université PARIS-SACLAY

CentraleSupélec

NNT : 20XXSACLXXXX

Thèse de doctorat

Periodic models and variations applied to health problems

Thèse de doctorat de Universidade Federal do Espírito Santo et de l'Université Paris-Saclay, préparée à UFES et CentraleSupélec

Ecole doctorale n°9 Sciences et technologies de l'information et de la communication (STIC)
Spécialité de doctorat : Traitement du signal et des images

Thèse présentée et soutenue à Vitória, Espírito Santo, Brésil, le 26 février 2019, par

M. PAULO ROBERTO PREZOTTI FILHO

Composition du Jury :

Mme Jane Meri Santos Professeure, Universidade Federal do Espírito Santo (PPGEA)	Président
M. Pierre Olivier Amblard Directeur de Recherche, CNRS	Rapporteur
M. Paulo Canas Rodrigues Assistant Professeur, Universidade Federal da Bahia	Rapporteur
M. Alexandre Renaux Maître de Conférences, Paris-Saclay	Examinateur
M. Valdério Anselmo Reisen Professeur, Universidade Federal do Espírito Santo (PPGEA)	Directeur de thèse
M. Pascal Bondon Directeur de Recherche, CNRS	Directeur de thèse

université PARIS-SACLAY
ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre : Modèles périodiques et variations appliqués aux problèmes de santé

Mots clés : Donées de comptage, périodicité, modèles INAR, modèles PINAR, modèles ZIP, saisonnalité.

Résumé : Ce manuscrit porte sur certaines extensions à des séries temporelles prenant des valeurs entières du modèle paramétrique périodique autorégressif établi pour des séries prenant des valeurs réelles. Les modèles que nous considérons sont basés sur l'utilisation de l'opérateur de Steutel et Van Harn (1979) et généralisent le processus autorégressif stationnaire à valeurs entières (INAR) introduit par Al-Osh & Alzaid (1987) à des séries de comptage périodiquement corrélées. Ces généralisations incluent l'introduction d'un opérateur périodique, la prise en compte d'une structure d'autocorrélation plus complexe dont l'ordre est supérieur à un, l'apparition d'innovations de variances périodiques mais aussi à inflation de zéro par rapport à une loi discrète donnée dans la famille des distributions exponentielles, ainsi que l'utilisation de covariables explicatives. Ces extensions enrichissent considérablement le domaine d'applicabilité des modèles de type INAR. Sur le plan théorique, nous établissons des propriétés mathématiques de nos modèles telles que l'existence, l'unicité, la stationnarité périodique de solutions aux équations définissant les modèles. Nous proposons trois méthodes d'estimation des paramètres des modèles dont une méthode des moments basée sur des équations du type Yule-Walker, une méthode des moindres carrés conditionnels, et une méthode du quasi maximum de vraisemblance (QML) basée sur la maximisation d'une vraisemblance gaussienne. Nous établissons la consistance et la normalité asymptotique de ces procédures d'estimation. Des simulations de type Monte Carlo illustrent leur comportement pour différentes tailles finies d'échantillon. Les modèles sont ensuite ajustés à des données réelles et utilisés à des fins de prédiction. La première extension du modèle INAR que nous proposons consiste à introduire deux opérateurs de Steutel et Van Harn périodiques, l'un modélisant les autocorrélations partielles d'ordre un sur chaque période et l'autre captant la saisonnalité périodique des données. Grâce à une représentation vectorielle du processus, nous établissons les conditions l'existence et d'unicité d'une solution périodiquement corrélées aux équations définissant le modèle. Dans le cas où les innovations suivent des lois de Poisson, nous étudions la loi marginale du processus. Nous ajustons ce modèle à des données de comptage journalières du nombre de personnes ayant reçu des antibiotiques pour le traitement de maladies respiratoires dans la région de Vitória au Brésil. Comme les affections respiratoires sont fortement corrélées au niveau de pollution atmosphérique et aux conditions climatiques, la structure de corrélation des nombres quotidiens de personnes recevant des antibiotiques montre, entre autres caractéristiques, une périodicité et un caractère saisonnier hebdomadaire. Nous étendons ensuite ce modèle à des données présentant des autocorrélations partielles périodiques d'ordre supérieur à un. Nous étudions les propriétés statistiques du modèle, telles que la moyenne, la variance, les distributions marginales et jointes. Nous ajustons ce modèle au nombre quotidien de personnes recevant du service d'urgence de l'hôpital public de Vitória un traitement pour l'asthme. Enfin, notre dernière extension porte sur l'introduction d'innovations suivant une loi de Poisson à inflation de zéro dont les paramètres varient périodiquement, et sur l'ajout de covariables expliquant le logarithme de l'intensité de la loi de Poisson. Nous établissons certaines propriétés statistiques du modèle et nous mettons en oeuvre la méthode du QML pour estimer ses paramètres. Enfin, nous appliquons cette modélisation à des données journalières du nombre de personnes qui se sont rendues dans le service d'urgence d'un hôpital pour des problèmes respiratoires, et nous utilisons comme covariable la concentration de polluant dans la même zone géographique.

Title : Periodic models and variations applied to health problems

Keywords : Count time series, periodicity, INAR Models, PINAR Models, ZIP Models, seasonality.

