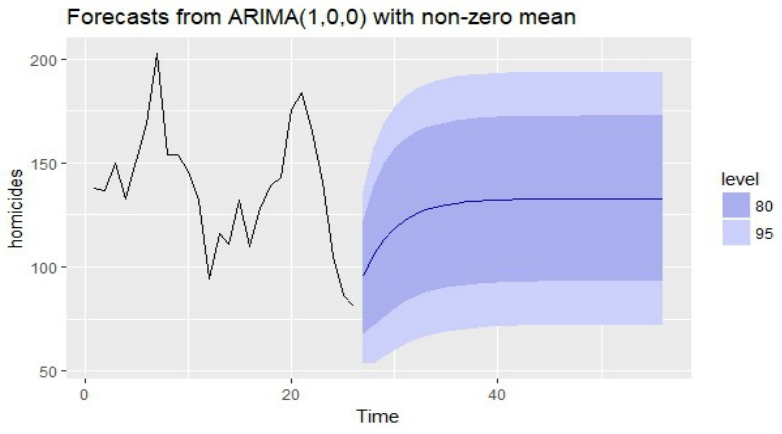

Diagram 4.7 Forecasts of homicides in Greece for the next 30 time units.

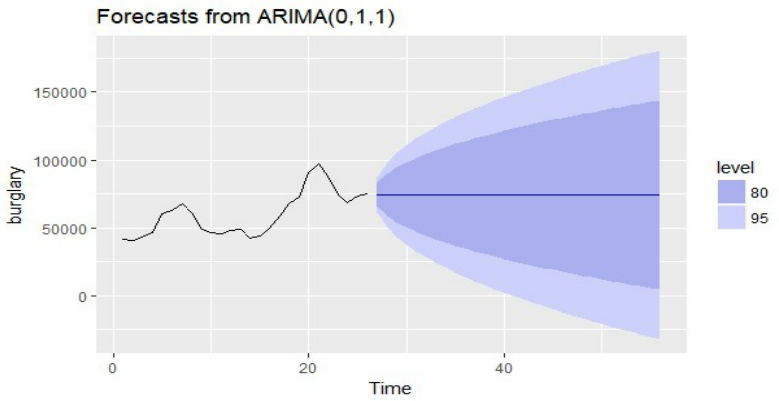

Diagram 4.8 Forecasts of burglaries in Greece for the next 30 time units.

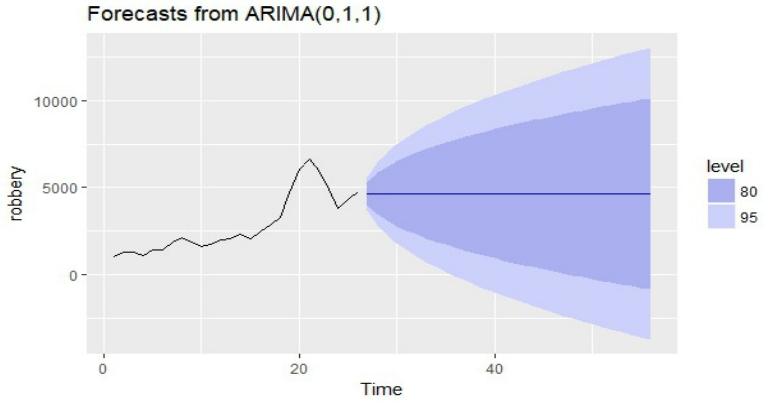

Diagram 4.9 Forecasts of robberies in Greece for the next 30 time units.

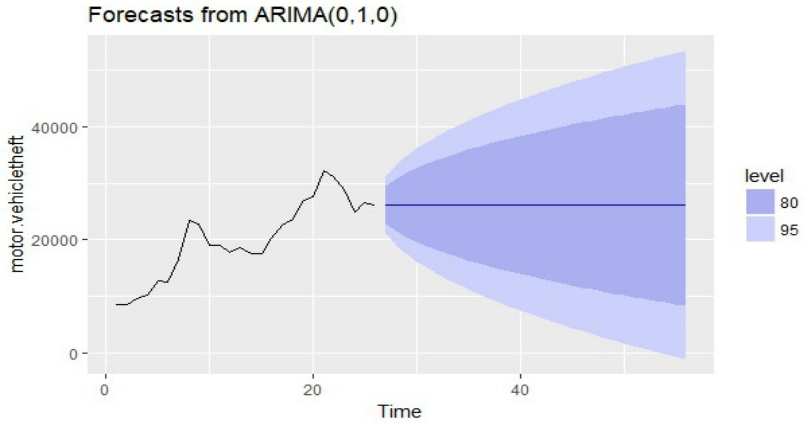

Diagram 4.10 Forecasts of motor and vehicle thefts in Greece for the next 30 time units.

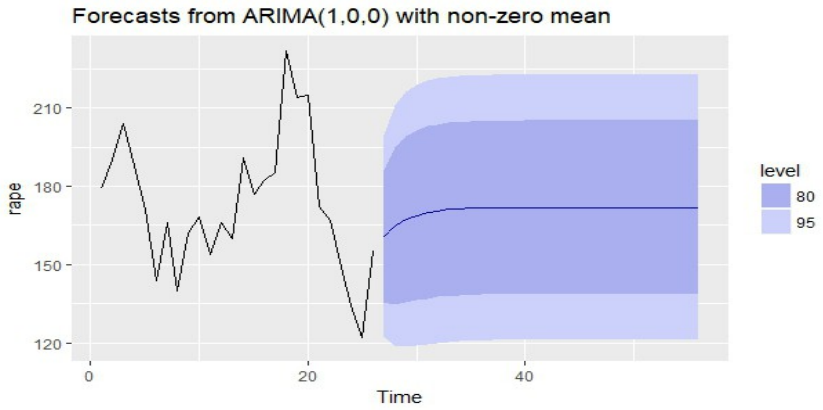

Diagram 4.11 Forecasts of rapes in Greece for the next 30 time units.

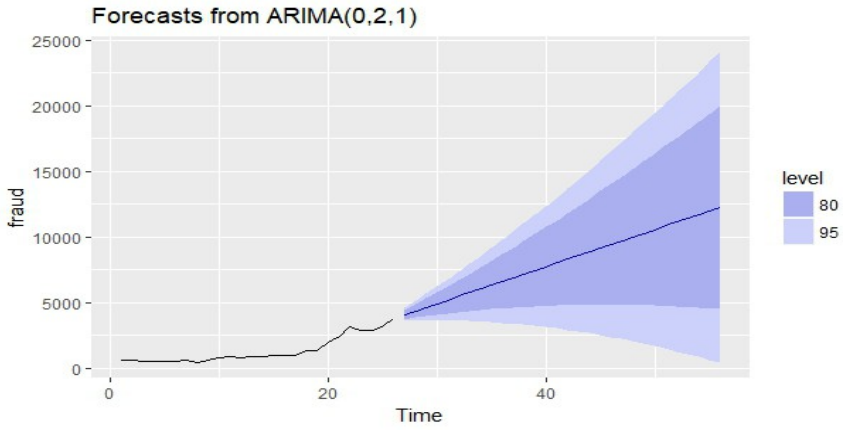

Diagram 4.12 Forecasts of frauds in Greece for the next 30 time units.

Regarding the time series for the variable “law about drugs”,with missing values for the years 1991-1999, to deal with this subject and substitute the missing values,the method of imputation was developed.The package “imputeTS” in R and the function “na.kalman”,were applied to perform replacement of missing values in the univariate time series.The package based on the information of the existed observations,managed to estimate values for replacing the missing data.

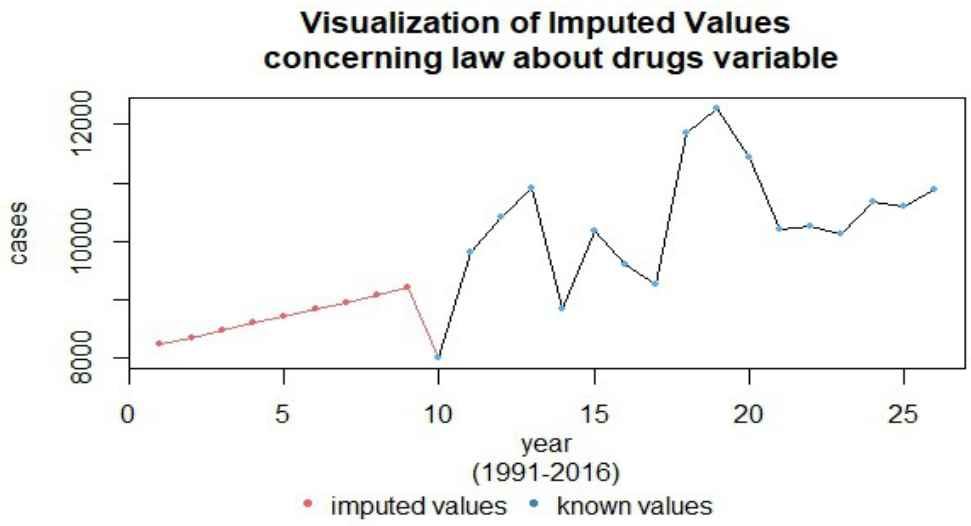

Diagram 4.13 Kalman imputation on 'law about drugs'missing data.

The new time series (imp) with the imputed values were capable to be analysed.The same methods as to the other variables was applied to get and test the stationary series proportionately. The fit of the model according to best AICC and the forecast diagram,are given below.

The predicted model with the lowest value of the AICC=414.8 (the Akaike criterion: AIC= `414.62)` is the ARIMA (0,1,0),which is random walk.

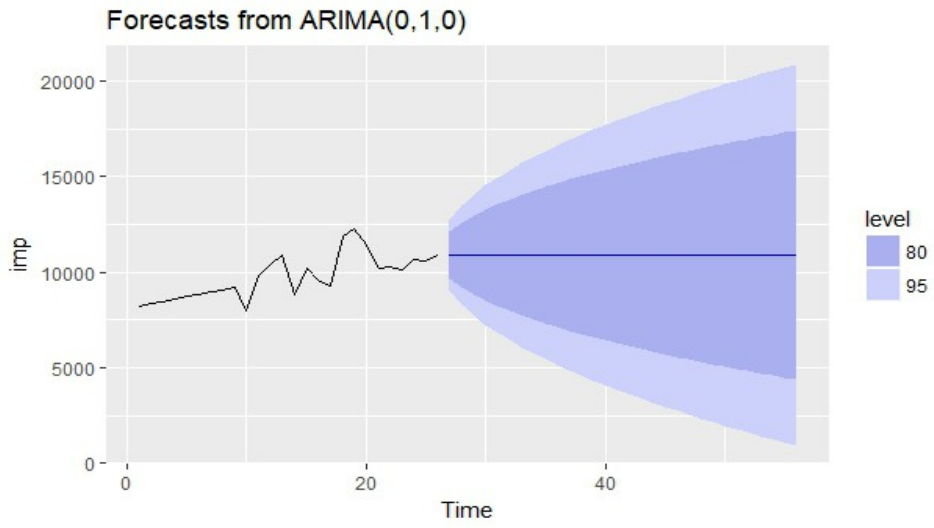

Diagram 4.14 Forecasts of the "law about drugs"variable with the imputed values, for the next 30 time units.

4.2.2 Multivariate Analysis.

An other approach in time series analysis is to try interpret the variables not one at a time,but all in the same time.The Vector Auto Regression (VAR) method is for forecasting multivariate time series using vectors to represent the relationship between variables and past values.In a VAR model, each variable is a linear function of the past values of itself and the past values of all the other variables.

A VAR model is able to understand and use the relationship between several variables. This is useful for describing the dynamic behavior of the data and also provides better forecasting results.

In this thesis,there has been an application of the “var” function of the package MTS in R,for a multivariate methodology of data interpretation.

Cross correlations are generally used when measuring information between two different time series.The range of the data is -1 to 1 such that the closer the cross-correlation value is to 1, the more closely the information sets are.In the results there were not enough cross-correlations in the data,therefore,the univariate analysis of the time series, results to be preferable,for that particular data set, than the multivariate method.

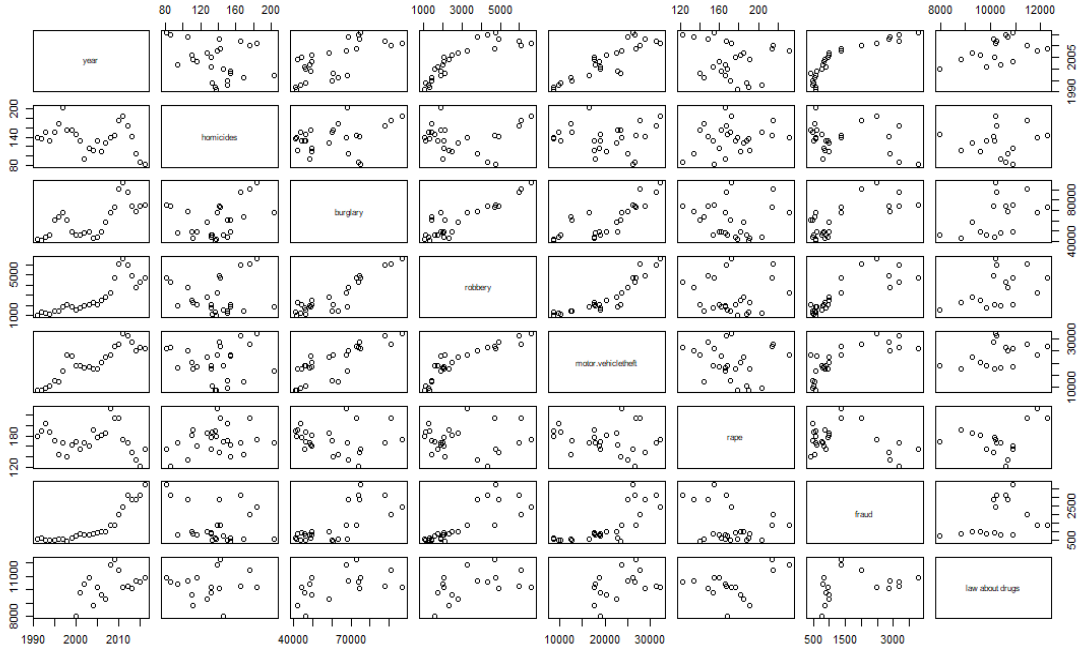

Diagram 4.15 Cross - Correlations of the variables in multivariate analysis.

