UNIVERSITY OF PIRAEUS

School of Finance and Statistics

Department of Statistics and Insurance Science

POSTGRADUATE PROGRAM IN

APPLIED STATISTICS

BIOSURVEILLANCE USING CONTROL CHARTS AND SCAN STATISTICS.

By

Mouzaki Georgia

MSc Dissertation

submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfilment of the requirements for the degree of Master of Science in Applied Statistics

Piraeus, Greece

November 2015

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ

ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

ΕΠΙΔΗΜΙΟΛΟΓΙΚΗ ΕΠΙΒΛΕΨΗ ΜΕ ΧΡΗΣΗ ΔΙΑΓΡΑΜΜΑΤΩΝ ΕΛΕΓΧΟΥ ΚΑΙ ΣΤΑΤΙΣΤΙΚΩΝ ΣΑΡΩΣΗΣ

Γεωργία Μουζάκη

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην *Εφαρμοσμένη Στατιστική*

Πειραιάς

Ιανουάριος 2015

To my family

Acknowledgements

Foremost, I would like to express the deepest appreciation to my advisor, S Bersimis for his continuous support, his patience and encouragement. His guidance helped me in all the time of research and writing of this thesis. Besides my advisor, I would like to thank the rest of my thesis committee, G. Tzabelas and E. Smirnakis, for their insightful comments and helpful suggestions. I would also like to thank my professors at Master of Applied Statistics who helped me enormously with their constant feedback throughout these two years. Special thanks to Dr Magda Gavana for providing the data that formed the basis of this thesis as well as for enlightening me the first glance of research. I would also like to thank my friends and my fellows for all the fun we had during the period of our studies and all the sleepless nights we were working together before deadlines. Last but not least, I owe more than thanks to my family for their support and encouragement through life.

Abstract

In the present study, is described a monitoring system which is developed to identify unusually large increases in time series of infectious disease counts (outbreak detection) compared to the expected number of cases. The designed two-phase monitoring system consists of (i) successful integration of count time series following generalized linear models, in order to provide dynamic forecasting of future expected disease counts and (ii) using SPC methods for tracking the forecast errors from the fitted models. Analysis for this study was illustrated on time-dependent count data which reflect the total number of infectious diseases from January 2005 through December 2012 in different geographical areas in Greece and were reported weekly to the Hellenic Center for Disease Control and Prevention (HCDCP). The systematic methodology that is developed, is capable of detecting aberrations in infectious disease patterns, facilitating a timely public health response and it can be generalized to other healthcare settings.

Περίληψη

Σκοπός της παρούσας εργασίας είναι η ανάπτυξη ενός συστήματος παρακολούθησης ικανό να ανιχνεύσει ασυνήθιστα μεγάλα ξεσπάσματα σε χρονοσειρές από κρούσματα λοιμωδών ασθενειών συγκριτικά με τον αναμενόμενο αριθμό κρουσμάτων. Το σχεδιασμένο σε δύο φάσεις σύστημα επιτήρησης αποτελείται από (i) προσαρμογή Γενικευμένων Γραμμικών Μοντέλων στις χρονοσειρές με σκοπό να εξασφαλίσει δυναμικές προβλέψεις των μελλοντικών αναμενόμενων κρουσμάτων και (ii) τη χρήση διαγραμμάτων ελέγχου του Στατιστικού Ελέγχου Ποιότητας για την παρακολούθηση των σφαλμάτων των προβλέψεων, όπως προκύπτουν από τα προσαρμοσμένα μοντέλα. Η ανάλυση της παρούσας εργασίας εφαρμόστηκε σε χρονικά εξαρτώμενες μετρήσεις, που αντικατοπτρίζουν το συνολικό αριθμό κρουσμάτων λοιμωδών ασθενειών που έχουν καταγραφεί σε διαφορετικές γεωγραφικές περιοχές στην Ελλάδα την περίοδο από τον Ιανουάριο του 2015 μέχρι το Δεκέμβριο του 2012 και εκθέτονται σε εβδομαδιαία βάση στο Κέντρο Ελέγχου και Πρόληψης Νοσημάτων (ΚΕΕΛΠΝΟ). Η μεθοδολογία που αναπτύχθηκε είναι ικανή να εντοπίζει αποκλίσεις από το μοτίβο που ακολουθούν οι λοιμώδεις ασθένειες συμβάλλοντας έτσι στην έγκαιρη ανταπόκριση της δημόσιας υγείας, καθώς επίσης μπορεί να γενικευτεί και σε άλλες εφαρμογές της φροντίδας υγείας.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The principal objective of Biosurveillance is the practice of monitoring data for the purpose of detecting disease outbreaks. In the last several years, there has been a shift towards surveillance systems that would provide early detection of diseases. Data collected from public health surveillance systems can provide important clues to the cause of diseases as well as assist in identification of important risk factors and detection of unusual observations in reports of infectious diseases or other conditions. Timely detection of disease outbreaks facilitates early public health response to minimize undue morbidity and mortality.

In monitoring, models that will be used should take into account that observations are nonnegative integers, while most of these applications involve relatively rare events which makes the use of the normal distribution questionable. Additionally, models should be designed to capture suitably the dependence among observations during time. Neglecting either of these two characteristics would lead to potentially serious misspecification.

Statistical process control (SPC) charts are among the most prevalent and valid methods for monitoring time series data in disease surveillance. The charts are very effective when monitored data meet the requirements of temporal independence, stationarity, and normality (i.i.d. random variables). However, when these assumptions are violated, the SPC charts will either fail to detect special cause variations or will alert frequently even in the absence of anomalies. Currently collected biosurveillance data contain predictable factors such as day-of-week effects, seasonal effects, holidays, autocorrelation, and global trends that cause the data to violate these assumptions.

In the present study, we will focus on monitoring time series of infectious disease counts. More specifically, the objective of this study is to develop a monitoring system to detect aberrations in reported data of infectious disease counts and provide a signal to alert public health practitioners to undertake timely public health action. In the designed monitoring system, the count time series were modeled with generalized linear models methodology, using likelihood-based estimation methods and the Poisson or Negative Binomial conditional distribution. The autocorrelative structure in the data is captured by the fitted model, and the subsequent forecast errors that are produced, can be monitored by the SPC charts. Thus, the objective of the designed monitoring system is not to track week-to-week changes in the

number of reported cases, but rather to identify any substantial deviation between the weekly observed numbers of reported cases and the corresponding expected ones, as determined from historical patterns. In other words, the control charts employ statistical limits to generate flags identifying deviations from historical data patterns and from the underlying stochastic process generating the observations (detect aberrations).

Analysis for this study was demonstrated by count time series which reflect the total number of infectious diseases from January 2005 through December 2012 and were recorded in different Primary Health Care Centers in Greece.

The present study is organized as follows. In chapter 2 the challenges of public health surveillance and some distinctive features of infectious diseases are discussed. Additionally, a more detailed description of the steps of the designed two-phase monitoring system is given and finally, a great number of models of time series count data which have been proposed in literature are reviewed, as well as approaches to control count processes using modified SPC control charts. In chapter 3 is included the theoretical background of the methods and models used for Phase I analysis which consists of model description, parameter estimation and model assessment. Chapter 4 introduces the methods used on the Phase II analysis of the monitoring system which correspond to forecasting and controlling forecast errors using Statistical Process Control (SPC) charts. Chapter 5 demonstrates the usage of the designed monitoring system and the methods described in Chapter 3 & Chapter 4, with real data examples of weekly infectious diseases reported cases in different Primary Health Care Centers during 2005-2012. Chapter 6 presents the concluding remarks and further research directions.

Chapter 2

Motivation of the study

2.1. Challenges of Public Health Surveillance.

Public Health Surveillance is defined as "the ongoing systematic collection, analysis, interpretation and dissemination of outcome-specific data used for planning, implementing, and evaluating public health practice"[2]. The establishment of the European Centre of Disease Control and Prevention (ECDC) and of the Greek Centre of Disease Control and Prevention (KEELPNO) support research programs to collect, review and disseminate health data at both European and national level.

Analysis of public health surveillance data aims to detect departures from historical patterns of disease frequency, in order to enable timely public health responses to decrease unnecessary morbidity and mortality. These data are generally collected in time sequence, at regular intervals and in an ongoing manner, thus, they often exhibit correlation, non-stationarity (in the mean and/or variance) and seasonality. Due to these special features of public health surveillance data and including the fact that data collection processes are distributed, where errors and delays are more likely to occur, detection of changes in public health data presents an analytic challenge.

An effective surveillance system has the following functions (World Health Organization[30]):

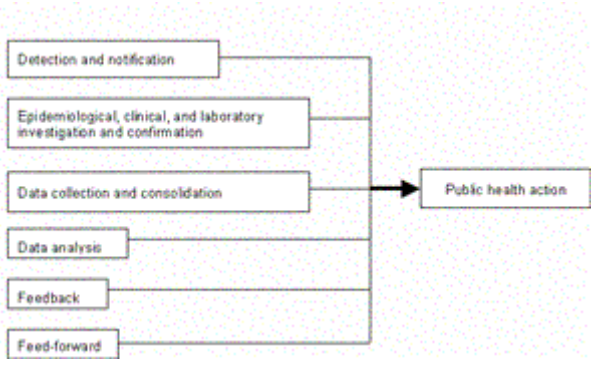


Figure 2.1. Functions of surveillance system.

- detection and notification of health events
- collection and consolidation of pertinent data
- investigation and confirmation (epidemiological, clinical and/or laboratory) of cases or outbreaks
- routine analysis and creation of reports
- feedback of information to those providing the data
- feed-forward (i.e. the forwarding of data to more central levels)
- reporting data to the next administrative level

It is more than clear that public health surveillance involves much more difficult tasks than industrial Statistical Process Control (SPC), where the measurement system and data collection processes are often validated and standardized upfront, thus, variability is reduced in the measurement process and SPC control charts are most frequently used with a narrower focus on a specific aspect of a manufacturing process. More specifically, in industrial quality control, and thus SPC, it is often reasonable to assume that:

- Statistical process control charts use requires observations to be independently and identically distributed (IID) random variables. Since the manufacturing process is controlled from the collecting samples stage, the in-control distribution is stationary and observations are independent so can be drawn from the process.
- Monitoring the process mean and standard deviation is usually sufficient.
- The asymptotic distributions of the statistics being monitored are known and thus can be used to design appropriate control charts.
- Shifts, when they occur, remain until they are detected and corrective action taken.
- Temporal detection is the critical problem.

However, in public health surveillance are violated many, if not all, of these assumptions:

- There is little to no control over disease incidence and thus the distribution of these data is usually non-stationary and observations (often daily or weekly counts) are autocorrelated.
- There is little information about what types of statistics are useful for monitoring—one is often looking for anything that seems unusual.
- Since individual observations are being monitored, the idea of asymptotic sampling distributions does not apply, and the data often contain significant systematic effects that must be accounted for.
- Outbreaks are transient, with disease incidence returning to its original state once an outbreak has run its course.
- Both spatial and temporal deviations are often critical.

Thus, in order to make timely, appropriate decisions, it is incumbent for public health practitioners to use all available data and choose analysis tools wisely, by taking into account all these special features of public health surveillance data.

2.2. Infectious diseases.

According to the World Health Organization, there is an infectious-disease crisis of global proportions. Infectious diseases are the largest killers of children and young adults, accounting for more than 13 million deaths per year (World Health Organization 1999b) and they remain a critical public health issue, as their epidemiology is changing, along with the burden they impose on humanity. While the changing, globalized world has brought significant advances for the prevention and control of diseases, it has increased the opportunities for emergence and spread of infectious diseases across the world. This is the result of dramatic increases in the volume and speed of international travel and trade in recent years. Thus, infectious-disease challenges have now become broader and more complex. On the other hand, budgetary and other constraints as a consequence of global economic crisis have in turn had a major impact on public health, requiring difficult decisions at the national, state, and local levels. In order to ensure that human health will not be affected negatively due to weakened public-health capacities, broad and well-coordinated collaborative efforts are required to determine the best use of limited resources. As a part of emerging healthcare decision problems, many researchers have studied how to detect and contain disease outbreaks, and our research is aligned with this trend.

Several of the infectious diseases' patterns (like the morbidity patterns of acute respiratory infections and influenza-like illness which are two of the most common infectious diseases) indicate climate sensitivity and seasonality since highest peaks of outbreaks occur in the late fall, winter and early spring. This disease pattern may result from increased likelihood of transmission due to indirect social or behavioral adaptations to the cold weather such as crowding indoors. Another possibility is that it may be attributed directly to pathogen sensitivities to climatic factors such as humidity.

The survival of the pathogenic organism outside a host depends on the characteristics of the environment, particularly temperature, humidity, exposure to sunlight, pH and salinity. Annual variation in climate can therefore result in annual or more complex peaks in disease incidence, depending on the influence of climatic variables, such as rainfall or cloud cover on the environment. This relationship depends on the type of environment (e.g. sewage, aerosol, droplets, etc.), and hence route of transmission.

2.3. Analysis Strategy

From the above mentioned, we can conclude that it is crucial to be developed an early warning system, capable of identifying public health emergencies, understand/monitor the epidemiology of a condition in order to set priorities and guide public health policy and strategies.

In this study, a two phase monitoring system was developed for controlling time series of infectious disease counts reported on a weekly basis. This monitoring system will take into account two challenging aspects of the data, the integer-valued and the autocorrelative structure of the data.

In the Phase I analysis a standard approach is to fit an appropriate model for count autocorrelated processes which will capture suitably the dependence among observations. A natural modification of the popular autoregressive moving average (ARMA) models for continuous variables is based on the assumption that the observation Y_t at time t is generated by a generalized linear model (GLM) conditionally on the past, choosing an adequate distribution for count data like the Poisson or Negative Binomial and the identity link function. Regression models for time series of counts are developed by adapting a latent process similar to the case of the ordinary generalized autoregressive conditional heteroscedasticity (GARCH). Likelihood based estimation methods are performed for model fitting.

The historical data will be used for fitting the model and techniques of model assessment will be analyzed briefly in the next Chapter. The response residuals of the fitted model will be monitored with an EWMA SPC chart in order to determine the control limits for the phase II analysis of the forecast errors.

In the phase II analysis, the successfully estimated glm fitted model for the historical data will be used in order to produce one-step-ahead predictions for the rest of the time series data. The new available observations will be used by re-estimating the model coefficients every time after the addition of each week's real observation. The forecast errors $e_t = Y_t - \hat{Y}_t$ arise from the difference between observed and expected from the model values. If the model has successfully captured the autocorrelation of the observations, then the forecast errors will be stationary and independent values with a mean close to zero. Stationarity and independence of observations over time are frequently reasonable assumptions in SPC upon

which control charting was first based. Therefore, to obtain these features in cases of autocorrelated data a standard approach is to plot the forecast errors on a control chart. The control charts will generate flags identifying high deviations from historical data patterns. More specifically, our interest is confined for aberrations beyond the upper limit of the control chart which correspond to observed outbreaks much higher than expected. The analysis will be implemented thoroughtly in Chapter 5 by using real data examples of reported infectious diseases events from 7 Primary Health Care Centers in different prefectures in Greece.

2.4. Literature overview.

Since recently there has been an increasing interest in modeling time series count of data, a considerable number of different approaches have been proposed in the literature. An important class of models for time series of counts is based on the thinning operator, like the integer autoregressive moving average (INARMA) models, which, in a way, imitate the structure of the common autoregressive moving average (ARMA) models (Weiß 2007[7], Weiß 2009[8]). A first attempt towards this direction was introduced by Alzaid and Al-Osh (1988)[9] as well as McKenzie (1985) who surveys various models based on ”binomial thinning”. In those models, the dependent variable y_t is assumed to be equal to the sum of an error term with some pre-specified distribution and the result of y_{t-1} draws from a Bernoulli which takes value 1 with some probability ρ and 0 otherwise. This guarantees that the dependent variable takes only integer values. The parameter ρ in that model is analogous to the coefficient on the lagged value in an AR(1) model. This model, called INAR(1), has the same autocorrelations as the AR(1) model of traditional time series analysis, which makes it its discrete counterpart. This family of models has been generalized to include integer valued ARMA processes as well as to incorporate exogenous regressors. The problem with this type of models is the difficulty in estimating them. Many models have been proposed and the emphasis was put more on their stochastic properties than on how to estimate them.

Discrete Autoregressive Moving Average (DARMA) models introduced by Jacobs and Lewis (1978a)[10], are models for time series count data with properties similar to those of ARMA processes found in traditional time series analysis. They are obtained by a probabilistic mixture of a sequence of independent identically distributed (i.i.d.) discrete random values.

random variables. One of the problems associated with these models seems to be the difficulty of estimating them as well. Franke and Seligmann (1993)[11] have used self excited threshold autoregressive (SETAR) models to describe the number of seizures in epileptic patients.

Markov chains are an alternative way of dealing with count data in time series. The method consists of defining transition probabilities between all the possible values that the dependent variable can take and determining, in the same way as in usual time series analysis, the appropriate order for the series. A prominent area of application for Markov chains is binary data. As soon as the number of values that the dependent variable takes gets too large, these models lose tractability. As a consequence, this method is only reasonable when there are very few possible values that the observations can take. One can also refer to MacDonald and Zucchini (1997)[12] in regard to the importance of discrete-valued time series.

At the present study, modeling of time series count data is primarily influenced by the generalized linear models (GLM) theory, see McCullagh and Nelder[13] for independent data and Kedem and Fokianos (2002)[28] for dependent data. An important special case of these models is the INGARCH model as described by Ferland R, Latour A, Oraichi D (2006)[14] which is an integer-valued analogue of the classical generalized autoregressive conditional heteroskedastic (GARCH) (p,q) model with Poisson deviates.

Christian H. Weiß (2007)[24] suggest approaches to control Poisson count processed from the class of INARMA models concentrating on the INAR(1) model. One of the approaches proposed, (Christian H. Weiß (2009)[25] is monitoring correlated processes of Poisson counts using a combination of the c – and a EWMA chart. Sung Won Han1 et al. (2009)[26] investigate and compare the performance of temporal scan statistics, CUSUM and EWMA when the observations follow the Poisson distribution. Using a simulation study, they showed that the Poisson CUSUM and EWMA charts generally outperformed the Poisson scan statistic methods, while EWMA charts outperformed the CUSUM charts in situations with a small shift and an early change in time and CUSUM charts were superior in dealing with a large shift with a later change in time. Finally, Christian H. Weiß & Murat Caner Testik (2015)[23] described thoroughly the Phase I analysis for monitoring time dependent data stemming from Poisson INAR(1) processes and proposed solutions to modify the method of moments, least squares and maximum likelihood for parameter estimation.

Chapter 3

Phase I Analysis

3.1. Time series and seasonality.

Many statistical methods relate to data which are independent, or at least uncorrelated. However, there are many practical situations where data might be correlated which occurs particularly when observations on a given system are gathered sequentially in time. These kind of data are called time series. A time series model provides a description of the random nature of the process that generated the sample of observations under study. The description is given not in terms of a cause-and-effect relationship, as would be the case in a regression model, but in terms of how that randomness is embodied in the process.

Seasonality is just a cyclical behavior that occurs on a regular calendar basis. Recognition of seasonality is important because it provides information about regularity in the series that can aid us in making a forecast. In a weekly time series observations Y_t exhibit annual seasonality, the data points in the series should show some degree of correlation with the corresponding data points which lead or lag by 52 weeks. In other words, we expect to see some degree of correlation between Y_t and Y_{t-52}.

3.2. Modelling count time series following generalized linear models.

Denote a count time series by $\{Y_t : t \in N\}$. The conditional mean $E(Y_t \mid F_{t-1})$ of the count time series is modeled by a latent mean process $\{ \lambda_t : t \in N \}$, such that $E(Y_t \mid F_{t-1}) = \lambda_t$. Denote by F_t the history of the joint process $\{Y_t, \lambda_t : t \in N\}$ up to time *t*. The models we are interested in, are of the general form:

$$g(\lambda_t) = \beta_0 + \sum_{k=1}^{p} \beta_k \tilde{g}(Y_{t-i_k}) + \sum_{l=1}^{q} a_l g(\lambda_{t-j_l}), \quad (1)$$

where $g : \mathbb{R}^+ \to \mathbb{R}$ is a link function and $\tilde{g} : \mathbb{R}^+ \to \mathbb{R}$ is a transformation function.

A useful extension of the model would be to include internal and/or external covariate effects, which does not consist part of the objective of the current study. An internal covariate effect propagates to future observations both by the regression on past observations and by the regression on past latent means, while an external covariate effect propagates to future

observations both by the regression on past observations but not directly by the regression on past latent means.

In the terminology of GLMs $\nu_t = g(\lambda_t)$ is called 'the linear predictor'. To allow for regression on arbitrary past observations of the response, define a set $P = \{i_1,i_2,\ldots,i_p\}$with $p \in \mathbb{N}_0$ and integers $0 < i_1 < i_{2\ldots} < i_q < \infty$. This enables us to regress on the lagged observations $Y_{t-i_1}, Y_{t-i_2},\ldots, Y_{t-i_p}$. Analogously, define a set $Q = \{j_1,j_2,\ldots,j_p\}$ with $q \in \mathbb{N}_0$ and integers $0 < j_1 < j_{2\ldots} < j_q < \infty$ for regression on lagged latent means $\lambda_{t-j_1}, \lambda_{t-j_2},\ldots, \lambda_{t-j_q}$. This more general case is covered by the theory for models with P={1,…,p}and Q ={1,…,q}, which are usually treated in the literature, by choosing p and q sufficiently large and setting unnecessary model parameters to zero.