Abstract : This manuscript deals with some extensions to time series taking integer values of the autoregressive periodic parametric model established for series taking real values. The models we consider are based on the use of the operator of Steutel and Van Harn (1979) and generalize the stationary integer autoregressive process (INAR) introduced by Al-Osh & Alzaid (1987) to periodically correlated counting series. These generalizations include the introduction of a periodic operator, the taking into account of a more complex autocorrelation structure whose order is higher than one, the appearance of innovations of periodic variances but also at zero inflation by relation to a discrete law given in the family of exponential distributions, as well as the use of explanatory covariates. These extensions greatly enrich the applicability domain of INAR type models. On the theoretical level, we establish mathematical properties of our models such as the existence, the uniqueness, the periodic stationarity of solutions to the equations defining the models. We propose three methods for estimating model parameters, including a method of moments based on Yule-Walker equations (YW), a conditional least squares method, and a quasi-maximum likelihood method (QML) based on the maximization of a Gaussian likelihood. We establish the consistency and asymptotic normality of these estimation procedures. Monte Carlo simulations illustrate their behavior for different finite sample sizes. The models are then adjusted to real data and used for prediction purposes. The first extension of the INAR model that we propose consists of introducing two periodic operators of Steutel and Van Harn, one modeling the partial autocorrelations of order one on each period and the other capturing the periodic seasonality of the data. Through a vector representation of the process, we establish the conditions of existence and uniqueness of a solution periodically correlated to the equations defining the model. In the case where the innovations follow Poisson's laws, we study the marginal law of the process. As an example of real-world application, we are adjusting this model to daily count data on the number of people who received antibiotics for the treatment of respiratory diseases in the Vitória region in Brazil. Because respiratory conditions are strongly correlated with air pollution and weather, the correlation pattern of the daily numbers of people receiving antibiotics shows, among other characteristics, weekly periodicity and seasonality. We then extend this model to data with periodic partial autocorrelations of order higher than one. We study the statistical properties of the model, such as mean, variance, marginal and joined distributions. We are adjusting this model to the daily number of people receiving emergency service from the public hospital of the municipality of Vitória for treatment of asthma. Finally, our last extension deals with the introduction of innovations according to a Poisson law with zero inflation whose parameters vary periodically, and on the addition of covariates explaining the logarithm of the intensity of the Poisson's law. We establish some statistical properties of the model, and we use the QML method to estimate its parameters. Finally, we apply this modeling to daily data of the number of people who have visited a hospital's emergency department for respiratory problems, and we use the concentration of a pollutant in the same geographical area as a covariate.

Título : Modelos periódicos e variações aplicados a problemas de saúde

Palavras-Chave : Séries temporais de contagem, periodicidade, Modelo INAR, Modelo PINAR, Modelo ZIP, sazonalidade.

Resumo : Este manuscrito trata de algumas extensões para séries temporais de valores inteiros do modelo paramétrico periódico autorregressivo estabelecido para séries temporais de valores reais. Os modelos considerados baseiam-se no uso do operador de Steutel e Van Harn (1979) e generalizam o processo autorregressivo de números inteiros estacionários (INAR) introduzidos por Al-Osh & Alzaid (1987) para séries de contagem periodicamente correlacionadas. Essas generalizações incluem a introdução de um operador periódico, a consideração de uma estrutura de autocorrelação mais complexa, cuja ordem é maior do que um, o aparecimento de inovações de variâncias periódicas, e também a inflação zero em relação a uma lei discreta dada na família de distribuições exponenciais, bem como o uso de covariáveis explicativas. Essas extensões enriquecem muito o domínio de aplicabilidade dos modelos do tipo INAR. No nível teórico, estabelecemos propriedades matemáticas de nossos modelos como a existência, a unicidade, e a estacionariedade periódica de soluções para as equações que definem os modelos. Propomos três métodos para estimar parâmetros de modelos, incluindo um método de momentos baseado nas equações de Yule-Walker, um método de mínimos quadrados condicionais e um método de quasi-máxima verossimilhança (QML) baseado na maximização de uma probabilidade Gaussiana. Estabelecemos a consistência e a normalidade assintótica desses procedimentos de estimativa. As simulações de Monte Carlo ilustram seus comportamentos para diferentes tamanhos de amostras finitas. Os modelos são então ajustados para dados reais e usados para fins de previsão. A primeira extensão do modelo INAR que propomos consiste na introdução de dois operadores periódicos de Steutel e Van Harn, o primeiro atua modelando as autocorrelações parciais de ordem um em cada período e o outro capturando a sazonalidade periódica dos dados. Através de uma representação vetorial do processo, estabelecemos as condições existência e unicidade de uma solução periodicamente correlacionada às equações que definem o modelo. No caso em que as inovações seguem as leis de Poisson, estudamos a lei marginal do processo. Como um exemplo de aplicação no mundo real, estamos ajustando este modelo aos dados diários de contagem do número de pessoas que receberam antibióticos para o tratamento de doenças respiratórias na região de Vitória, Brasil. Como as condições respiratórias estão fortemente correlacionadas com a poluição do ar e o clima, o padrão de correlação dos números diários de pessoas que recebem antibióticos mostra, entre outras características, a periodicidade semanal e a sazonalidade. Em seguida, estendemos esse modelo para dados com autocorrelações parciais periódicas de ordem maior que um. Estudamos as propriedades estatísticas do modelo, como média, variância, distribuições marginais e conjuntas. Ajustamos esse modelo ao número diário de pessoas com problema respiratório que receberam atendimento de emergência no pronto-atendimento da rede pública do município de Vitória. Finalmente, nossa última extensão trata da introdução de inovações de acordo com uma lei de Poisson com inflação zero cujos parâmetros variam periodicamente, e da adição de covariáveis explicando o logaritmo da intensidade da lei de Poisson. Estabelecemos algumas propriedades estatísticas do modelo e usamos o método QML para estimar seus parâmetros. Por fim, aplicamos essa modelagem aos dados diários sobre o número de pessoas que visitaram o departamento de emergência de um hospital por problemas respiratórios e usamos como covariável a série concentrações diárias de um poluente medido na mesma área geográfica.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France

PhD report:

PERIODIC MODELS AND VARIATIONS APPLIED TO HEALTH PROBLEMS

PhD student:
Paulo Roberto Prezotti Filho

Advisors:
Pascal Bondon
CentraleSupélec, L2S - France

Valdério Anselmo Reisen
Universidade Federal do Espírito Santo PPGEA-UFES- Brazil

January 7, 2019

Contents

List of Figures

List of Tables

List of abbreviations and / or acronyms

AAQMN	Automatic air quality monitoring network
ACF	Autocorrelation function
AIC	Akaike information criterion
AR	Autoregressive model
ARMA	Autoregressive and moving average model
BIC	Bayesian information criterion
Bin(n, α)	Binomial distribution with parameters $n \in \mathbb{N}$ and $\alpha \in [0, 1]$
CA	Cluster analysis
CCF	Cross-correlation function
CML	Conditional maximum likelihood
CO	carbon monoxide
ECOSOFT	Ecosoft consulting and environmental software
ES	Espírito Santo, Brazil
GAM	Generalized additive model
Geo(p)	Geometric distribution with parameter $p \in (0, 1]$
GINAR	Geometric integer autoregressive model
GVR	Greater Vitória region, ES, Brazil
HC	Hydrocarbons
I	Identity matrix
IBGE	Brazilian institute of geography and statistics
ICD	International classification of diseases
IEMA	State institute of environment and water resources
INAR	Integer autoregressive model
MA	Moving average
MCA	Multiple correspondence analysis
MGINAR	Multivariate GINAR
N	Set of positive integers
NO_2	Nitrogen dioxide
NO_X	Nitrogen oxide
O_3	Ozone
PACF	Partial autocorrelation function
PARMA	Periodic autoregressive and moving average model
PC	Periodically correlated processes
PCA	Principal component analysis
PeACF	Periodic autocorrelation function
PePACF	Periodic partial autocorrelation function
PINAR	Periodic integer autoregressive model
PM	Particulate matter
PM_{10}	Particulate matter particles with a diameter of 10 micrometers or less
$PM_{2.5}$	Particulate matter smaller than 2.5 micrometers in diameter
Poi(λ)	Poisson distribution with mean parameter $\lambda \in \mathbb{R}_+$
QML	Quasi-maximum likelihood
$\mathbb{R}$	Set of real numbers
$\mathbb{R}_+$	Set of non-negative real numbers
RAMQAr	Automatic air quality monitoring network of greater Vitória region
RMSE	Root mean squared error
SO_2	Sulfur dioxide
VAR	Vector autoregressive model
VARMA	Vectorial autoregressive model
WHO	World health organization
YW	Yule-Walker
$\mathbb{Z}$	Set of integers
$\mathbb{Z}_+$	Set of non-negative integers
ZINAR	Zero inflated integer autoregressive model
ZIP	Zero inflated Poisson

Chapter 1

Introduction

1 General introduction

A stochastic process is a family of random variables defined on a same probability space. The realizations of a stochastic process are functions of time, which represents the sample-path of the process, i.e., the values which are actually observed. A time series is a sequence of observations of the same random variable at different times, typically with constant period between observations. The time series is frequently used to refer both the data and the process of which it is a realization (Brockwell & Davis (2013)).

The realizations of a stochastic process can be uniformly varying, trending, noisy, integer-valued, or a mixture between these patterns. The count time series represents a specific sequence of counts, a number of times that one event occurs, for example, the daily number of persons that use some special public transport or the monthly number of motor vehicle accidents in a given region.

Statistical models are mathematical structures that seek ways to describe the generating process of stochastic series. Some interesting features of the time series to be investigated directly influence the shape of the model to be applied: the type of data and its probability distribution, stationarity, seasonality, structure of autoregressive autocorrelation, as well as external factors such as the use of covariates (exogenous variables) time-dependent or mathematical functions of these values.

The models proposed in this thesis were inspired by analysis of a group of real data sets related to health problems, and the daily mean concentration of some air pollutants. In fact, our first inspiration was raised after the data mining of the series of the daily number of people who got antibiotics for the treatment of respiratory infection from the public health system in the city of Vitória-ES, Brazil. This data set presents some characteristics that are easily observed in a certain group of real data; however, it becomes differentiated due to having all the characteristics present in a single time series. This is an integer-valued time series, periodically autocorrelated, with serial and seasonal dependence on its autoregressive structure.

The periodicity in this data set was expected. In the literature, several papers give scientific evidence that the effect of pollutants on the respiratory system determines a higher frequency of infections, compromises pulmonary function and increases the risk of developing allergic diseases (Baldacci et al. (2015)). The series related to health problems, especially related to respiratory diseases, are strongly correlated with air pollution levels and climatic conditions (Oudin et al. (2017), Caillaud et al. (2018)). The correlation structure of the daily number series of people receiving antibiotics shows, among other phenomena, periodicity and seasonality, which are often observed in series of daily average concentrations of atmospheric pollutants (Hies et al. (2000)).

The remaining health data sets are referring to the daily number of visits of people affected by respiratory diseases (asthma and rhinitis) to the public health hospital emergency service and the time series of counts referring to the daily number of people who got medicine based on salbutamol sulphate for the treatment of respiratory problems also from the public health hospital emergency service in the city of Vitória-ES, Brazil. These time series also present periodic autocorrelation characteristics. Besides that, they present an autoregressive structure with order larger than one and, for the time series of daily number of visits, it also presents a large number of zeros.

On the modeling of count time series, the Integer-valued Autoregressive model (INAR), introduced by Al-Osh & Alzaid (1987), appears as an alternative to the Poisson's models family. A special advantage of INAR models over the Poisson model is the close similarity to the continuous data modeling with the Box and Jenkins Autoregressive (AR) models. INAR model has the same additive structure of AR models instead of the multiplicative structure presented in Poisson models. This additive characteristic and the discreteness of the modeled process is proportioned by the Thinning Operator "$\circ$".