4.2.3 Regional Analysis.

4.2.3.1 Inference for count data.

Predominantly in this thesis lies the effort to give an insight to the crime pattern in Greece,studying closely a number of specific crimes taking place in a calendar year.For a more thoroughly investigation of the variables,we obtained data for the specific 14 regions that the country is divided to.As mentioned previously,we had to estimate the values of the missing data for the years 1991-2007 regarding all the variables of interest.(Chapter 3,paragraph 3.5.2)

We have tried to fit the most appropriate model in each time series for the distinct regions of the country.Two different methods have been aplied in order to explain the data and forecast future values.That is due to the small values of the crime observations appearing in the regional series.

As to the variables “burglary”,”robbery”,”motor.vehicletheft”,”fraud” and “lawaboutdrugs”,we have carried out the ARIMA theory,as we did in the univariate analysis.

For the regional time series where we confronted small number of observations,we have implemented the method for count data.The statistical package “tscount” of the software R,provided the application of the method.Based on the GLM models,providing the function”tsglm”,this technique,led to the best fit model in the “homicides” and “rape” time series regional data.

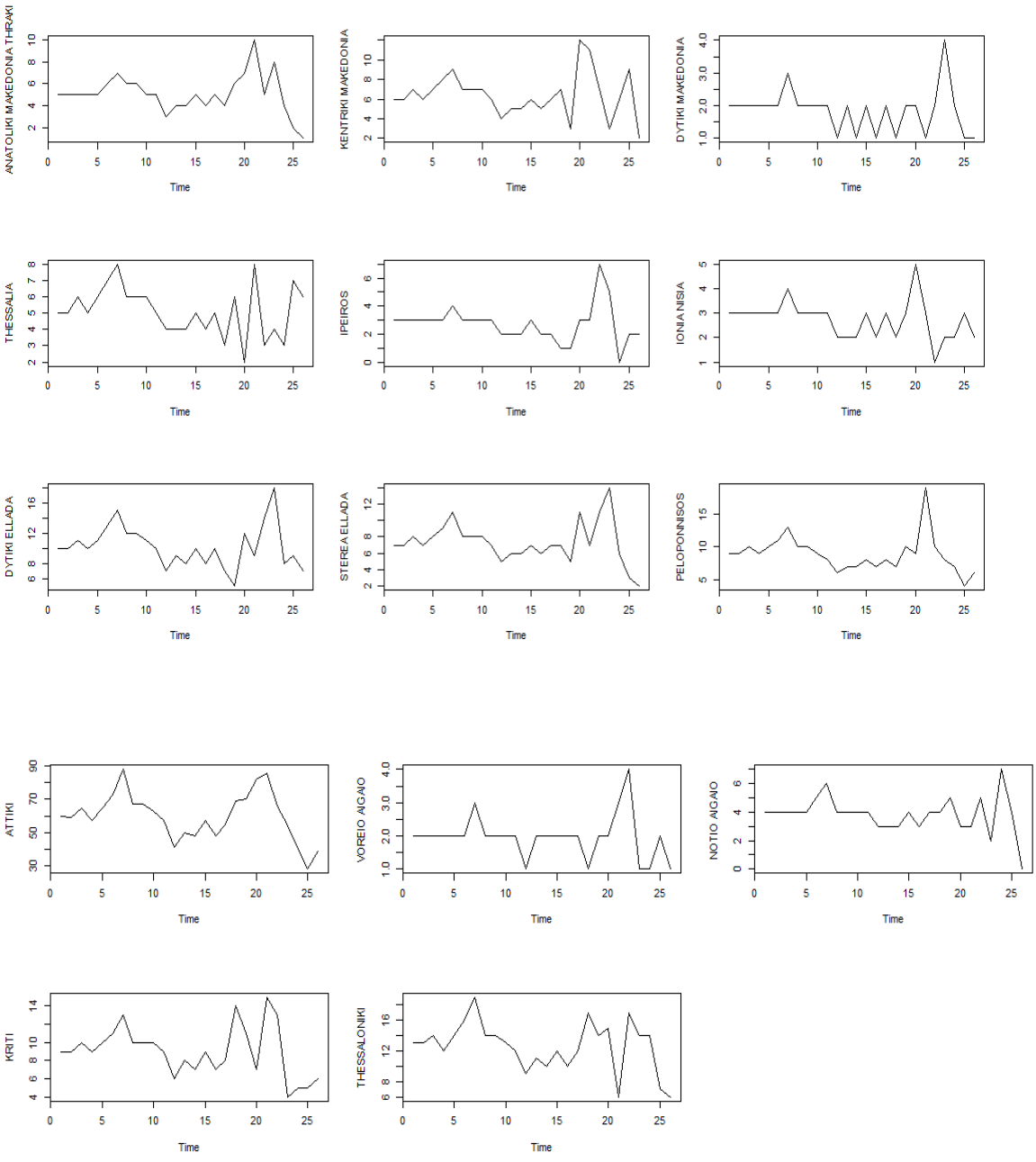

Diagram 4.16 Homicides in the 14 regions of Greece.

The 14 regions of Greece are shown in the **Diagram 4.1.6**,depicting the homicides during 1991-2016.We have produced ACF plots and fitted a model to this time series using the function tsglm.Following in the analysis we have fitted a Poisson model and we compared that fit of a Poisson with that of a Negative Binomial conditional distribution, specified by the argument distr.We have also included the two intervention effects detected by Fokianos and Fried (2010) in the model by suitably chosen covariates provided by the argument xreg.

The commands in the R environment are given below,ostensive presenting the analysis in one region(Western Greece) regarding the time series "homicides":

```
> #DYTIKI ELLADA:
> par(mfrow=c(2,2))
> plot.ts(`DYTIKI ELLADA`)
> mean(`DYTIKI ELLADA`)
[1] 10.23077
> var(`DYTIKI ELLADA`)
[1] 7.784615
> acf(`DYTIKI ELLADA`,lag.max=24)
> plot(table(`DYTIKI ELLADA`)/length(`DYTIKI ELLADA`),ylab="probs",xlab="values")
>points(0:55, dpois(0:55,lambda=mean(`DYTIKI ELLADA`)))
>rg_pois<-tsglm(`DYTIKI ELLADA`,model=list(past_obs=1),xreg=NULL,link=c("identity"),
distr="poisson")
> summary(rg_pois)
> rg_nbin<-tsglm(`DYTIKI ELLADA`,model=list(past_obs=1),xreg=NULL,link=c("identity"),
distr="nbinom")
> summary(rg_nbin)
```

We got warning messages in the process of the implemantation:

"Warning message:

In tsglm.distrfit(meanfit, distr = distr) :

The dispersion parameter of the negative binomial distribution cannot be estimated. This indicates that there is no or only very little overdispersion in the data. The Poisson distribution with argument 'distr' set to "poisson" was fitted instead. **"**

For that reason a Poisson model was fitted instead.[1] The procedure though,is demonstrated as follow:

The resulting fitted models rg_pois and rg_nbin have class‘tsglm’, for which the“summary” argument provides a detailed model summary and plot for diagnostic plots. The diagnostic plots can produced as follow:

```
> par(mfrow=c(2,2))
> #Diagnostic plots after model fitting to the regional crime data.
> acf(residuals(rg_pois),main="ACF of response residuals")
> #Marginal calibration plot
> marcal(rg_pois,ylim=c(-0.2,0.2),main="Marginal calibration")
> lines(marcal(rg_nbin,plot=FALSE),lty="dashed")
> #Probability integral transform histogram
> pit(rg_pois,ylim=c(0,1.5),main="PIT Poisson")
> pit(rg_nbin,ylim=c(0,1.5),main="PIT Negative Binomial")
```

The scoring rules for the two distributions are given:

```
> #scoring rules for the two distributions
> rbind(Poisson=scoring(rg_pois),NegBin=scoring(rg_nbin))
```

Finally,we predicted the future values:

```
> #Prediction
> rg_pois_pred<-predict(rg_pois,n.ahead=20,level=0.9,global=TRUE)
> rg_pois_pred$predrg_pois_pred$interval
```

1https://cran.r-project.org/web/packages/tscount/vignettes/tsglm.pdf (pg.18 and AppendixB.2.)

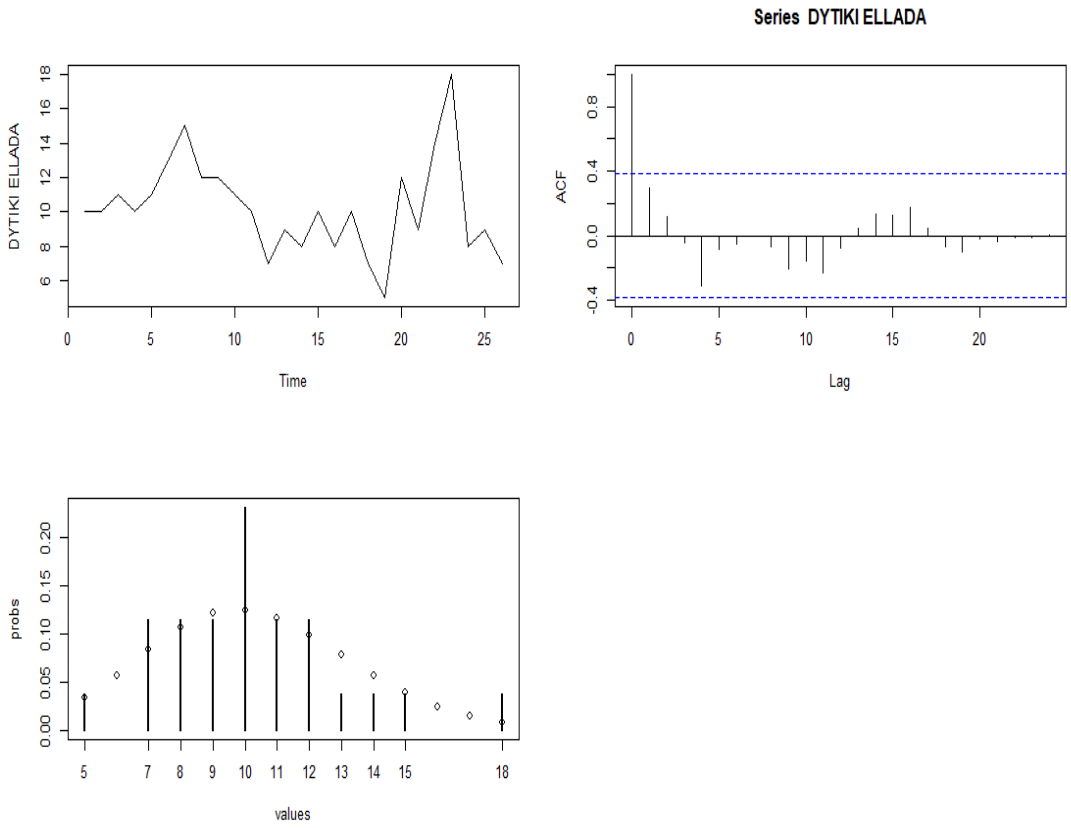

Diagram 4.17 Homicides in Dytiki Ellada and ACF plot of the time series.

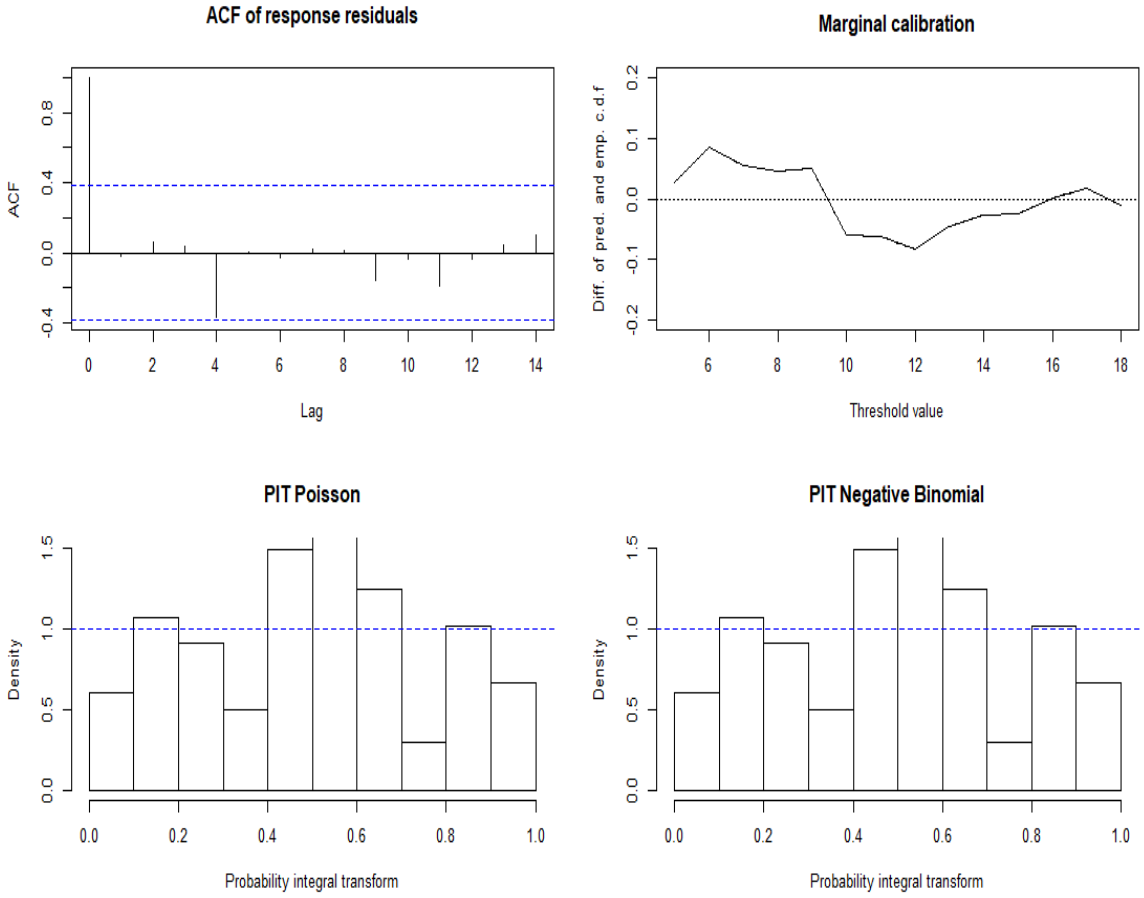

Diagram 4.18 Diagnostic plots after model fitting to the Dytiki Ellada homicides data.