If the link function g equals the identity as in present study, then $g(x) = \tilde{g}(x) = x$. In this case, considering that P={1,…,p} and Q ={1,…,q} model (1) takes the form:

$$\lambda_t = \beta_0 + \sum_{k=1}^{p} \beta_k Y_{t-k} + \sum_{l=1}^{q} a_l \lambda_{t-l} \quad (2)$$

The conditional distribution of the models can be Poisson or Negative Binomial.

Poisson assumption.

If Y_t given the past is Poisson distributed $Y_t|F_{t-1} \sim \text{Poisson}(\lambda_t)$, then we obtain an integer-valued GARCH model of order p and q, in short INGARCH(p,q). In that case,

$$P(Y_t = y|F_{t-1}) = \frac{\lambda_t^y \exp(-\lambda_t)}{y!}, \; y = 0, 1, \ldots$$

and the latent mean process is identical to the conditional variance of the observed process

$$E(Y_t|F_{t-1}) = VAR(Y_t|F_{t-1}) = \lambda_t .$$

INGARCH (p,q) is an integer-valued process $\{X_t : t \in \mathbb{Z}\}$ analogue of the generalized autoregressive conditional heteroskedastic (GARCH) (p,q) model with Poisson deviates (instead of the normal deviates) such that:

$$\begin{cases} Xt|Ft-1 \sim \text{Poisson}(\lambda t), & \forall t \in Z \\ \lambda t = \gamma_0 + \sum_{i=1}^{q} \gamma_i X_{t-i} + \sum_{j=1}^{p} \delta_j \lambda_{t-j} \end{cases}$$

where $\gamma_0 > 0$, $\gamma_i \geq 0$, $\delta_j \geq 0$, i= 1,...,q, j = 1,...,p.

The key insight of GARCH lies in the distinction between conditional and unconditional variances of the innovation process X_t. The term conditional implies explicit dependence on a past sequence of observations. The term unconditional is more concerned with long-term behavior of a time series and assumes no explicit knowledge of the past. The various GARCH models characterize the conditional distribution of X_t by imposing alternative parameterizations to capture serial dependence on the conditional variance of the innovations. The general GARCH (p,q) model for the conditional variance of innovations X_t is:

$$\begin{cases} \mathrm{Xt}|\mathrm{Ft}-1 \sim \mathrm{N}(0,\sigma_t^2), & \forall \mathrm{t} \in \mathrm{Z} \\ \sigma_t^2 = \gamma_0 + \sum_{i=1}^{q} \gamma_i {X_{t-i}}^2 + \sum_{j=1}^{p} \delta_j \sigma_{t-j}^2 \end{cases}$$

where $\gamma_0 > 0$, $\gamma_i \geq 0$, $\delta_j \geq 0$, i= 1,...,q, j = 1,...,p.

As an INGARCH(p,q) involves Poisson values, the conditional mean which happens to be also the conditional variance, depends on the past values of the series as well as on its own past values.

Negative Binomial assumption.

The Negative Binomial Distribution allows for a conditional variance larger than λ_t and is parameterized in terms of its mean with an additional dispersion parameter $\varphi \in (0, \infty)$.

Assuming that $Y_t|F_{t-1} \sim$ NegBin (λ_t, φ), and

$$\mathrm{P}(Y_t = \mathrm{y}|\mathrm{F}_{t-1}) = \frac{\Gamma(\varphi+y)}{\Gamma(y+1)\Gamma(\varphi)} \left(\frac{\varphi}{\varphi+\lambda_t}\right)^{\varphi} \left(\frac{\lambda_t}{\varphi+\lambda_t}\right)^{y}, \qquad \mathrm{y} = 0, 1, \ldots$$

the conditional variance increases quadratically with λ_t:

$$VAR(Y_t|F_{t-1}) = \lambda_t + (\lambda_t)^2/\varphi = \lambda_t + (\lambda_t)^2\sigma^2 ,$$

where $\sigma^2 = 1/\varphi$ is the overdispersion coefficient as it proportional to the extent of overdispersion of the conditional distribution. The limiting case of $\sigma^2 = 0$ corresponds to the Poisson distribution (no overdispersion). The Negative Binomial Distribution belongs to the class of mixed Poisson processes. The estimation procedure that we be used in the present study is not restricted to the Negative Binomial case, but to any mixed Poisson distribution. The Negative Binomial assumption is required though, for prediction intervals and model assessment.

Parameter Estimation.

For the parameter estimation will be followed the quasi conditional maximum likelihood (QML) estimation procedure. If the Poisson assumption holds true, then an ordinary ML estimator is obtained and under the mixed Poisson assumption a QML estimator is obtained.

The least squares method is used for estimating linear and nonlinear regressions and results in an approximation to the conditional mean function of the dependent variable. This approach is important and common in practice. However, this approach is still limited as it is not suitable for analyzing the dependent variables with special features, as it is unable to take data characteristics into account and it has little room to characterize other conditional moments, such as conditional variance, of the dependent variable. As far as a complete description of the conditional behavior of a dependent variable is concerned, it is desirable to specify a density function that admits specifications of different conditional moments and other distribution characteristics. The method of quasi-maximum likelihood (QML) is essentially the same as the ML method usually seen in statistics and econometrics textbooks. Maximum likelihood under normality is widely used in applications. The quasi-maximum likelihood estimator (QMLE) however, is applicable in a general class of dynamic models when a normal log-likelihood is maximized but the normality assumption is violated. It is conceivable though, that specifying a density function, while being more general and more flexible than specifying a function for conditional mean, is more likely to result in specification errors.

Regardless of the distributional assumption the parameter space for model (2) is given by

$$\Theta = \left\{ \theta \in \mathbb{R}^{p+q+1} : \beta_0 > 0, \beta_k \geq 0, \alpha_l \geq 0 \,, \quad \sum_{k=1}^{p} \beta_k + \sum_{l=1}^{q} \alpha_l < 1 \right\}$$

The intercept β_0 must be positive and all other parameters must be nonnegative to ensure positivity of the latent mean process. The other condition ensures that the fitted model has a stationary solution.

With the parameterization of the Negative Binomial distribution, the estimation of the regression parameters θ does not depend on the additional dispersion parameter φ. This allows to employ a quasi maximum likelihood approach based on the Poisson likelihood to estimate the regression parameters θ as it is described below. The nuisance parameter φ is then estimated separately in a second step.

The log-likelihood, score vector and information matrix are derived conditionally on pre-sample values of the time series and the latent mean process λ_t. Given the observations $y = (y_1,\ldots,y_n)$, the conditional quasi log-likelihood function, up to a constant, is given by

$$l(\theta) = \sum_{t=1}^{n} log\,(p_t\,(y_t\,;\theta)) \propto \sum_{t=1}^{n}(y_t \ln \lambda_t(\theta)) - \lambda_t(\theta)),$$

where $p_t\,(y;\theta) = P(Y_t = y|F_{t-1})$ is the p.d.f. of a Poisson distribution.

The quasi maximum likelihood (QML) estimator $\hat{\theta}_n$ of θ is the solution of the non-linear constrained optimization problem

$$\hat{\theta}_n = argmax_{\theta\in\Theta} l(\theta).$$

The latent mean process is regarded as a function $\lambda_t : \Theta \rightarrow \mathbb{R}^+$ and thus it is denoted by $\lambda_t(\theta)$ for all t. The conditional score function is the (p + q + 1) - dimensional vector given by

$$S_n(\theta) = \frac{\partial l(\theta)}{\partial \theta} = \sum_{t=1}^{n}\left(\frac{y_t}{\lambda_t(\theta)} - 1\right)\frac{\partial \lambda_t(\theta)}{\partial \theta}.$$

Finally, the conditional information matrix is given by

$$G_n\,(\theta;\,\sigma^2) = \sum_{t=1}^{n} COV\left(\frac{\partial l(\theta;Y_t)}{\partial \theta}\middle| F_{t-1}\right)$$
$$= \sum_{t=1}^{n}\left(\frac{1}{\lambda_t(\theta)} + \sigma^2\right)\left(\frac{\partial \lambda_t(\theta)}{\partial \theta}\right)\left(\frac{\partial \lambda_t(\theta)}{\partial \theta}\right)^{T}.$$

In the case of Poisson distribution it holds $\sigma^2=0$, $(G_n\,(\theta;\,0))$ and in the case of Negative Binomial distribution $\sigma^2 = 1/\varphi$. The dispersion parameter φ of the Negative Binomial distribution is estimated by solving the equation

$$\sum_{t=1}^{n}\frac{(Y_t-\hat{\lambda}_t)^2}{\hat{\lambda}_t(1+\hat{\lambda}_t/\hat{\varphi})} = n - m,$$

which is based on Pearson's χ^2 statistic. The variance parameter σ^2 is estimated by $\hat{\sigma}^2 = 1/\hat{\varphi}$. Inference for the regression parameters is based on the asymptotic normality of the QML

estimator, which has been shown by Christou V, Fokianos K (2014)[16].

3.3. Model Assessment.

In order to value the fit and predictive performance of the model used in analysis, diagnostic approaches originally developed for generalized linear models as well as for time series are proposed to be utilized. The observations of weekly counts during the years 2005-2011 are used for fitting the model and also for assessing the obtained fit. Within the class of count time series following generalized linear models it is desirable to assess the specification of the linear predictor and the conditional distribution.

Given the fitted values $\hat{\lambda}_t = \lambda_t(\hat{\theta})$ there are various types of residuals that can be used. For example, the most frequently used *Response* (or raw) residuals:

$$r_t = y_t - \hat{\lambda}_t ,$$

the standardized *Pearson* residuals:

$$r_t^P = \frac{(y_t - \hat{\lambda}t)}{\sqrt{\hat{\lambda}_t + \hat{\lambda}_t^2 \hat{\sigma}^2}} ,$$

or the more symmetrically distributed *Anscombe* residuals:

$$r_t^A = \frac{3\hat{\sigma}^2\left(\left(1+y_t/\hat{\sigma}^2\right)^{2/3} - \left(1+\hat{\lambda}_t/\hat{\sigma}^2\right)^{2/3}\right) + 3\left(y_t^{2/3} - \hat{\lambda}_t^{2/3}\right)}{2\left(\hat{\lambda}_t^2/\hat{\sigma}^2 + \hat{\lambda}_t\right)^{1/6}}.$$

The empirical autocorrelation function, ACF of residuals is a useful tool in order to test if they exhibit any remaining serial correlation or seasonality which has not been explained by the fitted model.

A plot of the residuals against time can reveal changes of the data generating process over time. A plot of squared (response) residuals r_t^2 against the corresponding fitted values $\hat{\lambda}_t$ can demonstrate 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.

Denote by $P_t(y) = P(Y_t \leq y|F_{t-1})$ the c.d.f., by $p_t(y) = P(Y_t = y|F_{t-1})$ the p.d.f., and by σ_t the standard deviation of the predictive distribution. Tools which follow the prequential principle, depending only on the realized observations and their respective

forecast distributions have been proposed for assessing the predictive performance for continuous data by (Gneiting et al. 2007)[17] and transferred to independent but not identically distributed count data by Czado, Gneiting, and Held (2009)[18].

Probabilistic forecasts of continuous variables take the form of predictive densities or predictive cumulative distribution functions. A diagnostic approach for evaluating the predictive performance that is based on maximizing the sharpness of the predictive distributions subject to calibration. Calibration refers to the statistical consistency between the distributional forecasts and the observations and is a joint property of the predictions and the events that materialize. Sharpness refers to the concentration of the predictive distributions, it is only a property of the forecasts and can be measured by the width of prediction intervals.

First we consider probabilistic forecasts (as opposed to point forecasts) of continuous and mixed discrete–continuous variables. In this situation, probabilistic forecasts take the form of predictive densities or predictive cumulative distribution functions (CDFs), and the diagnostic approach faces a challenge, in that the forecasts take the form of probability distributions whereas the observations are real valued.

Nature chooses a distribution G_t, which is the true data-generating process, and the forecaster picks a probabilistic forecast in the form of a predictive CDF, F_t. The outcome x_t is a random number with distribution G_t. We assume that that the forecaster's basis of information is at most that of nature. Hence, the forecaster would be ideal in the case of

$$F_t = G_t , \quad \text{for all t.}$$

In practice, the true distribution G_t remains hypothetical, and the predictive distribution F_t is an expert opinion that may or may not derive from a statistical prediction algorithm.

The predictive distributions need to be assessed only on the basis of the forecast–observation pairs (F_t, x_t). For implementing this, the use of the probability integral transform

$$\text{(PIT) value } p_t = F_t(x_t) ,$$

is proposed, which will follow a uniform distribution if the forecasts are ideal and the predictive distribution is correct.

For count data, a non-randomized PIT value for the observed value y_t and the predictive distribution $P_t(y)$ is defined by

$$\mathrm{Ft(u|y)} = \begin{cases} 0, & u \le P_t(y-1) \\ \frac{u - P_t(y-1)}{P_t(y) - P_t(y-1)} & , \; P_t(y-1) < u < P_t(y) \\ 1, & u \ge P_t(y) \end{cases}$$

The mean PIT is then given by

$$\bar{F}(\mathrm{u}) = \frac{1}{n}\sum_{t=1}^{n} F_t(u|y_t), \quad 0 \le u \le 1.$$

To check whether $\bar{F}(\mathrm{u})$ is the c.d.f. of a uniform distribution Czado et al.(2009)[18] propose plotting a histogram with H bins, where bin h has the height

$$f_i = \bar{F}(h/H) - \bar{F}((h-1)/H), h = 1, \dots, H.$$

By default H is chosen to be 10. A U-shape indicates underdispersion of the predictive distribution, whereas an upside down U-shape indicates overdispersion. Gneiting et al. (2007) point out that the empirical coverage of central, e.g., 90% prediction intervals can be read off the PIT histogram as the area under the 90% central bins.

Another tool that can be used to test if the predictions are appropriate is the marginal calibration which is defined as the difference of the average predictive c.d.f. and the empirical c.d.f. of the observations,

$$\frac{1}{n}\sum_{t=1}^{n} P_t(y) - \frac{1}{n}\sum_{t=1}^{n} 1(y_t \le y), \quad \forall y \in \mathbb{R}$$

By plotting the marginal calibration for y values, we can see whether the marginal distribution of the predictions resembles the marginal distribution of the observations, if their difference is close to zero. Major deviations from zero point at model deficiencies.

The uniformity of the PIT is a necessary condition for the forecaster to be ideal, although not sufficient. The PIT cannot distinguish the ideal forecaster between competitors. To address these limitations, a diagnostic approach is proposed to the evaluation of predictive performance that is based on maximizing the sharpness of the predictive distributions subject to calibration. The more concentrated the predictive distributions are, the sharper the forecasts, and subject to sufficient calibration, the sharper the better.

Proper scoring rules address calibration as well as sharpness and allow us to rank competing forecast procedures. Scoring rules assign numerical scores to probabilistic forecasts and form attractive summary measures of predictive performance, in that they address calibration and sharpness simultaneously. Denote a score for the predictive distribution P_t and the observation y_t by $s(P_t, y_t)$. A number of possible proper scoring rules is given in Table 3.1. The mean score for each corresponding model is given by

$$\sum_{t=1}^{n} s(P_t, y_t)/n$$

The model with the lowest score is preferable. Each of the different proper scoring rules captures different characteristics of the predictive distribution and its distance to the observed data.

Table 3.1: Different proper scoring rules, $s(P_t, y_t)$ and their definitions.

Scoring Rule	Definition
logarithmic score	$-log(p_t(y_t))$
quadratic score	$-2p_t(y_t) + \|\| p_t \|\|^2$
spherical score	$-p_t(y_t) / \|\| p_t \|\|$
ranked probability score	$\sum_{y=0}^{\infty}(P_t(y) - \mathbf{1}(y_t \leq y))^2$
Dawid-Sebastiani score	$(y_t - \lambda_t)2 /\sigma_t^2 + 2\, log(\sigma_t)$
normalized squared error score	$(y_t - \lambda_t)2 /\sigma_t^2$
squared error score	$(y_t - \lambda_t)^2$

- where $|| pt ||^2 = \sum_{y=0}^{\infty} p_t(y)^2$.

Chapter 4

Phase II Analysis

4.1. Forecasting.

In the Phase II of the analysis, future observations will be predicted based on the fitted GLM-type model (2) for time series of counts that was previously described. The conditional distribution of Y_{n+1} can be either a Poisson, either a Negative Binomial distribution with mean λ_{n+1}. In terms of the mean square error, the optimal predictor $\hat{Y}_{n+1}$ for Y_{n+1}, given the past F_n of the process up to time n, is the conditional expectation λ_{n+1} as was described in model (1).

An h-step-ahead prediction $\hat{Y}_{n+h}$ for $Y_{n+\ h}$ is obtained by recursive one-step-ahead predictions, where unobserved values $Y_{n+1},...,Y_{n+h-1}$ are replaced by their corresponding one-step-ahead prediction. The distribution of this h-step-ahead prediction $\hat{Y}_{n+h}$ is not known analytically but can be approximated numerically by simulation. The conditional expectation λ_{n+1} is substituted by its estimator $\hat{\lambda}_{n+1} = \lambda_{n+1}(\hat{\theta})$, which depends on the estimated model parameters $\hat{\theta}$. The dispersion parameter φ of the Negative Binomial distribution is replaced by its estimator $\hat{\varphi}$. Prediction intervals for $Y_{n+\ h}$ with a given coverage rate 1 - α are designed to cover the true observation Y_{n+h} with a probability of 1 - α. Simultaneous prediction intervals achieving a global coverage rate for $Y_{n+1},...,Y_{n+h}$ can be obtained by a Bonferroni adjustment of the individual coverage rates to 1- α/h each. The prediction intervals are based on B simulations of realizations $y_{(b)n+1},\ldots\ y_{(b)n+h}$ from the fitted model, b = 1,...,B. To obtain an approximate prediction interval for Y_{n+h} one can either use the empirical (α/2) – and (1- α/2) – quantile of the B simulations of y_{n+h} or find the shortest interval which contains at least [(1-α) B] of these observations.

In the present study however, the h-step-ahead predictions are computed as recursive 1-step-ahead predictions given all previous values, which are observations of the original time series. To be more specific, one-step-ahead (or else, one-week-ahead), rolling forecasts were developed for each of the 52 weeks of 2012 by retaining the same model form at each forecasting step, but re-estimating the model coefficients after the addition of each week's original value. As a result, by incorporating as much data as were available in the estimation

step, we ensure producing a more valid forecast. However, it should be pointed out that this could only be relevant if more than one observation ahead is to be predicted (h > 1).

4.2. Statistical Process Control (SPC) Methods

Statistical Process Control (SPC) is a group of tools and techniques used to determine the stability and predictability of a process. Graphical depictions of process output are plotted on Control Charts. Control charts are one of the most commonly used methods of SPC, which monitors the stability of a process. The main features of a control chart include the data points, a centerline (mean value), and upper and lower limits (bounds to indicate where a process output is considered "out of control").They visually display the fluctuations of a particular process variable and tests whether these variations fall within the specified process limits. The first Control Charts were developed by Walter Shewhart at Bell Labs in the 1920's. At this time, telephone technology was in its infancy with poor reliability. Shewhart used SPC to study variation and reduce special causes of failure which resulted in substantial increase of quality and reliability in phone service.

Statistical process control is frequently associated with the application of charting techniques for detecting shifts in mean or variability of a process. A process may either be classified as in control or out of control. The boundaries for these classifications are set by calculating the mean, standard deviation, and range of a set of process data collected when the process is under stable operation. Then, subsequent data can be compared to this already calculated mean, standard deviation and range to determine whether the new data fall within acceptable bounds. If a point falls beyond these critical boundaries, it is a signal that the process is statistically out of control, or that a statistical aberration has been identified. For good and safe control, subsequent data collected should fall within three standard deviations of the mean. Control charts build on this basic idea of statistical analysis by plotting the mean or range of subsequent data against time. Although calculation of control chart statistics is relatively easy, it is sometimes difficult to determine the most effective control charts and appropriate control limits for the specific monitoring problem.

SPC encompasses a much broader scope of activities including the design of sampling and inspection schemes, Pareto analysis experimental design and multivariate analysis. A basic assumption in traditional application of Statistical Process Control techniques is that the

observations from the processes under investigation are normally and independently distributed. The initial predictions for the process must be made while the process is assumed to be stable. Since future process quality will be compared to these predictions, they must be based off of a data set that is taken while the operation is running properly. Moreover, if data are not initially carefully and systematically recorded, especially at the point of manufacture or operation, they cannot be analyzed and put to use. Information recorded in a suitable way enables the magnitude of variations and trends to be observed. This allows conclusions to be drawn concerning errors, process capability, vendor ratings, risks, etc.

Control limits.

Shewhart found that control limits placed at three standard deviations from the mean in either direction provide an economical tradeoff between the risk of reacting to a false signal and the risk of not reacting to a true signal - regardless the shape of the underlying process distribution.

If the process has a normal distribution, 99.7% of the population is captured by the curve at three standard deviations from the mean. Stated another way, there is only a 1-99.7%, or 0.3% chance of finding a value beyond 3 standard deviations. Therefore, a measurement value beyond 3 standard deviations indicates that the process has either shifted or become unstable (more variability).

Assess Control.

After establishing control limits, the next step is to assess whether or not the process is in control (statistically stable over time). This determination is made by observing the plot point patterns and applying six simple rules to identify an out-of-control condition.

1. If one or more points falls outside of the upper control limit (UCL), or lower control limit (LCL). The UCL and LCL are three standard deviations on either side of the mean - see section A of the illustration below.

2. If two out of three successive points fall in the area that is beyond two standard deviations from the mean, either above or below - see section B of the illustration below.
3. If four out of five successive points fall in the area that is beyond one standard deviation from the mean, either above or below - see section C of the illustration below.
4. If there is a run of six or more points that are all either successively higher or successively lower - see section D of the illustration below.
5. If eight or more points fall on either side of the mean (some organization use 7 points, some 9)
6. If 15 points in a row fall within the area on either side of the mean that is one standard deviation from the mean - see section F of the illustration below.