The identification of periodically correlated process patterns, here simply treated as periodicity, is the subject of research and application in many areas of science, as discussed by Gardner et al. (2006). According to the author, many processes identified in nature arise from periodic phenomena. Although these processes are not necessarily periodic time-dependent functions, they generates random data whose statistical characteristics vary periodically with time. These processes are called cyclostationary.

The occurrence of PC processes is corroborated by real applications in many practical situations, see e.g. Sarnaglia et al. (2010), Basawa & Lund (2001) among others.

Even though many studies in the literature focus on periodically correlated processes, the vast majority are dedicated to the analysis and the applications for continuous data, with application of the PARMA model. Very little attention has been paid to the analysis of periodically correlated count series, excepting, for example, Monteiro et al. (2010), who introduced the periodic integer-valued autoregressive model of order 1, for Poisson distributed data, called PINAR$(1)_S$. The periodic structure of the series was also studied by Moriña et al. (2011), which presented a model based on two-order integer-valued autoregressive time series to analyze the number of hospital emergency service arrivals caused by diseases that present seasonal behavior. They

presented a variation of the INAR(1) and INAR(2) models, where the coefficients involved in the thinning operation are fixed parameters belonging to $(0;1)$, and the mean of the innovations, $\eta_t(\lambda_t)$, follow several Poisson distributed random variable with different means, such that $\lambda_t = \lambda_{t+S}$, where S represents the period of the periodic process observed in X_t.

Models that take into account the seasonal autocorrelation structure for INAR can be seen in the first-order seasonal structure introduced by Bourguignon et al. (2016) or on its extension, the subset INAR(p) process, which account both the first-order serial and seasonal correlations. The class of subset INAR models is investigated in the forthcoming paper Bondon et al. (2018).

Count time series presenting a large frequency of zeros are easily encountered in several areas of science. For example, in biomedical and public health domains, some types of rare diseases with low infection rates can be related to the daily number of hospital admittance's of affected patients. Usually in this type of series the large number of zeros, called zeros inflation, contrasts with high values, which can lead to bad statistical inference and offer spurious relations. Dealing with this kind of data, Yang et al. (2013) extended the classical Zero Inflated Poisson (ZIP) introduced by Lambert (1992) to accommodates count data series with an excess of zeros, autoregressive (AR) autocorrelation and time-dependent covariates regression framework.

The main contribution of this thesis is to propose extensions of periodic counting models to accommodate non-negative integer-valued time series data, which have periodic serial and seasonal autocorrelation characteristics, propose models to count time series periodically autocorrelated with autoregressive order larger than 1, and propose a model to count time series periodically autocorrelated with a large number of zeros. These contributions are presented in four papers, which are shown in the third, fourth and fifth chapters of this thesis.

In the first paper we present an innovative model for counting time series which presents periodic and seasonal autoregressive structure. The stationary condition of the process described by the model is discussed, and then we present the mean of the process and its probability distribution function. We present a likelihood-type method to estimate the parameters of the model, which are evaluated through a simulation study. An application to real data is conducted in order to demonstrate its usefulness.

The model introduced in the first paper is rigorously studied in the second paper. A comprehensive mathematical study of existence, uniqueness and stationary conditions is presented. Statistical properties of the model such as mean, variance, marginal and joint distributions are discussed. We present the probability distribution function of the process for sequences with innovations Poisson and Geometric distributed. A section is devoted to presenting some methods of estimation of the parameters, and their performances are investigated through a simulation study. The consistency and asymptotically normality of the estimators are proved. The model is applied to a real data, and a forecasting procedure is presented.

In the third paper, we present a model to periodically autocorrelated count time series with seasonal period S and autoregressive structure of order p. We discuss the properties of this model

based on time series with Poisson distributed innovations. In the fourth, we extend this model to periodically autocorrelated count time series with seasonal period S and autoregressive structure of order p, which present a large number of zeros. In this last paper, the use of covariates is proposed to study the relationship between health data and air pollution data.

This thesis is structured as follows: This introduction, the research objectives, the study region and the real data used in the applications of the proposed models are in Chapter 1. Chapter 2 is intended to review the literature regarding some statistical models and concepts that were essential for the development of the thesis. As mentioned before, Chapters 3, 4 and 5 refer to the original contributions and results of this thesis, described in the form of four articles. Next, the general conclusions and final comments are presented in Chapter 6, followed by the references used in this research. Lastly, Appendix A is devoted to the co-authoring papers, that served as background for dealing with the real data used in this thesis.

2 Goal and specific objectives

The Air-quality area derives interesting and challenging problems from the different point of views from the quality of life (health etc) and the science and technology, especially, in the probability and statistic fields. In this context, this research shows the praxis between these areas with special attention in proposing statistical models to explain the dynamic of count time series and to provide accurate forecasts with application in series observed in the health centers and from the Automatic Air-Quality Monitoring Network in the Great Vitória Region, ES, Brazil. Based on this direction, the main goal of this thesis is to propose extensions of the periodic integer autoregressive models for counting time series with serial and seasonal periodic autocorrelation structure, with innovations that follow the exponential family of probability distributions such as the Poisson, Geometric and Zero-Inflated Poisson. In the periodic model with zero-inflated data, the covariate concentrations of the pollutant Particulate Matter (PM_{10}) is considered as an explanatory variable. The model and estimation properties are derived and simulations are carried out to show the method estimation performances. The usefulness of the proposed models in real applications is also part of the present thesis. The data sets considered are; the daily number of people who got medicines (antibiotics, salbutamol sulphate based medications) for the treatment of respiratory infections from the public health system, the daily number of visits of people affected by respiratory diseases (asthma and rhinitis) to the public health emergency service and the pollutant concentration (PM_{10}).