The response residuals are identical for the two conditional distributions. Their empirical autocorrelation function,is shown in the **Diagram 4.18**(top left),does not exhibit any serial correlation which has not been taken into account by the models.**Diagram 4.18** (bottom left) shows that the PIT histogram resembles the uniform distribution, indicating that the Poisson distribution is quite adequate for model fitting.For that,the probabilistic calibration of the Poisson model is satisfactory. The marginal calibration plot[1], shown in **Diagram 4.18**(top right),depicts observations with no major deviations from zero,that would point to model deficiencies. We have also considered the scoring rules for the two distributions.The considered scoring rules are in favor of the Poisson distribution. Based on the PIT histograms and the results obtained by the scoring rules, we have decided for the Poisson model.

Table 3. Estimations about count data in Dytiki Ellada.

```
 >summary(rg_pois)
Call:
tsglm(ts = `DYTIKI ELLADA`, model = list(past_obs = 1), xreg = NULL,
    link = c("identity"), distr = "poisson")
Coefficients:
             Estimate  Std.Error  CI(lower)  CI(upper)
(Intercept)     6.942      2.522      2.000     11.885
beta_1          0.317      0.241     -0.155      0.789
Standard errors and confidence intervals (level =  95 %) obtained
by normal approximation.
Link function: identity
Distribution family: poisson
Number of coefficients: 2
Log-likelihood: -62.28403
AIC: 128.5681
BIC: 131.0843
QIC: 128.5752
```

```
> summary(rg_nbin)
Call:
tsglm(ts = `DYTIKI ELLADA`, model = list(past_obs = 1), xreg = NULL,
    link = c("identity"), distr = "nbinom")
Coefficients:
             Estimate  Std.Error  CI(lower)  CI(upper)
(Intercept)     6.942      2.522      2.000     11.885
beta_1          0.317      0.241     -0.155      0.789
Standard errors and confidence intervals (level =  95 %) obtained
by normal approximation.
Link function: identity
```

1https://cran.r-project.org/web/packages/tscount/vignettes/tsglm.pdf (p.g11 marginal calibration)

```
Distribution family: poisson
Number of coefficients: 2
Log-likelihood: -62.28403
AIC: 128.5681
BIC: 131.0843
QIC: 128.5752
```

>rbind(Poisson=scoring(rg_pois),NegBin=scoring(rg_nbin))

```
       logarithmic  quadratic  spherical rankprob   dawseb    normsq  sqerror
Poisson     2.39554 -0.1078044 -0.3295627 1.466985 2.971938 0.6501148
6.808523
NegBin      2.39554 -0.1078044 -0.3295627 1.466985 2.971938 0.6501148
6.808523
```

The coefficient beta_1 corresponds to regression on the previous observation,therefore,the fitted model for the number of new homicides Y_t in time period t ,is given by

$Y_t | F_{t-1} \sim Poisson(\lambda_t, \lambda_t)$ with $\lambda_t = 6.95 + 0.32\, Y_{t-1}$.

The beta_1 coefficient is very small and even slightly below the size of its approximate standard error, indicating that there is no notable dependence on the number of homicides in Dytiki Ellada of the previous year.

The predicted values are:

> rg_pois_pred<-predict(rg_pois,n.ahead=20,level=0.9,global=TRUE)

```
> rg_pois_pred$pred
 [1]  9.164024  9.850810 10.068771 10.137945 10.159898 10.166865 10.169076
10.169778
 [9] 10.170001 10.170071 10.170094 10.170101 10.170103 10.170104 10.170104
10.170104
[17] 10.170104 10.170104 10.170104 10.170104
> rg_pois_pred$interval
      lower upper
 [1,]     2    18
 [2,]     2    20
 [3,]     3    21
 [4,]     2    20
 [5,]     3    20
 [6,]     3    21
 [7,]     2    19
 [8,]     3    20
 [9,]     2    21
[10,]     2    20
[11,]     2    20
[12,]     2    21
[13,]     3    21
[14,]     2    23
[15,]     3    23
[16,]     2    22
[17,]     3    21
[18,]     2    20
[19,]     2    21
[20,]     3    21
```

Table 4. Predicted values for count data in Dytiki Ellada.

The same methodology has been applied for the other 13 districts of Greece concerning the variable“homicides”and the 14 districts regarding the variable“rapes”.

In the next Diagrams 4.19,4.20 it is portrayed the rape pattern in the specific 14 districts of Greece.

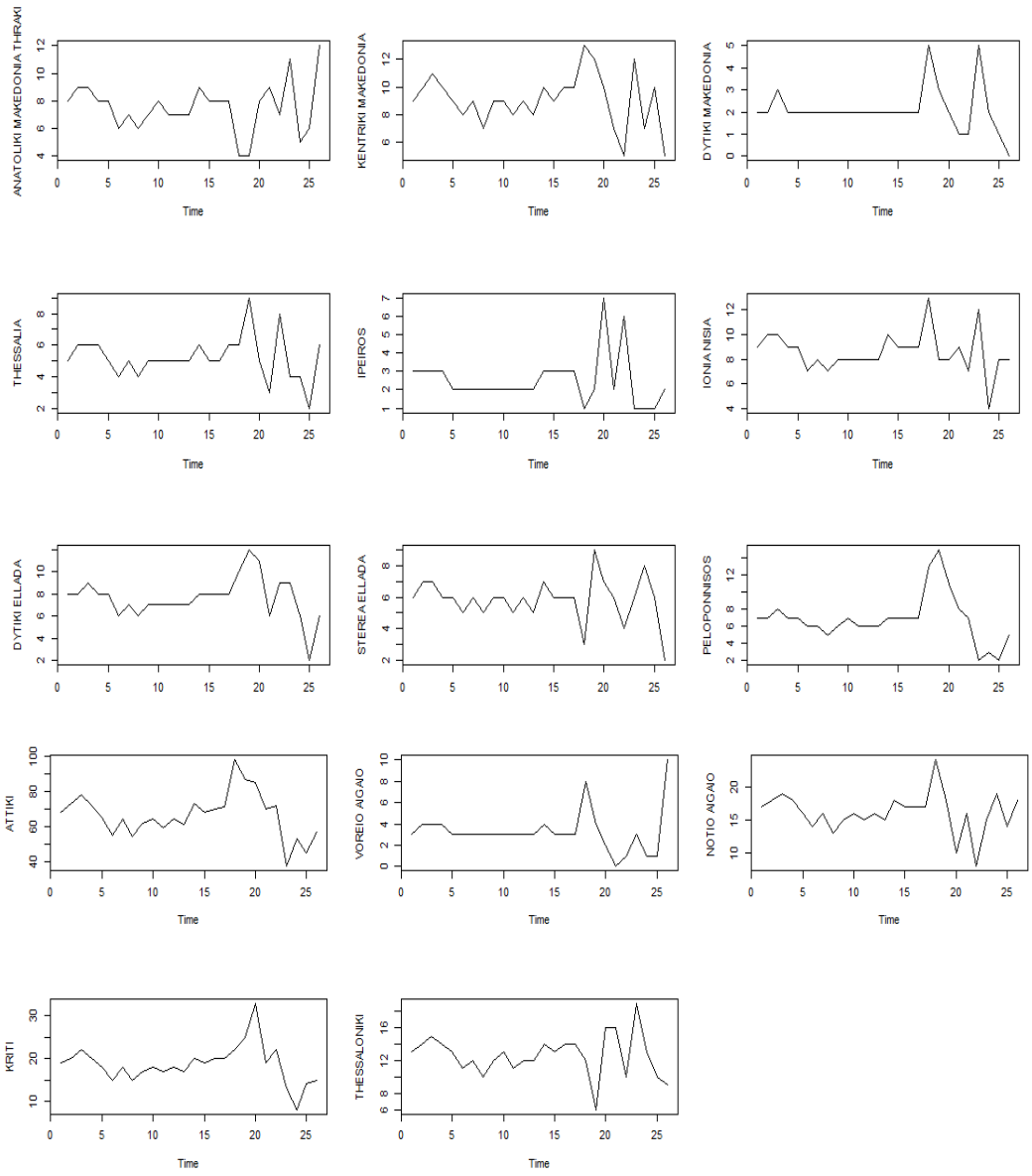

Diagram 4.19 Rapes in the regions of Greece,time series 1991-2016.

4.2.3.2 Inference with ARIMA Models Theory.

Next we have inteded to analyze the time series of the crimes “burglary” ,”robbery” , ”motor.vehicletheft”,”fraud” and “lawaboutdrugs”for the 14 districts of Greece,which have quite large numbers that can be treated with methods for continuous-valued data.Hence,we have proceeded with the implementation of the ARIMA theory for non stationary time series with a trend.

1. **Burglaries.**

We have analyzed the series for the 14 regions using the package “forecast” in the statistical environment of R.Firstly,we have created the plots of the regional series,so as to gain a better knowledge on the data.

Following on,we attempt to illustrate with an example for one region, the methodology that has been applied and the related predictions for the burglaries in the future.

For the region 'Anatoliki Makedonia (Eastern Macedonia),ACF plots have created in the series values,the log-values and the first time differentiated log values for stationarity purposes.To test the stationarity,ADF tests have been progressed,in order to proceed to the fit of a model with ARIMA theory.The fit of the model has been done with the “auto.arima” function of the “forecast” R package.The sellection of the best model has been done,evaluating the lowest AICc by the package.Finally,the forecast package provided the implementation of the prediction process with the “forecast” function.

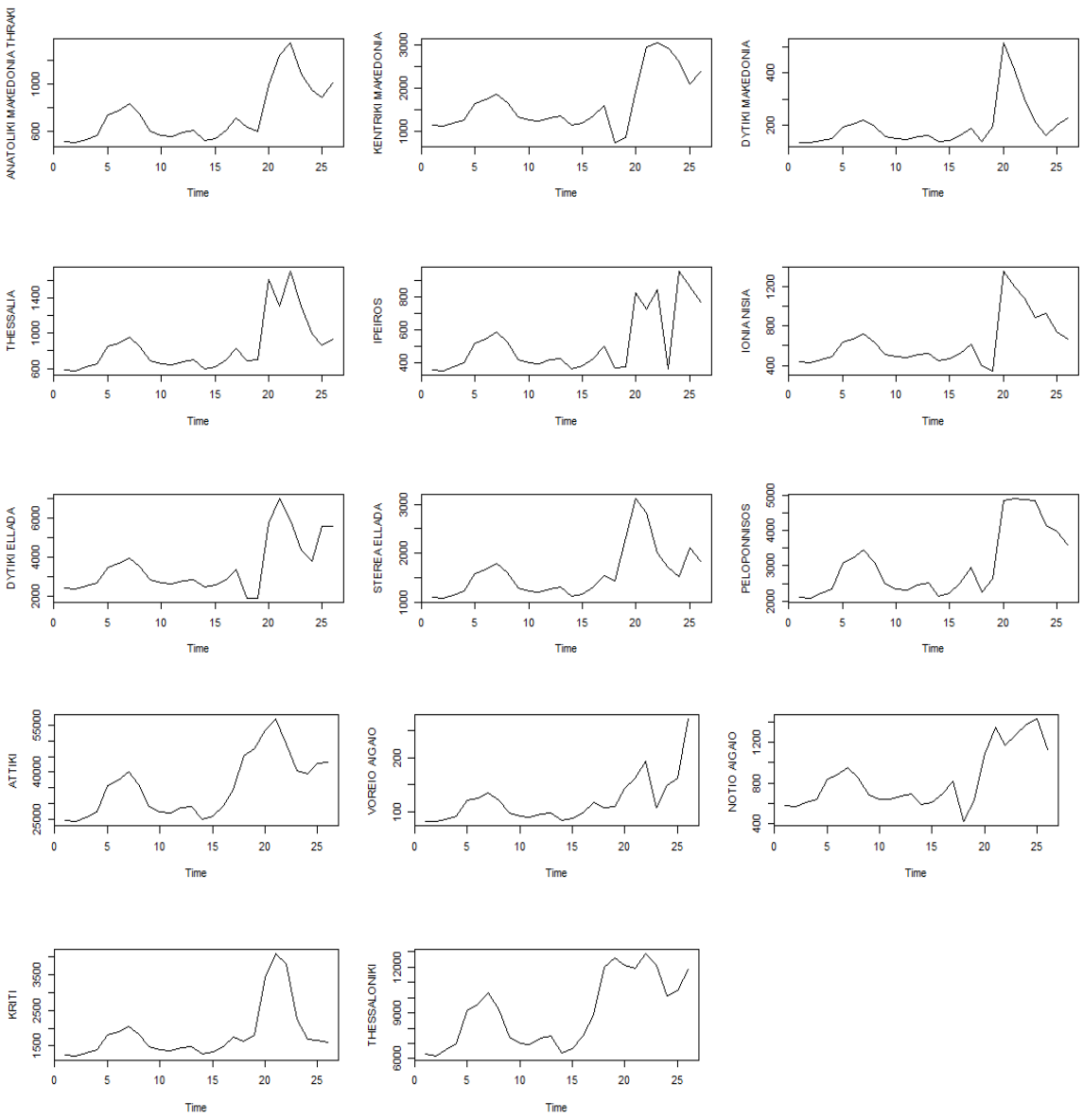

Diagram 4.20 Burglaries in Greece 1991-2016.