Statistical Process Control is based on the analysis of data, so the first step is to decide what data to collect. There are two categories of control chart distinguished by the type of data used: Variable or Attribute. Variable data comes from measurements on a continuous scale, such as: temperature, time, distance, weight. Attribute data is based on upon discrete distinctions such as good/bad, percentage defective, or number defective per hundred.

Two phase monitoring system.

In the literature, two distinct phases of control charting practice have been discussed, Woodall, W. H. and Montgomery, D. C. (1999)[29]. In Phase I, charts are used for retrospectively testing whether the process was in-control when the first subgroups were being drawn. In this phase, the charts are used as aids to the practitioner, in bringing a process into a state of statistical control. Once this is accomplished, the control chart is used to define what is meant by statistical control. This is referred as the retrospective use of control charts. In general, there is a lot more going on in this phase than just charting some data. During this phase the practitioner is studying the process very intensively. The data collected are then analyzed in an attempt to determine whether the data were collected from an in-control process.

In Phase II, control charts are used for testing whether the process remains in-control when future subgroups are drawn. In this phase, the charts are used for monitoring the process for

any change from an in-control state. At each sampling stage, should be evaluated if the state of the process has changed. The meaning of in-control, in this phase, is usually determined by the values of the process parameters e.g., the mean and standard deviation for univariate continuously distributed variables. The values of the parameters are either given to the practitioner or they are estimated from the historical data known to be under control from Phase I. Note that in this phase the data is not taken as being from an in-control process unless the data provide evidence against no change in the process. Using these data to define what is meant by the process being in-control might lead to use an out-of-control process to define a state of statistical in-control. Woodall (2000) states that much work, process understanding and process improvement is often required in the transition from Phase I to Phase II.

Average Run Length

When dealing with control charts, the two main objectives are:

1. How often will there be false alarms where we look for an assignable cause but nothing has changed. When a process is in-control, we want our chart to signal (false alarm) infrequently.
2. How quickly will be detected certain kinds of systematic changes, such as mean shifts. When a process is out-of-control, we want the chart to signal as soon as possible. In statistical terms we want the probability of the statistic computed to plot in-control if we are out-of-control to be as small as possible.

The most known measure for evaluating the performance of a chart, concerning the previous two objectives, is the average run length (ARL), which is based on the run length (RL) distribution. The number of observations, or samples, needed for a control chart to signal is a run length or alternatively one observation of the RL distribution. The mean of the RL distribution is the ARL, which is actually the average number of successive control charts points that will be plotted before we detect a point beyond the control limits. The average run length (ARL) is a typical method of comparing control charts. A measure similar to the ARL is the average time to signal (ATS), which is the average time needed for a control chart to signal. The ARL for Shewhart charts is given by

$$ARL = 1/ P(\text{a point plots outside the control limits}).$$

Exponentially weighted moving average (EWMA) control charts

In the proposed monitoring system, we considered the exponentially weighted moving average (EWMA) control charts. The Exponentially Weighted Moving Average (EWMA) was introduced by Roberts (1959) and it is used as the CUSUM chart to detect persistent shifts in a variable. It is a statistic for monitoring the process that averages the data in a way that gives less and less weight to data as they are further removed in time, by smoothing a series of values based on a moving average with weights which decay exponentially. It is a type of moving mean chart in which an 'exponentially weighted mean' is calculated each time a new result becomes available. These charts perform well in situations with small shift and an early change in time thus, they can be used to detect small and permanent variation on the mean of the process. Its performance is generally similar to the performance of the CUSUM chart and its ability is to signal faster than the Shewhart charts for small and moderate shifts but not that fast for large shifts.

The Exponentially Weighted Moving Average (EWMA) chart depends on the smoothing parameter λ, which controls the weighting scheme applied. By default it is commonly set at approximately 0.2, but it can be modified and determined subjectively. The greater the value of λ, the smaller is the influence of the data in the more distant past. EWMA combines historical data to give less weight to data as they get older:

$$Z_i = \lambda x_i + (1 - \lambda) Z_{i-1} ,$$

where Z_i is the statistic at time i (the new EWMA), Z_{i-1} is the statistic at time i-1 (the old EWMA), λ is the EWMA weighting parameter and x_i is the observed value at time t (the new observation) and Z_0 is the initial value.

The EWMA control chart can be designed to resemble the performance of the Shewhart control chart by the selection of λ in the interval (0, 1). Since the EWMA is a weighted average of observations, it is less sensitive to the normality assumption and, therefore, provides more flexibility in its application to monitoring problems. when λ = 1 it is actually the $\bar{X}$ chart. As a starting value, instead of the in-control process mean, we can use the target value. The control limits of this chart are:

$$\text{UCL} = \mu + L\frac{\sigma}{\sqrt{n}}\sqrt{\left(\frac{\lambda}{2-\lambda}\right)[1-(1-\lambda)^{2i}]}$$

$$\text{LCL} = \mu - L\frac{\sigma}{\sqrt{n}}\sqrt{\left(\frac{\lambda}{2-\lambda}\right)[1-(1-\lambda)^{2i}]} \quad ,$$

where L is a constant used to specify the width of the control limits, μ is the mean of the process and $\frac{\sigma}{\sqrt{n}}\sqrt{\left(\frac{\lambda}{2-\lambda}\right)[1-(1-\lambda)^{2i}]}$ the standard deviation of Z_i when the process is in-control.

In case the EWMA chart is used for some time, instead of the above control limits we may use their limiting values

$$\text{UCL} = \mu + L\frac{\sigma}{\sqrt{n}}\sqrt{\left(\frac{\lambda}{2-\lambda}\right)}$$

$$\text{LCL} = \mu - L\frac{\sigma}{\sqrt{n}}\sqrt{\left(\frac{\lambda}{2-\lambda}\right)}$$

The main features of the EWMA chart are the same as the ones for the CUSUM except of the optimality. The computation of its run length distribution and the ARL can be done by the exact way using integral equations. The ARL L(u) of a two-sided EWMA chart for the mean given that the EWMA starts at u is computed through the relation

$$\text{L(u)} = 1 + \frac{1}{\lambda}\int_{-h}^{h} f\left(\frac{y-(1-\lambda)^{u}}{\lambda}\right)\text{L(y) dy.}$$

where y_i are assumed to be independent, identically distributed observations with probability density function f( ), h is the upper control limit and -h the lower control limit. This can be explained as follows; if for the first observation y_1, we have that

$|(1-\lambda u) + \lambda y_1| > h$, then we have a signal. On the other hand, if this relation does not hold, the run length continues to move from $(1-\lambda u) + \lambda y_1$ and $L((1-\lambda u) + \lambda y_1)$ stands for the additional run length. The approximation method of the Markov chain is the other alternative. The ARL in this case is computed by

$$ARL = (\boldsymbol{I} - \boldsymbol{R})^{-1}\,\mathbf{1},$$

where **I** is the identity matrix, **1** is a vector of unities and **R** is a submatrix of the transition probability matrix P, where

$$\mathrm{P} = \begin{bmatrix} \boldsymbol{R} & (\mathbf{I}-\mathbf{R})\mathbf{1} \\ 0^{T} & 1 \end{bmatrix}.$$

Chapter 5

Analysis implementation

5.1. Data

Analysis for this study was performed on data from the Hellenic Center for Disease Control and Prevention (HCDCP / KEELPNO). HCDCP is a private law entity established with Law 2071/92 and has operated since 1992. It is supervised and funded directly by the Ministry of Health and Social Solidarity. Data were collected in the frame of the Doctoral thesis of Magda Gavana[1], the study of which, had as main purpose to design, implement and evaluate a network for the epidemiological surveillance of infectious diseases in Primary Health Care. The sampling frame of the study was the sentinel network constituted by General Practitioners and Pediatricians working in the National Health System (ESY) primarily in agricultural areas of Greece. Data used in the analysis implementation, are weekly counts of infectious diseases which were collected from January 2005 through December 2012. The infectious diseases that were recorded in the sentinel network were selected according to their impact on public health and namely they are: acute respiratory infections, influenza like syndrome, chickenpox, herpes zoster, gastroenteritis, whooping cough, measles, mumps and rubella.

5.2 Missing values imputation

Incomplete datasets may lead to results that are different from those that would have been obtained from a complete dataset. The major problems that may arise when dealing with incomplete data is loss of information and efficiency, bias due to systematic differences between observed and unobserved data, as well as several complications related to data handling, computation and analysis, due to the irregularities in data structure. In the time series used in analysis the percentage of missing values was approximately 5% - 11%. Since data used in analysis are count time series with dependence among observations, Interpolation Technique was preferred to fill the missing values. This is a safe choice, in order not to distort the autocorrelation structure of the series. With interpolation we produce missing numbers

which are in between of the two already existing values. For example, we interpolate the missing count of diseases for week 31, using the counts of weeks 30 and 32.

The simplest form of interpolation is to connect two data points with a straight line. This technique is called linear interpolation.

The equation of the linear interpolation function is (Chapra and Canale, 1998)[4]:

$$f(x) = f(x_0) + \frac{f(x_1) - f(x_0)}{x_1 - x_0}(x - x_0),$$

where x is the independent variable, x_1 and x_o are known values of the independent variable and f(x) is the value of the dependent variable for a value x of the independent variable.

In order not to violate the integer value structure of the time series we filled the missing values with the rounded interpolated values.

5.3. Phase I analysis.

The monitoring system as was described in previous chapters, will be applied to the time series of weekly reported cases of infectious diseases in the National Primary Health Care Center of Kiato, prefecture of Korinthia, Greece, between 2005 and 2012. The analysis was performed using the R software package.

Number of infectious disease cases reported every week (2005-2012), shown in Figure 5.1., indicate to exhibit cyclic seasonal patterns, as well as the empirical autocorrelation function in Figure 5.2. shows clearly dependence among observations during time.

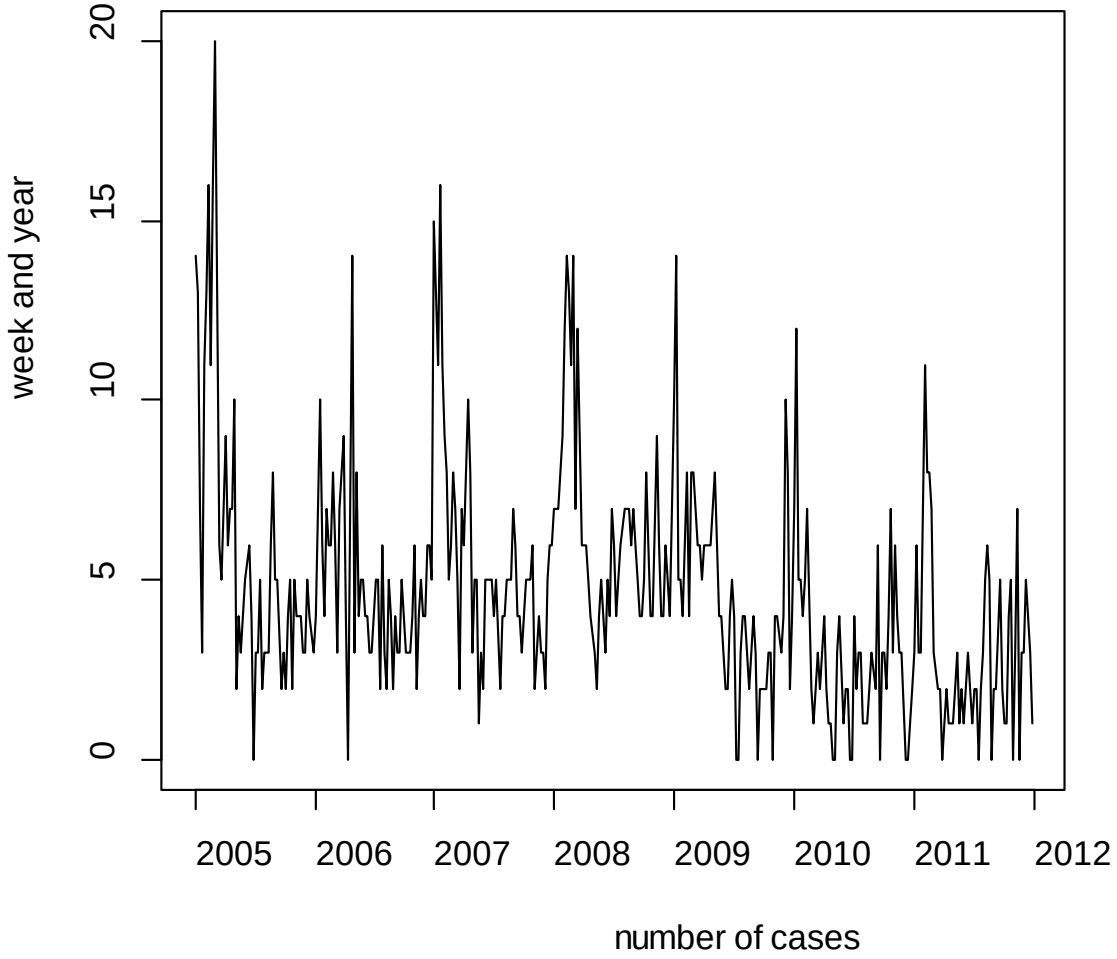


Figure 5.1. Number of infectious disease cases reported every week (2005-2012) in Primary Health Care Center of Kiato.

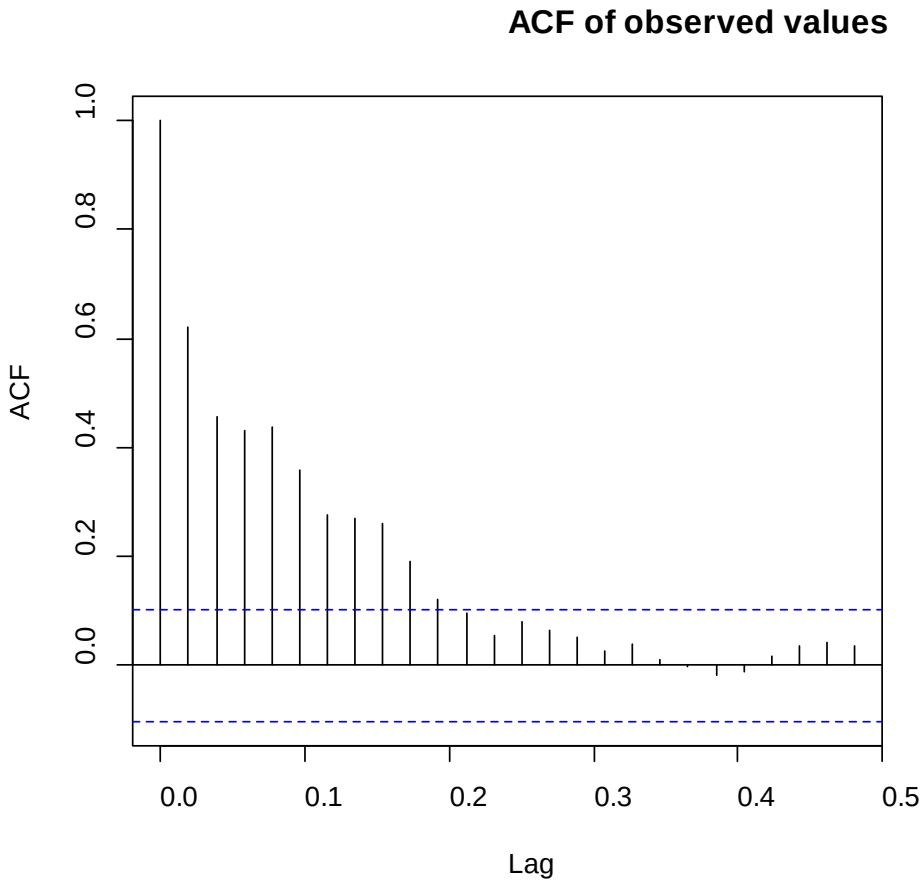


Figure 5.2. Autocorrelation function of observed values.

In Figure 5.3. the reported cases are plotted for each year (2005-2012) separately. Plots indicate that infectious diseases seem to follow the expected seasonality, with higher rates mainly in winter and spring (December to May) and reduced activity during the summer and autumn months.

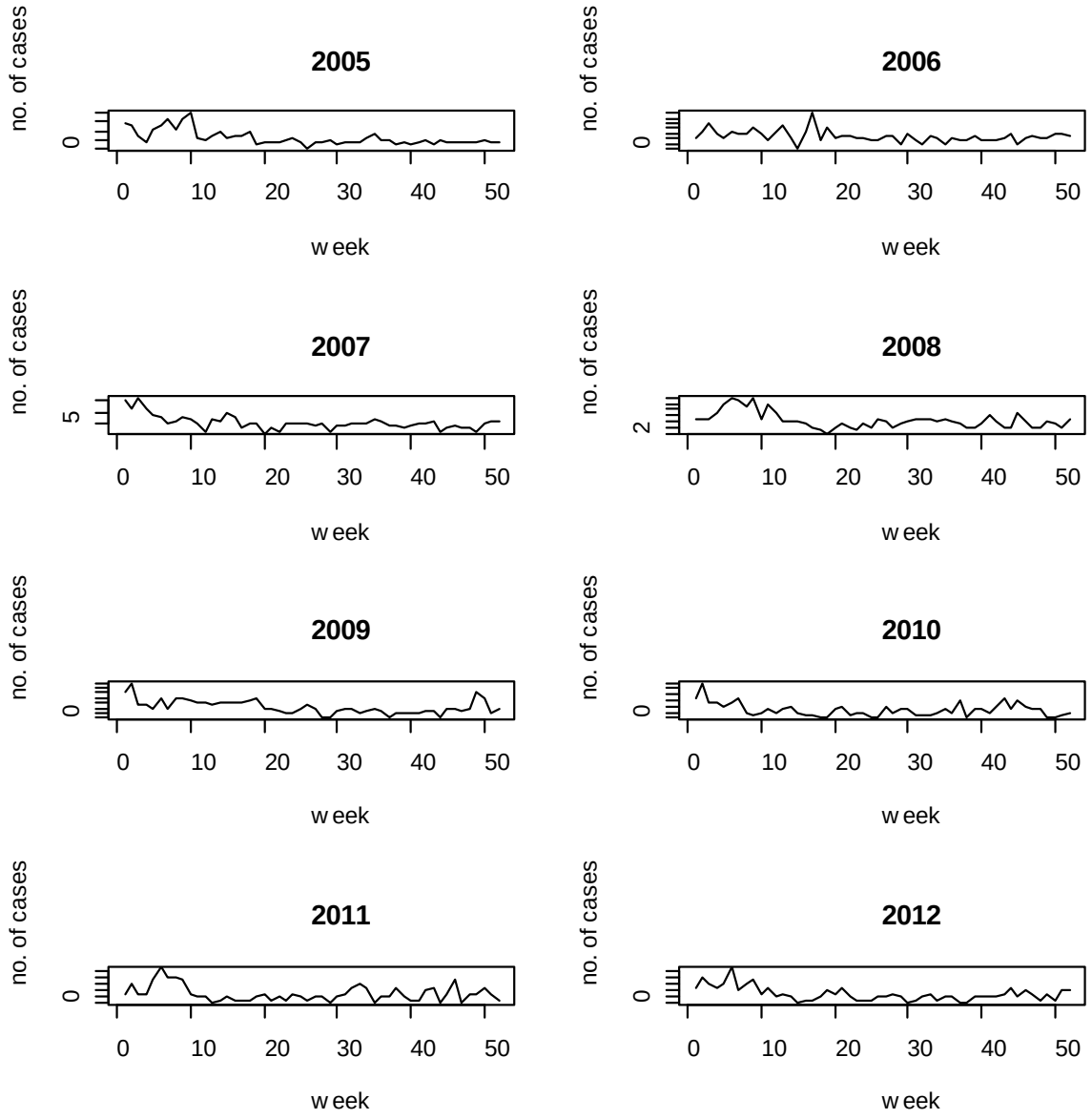


Figure 5.3. Number of infectious disease cases reported every week for each year separately (2005-2012).

Specifically, the highest peak (outbreak) in reported cases occurred in April for the year 2006 and between the months January-March for the rest of the years (Table 5.1.).

Table 5.1. Maximum number of reported cases per year.

Year	Week	Max
2005	10	20
2006	17	14
2007	3	16
2008	6 & 9	14
2009	2	14
2010	2	12
2011	6	11
2012	6	11

The overall morbidity during the study period appears increased in winter and spring, following the corresponding morbidity of acute respiratory infections and influenza-like illness, which comprises the biggest part of it.

The histogram in Figure 5.4. reveals that the counts are very small with more than 65% less than 6 and therefore, methods for count data are to be preferred. The general mean is 4.48 and the variance is 9.55.

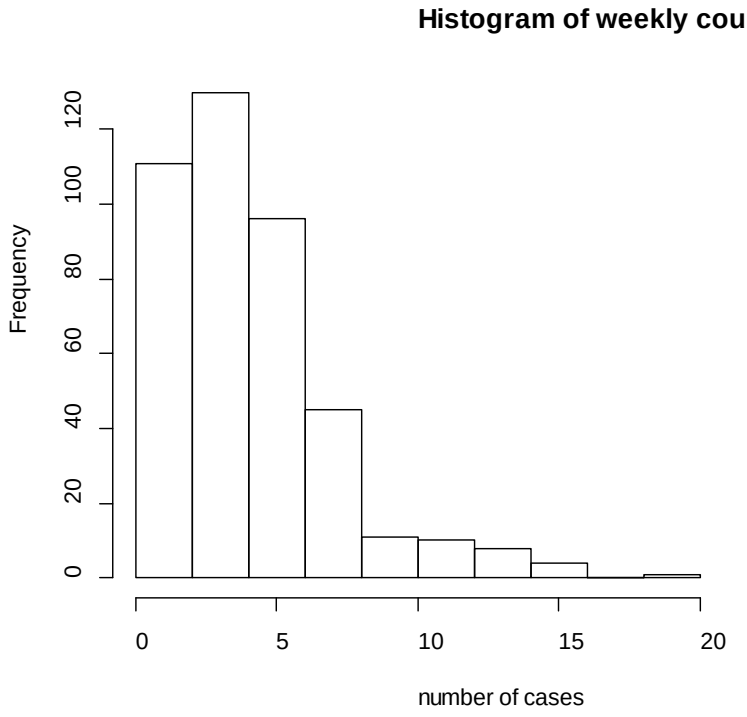


Figure 5.4. Histogram of infectious disease reported cases (2005-2012).