Specific objectives

- To propose the PINAR$(1, 1_S)$ model for count time series that presents both serial and seasonal periodic autocorrelation structures and to derive the model and estimation properties.

- To fit the PINAR$(1, 1_S)$ model to a real data set (health data) to describe the dynamic of the data and to compute accurate forecasts.
- To extend PINAR$(1, 1_S)$ model to the PINAR$(p)_S$ model and its model and estimation properties.
- To fit the PINAR$(p)_S$ model to a real data set (health data).
- To extend PINAR$(p)_S$ model to the regression ZIP-PINAR$(p)_S$ model with the use of explanatory covariates.
- To apply the regression ZIP-PINAR$(p)_S$ model to evaluate and describe the relationship between the response variable (health data) and the explanatory covariates (air pollutant concentration data).

3 Region of study

Vitória is the capital of the state of Espírito Santo, and with the municipalities of Cariacica, Fundão, Guarapari, Serra, Viana, and Vila Velha it integrates a geographic area of urbanization and high industrial development, denominated Region of Greater Vitória (RGV) (Figure 1.1). It is located on the coast, at 12 meters of altitude. The climate is tropical humid, with average temperature varying from 24.4°C to 34.4°C. The RGV has 55% to 65% of the potentially polluting industrial activities installed in Espírito Santo, such as: Steel, Pelletizing, Quarry, Cement, Food industry, Asphalt Plant, among others (IEMA (2015)). The municipality of Vitória concentrates activities with high polluting potential within the urban network and, in recent years, there has been a significant increase in the number of motor vehicles. According to the Brazilian Institute of Geography and Statistics (IBGE), the municipality has a population of 327 801 inhabitants distributed in an area of 98 194 Kms (IBGE (2010)).

The monitoring of air quality in the RGV is carried out by the State Institute of Environment (IEMA), which has an Automatic Air Quality Monitoring Network (RAMQAr). In the municipality of Vitória, 3 monitoring stations are located (Enseada do Suá, Downtown Vitória and Jardim Camburi), see Figure 1.2. We are specifically interested in the Enseada do Suá station, because it is located at a strategic point of Greater Victoria and covers a large area. In addition, it is directly influenced by the industrial emissions and by the mobile sources that converge to that area. (IEMA (2015)).

The study area will be the area covered by the Health Unit of Praia do Suá, composed of the neighborhoods of the Enseada do Suá, Praia do Suá, Bento Ferreira, Santa Helena and Ilha do Boi. The criteria used for the selection of the area were that these sites are less than 2 km from the air quality monitoring station of Enseada do Suá Station and this station is the only one in the municipality that measures all pollutants monitored by RAMQAr, including PM_{10} - particles with a diameter of less than 10 mm (IEMA (2013)).