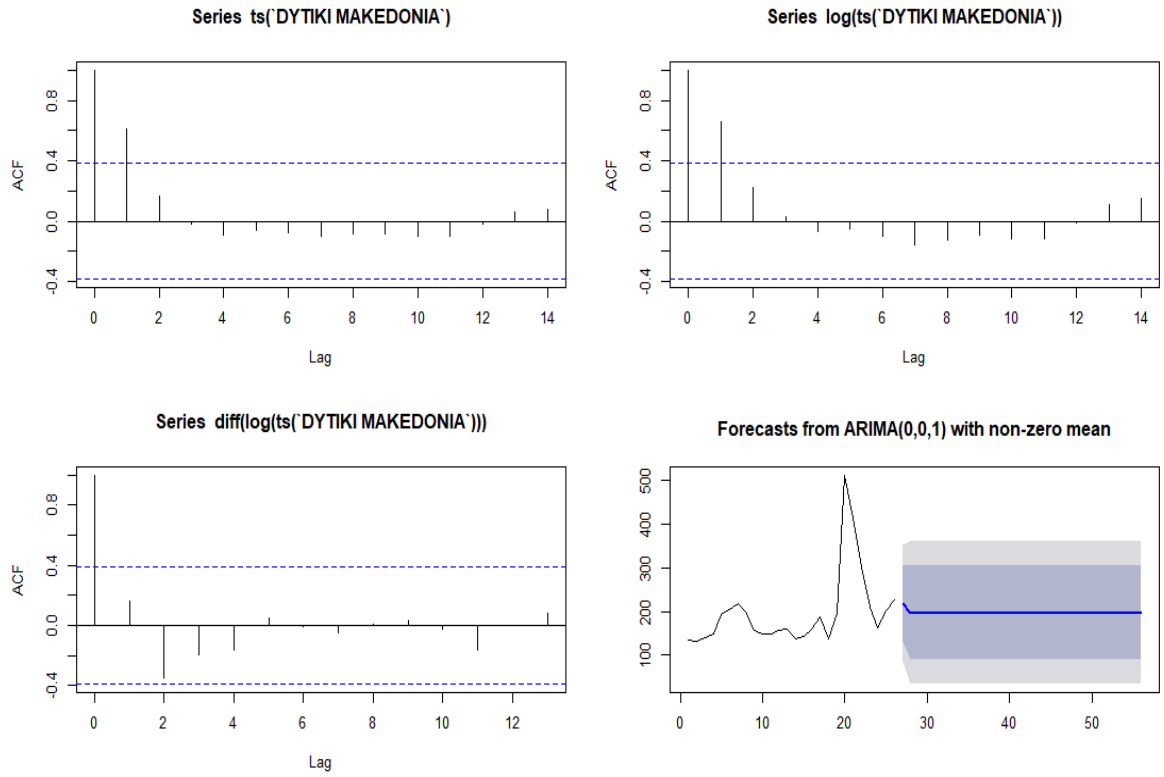

Diagram 4.21. ACF plots of the burglary time series,
log-values of the series and the first differences of the log-values
of the series and forecasts,in Dytiki Makedonia.(Western Macedonia.)

We demonstrate the p-values,as part of the R code output of the ADF tests for evaluating the series stationarity:

```
Augmented Dickey-Fuller Test
 >data:  ts(`DYTIKI MAKEDONIA`)
Dickey-Fuller = -2.5548, Lag order = 0, p-value = 0.361
alternative hypothesis: stationary
 >data:  log(ts(`DYTIKI MAKEDONIA`))
Dickey-Fuller = -2.3573, Lag order = 0, p-value = 0.4363
 >data:  diff(log(ts((`DYTIKI MAKEDONIA`))))
Dickey-Fuller = -3.8759, Lag order = 0, p-value = 0.03029
alternative hypothesis: stationary
```

Table 5. ADF tests for burglaries in Dytiki Makedonia

The first differences of the log-values are stationary series,since we reject the null hypothesis (p-value=0.03029) of the presence of a unit root in the time series sample.

The prediction of the model is given from the output below,evaluated with the smallest AICC=299.17.It is an ARIMA (0,0,1) with non -zero mean,or a MA(1) model (e.g a moving average model).

The selected model with the smallest AICC=298.08 would be: $Y_t = c + ma(1)\varepsilon_t - _1$. or $Y_t = 197.13 + 0.715\,\varepsilon_t - _1$.

where ε_t is white noise, with $sd = \sqrt{\hat{s}^2} = \sqrt{4668} = 68.32$.

```
> summary(arima_forecast)
Forecast method: ARIMA(0,0,1) with non-zero mean
Model Information:
Series: DYTIKI MAKEDONIA
ARIMA(0,0,1) with non-zero mean
Coefficients:
         ma1      mean
      0.7154  197.1299
s.e.  0.1364   21.7453
sigma^2 estimated as 4668:  log likelihood=-146.04
AIC=298.08   AICc=299.17   BIC=301.86
Error measures:
                     ME      RMSE      MAE       MPE      MAPE       MASE
ACF1
Training set 0.5926142 65.64574 39.1135 -7.06151 18.67045 0.8938187
0.1602182
```

Table.6 Estimated coefficients for the time series “burglary”in Dytiki Makedonia.

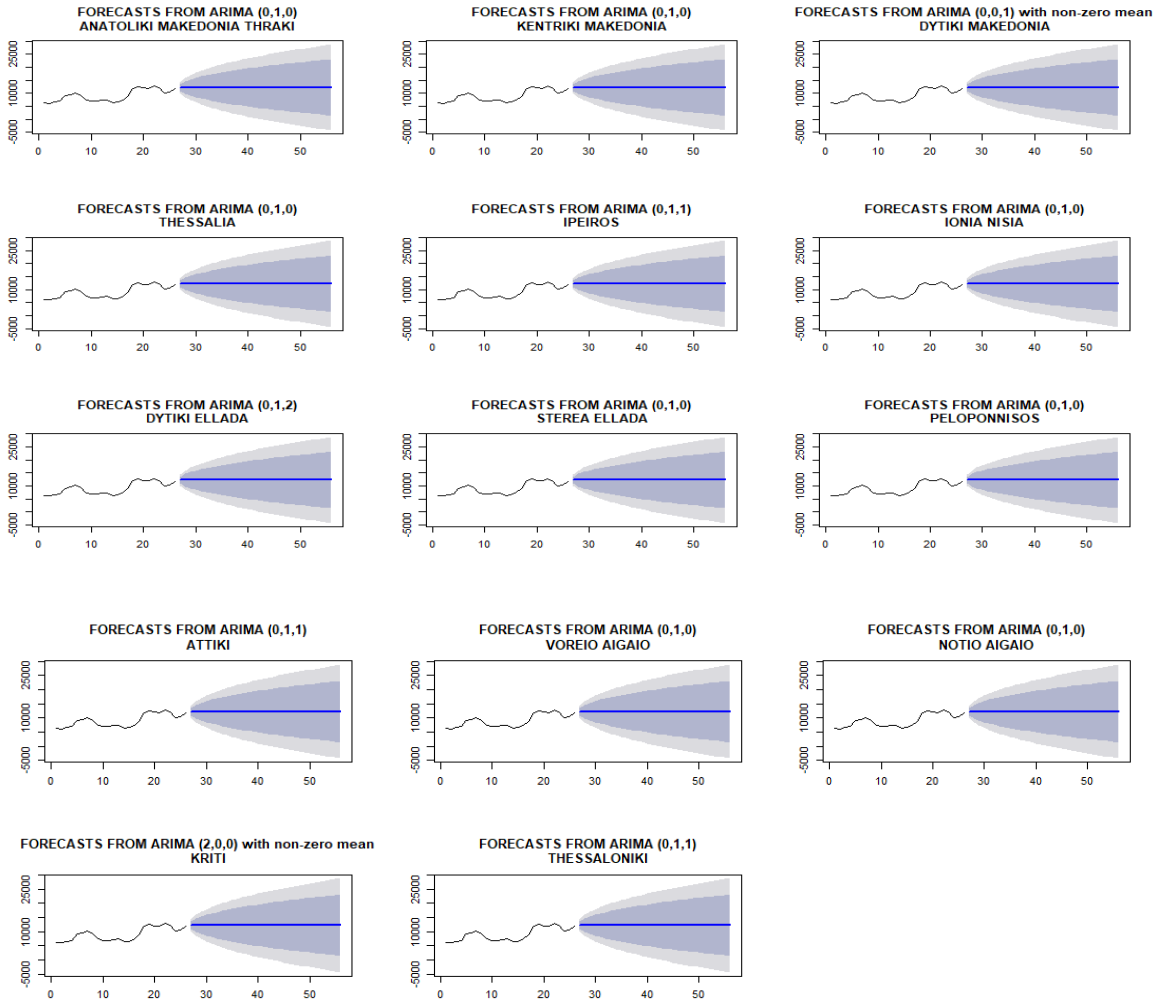

Diagram 4.22 Forecasts about burglaries in the districts of Greece

The sellected models from “auto.arima” function in the “forecast” package in R, for **burglaries**,based on the smallest AICC criterion are given in the table below :

REGIONS	ARIMA Models	AICC criterion
ANATOLIKI MAKEDONIA THRAKI	ARIMA(0,1,0)	317.04
KENTRIKI MAKEDONIA	ARIMA(0,1,0)	371.72
DYTIKI MAKEDONIA	ARIMA(0,0,1) with non-zero mean	299.17
THESSALIA	ARIMA(0,1,0)	348.58
IPEIROS	ARIMA(0,1,1)	329.04
IONIA NISIA	ARIMA(0,1,0)	344.09
DYTIKI ELLADA	ARIMA(0,1,2)	414.69
STEREA ELLADA	ARIMA(0,1,0)	366.88
PELOPONNISOS	ARIMA(0,1,0)	389.69
ATTIKI	ARIMA(0,1,1)	490.8
VOREIO AIGAIO	ARIMA(0,1,0)	246.87
NOTIO AIGAIO	ARIMA(0,1,0)	330.22
KRITI	ARIMA(2,0,0) with non-zero mean	394.46
THESSALONIKI	ARIMA(0,1,1)	422.44

Table 7. ARIMA models and AICC about burglaries time series.

Following the same procedure,we demonstrate the regional time series plots,forecasts with the related ARIMA models and the corresponding AICC criteria of the remaining response variables,namely “robbery”, “motor.vehicletheft”, “fraud”, “lawaboutdrugs”.

2.Robberies.

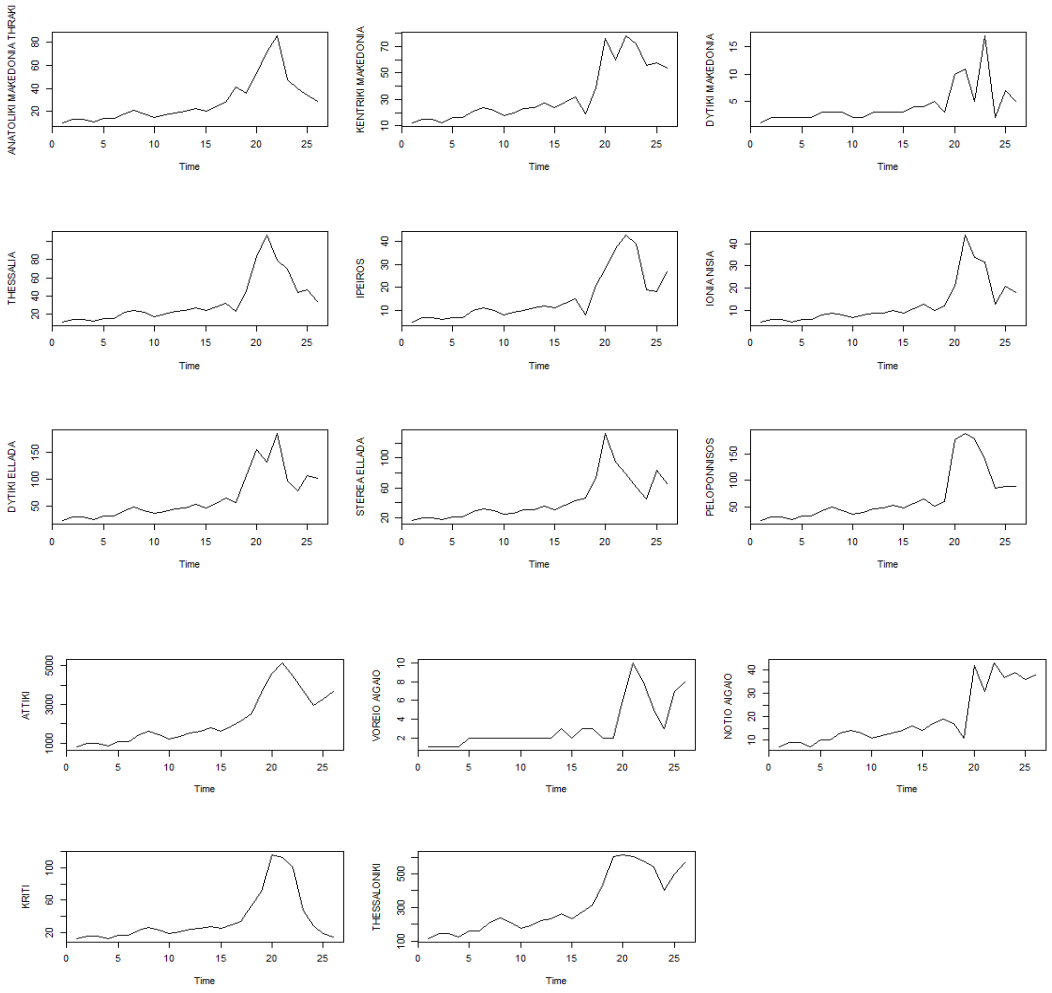

Diagram 4.23 Robberies time series in the region of Greece ,1991-2016.