For the Phase I analysis, was developed a generalized linear model (GLM) using the training data from January (week 1) of 2005 through December (week 52) of 2011, in order to produce forecasted weekly values for all 52 weeks of 2012. Model was fitted using the R Package 'tscount'[19].

Model (2) was fitted with the identity link function. For taking into account short range serial dependence a first order autoregressive term was included, while seasonality is captured by regressing on the unobserved conditional mean 52 weeks (one year) back in time (52^{nd} order autoregressive term). Observations were modeled conditionally on the past information using both, a Poisson and a Negative Binomial conditional distribution in order to compare their fitness.

The response residuals, which are shown in Figure 5.5. are identical for the two conditional distributions and they seem to be independently distributed. Figure 5.6. indicates that the empirical autocorrelation function of response residuals does not exhibit any significant remaining serial correlation or seasonality which is not described by the models.

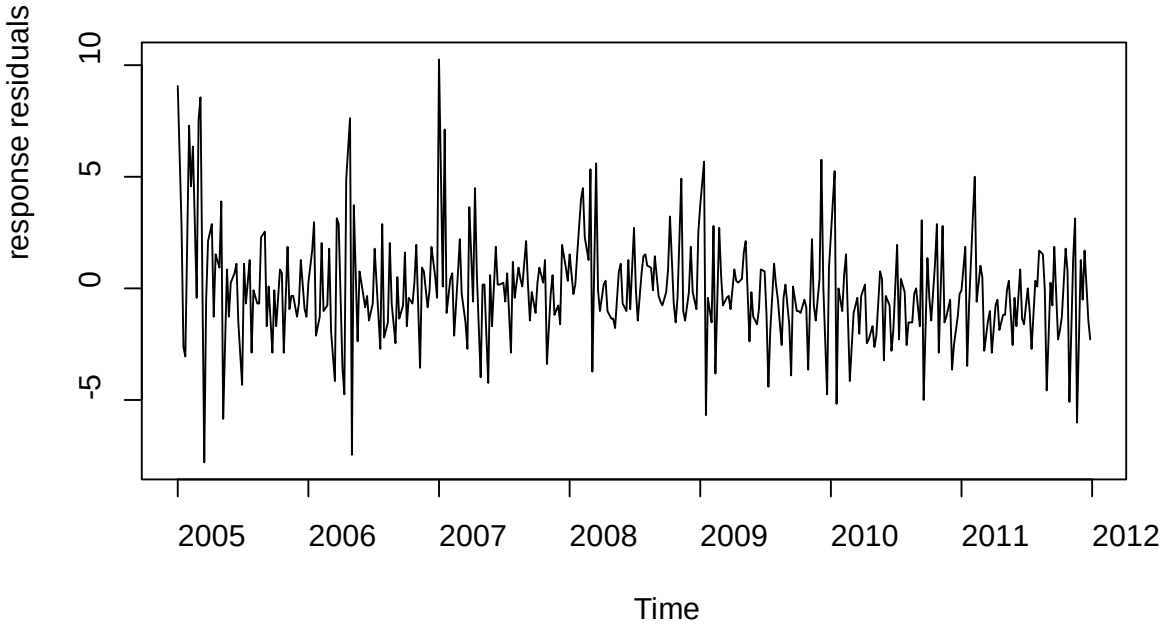


Figure 5.5. Response residuals of the fitted models over time.

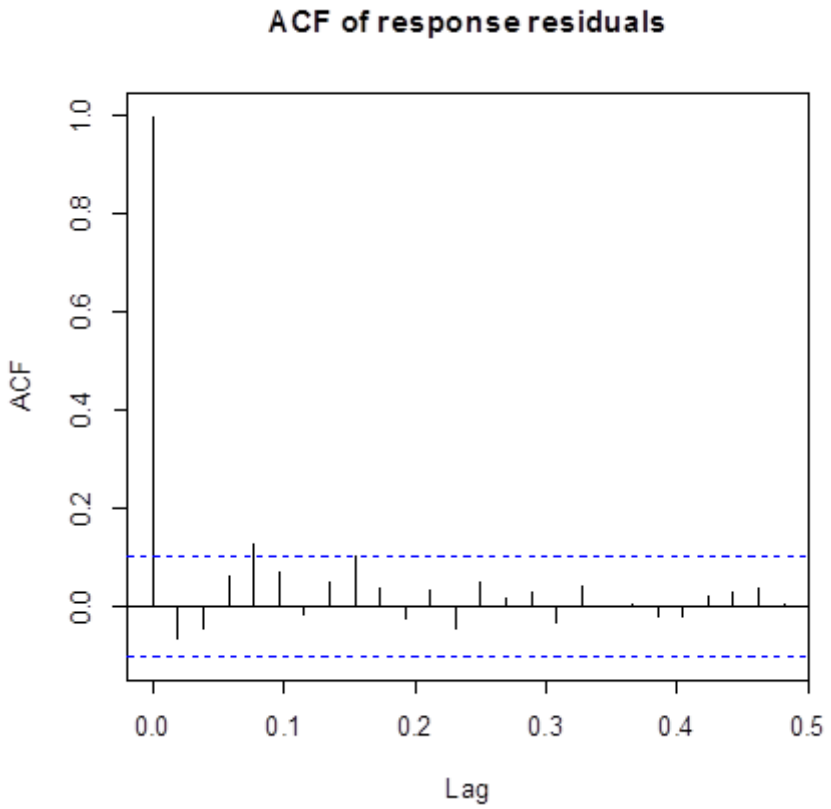


Figure 5.6. Autocorrelation function of response residuals.

The cumulative periodogram can be used as well to test the residuals in the fitted model. Below are displayed the Pearson residuals over time and their cumulative periodogram.

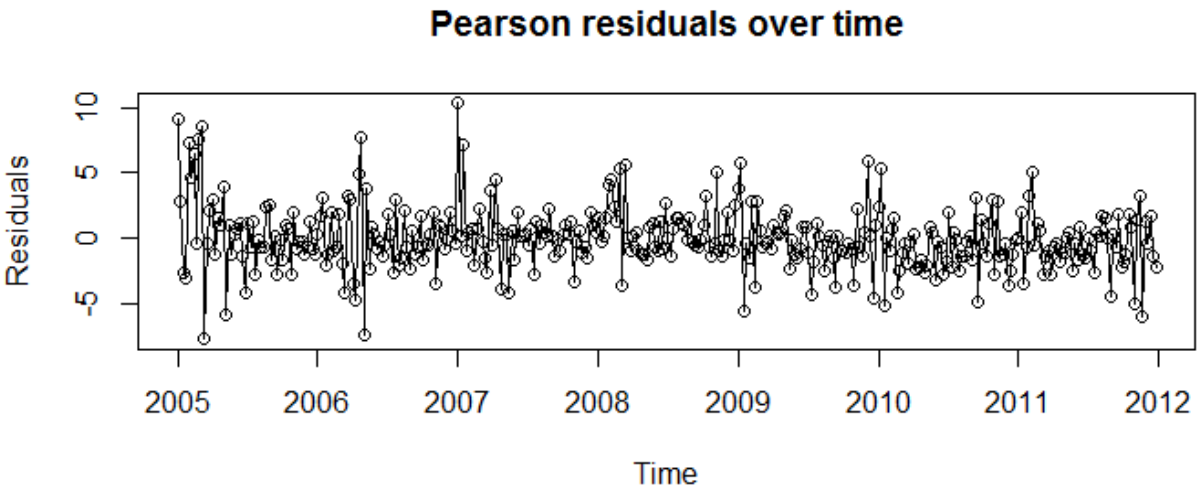


Figure 5.7. Pearson residuals of the fitted models over time.

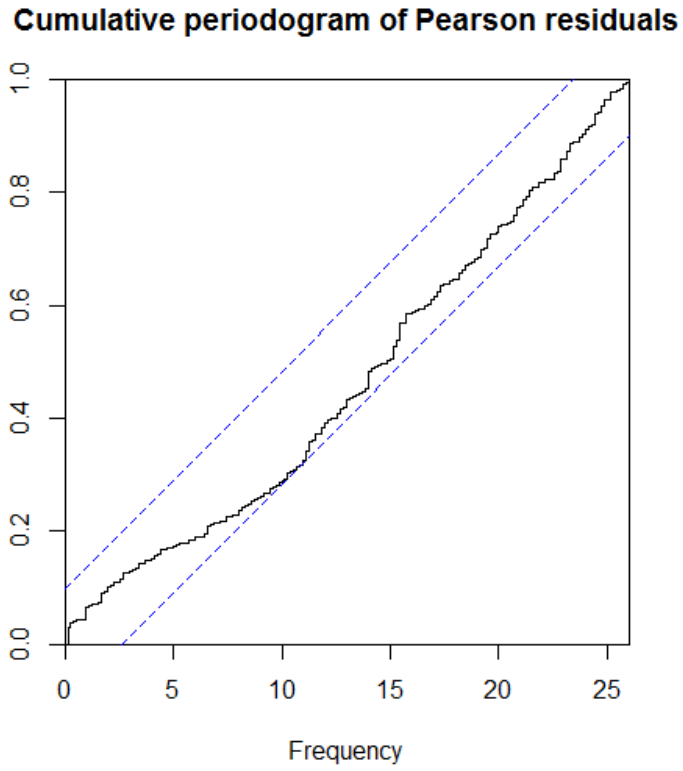


Figure 5.8. Cumulative periodogram for Pearson residuals.

For a purely random series, the cumulative periodogram of the residuals should increase linearly from (0, 0) to (0.5, 1). As it is shown in Figure 5.8. the plot is an approximate straight line.

The parameter estimates of the two fitted models which are shown analytically in Tables 5.2. and 5.3., are identical with slightly different standard errors.

Table 5.2. Parameter estimates and st. errors using Poisson conditional distribution.

Parameter	β_0	β_1	α_{52}
Estimated value	1.031	0.593	0.186
Standard error	0.2862	0.0412	0.0662

Table 5.3. Parameter estimates and st. errors using Neg. Binom. conditional distribution.

Parameter	β_0	β_1	α_{52}	σ^2
Estimated value	1.0311	0.5933	0.1859	0.0323
Standard error	0.3069	0.0450	0.0714	NA

The degree of overdispersion seems to be small, since the estimated overdispersion coefficient $\sigma^2 = 0.0323$ is close to zero which indicates that Y_t given the past seem to be Poisson distributed: $Y_t|F_{t-1} \sim \text{Poisson}(\lambda_t)$. However, no analytical approximation for its standard error is available.

The probability integral transform (PIT) will follow a uniform distribution if the predictive distribution is correct. Figure 5.9. indicates that PIT histograms corresponding to Poisson and Negative Binomial distributions do not differ much. However, the PIT histogram which corresponds to Poisson distribution seems to approach uniformity slightly better and points out that the probabilistic calibration of the predictive distribution is quite satisfactory.

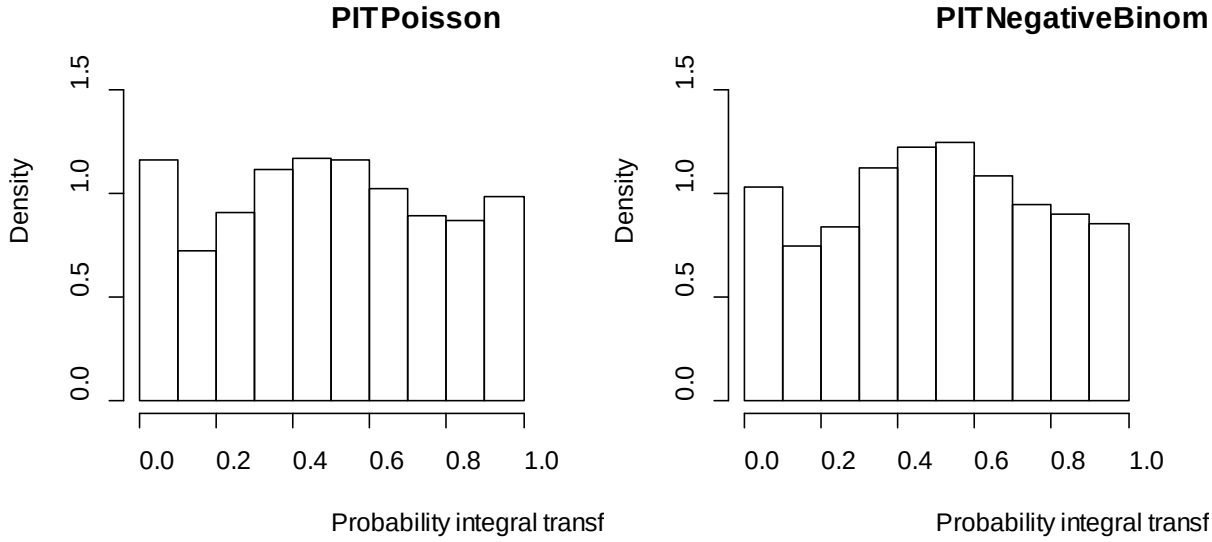


Figure 5.9. PIT histograms for Poisson and Negative Binomial distributions.

From the calibration plot given in Figure 5.10. we don't get an explicit image of which model is more efficient.

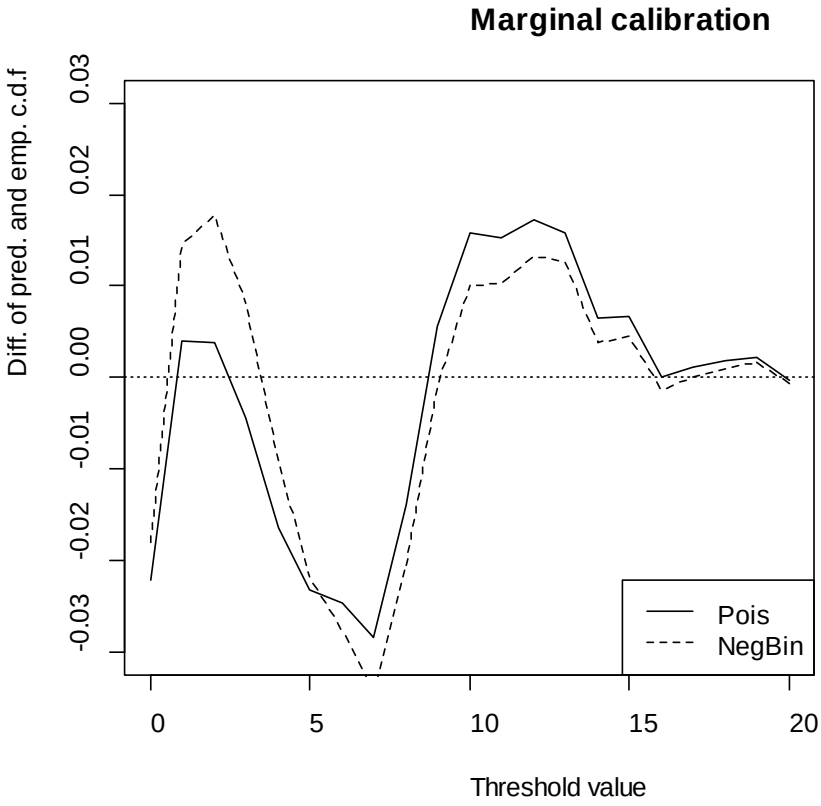


Figure 5.10. Marginal calibration plot for Poisson and Negative Binomial distributions.

As a simultaneous assessment of calibration and sharpness we will consider the tool of scoring rules for the two distributions. In Table 4.4. are given different proper scoring rules for the two models and as expected, they do not differ significantly. However, all of the scoring rules are slightly in favor of Negative Binomial distribution.

Table 5.4. Scoring rules for Poisson and Negative Binomial distribution models.

	logarithmic	**quadratic**	**spherical**	**rankprob**	**dawseb**	**normsq**	**sqerror**
Poisson	2.204577	-0.1413396	-0.3734236	1.264342	2.627560	1.1518377	5.91289
NegBin	2.194683	-0.1399605	-0.3716576	1.263299	2.608627	0.9917582	5.91289

Based on the assessment of the two models, the model which corresponds to Negative Binomial distribution with a very small degree of overdispersion will be used for fitting and forecasting. However, the Negative Binomial assumption is only required for the model assessment and in case of evaluating prediction intervals. The fitted values $\hat{\lambda}_t = \lambda_t(\hat{\theta})$ do not depend on the chosen distribution (Poisson or Binomial) since the mean is the same regardless the response distribution.

The plot of observed versus fitted values from the model (Figure 5.11) indicates the model provides an excellent fit to the data. The red dashed line represents the fitted values.

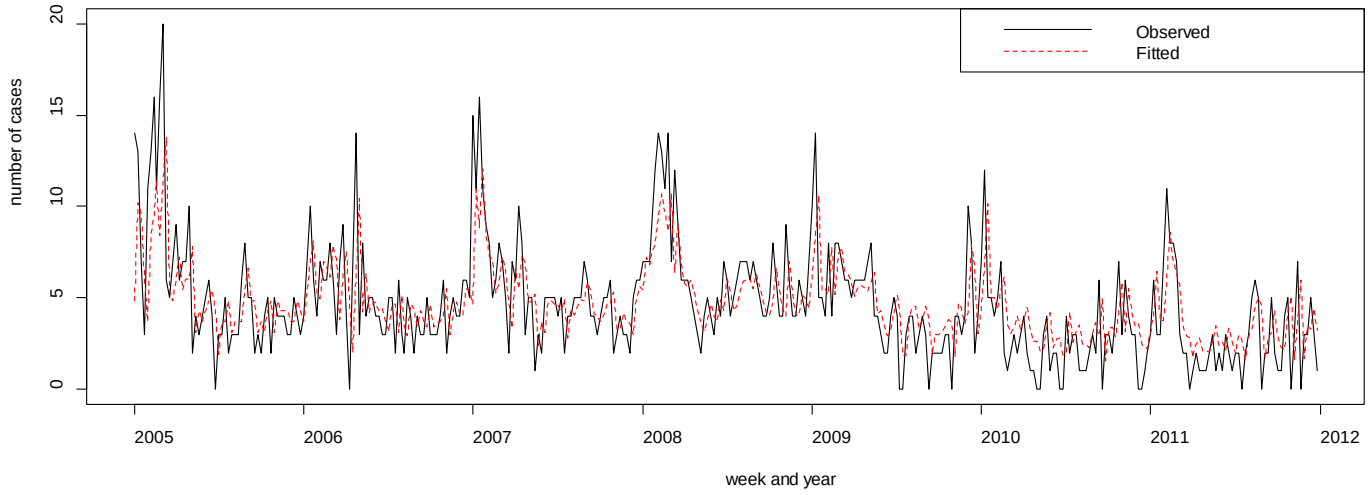


Figure 5.11. Expected and actual number of reported cases of infectious diseases by week (2005-2011).

From the model's assessment, we can conclude that the fitted model manages to capture successfully the autocorrelative structure in the data. In the next session, this model will be used to produce subsequent forecast errors for the next year (2012), which will be monitored by an EWMA chart in order to detect any possible deviations from historical data patterns and from the underlying stochastic process generating the observations.

The residuals from the fitted model using the training data (2005-2011) are plotted in Figure 5.12. with an EWMA chart in order to determine the control chart's limits that will be used for the Phase II analysis. Upper and lower control limits of the EWMA control chart are set at ± 3 standard deviations from the overall average level of residuals, which is

approximately: - 0,037. The value of λ (the weight assigned to the current observation) is set at 0,2.