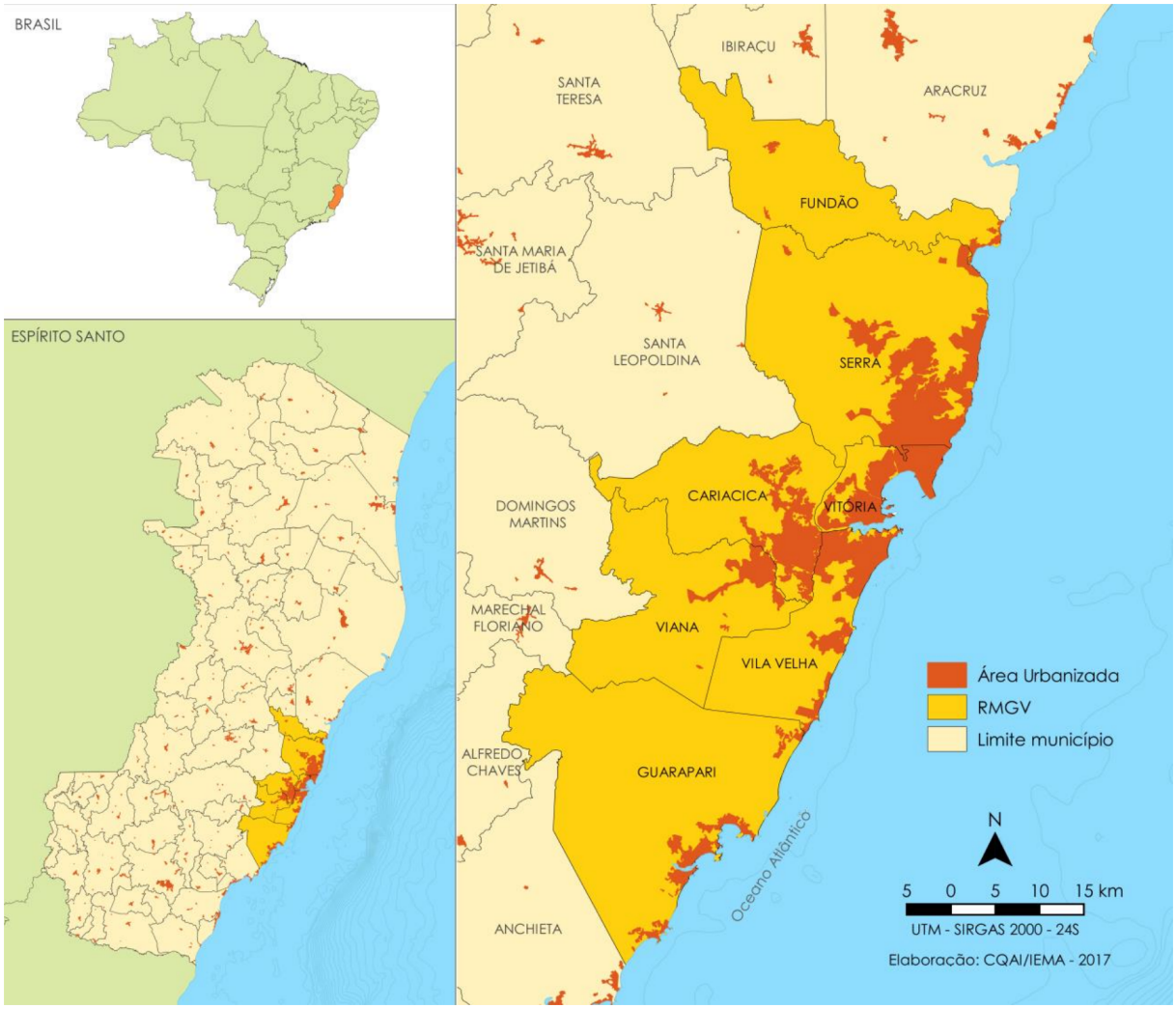


Figure 1.1: Location of the Metropolitan area of Region of Greater Vitória.

4 Health and pollution variables

The models introduced in this thesis were applied to four real count time series related to health data. The first set, used in the application of the PINAR$(1, 1_S)$ model, is related to the daily number of people who received antibiotics from the public health care system in the hospital emergency service of the Praia do Suá, at the region of Vitória, Brazil. The observed series corresponds to the period of May 26, 2013, to September 07, 2015, resulting in 834 daily observations.

The second real data set was used in the application of the PINAR$(p)_S$ model, and is the time series of counts referring to the daily number of people who got medicine based on salbutamol sulphate for the treatment of respiratory problems from the public health care system in the hospital emergency service of the region of Vitória-ES, Brazil. The series covers from January 03, 2013 to July 18, 2017, resulting in 1659 daily observations. Both of these data sets were obtained from the Drug Dispensing Data Registration Network of the municipal health secretariat of Vitória.

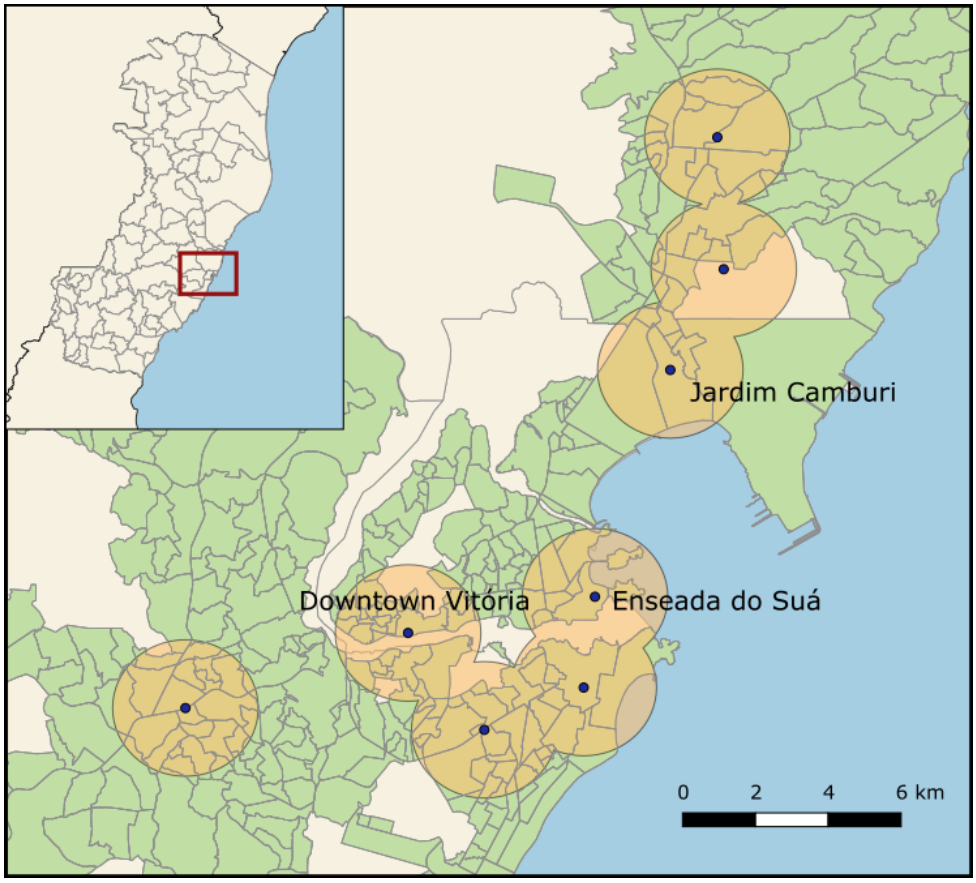


Figure 1.2: Location of the monitoring stations of Vitória.

The third set is about the time series of counts referring to the daily number of visits of children with respiratory problems (International Classification of Diseases ICD-10) in the emergency service of the public health care system of the region of Vitória-ES. This data was used as a second application of the PINAR$(1, 1_S)$ model. The period of the study covers from June 26, 2013 to April 7, 2016, resulting in 1022 daily observations. The last data set, used on the application of Regression ZIP-PINAR$(p)_S$ model, is the time series of counts referring to the daily number of hospital emergency service visits of people with respiratory airway diseases, classified according to International Classification of Diseases (ICD-10, j31 and j45). The data selected were the people of any age group who visited the hospital emergency service in Vitória-ES city , specifically those living in the neighborhoods of Praia do Suá, Enseada do Suá, Bento Ferreira and Ilha do Boi. These data sets were obtained from the network records system Welfare (*Bem-Estar Network*) of the municipality.

For the application to the real data set of the Regression ZIP-PINAR$(p)_S$ model, the use of the daily average of concentration levels of Particulate Matter (PM_{10}) was proposed as covariate of the modeling of the health data. These measures were obtained from the IEMA, with data collected in Enseada do Suá Station, Vitória-ES, Brazil, belonging to the Automatic Network of Air Quality Monitoring (RAMQAr). The data collection comprised a 24-hour period, which began in the first half hour of the day. The average of 24 hours was considered.