REGIONS	ARIMA Models	AICC
ANATOLIKI MAKEDONIA THRAKI	ARIMA(0,1,0)	190
KENTRIKI MAKEDONIA	ARIMA(0,1,0)	192.66
DYTIKI MAKEDONIA	ARIMA(2,1,0)	131.55
THESSALIA	ARIMA(0,1,0)	202.31
IPEIROS	ARIMA(0,1,0)	163.15
IONIA NISIA	ARIMA(0,1,0)	169.54
DYTIKI ELLADA	ARIMA(0,1,0)	237.45
STEREA ELLADA	ARIMA(0,1,0)	219.42
PELOPONNISOS	ARIMA(0,1,0)	239.48
ATTIKI	ARIMA(0,1,2)	371.56
VOREIO AIGAIO	ARIMA(0,1,0)	99.22
NOTIO AIGAIO	ARIMA(1,1,0)	167.1
KRITI	ARIMA(0,1,2)	200.66
THESSALONIKI	ARIMA(0,1,0)	278.63

Table 8. ARIMA models and AICC for the robberies time series.

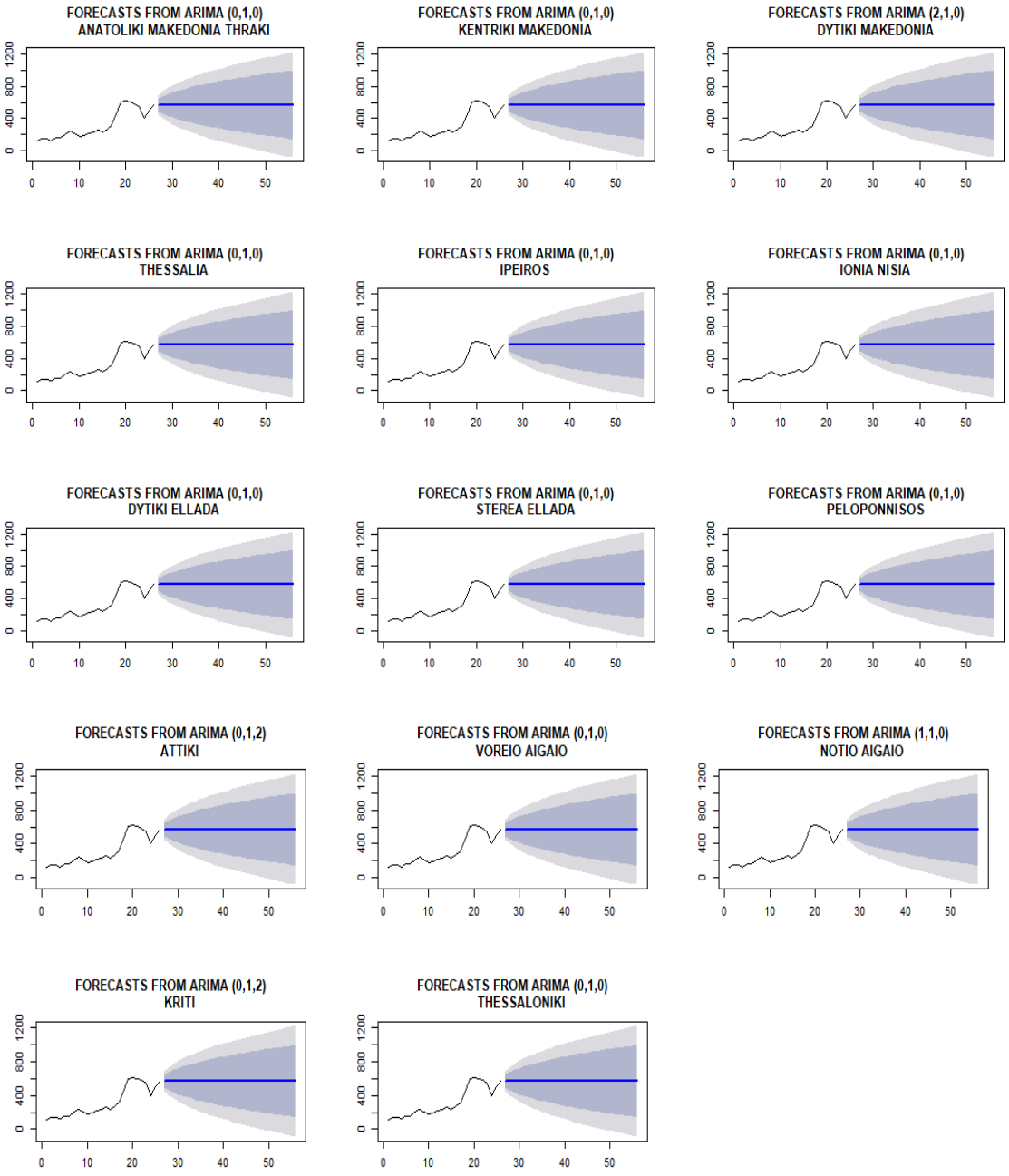

Diagram 4.24 Robberies forecast plots .

3. Motor and vehicle thefts.

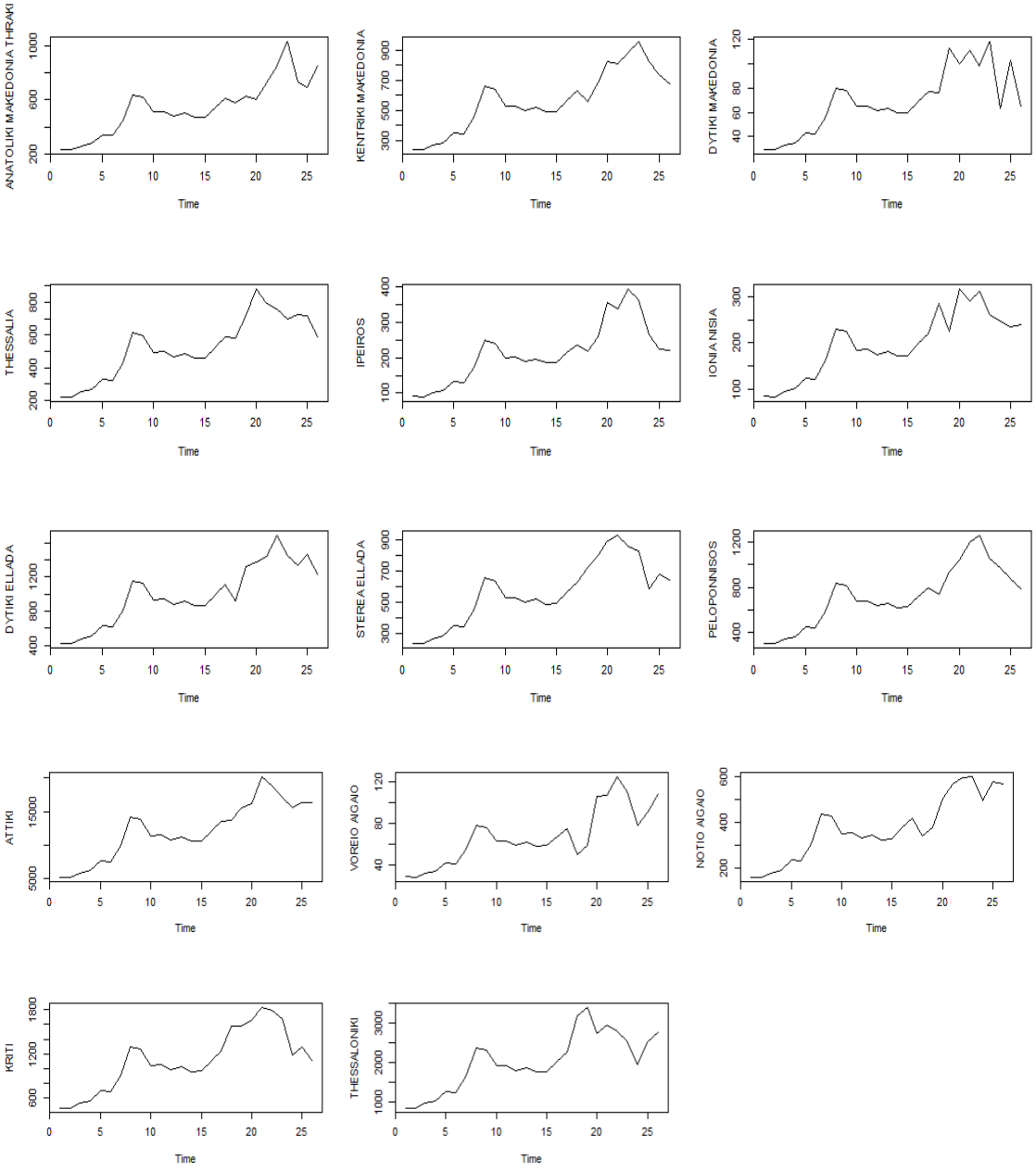

Diagram 4.25 Motor and vehicle theft time series in the 14 districts of Greece.

REGIONS	ARIMA Models	AICC
ANATOLIKI MAKEDONIA THRAKI	ARIMA(0,1,0)	304.13
KENTRIKI MAKEDONIA	ARIMA(0,1,0)	292.42
DYTIKI MAKEDONIA	ARIMA(2,1,0)	211.73
THESSALIA	ARIMA(0,1,0)	290.12
IPEIROS	ARIMA(0,1,0)	257.17
IONIA NISIA	ARIMA(0,1,0)	250.86
DYTIKI ELLADA	ARIMA(0,1,0)	327.77
STEREA ELLADA	ARIMA(0,1,0)	295.12
PELOPONNISOS	ARIMA(0,1,0)	306.3
ATTIKI	ARIMA(0,1,0)	442.48
VOREIO AIGAIO	ARIMA(0,1,0)	210.55
NOTIO AIGAIO	ARIMA(0,1,0)	275.5
KRITI	ARIMA(0,1,0)	331.42
THESSALONIKI	ARIMA(0,1,0)	367.4

Table 9.ARIMA models and AICC for the motor.vehicle theft time series.

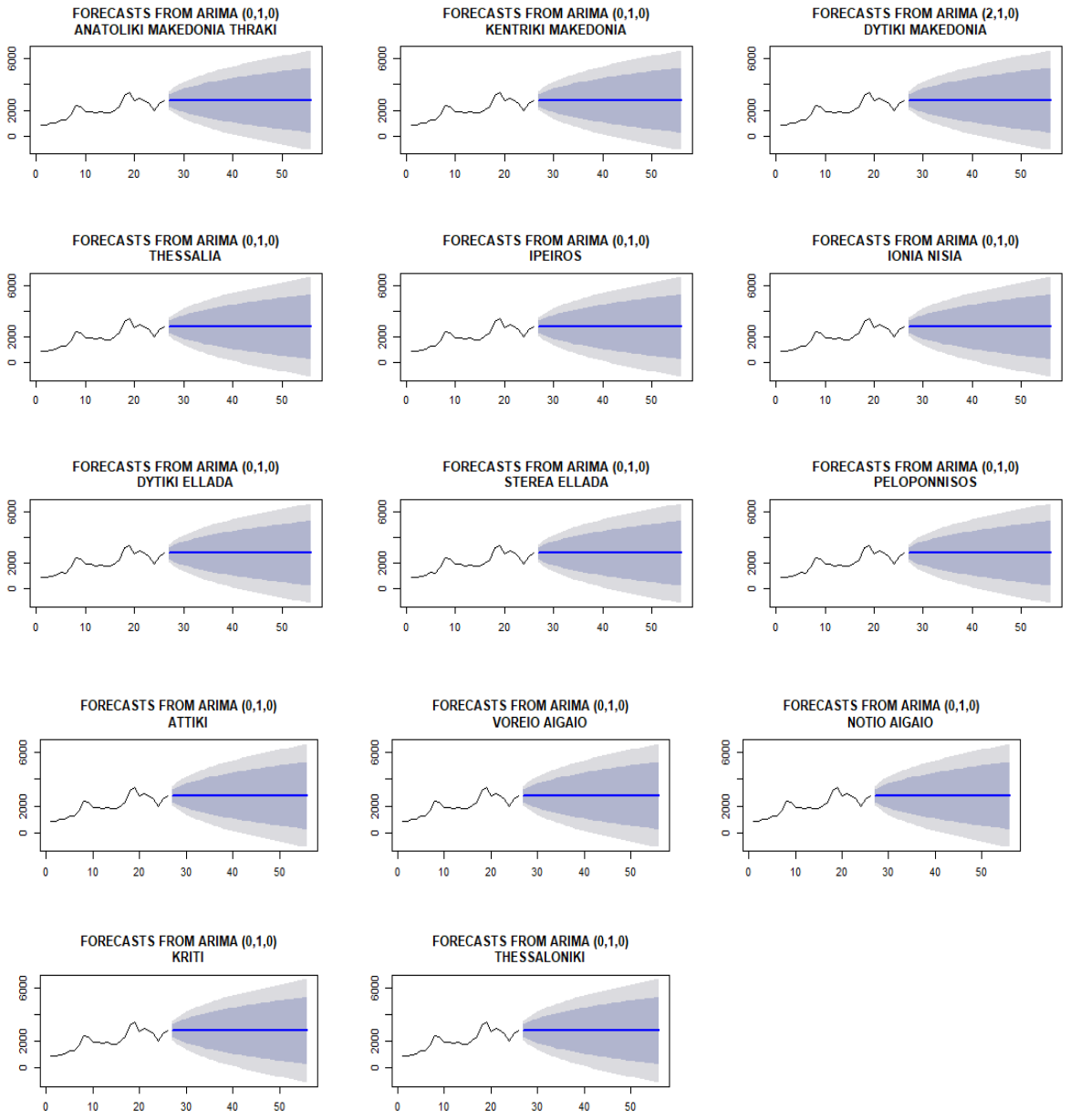

Diagram 4.26 Forecasts in Motor and vehicle thefts in the districts of Greece.