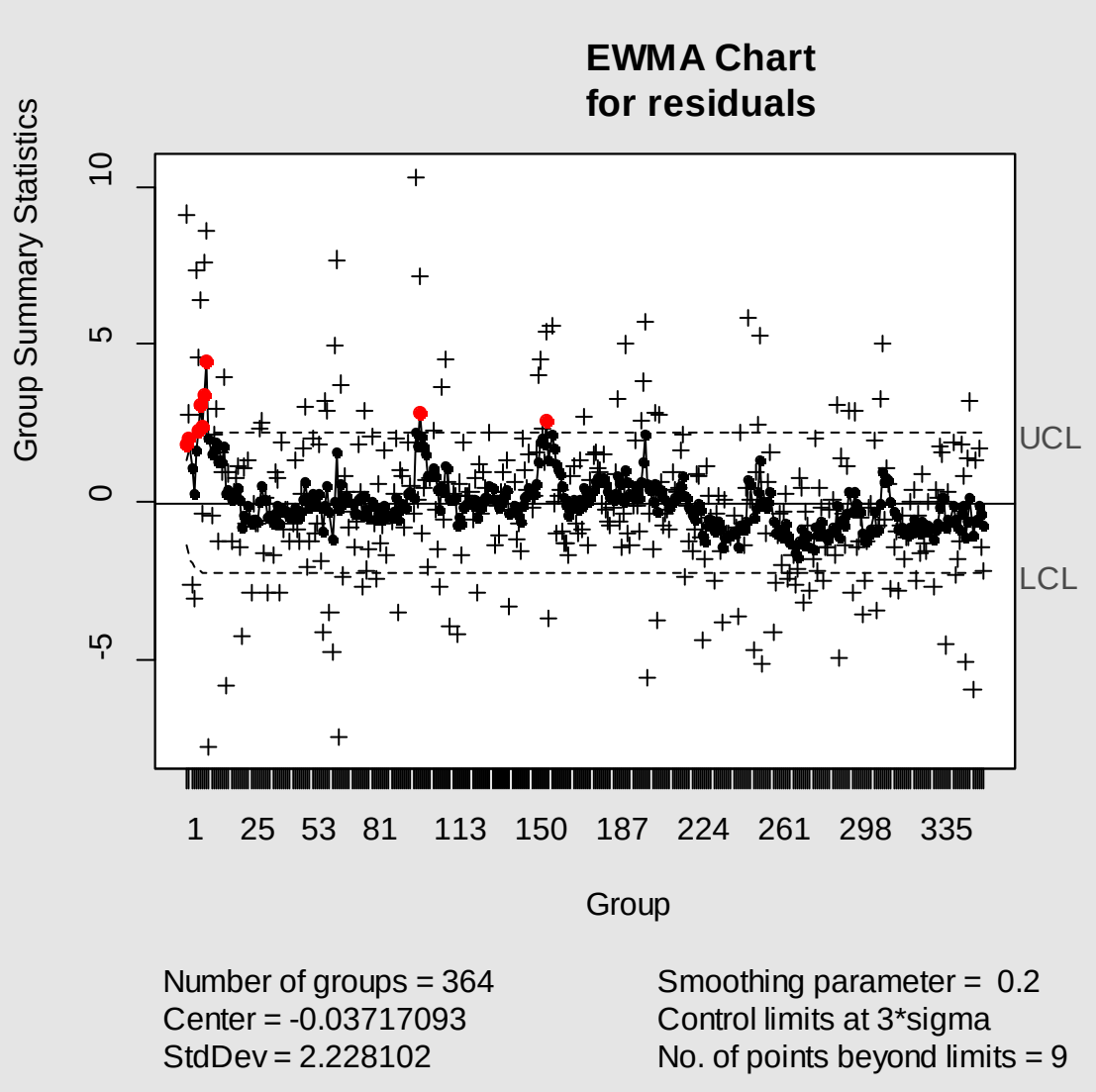


Figure 5.12. EWMA chart monitoring residuals from the fitted model using training data (2005-2011).

The percentage of observations exceeding the upper control limit is approximately 2,5% (9 points) therefore, the control limits will not be much inflated by the existing outliers.

5.4. Phase II analysis.

The estimated model in the previous session, will now be used to forecast weekly reports for 2012. One-week-ahead, rolling forecasts will be predicted for each of the 52 weeks using the same model form, but re-estimating the model coefficients after the addition of each week's real observation.

For the phase II analysis, EWMA SPC chart will be implemented to monitor and identify aberrations in the forecast errors for 2012 data. The one-step ahead forecast error is the difference between the observed value and the prediction of the series at time *t*:

$$e_t = y_t - \hat{y}_{t|t-1}.$$

In monitoring, we are only interested to identify which forecast errors exceed the upper limit. In other words, we aim to identify cases where the observed count of diseases is much higher than the expected from the model.

Based on the model fitted to the training data (2005-2011), the number of reported cases of infectious diseases (by week) for the next year were predicted. In Figure 5.13 are represented the forecast errors that occurred for 2012.

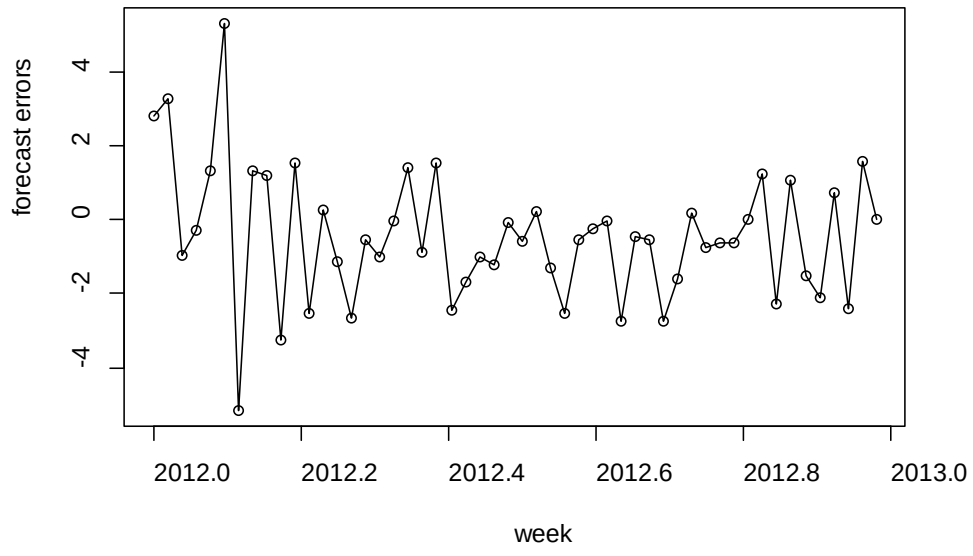


Figure 5.13 Forecast errors for 2012.

A graphical representation of the observed values and the forecasts is given in Figure 5.14. The blue dashed line represents the forecasts for 2012, which appear to approximate the observed values quite efficiently.

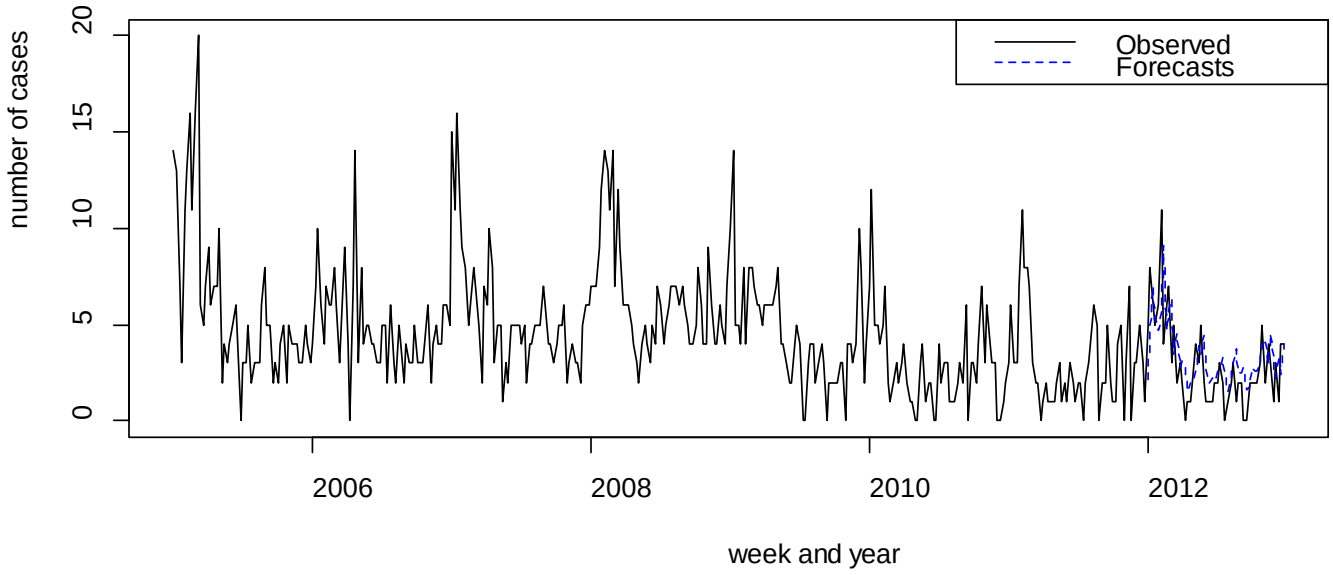


Figure 5.14. Actual number of reported cases of infectious diseases (2005-2012) and forecasts for 2012 according to the fitted model.

The EWMA chart with the smoothing parameter λ set at 0.2 and control limits at 3*sigma as they were estimated in the Phase I analysis, is shown in Figure 5.15. From the monitoring, no forecast error appears to exceed the upper control limit, which implies that there is no signal of a potential change in the pattern of infectious diseases' incidence.

The highest peak corresponds to the 6th week (February) of 2012. During this week, there was an outbreak of 11 reported cases of infectious diseases and it was the maximum observed count of the year 2012. However, a high number of disease counts was expected from the fitted model at this time point, thus, the forecast error does not exceed the upper control limit.

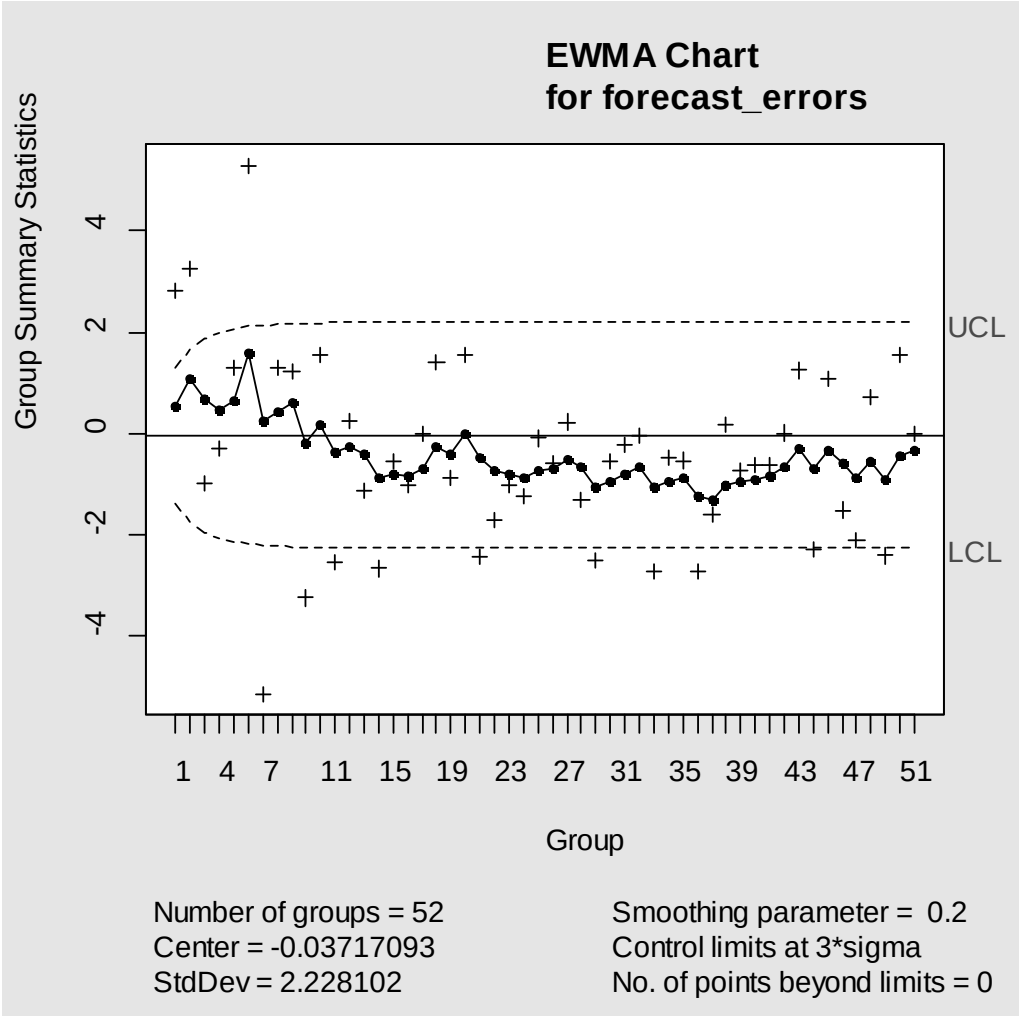


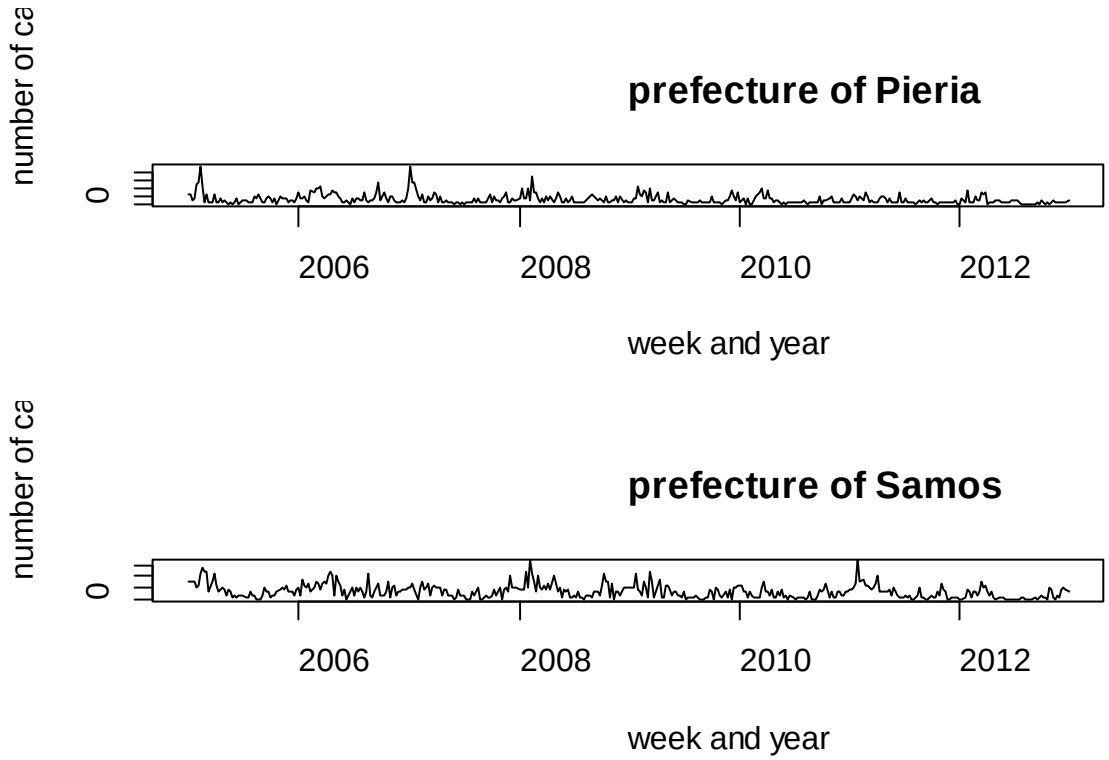
Figure 5.15 EWMA chart monitoring forecast errors for 2012.

5.5. The usage of the monitoring system demonstrated by different data examples.

In this section, the application of the monitoring system is illustrated by data collected in different areas of Greece. More specifically, in the analysis will be used weekly reported cases of infectious diseases between 2005 and 2012, collected by General Practitioners in Primary Health Care Centers of the below areas:

1. Aiginio, prefecture of Pieria
2. Mutilinioi, prefecture of Samos
3. Polykastro, prefecture of Kilkis
4. Kalabaka, prefecture of Trikala
5. Dikaia, prefecture of Evros
6. Epanomi, prefecture of Thessaloniki

In Figure 5.16 are represented the time series of disease counts originated in the above mentioned Health Care Centers in Greece (1-6) which indicate to exhibit cyclic seasonal patterns.



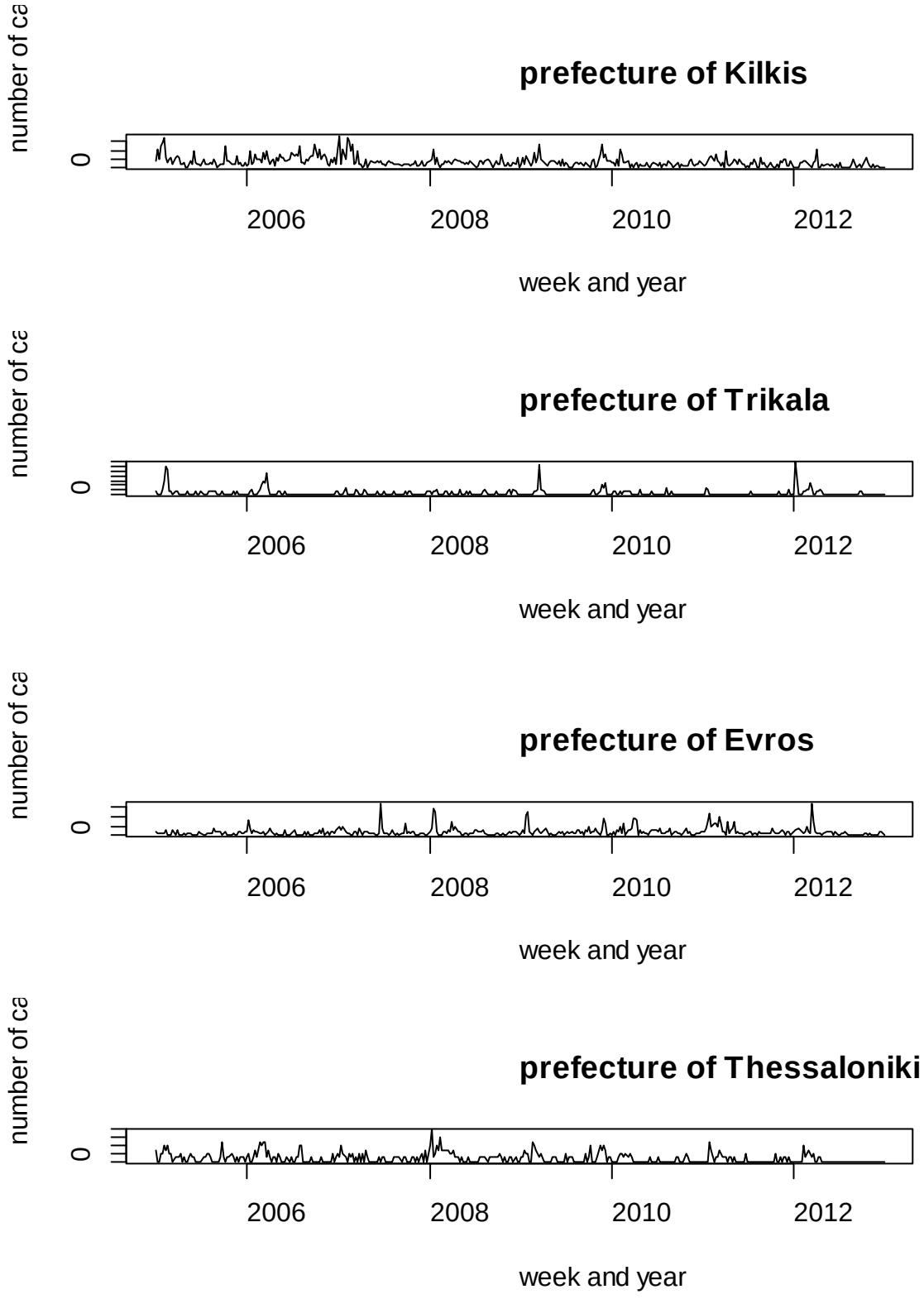


Figure 5.16 Number of infectious disease cases reported every week (2005-2012) in Primary Health Care Centers 1-6.

The generalized linear model (GLM), as was described in the previous session, will be applied using the training data (2005-2011) from the 6 different Health Care Centers in order to produce forecasted weekly values for all 52 weeks of 2012. For taking into account short range serial dependence a first order autoregressive term was included and a 52^{nd} order autoregressive term for seasonality. In all cases, observations were modeled conditionally on the past information using the Negative Binomial conditional distribution. In Table 5.5. are summarized the parameter estimates and the standard errors as they occurred from the fitted models.

Table 5.5. Parameter estimates and st. errors using Neg. Binom. conditional distribution per Health Care Center (1-6).

	Parameter	β_0	β_1	α_{52}	σ^2
1.	Estimated value	0.852	0.438	0.319	0.176
	Standard error	0.2716	0.0578	0.0942	NA
2.	Estimated value	0.00805	0.42665	0.57137	0.13810
	Standard error	0.1390	0.0461	0.0669	NA
3.	Estimated value	1.422	0.399	0.202	0.306
	Standard error	0.4204	0.0616	0.1226	NA
4.	Estimated value	0.1808	0.5609	0.0776	1.3408
	Standard error	0.0396	0.1391	0.0712	NA
5.	Estimated value	0.332	0.364	0.448	0.758
	Standard error	0.1927	0.0773	0.1410	NA
6.	Estimated value	0.179	0.403	0.396	0.365
	Standard error	0.0734	0.0645	0.1059	NA

The response residuals of the fitted models, which are shown in Figure 5.17, seem to be uncorrelated with a mean close to zero. In figures 5.18 a & 5.18 b. is represented the empirical autocorrelation function of the observed values (left side) as well as the empirical autocorrelation function of response residuals.

The autocorrelation function of time series shows clearly dependence among observations during time, while the empirical autocorrelation function of response residuals does not

exhibit any significant remaining serial correlation or seasonality which is not described by the models. The lack of correlation suggests the there is not significant information left in the residuals which should be used in computing forecasts.