Chapter 2

Overview of some models and properties

The development of the models presented in this thesis, were based on the study of some known models of the literature. In what follows, we present some of these models.

1 The integer autoregressive models (INAR)

1 .1 Introduction

A time series is a set of data evolving randomly over time, in which the order of the data is fundamental. When a time series of counting is observed, one of the main purposes is to find a model that conforms to it as faithfully as possible, and is able to describe the dynamics of the information observed.

Time series of non-negative integer values, also referred to as count series, are naturally present in our daily lives, usually associated with counting processes in a given time interval. As examples of the application of these models, we have some phenomena, such as the daily number of visits to hospital emergency service by people with a certain illness, the monthly number of work accidents in a certain company or country, the weekly number of defects in the products from an assembly line of an industry, the daily number of medicines dispensed to the patients of an hospital emergency service due to a specific disease, among other several examples of counting of cases that occur during a certain period of time. Given the frequent presence of these phenomena, the interest and the need to study modeling methods for these time series of counting have arisen and such methods are now an emerging field of science.

There are several classes of models proposed in the literature for the analysis of a time series, In the general class of linear models, the Box-Jenkins autoregressive linear models are widely used to model stationary dependent time series under the Gaussianity hypothesis. However, this assumption is inadequate for modeling non-negative integer-valued processes, called the

counting processes.

The count time series analog of an autoregression is the integer-valued autoregressive INAR class of models. The INAR models, initially introduced by the INAR(1) model in Al-Osh & Alzaid (1987), appears as an alternative to the well-known Poisson model family for modeling count time series, see, e.g., Fokianos et al. (2009). These models are based on the thinning operator, see Steutel & Van Harn (1979). In what follows, the thinning operator will be defined based on the Binomial distribution (for alternative thinning concepts see, for example, Weiß (2008)).

1 .2 The binomial thinning operator

The binomial thinning operator $\alpha\circ$ for a random variable (r.v.) Y is defined as

$$\alpha \circ Y = \sum_{i=1}^{Y} U_i(\alpha),$$

where Y is a $\mathbb{Z}_+$-valued r.v., $\alpha \in [0,1]$ and $\{U_i(\alpha)\}_{i\in\mathbb{Z}_+}$ is a sequence of independent identically distributed (i.i.d.) r.v.'s which are Bernoulli distributed with parameter α. It is assumed that the sequence $\{U_i(\alpha)\}_{i\in\mathbb{Z}_+}$ is mutually independent of Y. Note that the empty sum is set to 0 if $Y = 0$. The sequence $\{U_i(\alpha)\}_{i\in\mathbb{Z}_+}$ is called a counting sequence. Remark that the probability of success in the thinning is $\mathsf{P}(U_i(\alpha) = 1) = \alpha$ and, conditionally on Y, $\alpha \circ Y \sim \mathrm{Bin}(Y, \alpha)$.

The special assignment of this operator is to "replace" the usual scalar multiplication, making it possible for the result of the binomial thinning operator between a real $\alpha \in [0,1]$ and a non-negative integer value to still be a non-negative integer value. We present here some properties of thinning based count time series models (da Silva & Oliveira (2004)). Let X and Y be non-negative integer-valued random variables. Then for any $\alpha, \beta \in [0,1]$, we have that

(i) $0 \circ X = 0$

(ii) $1 \circ X = X$

(iii) $\alpha \circ (\beta \circ X) \stackrel{d}{=} (\alpha\beta) \circ X$ where $\stackrel{d}{=}$ stands for equal in distribution.

(iv) $\mathsf{E}(\alpha \circ X) = \alpha\mathsf{E}(X)$

(v) $\mathsf{E}(\alpha \circ X)^2 = \alpha^2\mathsf{E}(X^2) + \alpha(1-\alpha)\mathsf{E}(X)$

(vi) $\mathsf{E}(\alpha \circ X)^3 = \alpha^3\mathsf{E}(X^3) + 3\alpha^2(1-\alpha)\mathsf{E}(X^2) + \alpha(1-\alpha)(1-2\alpha)\mathsf{E}(X)$

(vii) $\mathsf{E}(X(\alpha \circ Y)) = \alpha\mathsf{E}(XY)$

(viii) $\mathsf{E}(X(\alpha \circ Y)^2) = \alpha^2\mathsf{E}(XY^2) + \alpha(1-\alpha)\mathsf{E}(XY)$

(ix) if X and Y are independent, then $\mathsf{E}((\alpha \circ X)(\beta \circ Y)) = \alpha\beta\mathsf{E}(X)\mathsf{E}(Y)$

(x) $\mathsf{E}((\alpha \circ X)(\beta \circ Y)) = \alpha\beta\mathsf{E}(XY)$ if the counting series of $\alpha \circ X$ and $\beta \circ Y$ are independent, and independent of X and Y.

(xi) $\mathsf{E}((\alpha \circ X)^2(\beta \circ Y)) = \alpha^2\beta\mathsf{E}(X^2Y) + \alpha(1-\alpha)\beta\mathsf{E}(XY)$ if the counting series of $\alpha \circ X$ and $\beta \circ Y$ are independent, and independent of X and Y.

(xii) $\mathsf{E}(XY(\beta \circ Z)) = \beta\mathsf{E}(XYZ)$

(xiii) $\mathsf{E}(X(\beta \circ Y)(\gamma \circ Z) = \beta\gamma\mathsf{E}(XYZ)$ if the counting series of $\beta \circ Y$ and $\gamma \circ Z$ are independent, and independent of X, Y and Z.