4. Fraud.

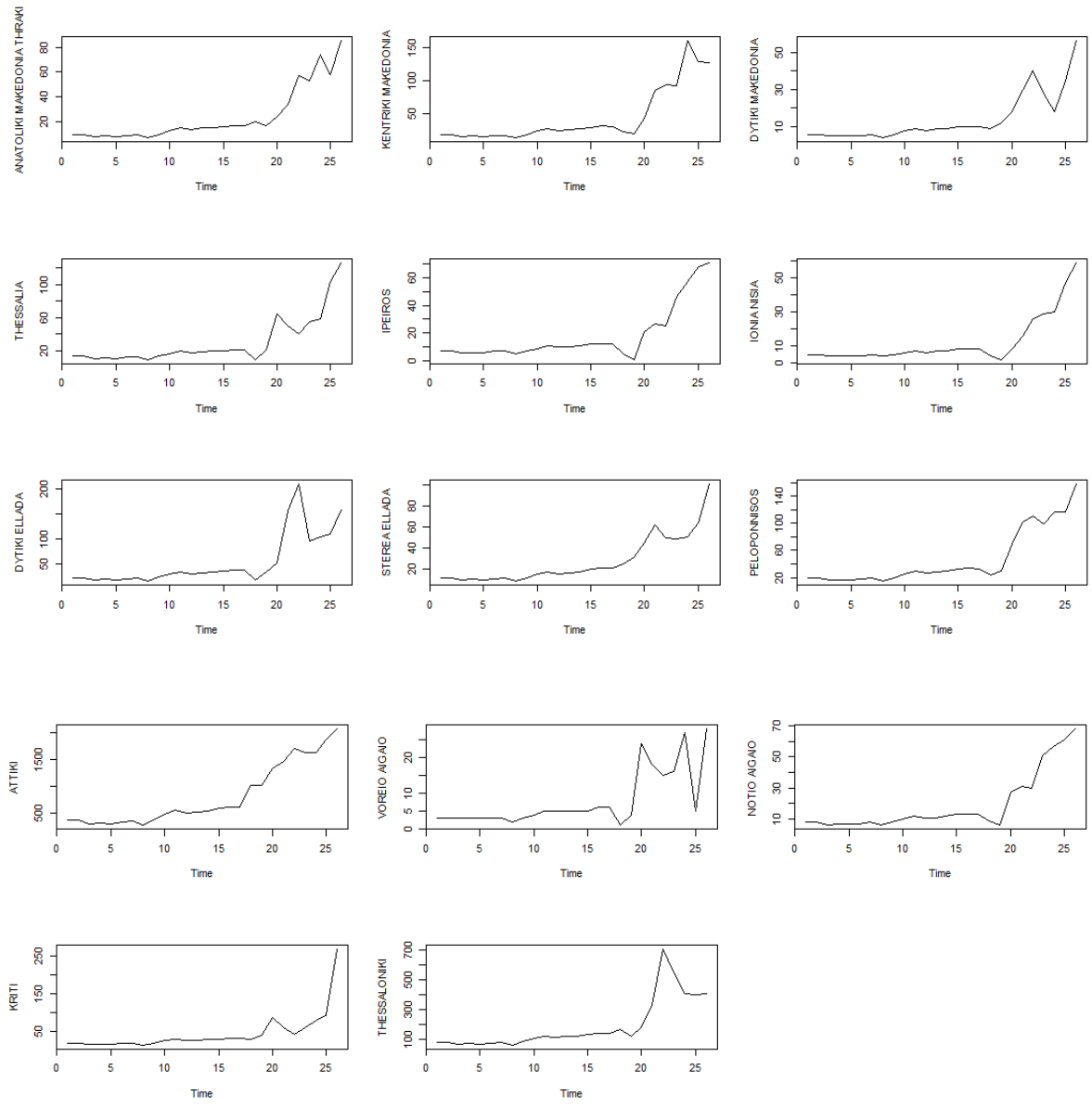

Diagram4.27 Fraud in the regions of Greece.

REGIONS	ARIMA Models	AICC
ANATOLIKI MAKEDONIA THRAKI	ARIMA(0,2,2)	176.54
KENTRIKI MAKEDONIA	ARIMA(0,1,0)	217.63
DYTIKI MAKEDONIA	ARIMA(0,1,1)	160.73
THESSALIA	ARIMA(0,1,0) with drift	207.23
IPEIROS	ARIMA(0,2,1)	163.34
IONIA NISIA	ARIMA(0,2,1)	139.78
DYTIKI ELLADA	ARIMA(0,1,0)	250.57
STEREA ELLADA	ARIMA(0,1,1)	182.84
PELOPONNISOS	ARIMA(0,1,0) with drift	204.38
ATTIKI	ARIMA(0,2,1)	305.19
VOREIO AIGAIO	ARIMA(0,1,1) with drift	168.32
NOTIO AIGAIO	ARIMA(0,2,1)	160.73
KRITI	ARIMA(0,1,0)	254.77
THESSALONIKI	ARIMA(0,1,0)	300.28

Table 10. ARIMA models and AICC for the fraud time series.

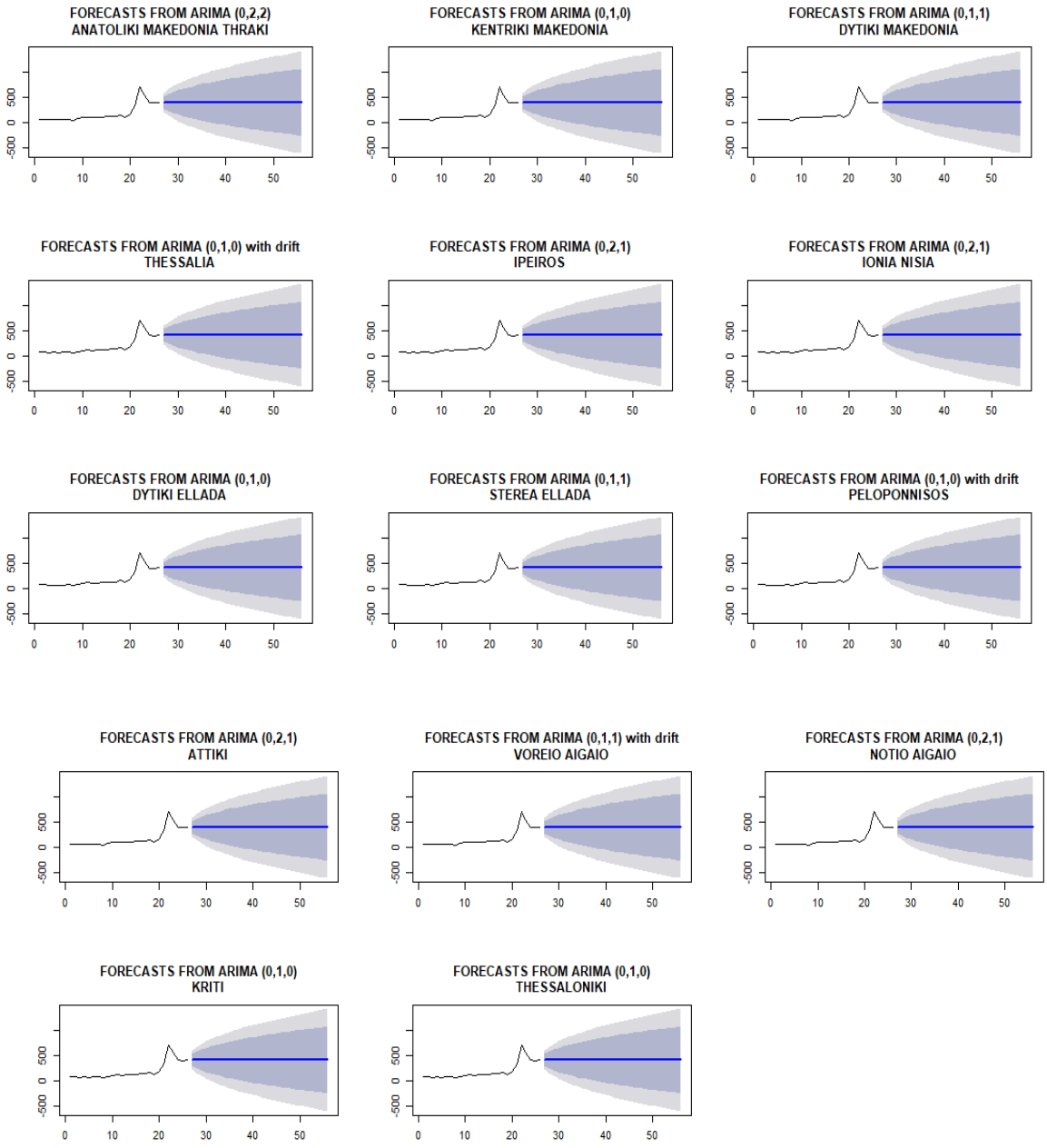

Diagram 4.28. Forecasts in fraud in the districts of Greece.

5.Law about drugs.

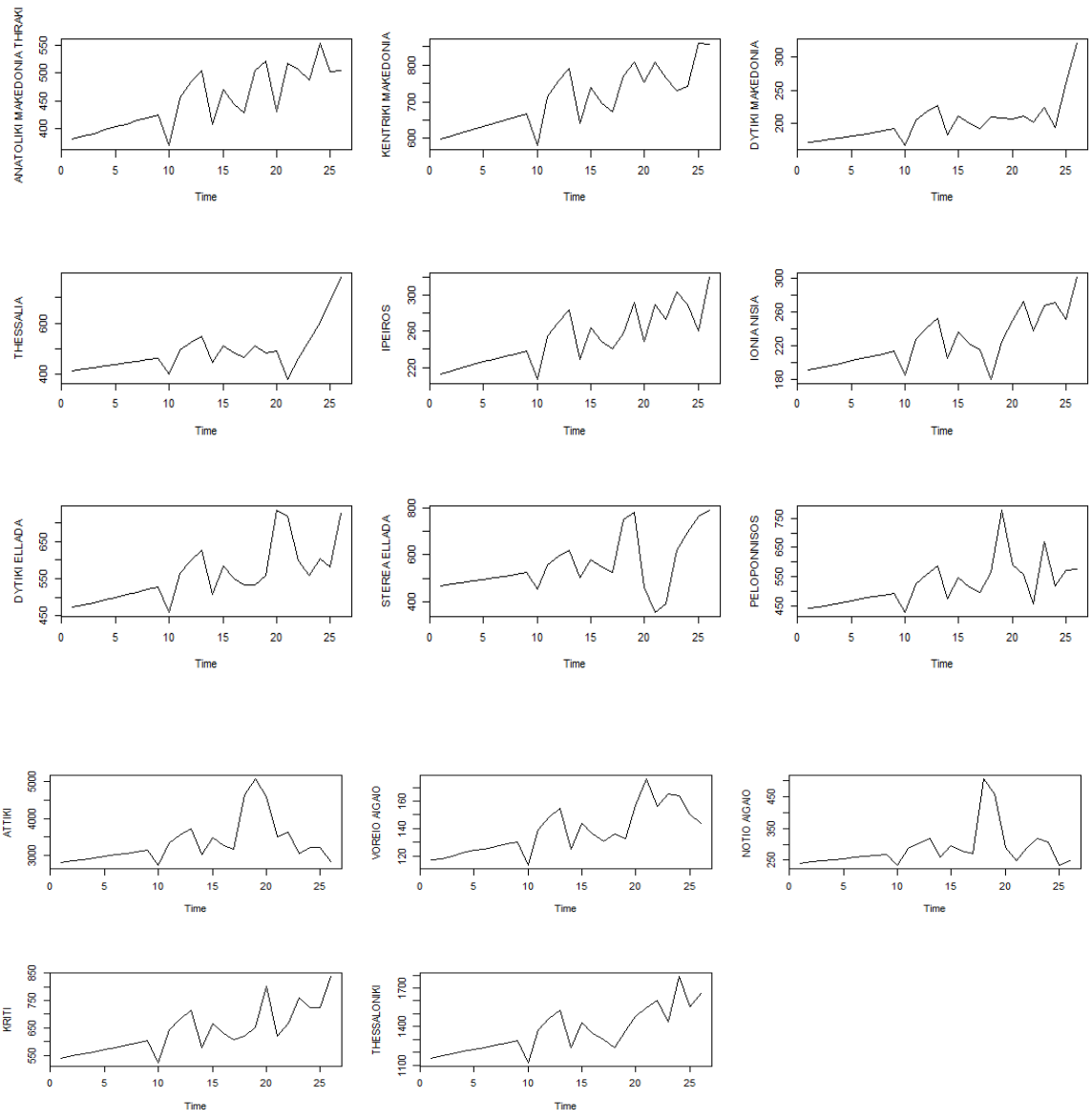

Diagram 4.29.Crimes regarding the law about drugs in the regions of Greece.