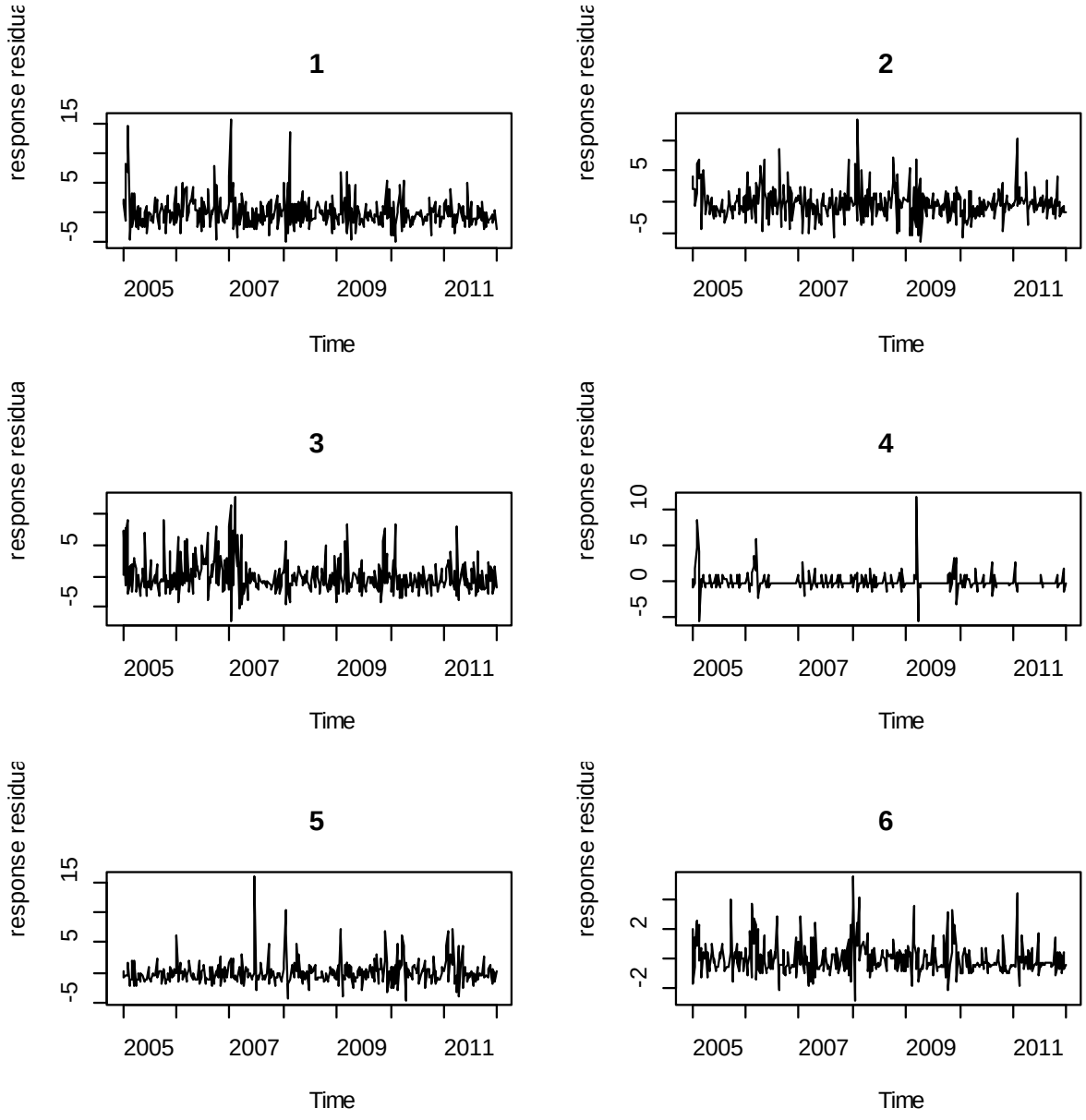


Figure 5.17 Response residuals of the fitted models for Primary Health Care Centers (1-6) over time.

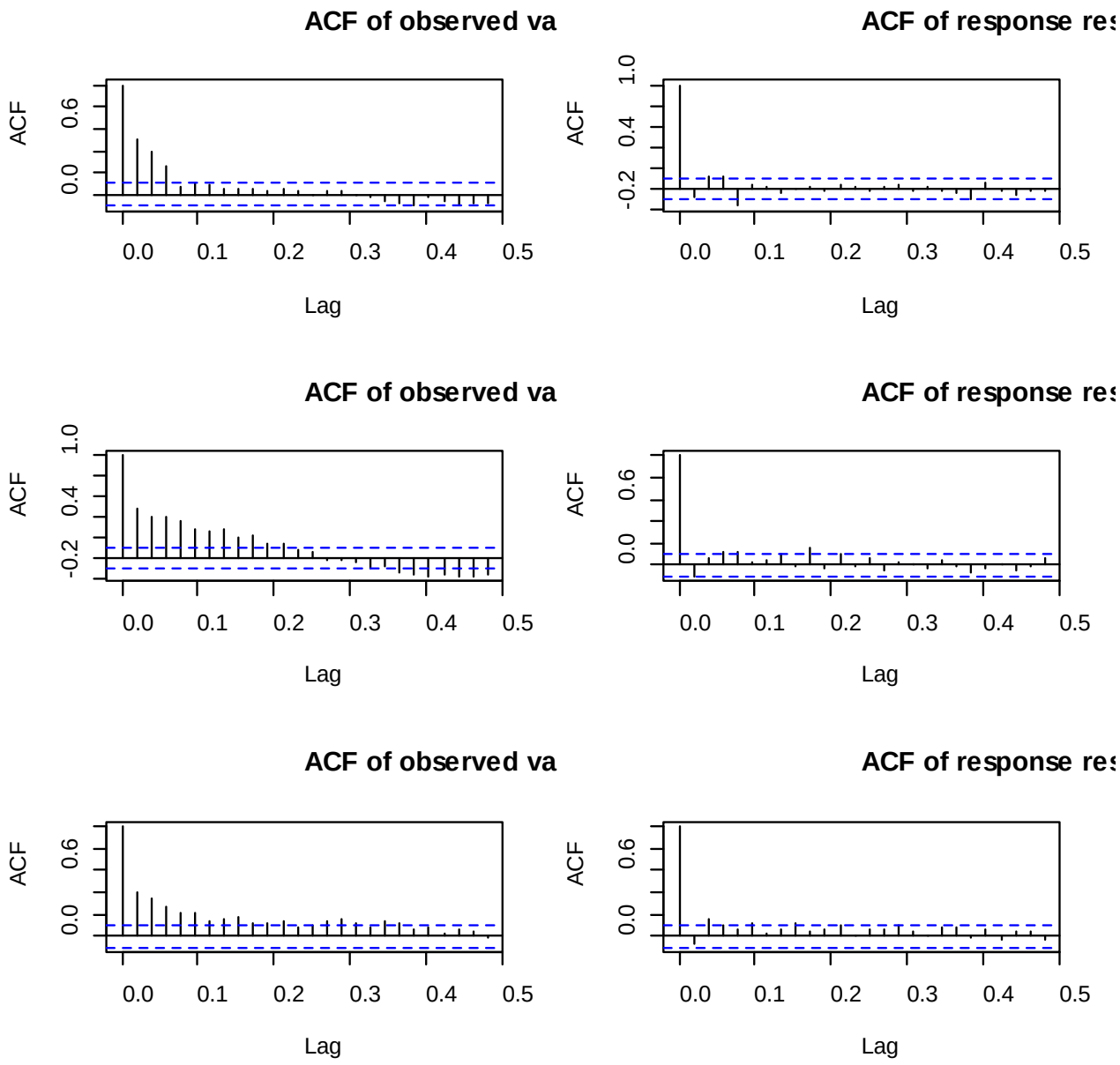


Figure 5.18 a. Autocorrelation function of the observed values and response residuals for Primary Health Care Centers (1-3)

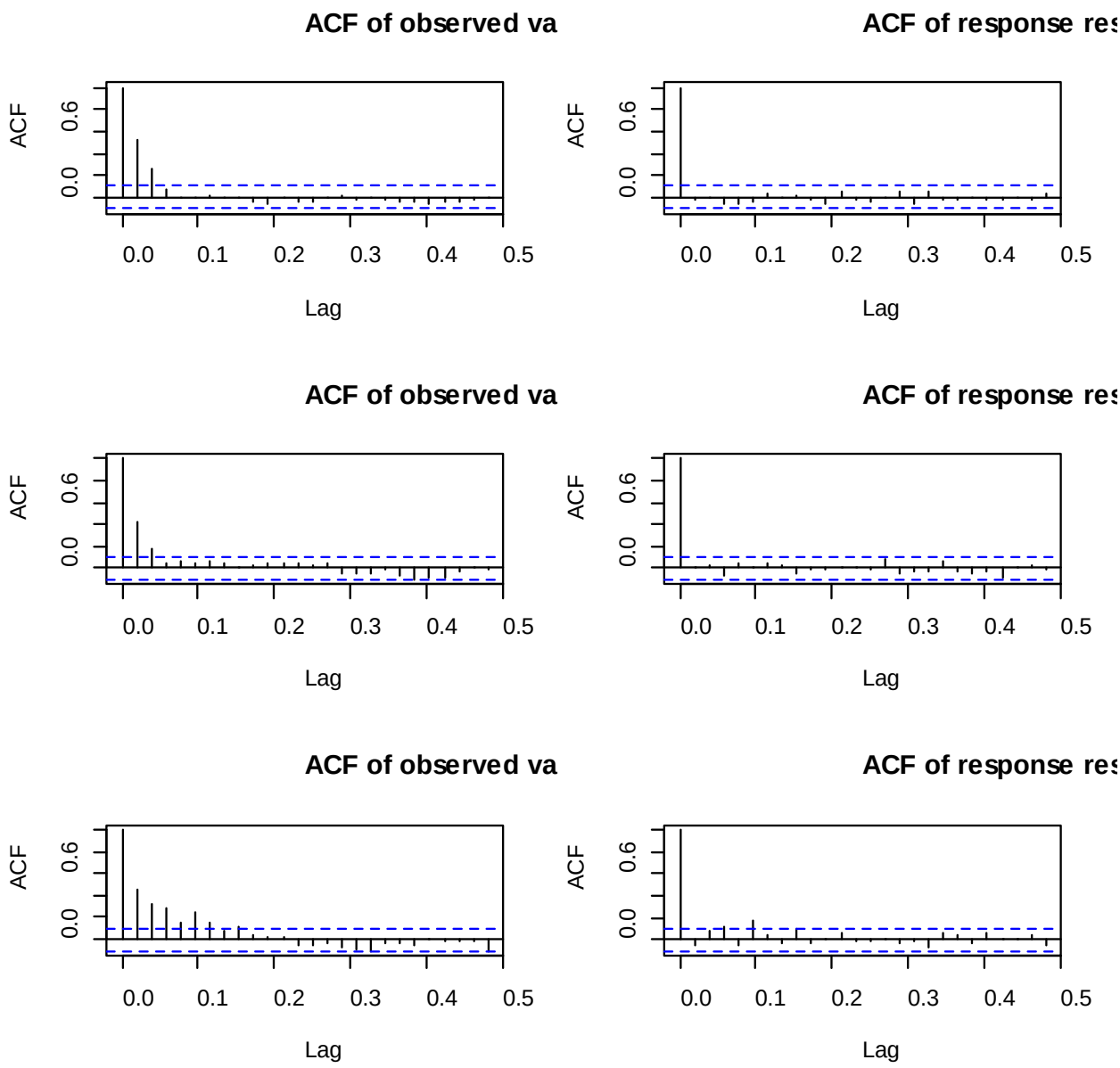


Figure 5.18 b. Autocorrelation function of the observed values and response residuals for Primary Health Care Centers (4-6).

The graphical representations (Figures 5.19 a. & 5.19 b.) of observed versus fitted values from each one of the models (using data from the 6 Primary Health Care Centers), indicate that models provide quite satisfactory fit to the data. The red dashed line represents the fitted values.

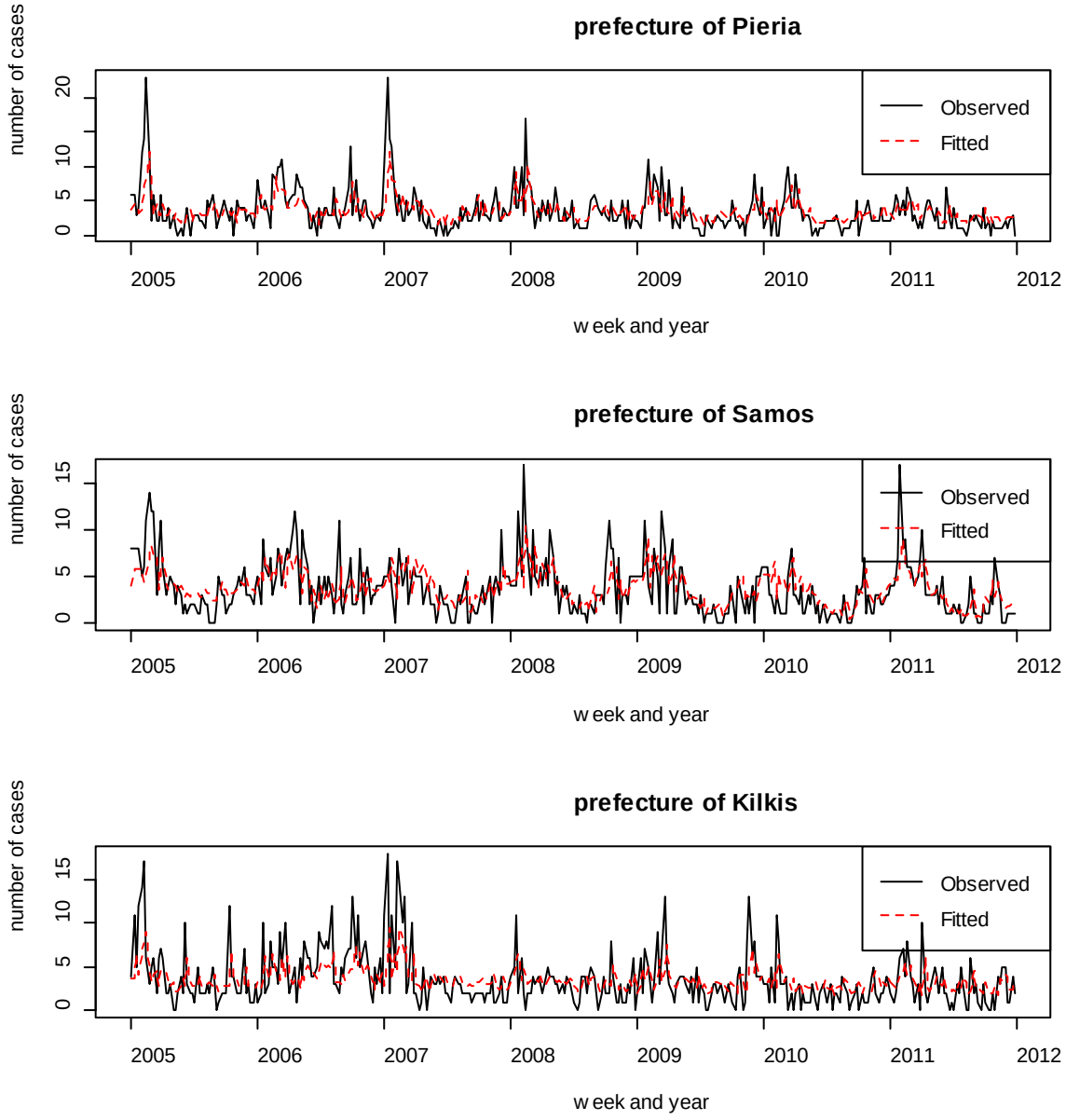


Figure 5.19 a. Expected and actual number of reported cases of infectious diseases by week (2005-2011) for Primary Health Care Centers (1-3).

.

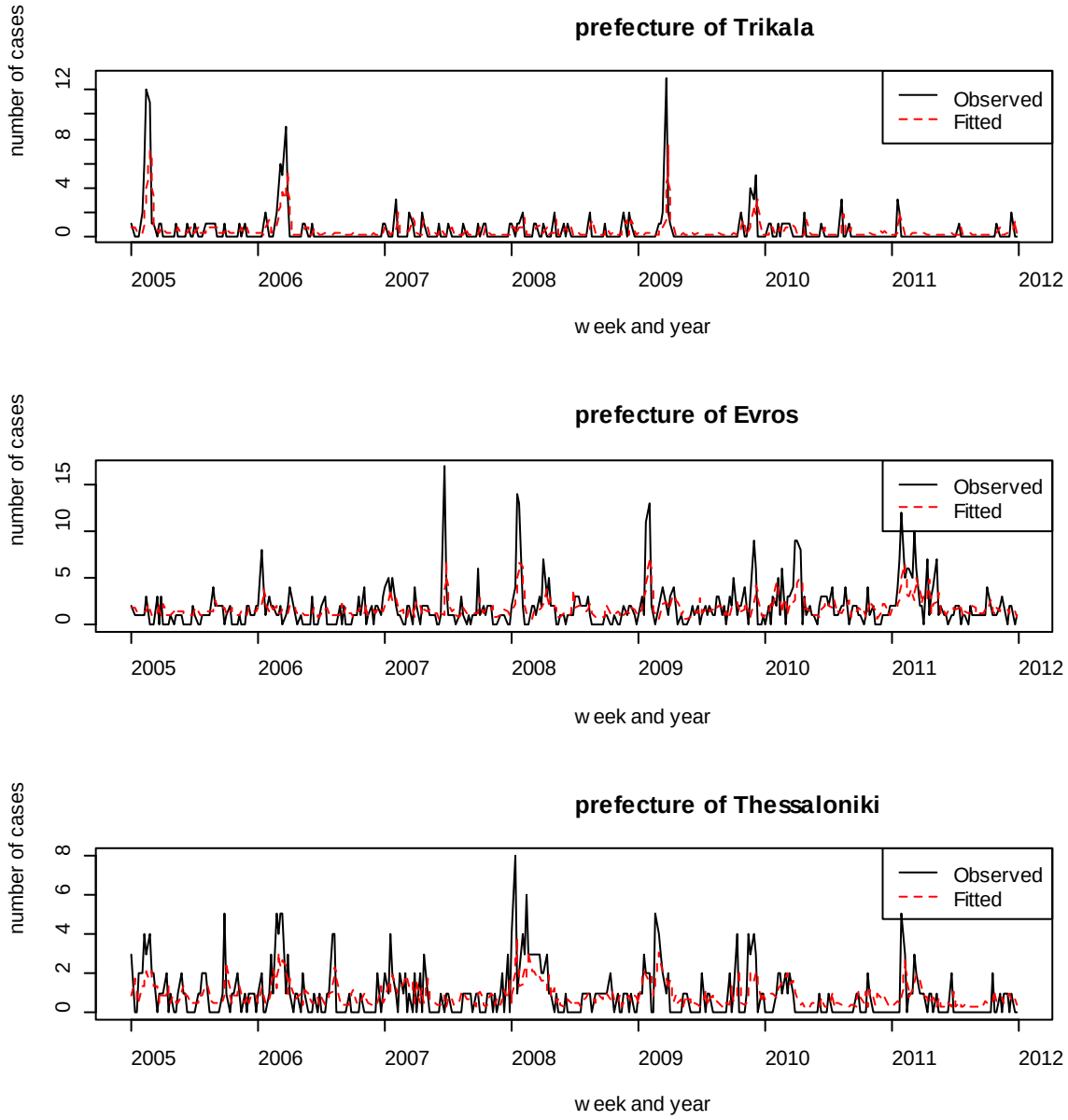


Figure 5.19 b. Expected and actual number of reported cases of infectious diseases by week (2005-2011) for Primary Health Care Centers (4-6).

Using the fitted models will be predicted one-week-ahead, rolling forecasts for each of the 52 weeks of 2012, by re-estimating the model coefficients after the addition of each week's real observation. A graphical representation of the observed values and the forecasts is given in Figures 5.20 a. & 5.20 b. The blue dashed line represents the forecasts for 2012, which in a general view appear to approximate the observed values of the 6 Primary Health Care Centers quite efficiently.

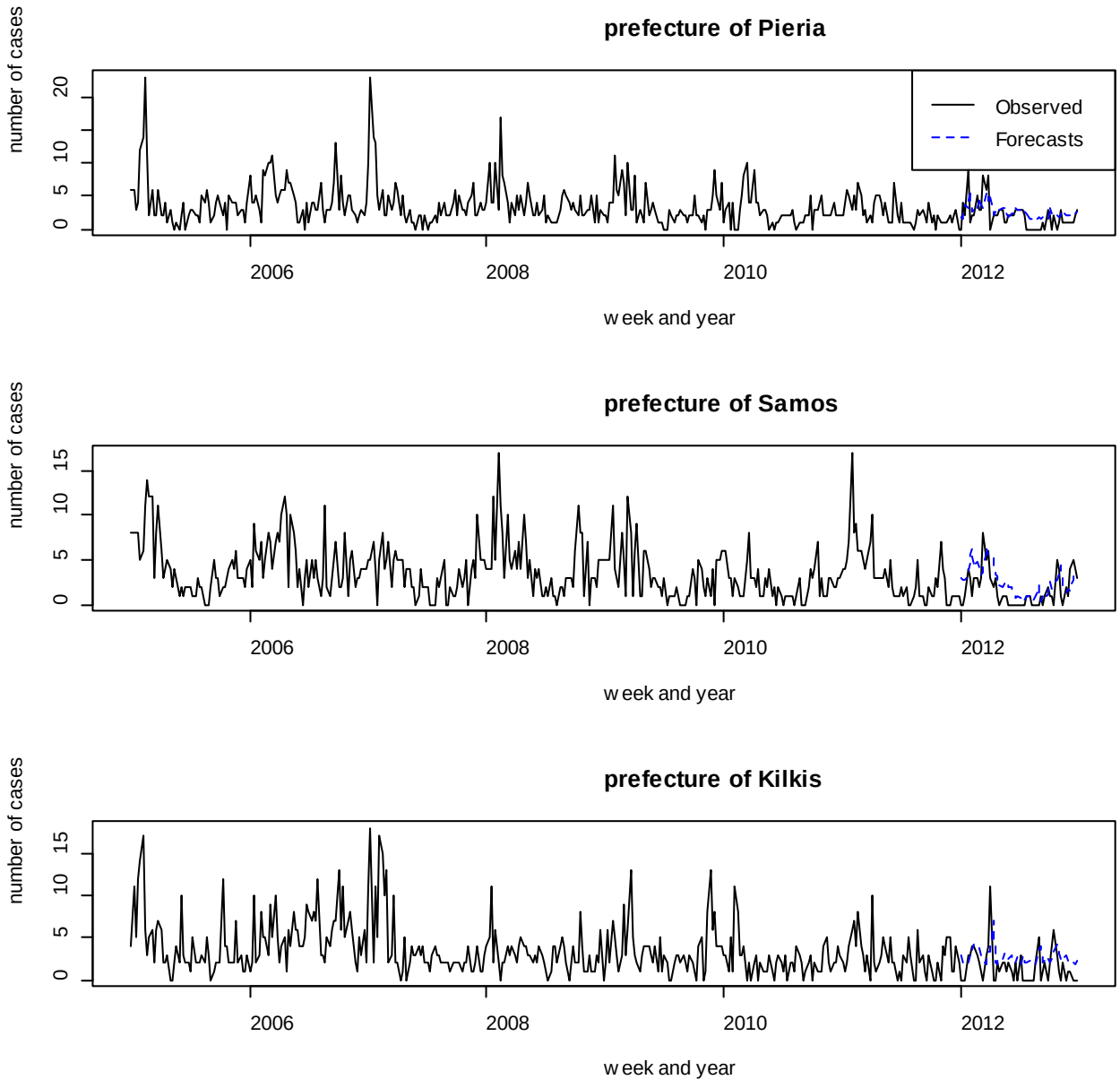


Figure 5.20 a. Actual number of reported cases of infectious diseases (2005-2012) and forecasts for 2012 according to the fitted models for Primary Health Care Centers (1-3).