REGIONS	ARIMA Models	AICC
ANATOLIKI MAKEDONIA THRAKI	ARIMA(2,1,0)	259.04
KENTRIKI MAKEDONIA	ARIMA(0,1,0)	279.69
DYTIKI MAKEDONIA	ARIMA(0,1,0)	233.02
THESSALIA	ARIMA(0,1,0)	274.14
IPEIROS	ARIMA(0,1,1)	231.05
IONIA NISIA	ARIMA(0,1,0)	234.59
DYTIKI ELLADA	ARIMA(0,1,0)	283.29
STEREA ELLADA	ARIMA(0,0,1) with non-zero mean	315.26
PELOPONNISOS	ARIMA(0,1,1)	289.19
ATTIKI	ARIMA(1,0,0) with non-zero mean	397.49
VOREIO AIGAIO	ARIMA(0,1,0)	200.94
NOTIO AIGAIO	ARIMA(0,0,1) with non-zero mean	285.16
KRITI	ARIMA(2,1,0)	284.25
THESSALONIKI	ARIMA(1,1,0)	315.17

Table 11. ARIMA models and AICC for the law about drugs time series.

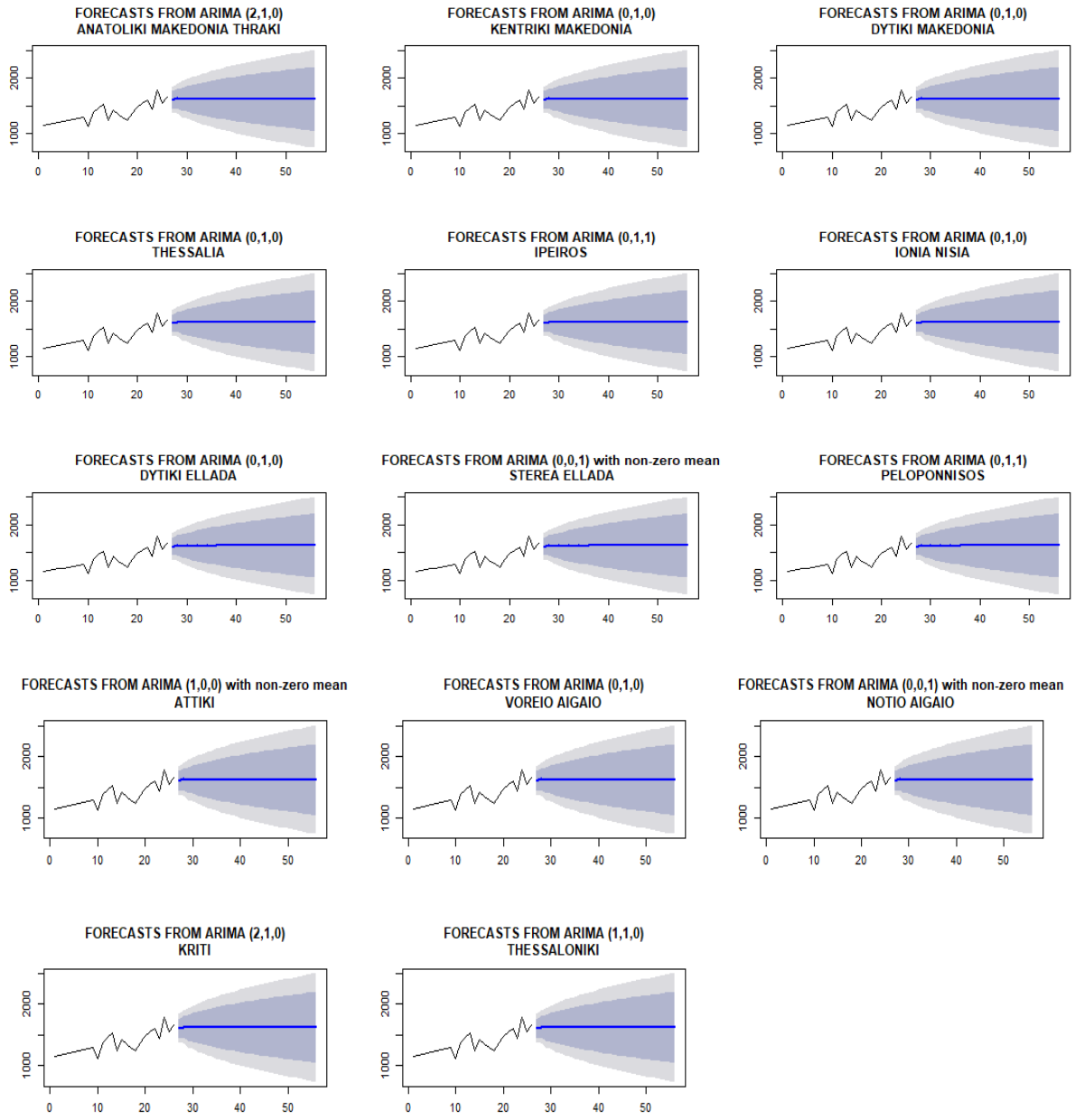

Diagram 4.30 Forecasts for the law about drugs in the regions of Greece.

CHAPTER 5.

DISCUSSION.

We have attempted to contribute in the better understanding of the crime in Greece,analyzing specific crimes that affect society in numerous ways. Firstly,we have obtained administrative official data, regarding seven most common crimes for the period of time 1991-2016. The sets of the observations refer to the total Greece's population and the fourteen regions of the country respectively. There were severe breaks in the time series, hence we estimated the missing values in order to proceed to the analysis.

Autoregressive Moving Average (ARIMA) model theory for non stationary time series with trend, has been applied to the resulting stationarized time series, based on the "forecast"package in R and the function "auto.arima", in order to understand the data better and predict future points. Specifically, the Box-Jenkins methodology was used in the univariate time series analysis of the total population and for the regional data also.An approach of a relatively new method for count data, based on the "tscount" package of R, was implemented in the regional series regarding homicides and rapes, due to the small number of the observations in the districts of Greece. A multivariate analysis,with the Rpackage "MTS" and the function "VAR", has also been applied to the territorial series,so as to gain more robust results.The absence of a sufficient number of cross-correlations,led us to continue with the univariate methods.

The methods we have applied, provided us with results that managed to capture the autocorrelation in the series and model it in a sufficient extend.Undoubtedly,more extended time series would have provide better results,since data before 1991 were not listed in any database.The predicted models managed to explain adequately the time series and provided with predictions that can contribute to the comprehension of the crime profile.According to the forecasting models,we do not expect significant change in the number of crimes.The forecasts deteriorate with increasing horizons, hence more short run forecasts could perform better.

APPENDIX.

```
Forecast method: ARIMA(1,0,0) with non-zero mean
Model Information:
Series: homicides
ARIMA(1,0,0) with non-zero mean
Coefficients:
         ar1      mean
      0.7270  132.8364
s.e.  0.1403   13.7145
sigma^2 estimated as 455.8:  log likelihood=-115.82
AIC=237.63   AICc=238.72   BIC=241.4
Error measures:
                     ME      RMSE       MAE        MPE     MAPE       MASE
ACF1
Training set -0.206947 20.51243 16.96606 -2.724707 13.15689 0.9281217
0.147894
Forecasts:
   Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
27       95.15177 67.79063 122.5129 53.30652 136.9970
28      105.43998 71.61253 139.2674 53.70536 157.1746
29      112.91944 76.13068 149.7082 56.65589 169.1830
30      118.35694 80.09555 156.6183 59.84120 176.8727
31      122.30995 83.29270 161.3272 62.63822 181.9817
32      125.18377 85.77288 164.5947 64.91002 185.4575
33      127.27300 87.65566 166.8904 66.68350 187.8625
34      128.79186 89.06583 168.5179 68.03614 189.5476
35      129.89606 90.11271 169.6794 69.05267 190.7395
36      130.69880 90.88519 170.5124 69.80913 191.5885
37      131.28239 91.45279 171.1120 70.36827 192.1965
38      131.70666 91.86861 171.5447 70.77962 192.6337
39      132.01509 92.17258 171.8576 71.08123 192.9490
40      132.23932 92.39445 172.0842 71.30185 193.1768
41      132.40234 92.55622 172.2485 71.46296 193.3417
42      132.52085 92.67407 172.3676 71.58046 193.4612
43      132.60700 92.75988 172.4541 71.66609 193.5479
44      132.66964 92.82233 172.5169 71.72844 193.6108
45      132.71517 92.86777 172.5626 71.77382 193.6565
46      132.74827 92.90082 172.5957 71.80685 193.6897
47      132.77234 92.92486 172.6198 71.83087 193.7138
48      132.78984 92.94234 172.6373 71.84835 193.7313
49      132.80256 92.95505 172.6501 71.86105 193.7441
50      132.81180 92.96429 172.6593 71.87030 193.7533
51      132.81852 92.97101 172.6660 71.87701 193.7600
52      132.82341 92.97590 172.6709 71.88190 193.7649
53      132.82696 92.97945 172.6745 71.88545 193.7685
54      132.82955 92.98203 172.6771 71.88803 193.7711
55      132.83142 92.98391 172.6789 71.88991 193.7729
56      132.83279 92.98528 172.6803 71.89128 193.7743
```

```
Forecast method: ARIMA(0,1,1)
Model Information:
Series: burglary
ARIMA(0,1,1)
Coefficients:
         ma1
      0.4935
s.e.  0.1750
sigma^2 estimated as 45022318:  log likelihood=-255.39
AIC=514.77   AICc=515.32   BIC=517.21
Error measures:
                    ME      RMSE       MAE       MPE      MAPE       MASE
ACF1
Training set 844.1592 6446.632 4860.126 1.232984 7.797316 0.8092818
0.05154705
Forecasts:
   Point Forecast     Lo 80     Hi 80       Lo 95     Hi 95
27       74353.74 65754.700  82952.78  61202.6426  87504.84
28       74353.74 58897.730  89809.75  50715.8129  97991.67
29       74353.74 54258.144  94449.34  43620.1780 105087.30
30       74353.74 50504.679  98202.80  37879.7483 110827.73
31       74353.74 47266.428 101441.05  32927.2713 115780.21
32       74353.74 44375.962 104331.52  28506.6839 120200.80
33       74353.74 41740.676 106966.81  24476.3624 124231.12
34       74353.74 39302.967 109404.51  20748.2082 127959.27
35       74353.74 37024.108 111683.37  17262.9940 131444.49
36       74353.74 34876.580 113830.90  13978.6332 134728.85
37       74353.74 32839.996 115867.49  10863.9474 137843.53
38       74353.74 30898.756 117808.73   7895.0761 140812.41
39       74353.74 29040.603 119666.88   5053.2764 143654.20
40       74353.74 27255.702 121451.78   2323.5068 146383.97
41       74353.74 25536.019 123171.46   -306.5221 149014.00
42       74353.74 23874.887 124832.59  -2847.0047 151554.49
43       74353.74 22266.703 126440.78  -5306.5087 154013.99
44       74353.74 20706.707 128000.77  -7692.3170 156399.80
45       74353.74 19190.810 129516.67 -10010.6821 158718.16
46       74353.74 17715.470 130992.01 -12267.0194 160974.50
47       74353.74 16277.597 132429.88 -14466.0562 163173.54
48       74353.74 14874.474 133833.01 -16611.9483 165319.43
49       74353.74 13503.696 135203.79 -18708.3720 167415.85
50       74353.74 12163.125 136544.36 -20758.5985 169466.08
51       74353.74 10850.848 137856.63 -22765.5535 171473.03
52       74353.74  9565.145 139142.34 -24731.8665 173439.35
53       74353.74  8304.464 140403.02 -26659.9109 175367.39
54       74353.74  7067.400 141640.08 -28551.8376 177259.32
55       74353.74  5852.672 142854.81 -30409.6036 179117.08
56       74353.74  4659.113 144048.37 -32234.9949 180942.48
```


```
Forecast method: ARIMA(0,1,1)
Model Information:
Series: robbery
ARIMA(0,1,1)
Coefficients:
         ma1
      0.6039
s.e.  0.1667
sigma^2 estimated as 243441:  log likelihood=-190.22
AIC=384.45   AICc=384.99   BIC=386.88
Error measures:
                    ME      RMSE       MAE       MPE      MAPE       MASE
ACF1
Training set 86.13633 474.0408 366.4034 2.957163 13.14976 0.8037982
0.08430754
Forecasts:
   Point Forecast      Lo 80     Hi 80      Lo 95     Hi 95
27       4642.463 4010.14929  5274.777  3675.4225  5609.504
28       4642.463 3447.34877  5837.578  2814.6933  6470.233
29       4642.463 3075.05254  6209.874  2245.3154  7039.611
30       4642.463 2775.57973  6509.347  1787.3113  7497.615
31       4642.463 2517.90876  6767.018  1393.2375  7891.689
32       4642.463 2288.27347  6996.653  1042.0406  8242.886
33       4642.463 2079.12806  7205.799   722.1803  8562.746
34       4642.463 1885.80494  7399.122   426.5181  8858.409
35       4642.463 1705.17832  7579.748   150.2736  9134.653
36       4642.463 1535.03336  7749.893  -109.9407  9394.867
37       4642.463 1373.73287  7911.194  -356.6285  9641.555
38       4642.463 1220.02610  8064.901  -591.7027  9876.629
39       4642.463 1072.93194  8211.995  -816.6638 10101.591
40       4642.463  931.66394  8353.263 -1032.7146 10317.641
41       4642.463  795.58019  8489.346 -1240.8367 10525.763
42       4642.463  664.14865  8620.778 -1441.8439 10726.771
43       4642.463  536.92251  8748.004 -1636.4195 10921.346
44       4642.463  413.52218  8871.404 -1825.1440 11110.071
45       4642.463  293.62199  8991.305 -2008.5156 11293.442
46       4642.463  176.93998  9107.987 -2186.9653 11471.892
47       4642.463   63.23014  9221.697 -2360.8695 11645.796
48       4642.463  -47.72370  9332.650 -2530.5587 11815.485
49       4642.463 -156.11272  9441.039 -2696.3254 11981.252
50       4642.463 -262.10698  9547.034 -2858.4297 12143.356
51       4642.463 -365.85851  9650.785 -3017.1040 12302.031
52       4642.463 -467.50394  9752.431 -3172.5572 12457.484
53       4642.463 -567.16653  9852.093 -3324.9780 12609.905
54       4642.463 -664.95799  9949.885 -3474.5371 12759.464
55       4642.463 -760.97991 10045.907 -3621.3900 12906.317
56       4642.463 -855.32500 10140.252 -3765.6784 13050.605
```


```
Forecast method: ARIMA(0,1,0)
Model Information:
Series: motor.vehicletheft
ARIMA(0,1,0)
sigma^2 estimated as 6469939:  log likelihood=-231.51
AIC=465.01   AICc=465.19   BIC=466.23
Error measures:
                    ME     RMSE      MAE      MPE     MAPE      MASE      ACF1
Training set 677.5605 2494.212 1815.945 3.53557 8.723616 0.9617131 0.2078365
Forecasts:
   Point Forecast     Lo 80    Hi 80      Lo 95    Hi 95
27          26180 22920.236 29439.76 21194.6208 31165.38
28          26180 21569.997 30790.00 19129.6092 33230.39
29          26180 20533.923 31826.08 17545.0700 34814.93
30          26180 19660.472 32699.53 16209.2416 36150.76
31          26180 18890.946 33469.05 15032.3533 37327.65
32          26180 18195.241 34164.76 13968.3648 38391.64
33          26180 17555.475 34804.53 12989.9265 39370.07
34          26180 16959.994 35400.01 12079.2183 40280.78
35          26180 16400.707 35959.29 11223.8625 41136.14
36          26180 15871.720 36488.28 10414.8468 41945.15
37          26180 15368.585 36991.41  9645.3678 42714.63
38          26180 14887.845 37472.15  8910.1399 43449.86
39          26180 14426.753 37933.25  8204.9597 44155.04
40          26180 13983.079 38376.92  7526.4192 44833.58
41          26180 13554.987 38805.01  6871.7095 45488.29
42          26180 13140.943 39219.06  6238.4833 46121.52
43          26180 12739.648 39620.35  5624.7551 46735.24
44          26180 12349.992 40010.01  5028.8275 47331.17
45          26180 11971.017 40388.98  4449.2360 47910.76
46          26180 11601.891 40758.11  3884.7065 48475.29
47          26180 11241.884 41118.12  3334.1226 49025.88
48          26180 10890.350 41469.65  2796.4989 49563.50
49          26180 10546.720 41813.28  2270.9614 50089.04
50          26180 10210.482 42149.52  1756.7297 50603.27
51          26180  9881.179 42478.82  1253.1041 51106.90
52          26180  9558.399 42801.60   759.4543 51600.55
53          26180  9241.768 43118.23   275.2099 52084.79
54          26180  8930.949 43429.05  -200.1470 52560.15
55          26180  8625.632 43734.37  -667.0885 53027.09
56          26180  8325.536 44034.46 -1126.0463 53486.05
```


```
Forecast method: ARIMA(1,0,0) with non-zero mean
Model Information:
Series: rape
ARIMA(1,0,0) with non-zero mean
Coefficients:
         ar1      mean
      0.6503  171.8790
s.e.  0.1412    9.9281
sigma^2 estimated as 387.5:  log likelihood=-113.6
AIC=233.2   AICc=234.3   BIC=236.98
Error measures:
                     ME      RMSE       MAE        MPE     MAPE       MASE
ACF1
Training set -0.2439383 18.91222 15.37534 -1.389839 9.195387 0.8775881
0.004342113
Forecasts:
   Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
27       160.9023 135.6756 186.1289 122.3214 199.4831
28       164.7406 134.6487 194.8325 118.7191 210.7621
29       167.2367 135.3096 199.1639 118.4084 216.0651
30       168.8600 136.1877 201.5323 118.8921 218.8280
31       169.9157 136.9333 202.8981 119.4735 220.3579
32       170.6022 137.4896 203.7148 119.9608 221.2436
33       171.0487 137.8811 204.2162 120.3233 221.7741
34       171.3390 138.1482 204.5298 120.5781 222.0999
35       171.5278 138.3272 204.7284 120.7519 222.3037
36       171.6506 138.4459 204.8553 120.8684 222.4328
37       171.7304 138.5240 204.9369 120.9455 222.5154
38       171.7824 138.5752 204.9896 120.9963 222.5684
39       171.8161 138.6086 205.0237 121.0296 222.6027
40       171.8381 138.6304 205.0458 121.0514 222.6249
41       171.8524 138.6447 205.0601 121.0656 222.6392
42       171.8617 138.6539 205.0694 121.0748 222.6485
43       171.8677 138.6600 205.0755 121.0808 222.6546
44       171.8716 138.6639 205.0794 121.0848 222.6585
45       171.8742 138.6664 205.0820 121.0873 222.6611
46       171.8759 138.6681 205.0836 121.0890 222.6628
47       171.8769 138.6692 205.0847 121.0900 222.6638
48       171.8776 138.6699 205.0854 121.0908 222.6645
49       171.8781 138.6703 205.0859 121.0912 222.6650
50       171.8784 138.6706 205.0862 121.0915 222.6653
51       171.8786 138.6708 205.0864 121.0917 222.6655
52       171.8787 138.6710 205.0865 121.0918 222.6656
53       171.8788 138.6710 205.0866 121.0919 222.6657
54       171.8789 138.6711 205.0866 121.0920 222.6657
55       171.8789 138.6711 205.0867 121.0920 222.6658
56       171.8789 138.6711 205.0867 121.0920 222.6658
```

```
Forecast method: ARIMA(0,2,1)
Model Information:
Series: fraud
ARIMA(0,2,1)
Coefficients:
          ma1
      -0.7961
s.e.   0.1226
sigma^2 estimated as 65049:  log likelihood=-167.04
AIC=338.08   AICc=338.65   BIC=340.44
Error measures:
                    ME      RMSE       MAE       MPE      MAPE       MASE
ACF1
Training set 55.8123 239.8819 162.0677 3.159744 11.86791 0.8885293
-0.02399691
Forecasts:
   Point Forecast    Lo 80     Hi 80     Lo 95     Hi 95
27       4072.329 3745.472  4399.185 3572.4449  4572.213
28       4354.658 3843.120  4866.195 3572.3277  5136.987
29       4636.986 3948.954  5325.018 3584.7322  5689.240
30       4919.315 4052.790  5785.840 3594.0789  6244.551
31       5201.644 4151.412  6251.875 3595.4535  6807.834
32       5483.973 4243.558  6724.387 3586.9226  7381.023
33       5766.301 4328.709  7203.894 3567.6937  7964.909
34       6048.630 4406.681  7690.580 3537.4851  8559.775
35       6330.959 4477.452  8184.466 3496.2650  9165.653
36       6613.288 4541.086  8685.489 3444.1288  9782.446
37       6895.616 4597.688  9193.545 3381.2382 10409.995
38       7177.945 4647.385  9708.505 3307.7876 11048.103
39       7460.274 4690.314 10230.234 3223.9860 11696.562
40       7742.603 4726.615 10758.591 3130.0468 12355.159
41       8024.931 4756.425 11293.438 3026.1817 13023.681
42       8307.260 4779.880 11834.641 2912.5973 13701.923
43       8589.589 4797.110 12382.068 2789.4933 14389.685
44       8871.918 4808.242 12935.594 2657.0617 15086.774
45       9154.247 4813.395 13495.098 2515.4864 15793.007
46       9436.575 4812.684 14060.467 2364.9432 16508.207
47       9718.904 4806.219 14631.590 2205.5999 17232.208
48      10001.233 4794.104 15208.362 2037.6164 17964.849
49      10283.562 4776.440 15790.683 1861.1457 18705.977
50      10565.890 4753.322 16378.459 1676.3336 19455.447
51      10848.219 4724.840 16971.598 1483.3193 20213.119
52      11130.548 4691.083 17570.013 1282.2360 20978.860
53      11412.877 4652.133 18173.621 1073.2111 21752.542
54      11695.205 4608.069 18782.341  856.3663 22534.044
55      11977.534 4558.969 19396.099  631.8186 23323.250
56      12259.863 4504.906 20014.820  399.6798 24120.046
```

```
Forecast method: ARIMA(0,1,0)
Model Information:
Series: imp [1]
ARIMA(0,1,0)
sigma^2 estimated as 861999:  log likelihood=-206.31
AIC=414.62   AICc=414.8   BIC=415.84
Error measures:
                    ME      RMSE       MAE        MPE     MAPE       MASE
ACF1
Training set 102.496 910.4092 599.4303 0.6596396 6.00453 0.9620468
-0.2230865
Forecasts:
   Point Forecast    Lo 80    Hi 80     Lo 95    Hi 95
27          10892 9702.158 12081.84 9072.2932 12711.71
28          10892 9209.309 12574.69 8318.5460 13465.45
29          10892 8831.133 12952.87 7740.1754 14043.82
30          10892 8512.315 13271.68 7252.5865 14531.41
31          10892 8231.432 13552.57 6823.0120 14960.99
32          10892 7977.493 13806.51 6434.6469 15349.35
33          10892 7743.973 14040.03 6077.5084 15706.49
34          10892 7526.618 14257.38 5745.0920 16038.91
35          10892 7322.473 14461.53 5432.8797 16351.12
36          10892 7129.388 14654.61 5137.5820 16646.42
37          10892 6945.740 14838.26 4856.7154 16927.28
38          10892 6770.265 15013.73 4588.3509 17195.65
39          10892 6601.963 15182.04 4330.9540 17453.05
40          10892 6440.018 15343.98 4083.2807 17700.72
41          10892 6283.761 15500.24 3844.3060 17939.69
42          10892 6132.631 15651.37 3613.1729 18170.83
43          10892 5986.155 15797.85 3389.1568 18394.84
44          10892 5843.927 15940.07 3171.6380 18612.36
45          10892 5705.598 16078.40 2960.0821 18823.92
46          10892 5570.863 16213.14 2754.0240 19029.98
47          10892 5439.458 16344.54 2553.0560 19230.94
48          10892 5311.145 16472.86 2356.8187 19427.18
49          10892 5185.717 16598.28 2164.9929 19619.01
50          10892 5062.987 16721.01 1977.2939 19806.71
51          10892 4942.789 16841.21 1793.4662 19990.53
52          10892 4824.971 16959.03 1613.2797 20170.72
53          10892 4709.398 17074.60 1436.5263 20347.47
54          10892 4595.946 17188.05 1263.0169 20520.98
55          10892 4484.503 17299.50 1092.5792 20691.42
56          10892 4374.965 17409.03  925.0556 20858.94
```

1 : variable 'imp' : imputed values for the 'law about drugs' variable.

REFERENCES.

1. EUROSTAT. Crime and Criminal Justice Statistics - Methodological guide for users 2017 Version (updated May 2017).

2. EUROSTAT. EU guidelines for the International Classification of Crime for Statistical Purposes 2017 edition.

3. EUROSTAT. https://ec.europa.eu/eurostat/data/database

4. Ministry of Citizen Protection,Hellenic Police http://www.astynomia.gr/index.php

5. Peter J. Brockwell Richard A. Davis (2002). Introduction to Time Series and Forecasting Second Edition.

6. James D. Hamilton (1994). Time Series Analysis.

7. Δημέλη Σ. (2013), Σύγχρονες Μέθοδοι Ανάλυσης Χρονολογικών Σειρών, Εκδόσεις ΟΠΑ,Αθήνα

8. Tobias Liboschik, Konstantinos Fokianos, Roland Fried (2017). tscount: An R Package for Analysis of Count Time Series Following Generalized Linear Models. Journal of Statistical Software (November 2017), Volume 82, Issue 5.

9. Ruey S. Tsay (May 2013). Multivariate Time Series Analysis in R. Booth School of Business, University of Chicago

10. Ruey S. Tsay and David Wood (October 2018 ,Version1.0), All-Purpose Toolkit for Analyzing Multivariate Time Series (MTS) and Estimating Multivariate Volatility Models.


UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