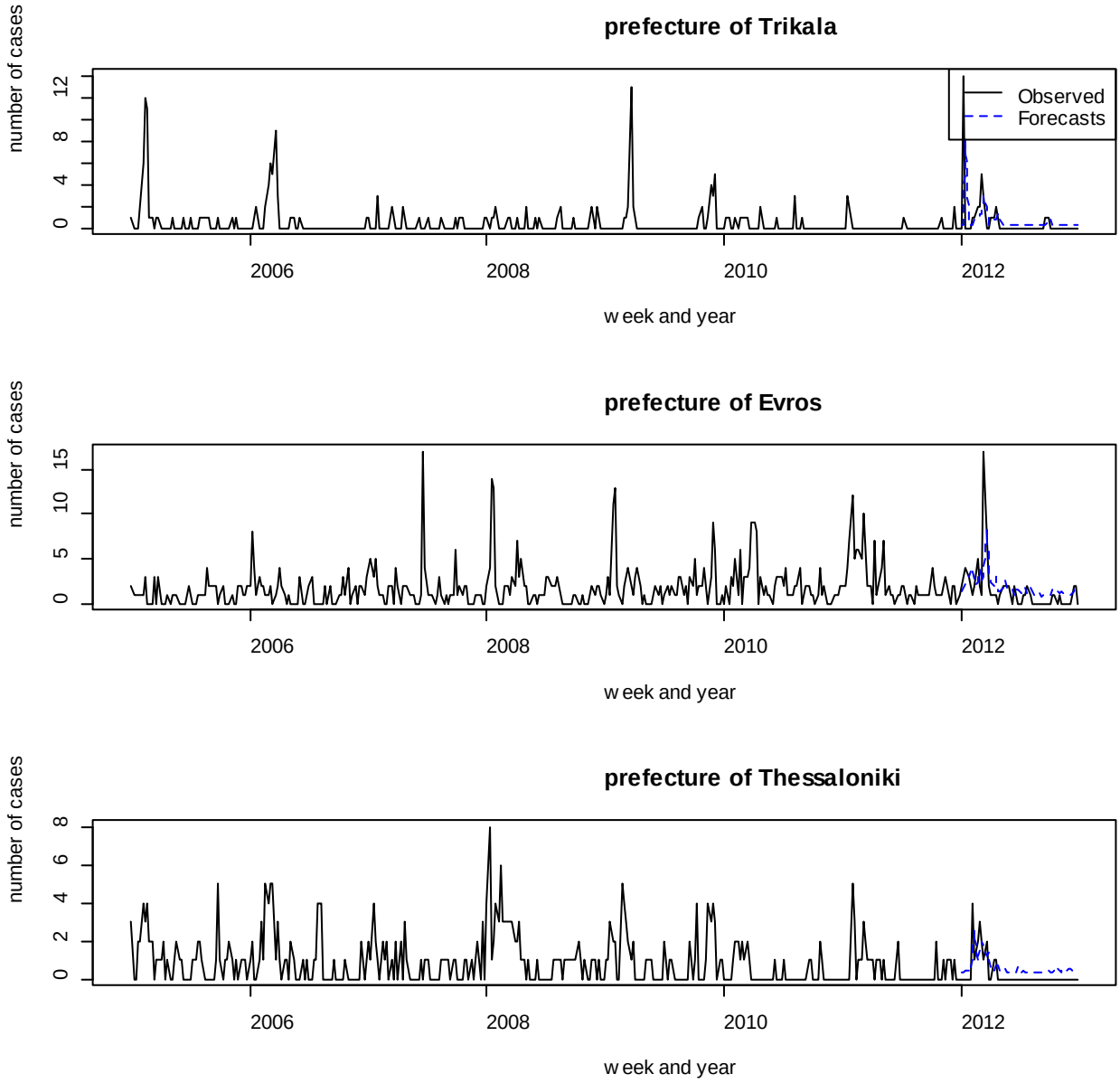


Figure 5.20 b. Actual number of reported cases of infectious diseases (2005-2012) and forecasts for 2012 according to the fitted models for Primary Health Care Centers (4-6).

The one-step ahead forecast errors for all 52 weeks of 2012 , $e_t = y_t - \hat{y}_{t|t-1}$, are shown in Figure 5.21. For all Primary Health Care Centers, the variance of the difference between the observed value and the prediction of the series at time *t*, seems to be larger during the winter months and early in spring.

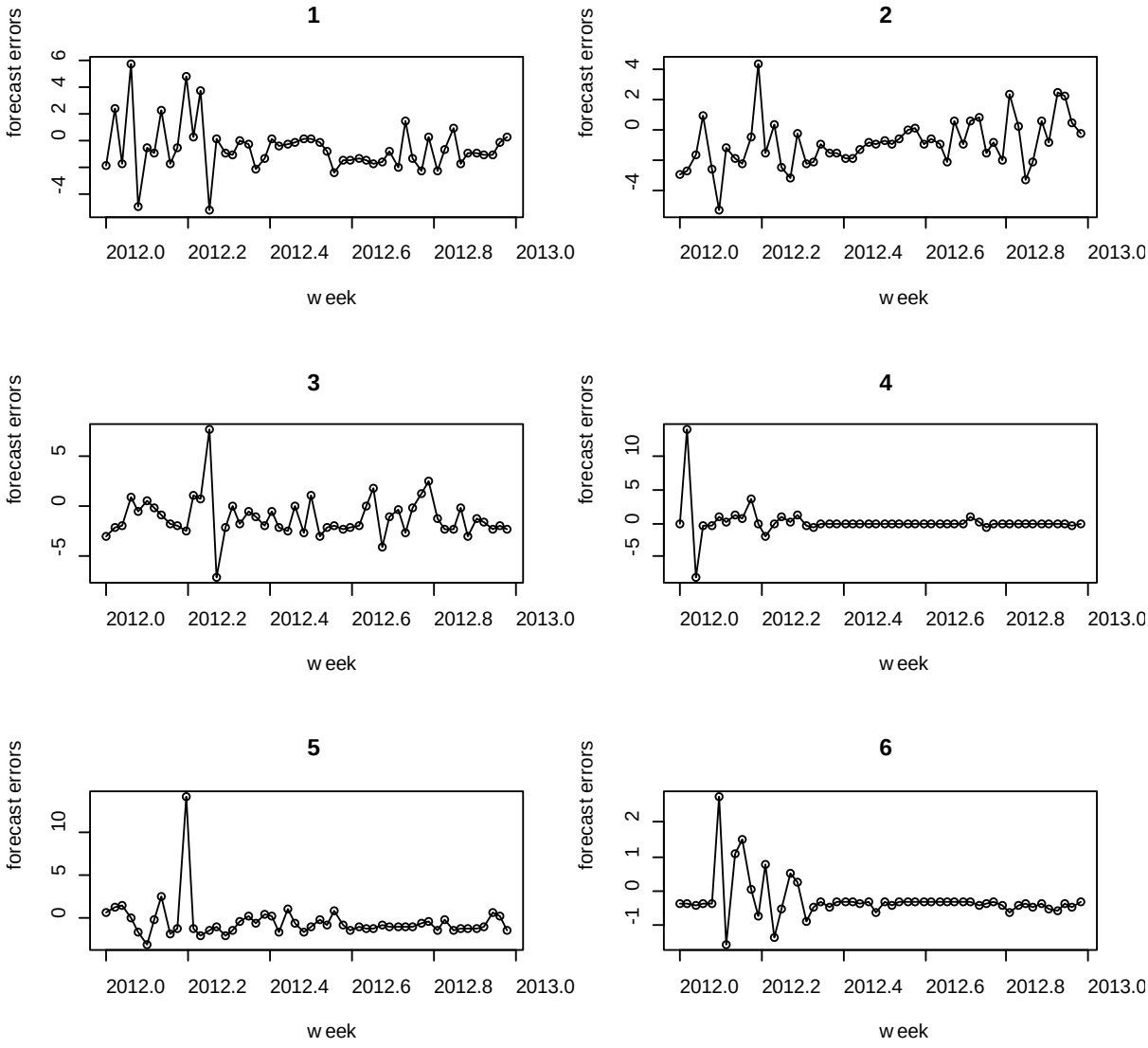


Figure 5.21 Forecast errors (2012) for Primary Health Care Centers (1-6).

The residuals from the fitted models using the training data (2005-2011) for Primary Health Care Centers (1-6) are plotted with EWMA charts in the phase I analysis (Figures 5.22 a. – 5.27 a.), while in Figures 5.22 b. – 5.27 b. are plotted the forecast errors for 2012 using the control chart's limits as they were determined in the Phase I analysis.

Upper and lower control limits of the EWMA control chart are set at ± 3 standard deviations from the overall average level of residuals and the value of λ is set at 0,2. Phase I analysis results using the EWMA charts are summarized in Table 5.6.

The percentage of observations exceeding the upper control limit, as occurred from the below charts, is between 1,9% - 4,6% thus, we don't expect the control limits to be much inflated by the existing outliers.

Table 5.6. Phase I analysis results.

Primary Health Care Center	**average level of residuals**	**LCL**	**UCL**	**No of points beyond limits**
1	-0.04171	-2.498850	2.415429	9
2	-0.1686	-2.528785	2.191593	7
3	-0.01344	-2.627134	2.600251	13
4	-0.005063	-0.712067	0.7019415	13
5	0.06733	-1.585577	1.720229	16
6	-0.01544	-0.983798	0.9529105	17

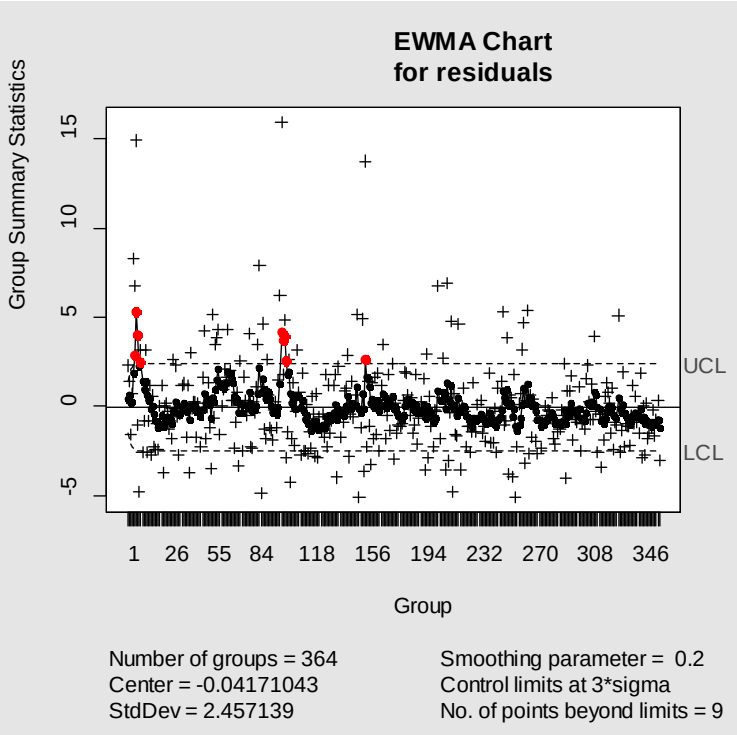


Figure 5.22 a. EWMA chart monitoring residuals from the fitted model using training data (2005-2011) Primary Health Care Center of Aiginio, prefecture of Pieria.

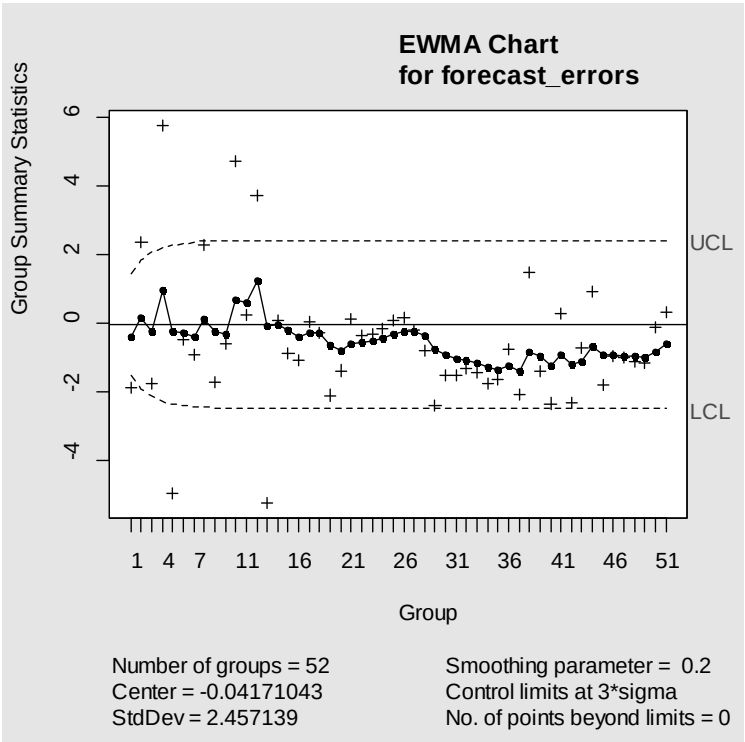


Figure 5.22 b. EWMA chart monitoring forecast errors for 2012. Primary Health Care Center of Aiginio, prefecture of Pieria.

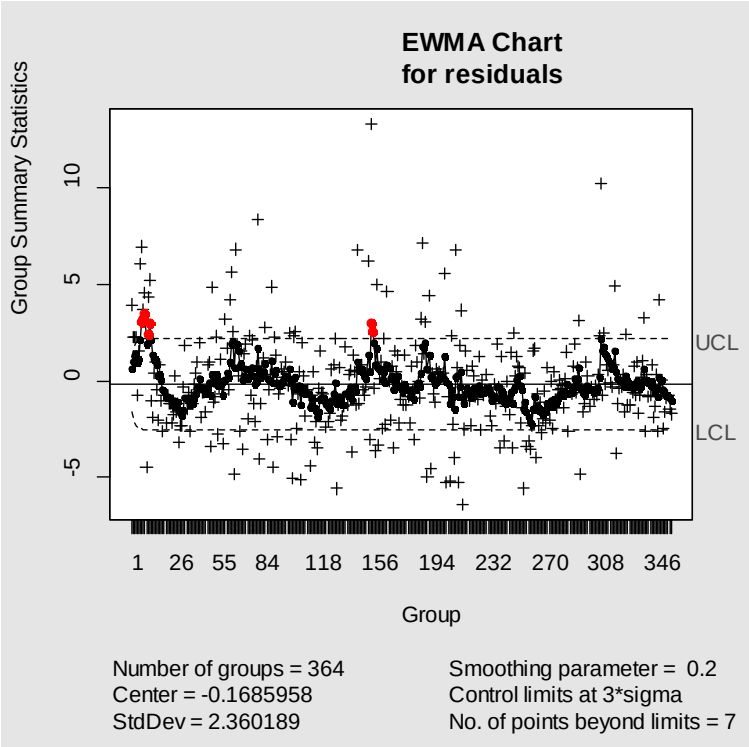


Figure 5.23 a. EWMA chart monitoring residuals from the fitted model using training data (2005-2011) Primary Health Care Center of Mutilinioi, prefecture of Samos.

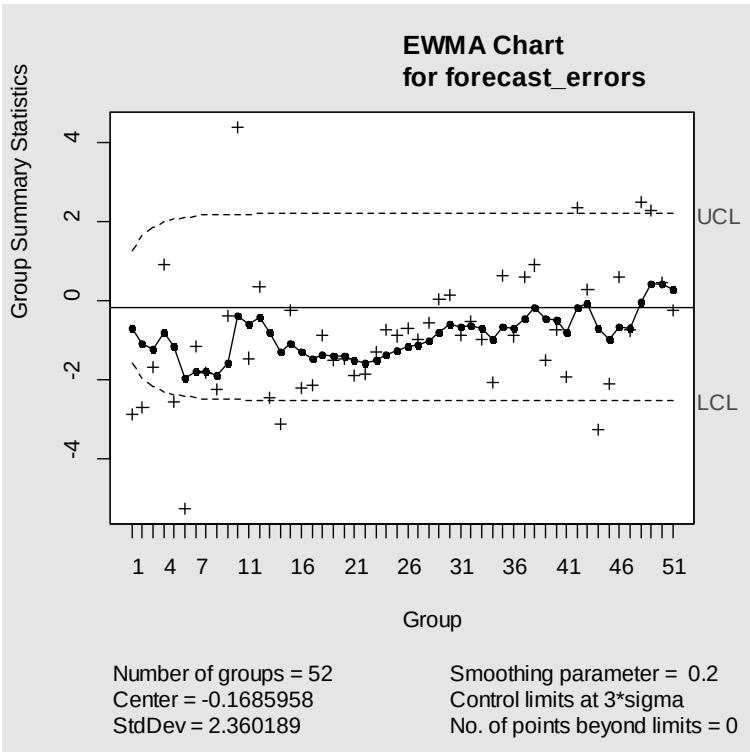


Figure 5.23 b. EWMA chart monitoring forecast errors for 2012. Primary Health Care Center of Mutilinioi, prefecture of Samos.

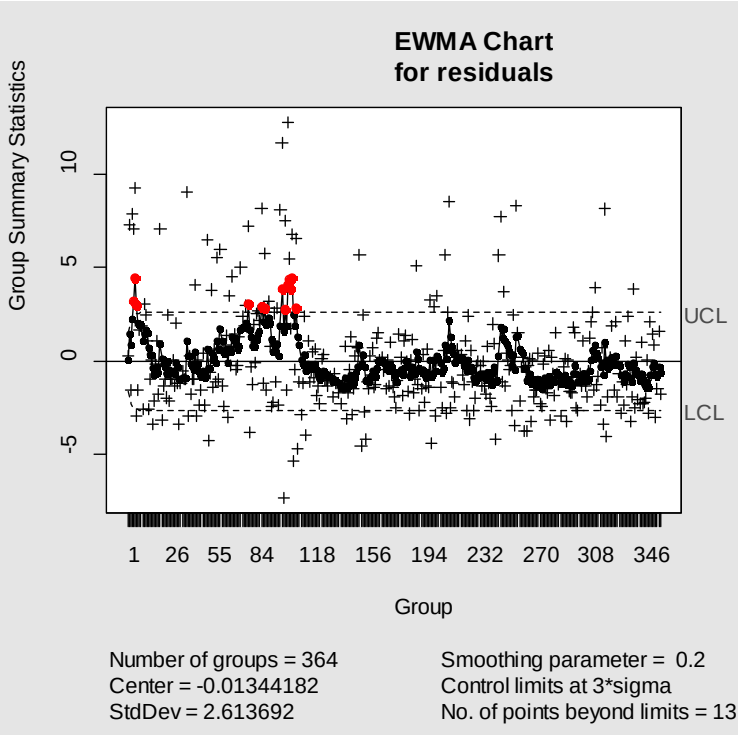


Figure 5.24 a. EWMA chart monitoring residuals from the fitted model using training data (2005-2011) Primary Health Care Center of Polykastro, prefecture of Kilkis.

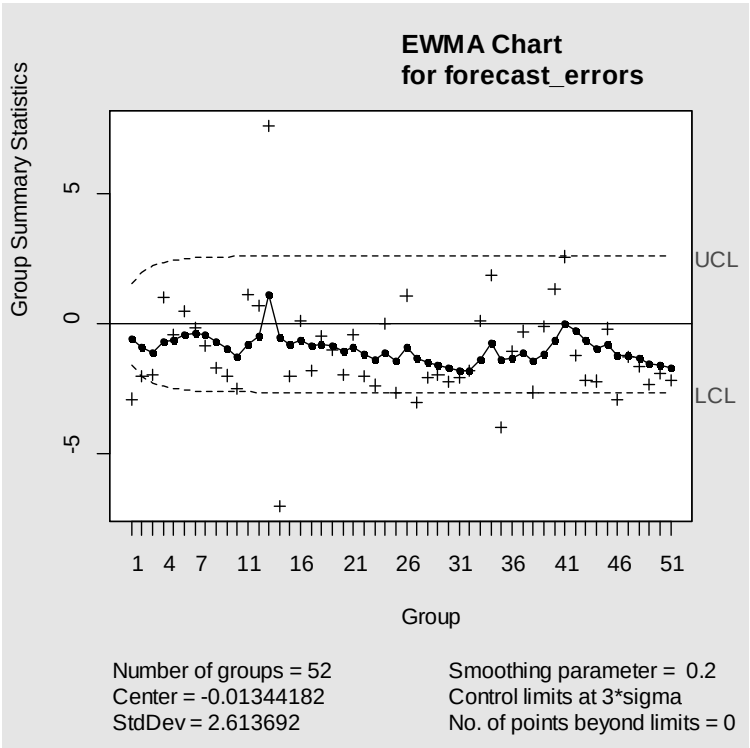


Figure 5.24 b. EWMA chart monitoring forecast errors for 2012. Primary Health Care Center of Polykastro, prefecture of Kilkis.

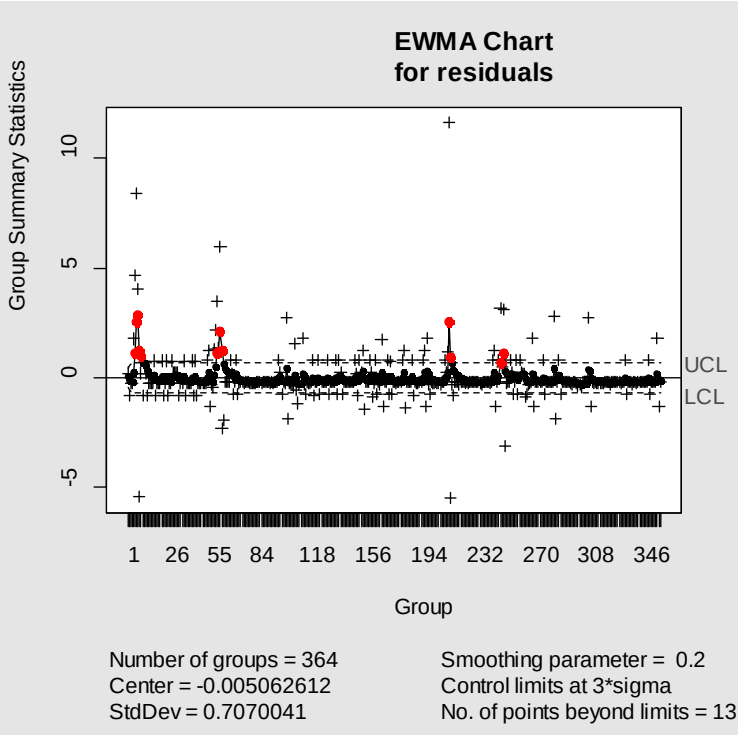


Figure 5.25 a. EWMA chart monitoring residuals from the fitted model using training data (2005-2011) Primary Health Care Center of Kalabaka, prefecture of Trikala.

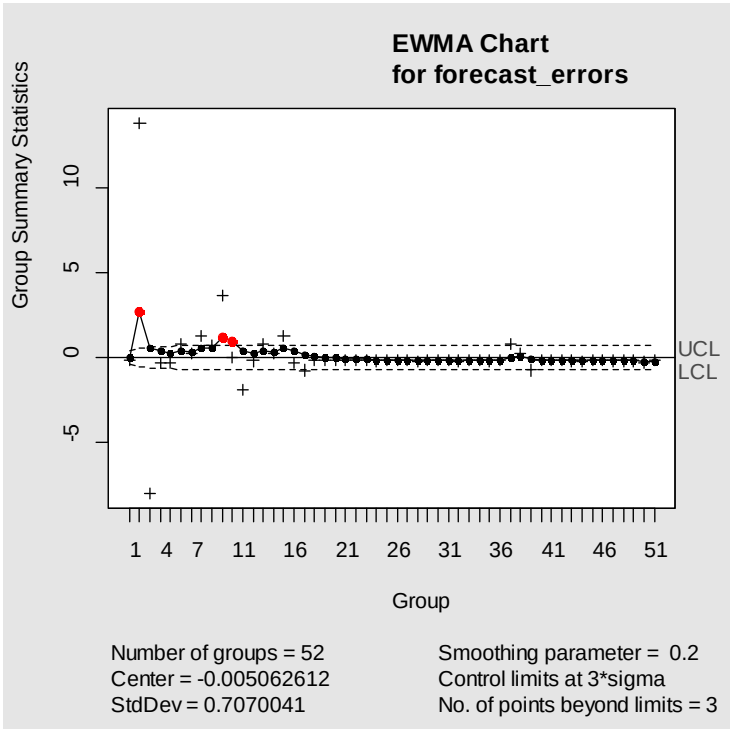


Figure 5.25 b. EWMA chart monitoring forecast errors for 2012. Primary Health Care Center of Kalabaka, prefecture of Trikala.

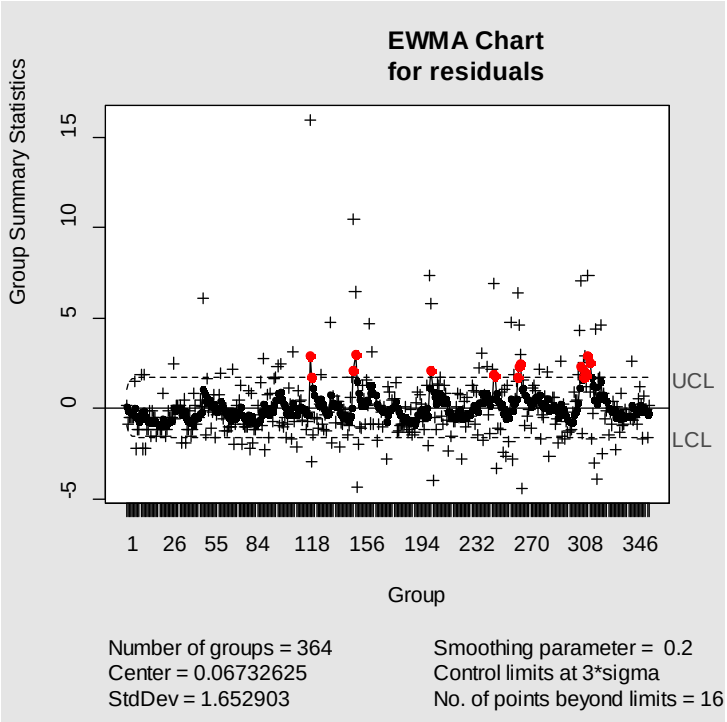


Figure 5.26 a. EWMA chart monitoring residuals from the fitted model using training data (2005-2011) Primary Health Care Center of Dikaia, prefecture of Evros.

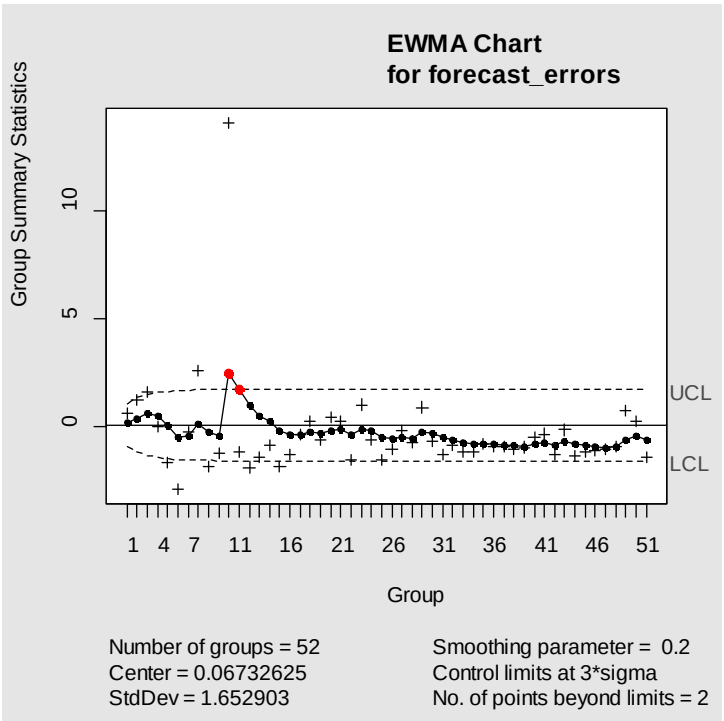


Figure 5.26 b. EWMA chart monitoring forecast errors for 2012. Primary Health Care Center of Dikaia, prefecture of Evros.

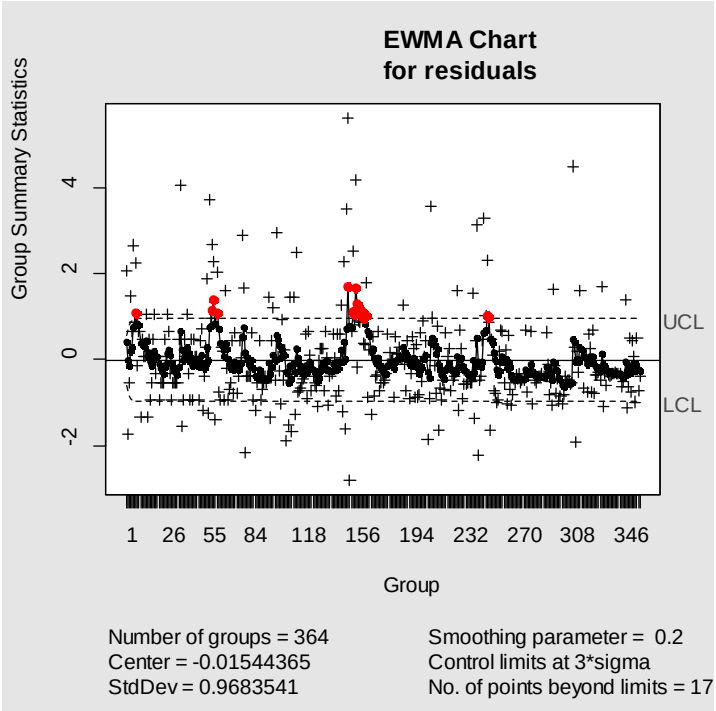


Figure 5.27 a. EWMA chart monitoring residuals from the fitted model using training data (2005-2011) Primary Health Care Center of Epanomi, prefecture of Thessaloniki.

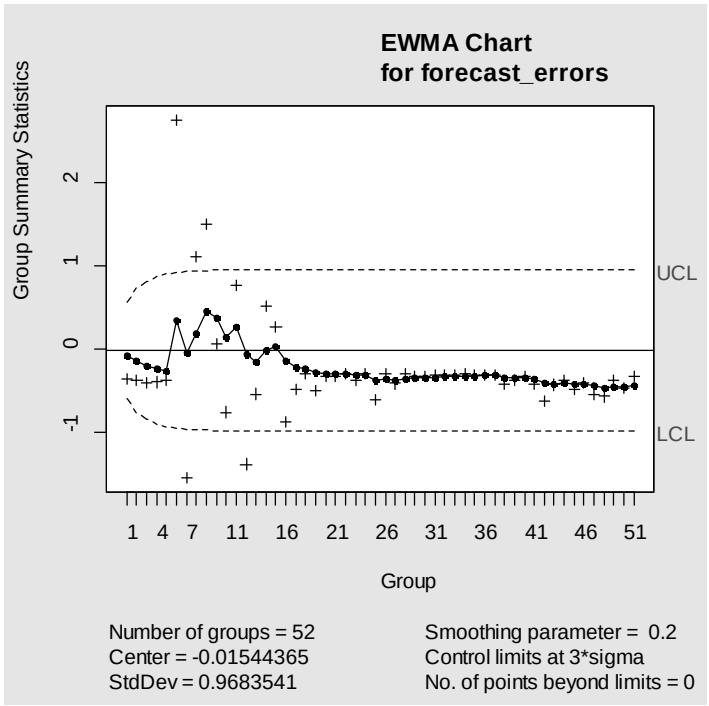


Figure 5.27 b. EWMA chart monitoring forecast errors for 2012. Primary Health Care Center of Epanomi, prefecture of Thessaloniki.

From the monitoring, 3 forecast errors were identified exceeding the upper control limit from reported data in Primary Health Care Center of Kalabaka, prefecture of Trikala, as well as 2 forecast errors from reported data in Primary Health Care Center of Dikaia, prefecture of Evros.

In the first case, the first significantly high value was detected in the 2nd week of January, where the corresponding observed reported cases where 14. This was a big outburst, considering that most of the reported cases in this Health Center were 0 or 1 per week. The expected value from the model at this time point was approximately 0.195. The other two significantly high values were detected during March (10th and 11th week).

In the second case, both significantly high forecast errors were detected during March (11th and 12th week), where the corresponding observed reported cases where 17 and 7 respectively.

The monitoring system detected no significantly high values for all the rest Primary Health Care Centers which implies that there is no signal of a potential change in the pattern of infectious diseases in 2012.

Chapter 6

Conclusion

A critical aspect of health surveillance is to detect a change in the incidence of natural outbreaks and issue an alarm as soon as possible so that appropriate actions are taken. Thus, the timely detection of increases in the rate of unusual events is an important objective in public health and healthcare surveillance. In the present study, was introduced a monitoring system which aids in identification of unusual observations in reported data on infectious disease counts by providing early warning signals.

The monitoring system was adapted for count autocorrelated processes and it was conducted in two phases. For the phase I analysis was employed the generalized linear methodology (GLM) using likelihood- based estimation and the Poisson or Negative Binomial distribution for modelling the observations conditionally on the past information. Model was fitted using the training data (2005-2011) of reported infectious diseases events from 7 Primary Health Care Centers in different prefectures in Greece. In all cases, Negative Binomial conditional distribution was assumed for the analysis. However, both conditional distributions provided the same results, since the estimation of the regression parameters does not depend on the additional dispersion parameter φ of the Negative Binomial distribution. Assumption of the distribution was required mostly for the model assessment, where all of the assessment techniques seemed to be slightly in favor of Negative Binomial distribution. In all cases, the fitted models take into account short range serial dependence by a first order autoregressive term as well as seasonality, by regressing on the unobserved conditional mean 52 weeks (one year) back in time and the results indicate the model successfully captures the autocorrelative structure, providing a satisfactory fit to the data.

In the second phase of analysis, fitted models using the training data were used in order to produce one-week-ahead, rolling forecasts for each of the 52 weeks of the next year (2012). The real observations of 2012 were used by re-estimating the model coefficients every time after the addition of each week's real observation. The subsequent forecast errors that were produced using the model, were monitored by EWMA charts and the control limits were predetermined from the response residuals which occurred from the fitted model in phase I analysis. The percentage of residuals exceeding the upper control limit in phase I was between 1,9% - 4,6% thus, the control limits were not much inflated by the existing outliers.

From the monitoring results, in five out seven Primary Health Care Centers, no forecast errors were detected exceeding the upper control limit. Thus, there was no signal of a potential change in the pattern of infectious diseases in 2012. In Primary Health Care Center of Kalabaka (prefecture of Trikala), 3 forecast errors were identified exceeding the upper control limit, where the first significantly high value was detected in the 2^{nd} week of January and the other two significantly high values were detected during March (10^{th} and 11^{th} week). In Primary Health Care Center of Dikaia (prefecture of Evros), two significantly high forecast errors were identified, both detected during March (11^{th} and 12^{th} week).

However, monitoring results varied by Health Care Center do not provide evident relationships between statistically high values among the different areas in Greece. The direct comparison of monitoring results would not be prudent in this case, since analysis was applied to reported number of cases of diseases rather than to reported rates of diseases and therefore, models have not taken account of differences in population size for the areas. The present study aims in producing a monitoring system for the challenging case of autocorrelated count data. For further research, the focus of analyses could shift to comparison among reporting areas with different population sizes. In that case, analyses of rates would be more appropriate. The analysis described in the study, however, could readily be applied to rates, as well as it could provide insight into etiology or risk factors of disease by including covariate effects in the models.

The monitoring system described in this study is an automated, flexible one, which shows promise of assisting the public health community in the frame of timely detection of disease outbreaks and facilitating early public health response. However, there remain numerous opportunities for development, application and evaluation of quantitative methods to aid in identifying outbreaks, sentinel public health events and aberrations in disease data, and, thus, facilitate timely actions to decrease unnecessary morbidity and mortality.

References.

1. Γαβανά Μαγδαληνή (2009). "Σχεδιασμός, Εφαρμογή και Αξιολόγηση Ενός Δικτύου Επιδημιολογικής Επιτήρησης Νοσημάτων Μέσω Παρατηρητών Νοσηρότητας στην Πρωτοβάθμια Φροντίδα Υγείας." Διδακτορική Διατριβή (Ιατρική σχολή του Αριστοτέλειου Πανεπιστημίου Θεσσαλονίκης).

2. Thacker, S. B. and Berkelman, R. L. (1988) "Public health surveillance in the United States", Epidemiologic Review,10, 164-190.

3. Porta M, ed. Dictionary of epidemiology. 5th ed. (2008) International Epidemiological Association. New York, NY: Oxford University Press.

4. Chapra, S.C. and Canale, R.P., (1998) Numerical Methods for Engineers. Singapore: McGraw-Hill.

5. Hawthorne, G. and Elliot, P. (2005) Imputing Cross-Sectional Missing Data: Comparison of Common Techniques. Australian and New Zealand Journal of Psychiatry, 39, p. 583-590.

6. Montgomery DC. (2005) Introduction to Statistical Quality Control (5th edn). Wiley: New York.

7. C.H.Weiß. (2007) Controlling correlated processes of Poisson counts, Qual. Reliab. Eng. Int. 23 pp. 741–754.

8. C.H.Weiß. (2009) EWMA monitoring of correlated processes of Poisson counts, Qual. Technol. Quant. Manage. pp. 137–153.

9. Alzaid, A. and M. Al-Osh (1988). First-order integer-valued autoregressive (INAR (1)) process: distributional and regression properties. Statistica Neerlandica 42,53–61.

10. Jacobs, P. A. and Lewis, P. A. W. (1978a) Discrete time series generated by mixtures. I: correlation and ruins properties. J. Roy. Statist. Soc. B 40, 94-105.

11. Franke, J. and Seligmann, T. (1993) Conditional maximum likelihood estimates for INAR (1) processes and their application to modelling epileptic seizure counts. In Developments in Time Series Analysis (ed. T. S. Rao). 310–30. New York: Chapman and Hall.

12. MacDonald, I. L. and Zucchini, W. (1997). Hidden Markov and Other Models for Discrete-valued Time Series. Boca Raton, LA: Chapman & Hall.

13. P. McCullagh, John A. Nelder (1989) Generalized Linear Models, Second Edition. ISBN 0-412-31760-5 CRC Press 99-13896.

14. Ferland R, Latour A, Oraichi D (2006). Integer-Valued GARCH Process. Journal of Time Series Analysis, 27(6), 923–942. ISSN 1467-9892.doi:10.1111/ j.1467-9892.2006.00496.x.

15. Fokianos K (2011). Some Recent Progress in Count Time Series. Statistics: A Journal of Theoretical and Applied Statistics,45(1), 49. ISSN 0233-1888. doi:10.1080/02331888.2010.541250.

16. Christou V, Fokianos K (2014). Quasi-Likelihood Inference for Negative Binomial Time Series Models. Journal of Time Series Analysis,35(1), 55–78. ISSN 1467-9892. doi:10.1111/jtsa.12050.

17. Gneiting T, Balabdaoui F, Raftery AE (2007). Probabilistic Forecasts, Calibration and Sharpness. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 69(2), 243–268. ISSN 1467-9868.doi:10.1111/j.1467-9868.2007.00587.x.

18. Czado C, Gneiting T, Held L(2009). Predictive Model Assessment for Count Data. Biometrics, 65(4), 1254–1261. ISSN 1541-0420.doi:10.1111/j.1541-0420.2009.01191.x.

19. http://tscount.r-forge.r-project.org R Package'tscount'

20. Fokianos K, Liboschic T, Fried R. (2015) tscount: An R Package for Analysis of Count Time Series Following Generalized Linear Models

21. Williamson GD1, Weatherby Hudson G. (1999) A monitoring system for detecting aberrations in public health surveillance reports. Statist. Med.18, 3283-3298.

22. Ronald D. Fricker Jr (2011) Some methodological issues in biosurveillance. Statist. Med.30, 403–415, doi:10.1002/sim.3880.

23. Christian H. Weiß & Murat Caner Testik (2015) On the Phase I analysis for monitoring time-dependent count processes. IIE Transactions DOI:10.1080/0740817X.2014.952850

24. Christian H. Weiß (2007) Controlling correlated processes of Poisson counts. Quality and Reliability Engineering International 23, Issue 6,741–754, doi: 10.1002/qre.875

25. Christian H. Weiß (2009) EWMA Monitoring of Correlated Processes of Poisson Counts QTQM Vol. 6, No. 2, pp. 137-153

26. Sung Won Han1, Kwok-Leung Tsui1, Bancha Ariyajunya, Seoung Bum Kim (2009) A comparison of CUSUM, EWMA, and temporal scan statistics for detection of increases in poisson rates. Qual. Reliab. Engng. 26, 279-289 DOI:10.1002/qre.1056.

27. https://cran.r-project.org/web/packages/qcc/qcc.pdf R Package "qcc".

28. Kedem B, Fokianos K (2002). Regression Models for Time Series Analysis. Wiley series in probability and statistics. Wiley-Interscience, Hoboken. ISBN 0-471-36355-3.

29. Woodall, W. H. and Montgomery, D. C. (1999), Research Issues and Ideas in Statistical Process Control, Journal of Quality Technology, 31(4), 376-386.

30. http://www.who.int/en/ WHO World Health Organization.