(xiv) $\mathsf{E}((\alpha \circ X)(\beta \circ Y)(\gamma \circ Z)) = \alpha\beta\gamma\mathsf{E}(XYZ)$ if the counting series of $\alpha \circ X$, $\beta \circ Y$ and $\gamma \circ Z$ are independent, and independent of X, Y and Z.

(xv) $\mathrm{Var}(\alpha \circ X) = \alpha^2 \mathrm{Var}(X) + \alpha(1-\alpha)\mathsf{E}(X)$.

(xvi) $\mathrm{Cov}(\alpha \circ X, X) = \alpha \mathrm{Var}(X)$.

(xvii) $\mathrm{Cov}(\alpha \circ X, \beta \circ Y) = \alpha\beta \mathrm{Cov}(X, Y)$

Proofs.

The proof of properties (i), (ii) and (iii) follow from the definition; properties (iv), (v), (ix) and (x) are from Al-Osh & Alzaid (1987), Du & Li (1991) and Franke & Seligmann (1993); properties (vi), (vii), (viii), (xi), (xii) follow from the definition of $\circ$ and the conditional expectation.

(xv) From the Köenig formula

$$\begin{aligned}\mathrm{Var}(\alpha \circ X) &= \mathsf{E}[(\alpha \circ X)^2] - [\mathsf{E}(\alpha \circ X)]^2 \\ &= \alpha^2\mathsf{E}[(X^2) + \alpha(1-\alpha)\mathsf{E}(X) - [\alpha\mathsf{E}(X)]^2 \\ &= \alpha^2 \mathrm{Var}(X) + \alpha(1-\alpha)\mathsf{E}(X)\end{aligned}$$

(xvi) $\mathrm{Cov}(\alpha \circ X, X) = \mathsf{E}[(\alpha \circ X)X)] - [\mathsf{E}(\alpha \circ X)\mathsf{E}(X)$, where

$$\mathsf{E}[(\alpha \circ X)X] = \mathsf{E}\left[\mathsf{E}\left[\left(\sum_{j=1}^{X} Y_j\right) X|X\right]\right] = \mathsf{E}\left[X\sum_{j=1}^{X}\mathsf{E}\left(Y_j|X\right)\right] = \mathsf{E}\left[X\sum_{j=1}^{X}\mathsf{E}\left(Y_j\right)\right] = \alpha\mathsf{E}\left(X^2\right).$$

(xviii) $\mathrm{Cov}(\alpha \circ X, \beta \circ Y) = \mathsf{E}[(\alpha \circ X)(\beta \circ Y)] - \mathsf{E}(\alpha \circ X)\mathsf{E}(\beta \circ Y)$.

For more details on thinning based count time series models see, e.g., Scotto et al. (2015) in the univariate and Latour (1997) in the multivariate case, respectively.

1 .3 The INAR(1) model

The model for non-negative integer-valued time series with autoregressive order equal to 1, the INAR(1) model, was introduced by Al-Osh & Alzaid (1987). This model is a widely used tool for the modeling of counting processes. $\{Y_t\}_{t\in\mathbb{Z}}$ is an INAR(1) process if it satisfies the following stochastic recursion

$$Y_t = \alpha \circ Y_{t-1} + \varepsilon_t, \tag{2.1}$$

where $\alpha \in [0,1]$ is the autoregressive coefficient and $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ is a sequence of $\mathbb{Z}_+$-valued i.i.d.r.v.'s, with finite mean $\mathsf{E}(\varepsilon_t) = \lambda$ and finite variance $\mathrm{Var}(\varepsilon_t) = \sigma^2 > 0$. As pointed out by Al-Osh & Alzaid (1987) and Du & Li (1991), if $0 \leq \alpha < 1$, then (2.1) has a unique second-order stationary solution Y_t and ε_t is independent of Y_{t-1} and $\alpha \circ Y_{t-1}$.

As can be seen, for each time t, Y_t in (2.1) has two random components; the immigration of the immediate past Y_{t-1} with survival probability α and the elements which entered in the system in the interval (t-1, t], which define the innovation term ε_t for all $t \in \mathbb{Z}$.

If $0 \leq \alpha < 1$, we have :

(i) $\mathsf{E}[Y_t] = \frac{\lambda}{1-\alpha}$ and $\mathsf{E}[Y_t^2] = \frac{\alpha\lambda+\sigma^2}{1-\alpha^2} + \frac{\lambda^2}{(1-\alpha)^2}$.

(ii) For all $h \in \mathbb{Z}$,

$$\gamma(h) = \mathrm{Cov}(Y_{t-h}, Y_t) = \mathrm{Cov}(Y_{t-h}, \alpha^h \circ Y_{t-h}) + \mathrm{Cov}(Y_{t-h}, \sum_{j=0}^{h-1} \alpha^j \circ \varepsilon_{t-j})$$
$$= \alpha^h \,\mathrm{Var}(Y_{t-h}) + \sum_{j=0}^{h-1} \alpha^j \,\mathrm{Cov}(Y_{t-h}, \varepsilon_{t-j}) = \alpha^h \gamma(0).$$

(iii) The representation for the marginal distribution of the INAR(1) model expressed in terms of the innovation sequence is given by $Y_t \stackrel{d}{=} \sum_{j=0}^{\infty} \alpha^j \circ \varepsilon_{t-j}$.

Proof: (i) follows from the definition, (ii) and (iii) are due to Al-Osh & Alzaid (1987, eqns. (3.3) and (2.2)), respectively.

Since the counting r.v. and the immigration process involved in (2.1) are mutually independent, the conditional probability distribution of the INAR(1) process is given by

$$\mathsf{P}(Y_t = y_t | Y_s = y_s,\ s = 1, \ldots, t-1) = \mathsf{P}(Y_t = y_t | Y_{t-1} = y_{t-1}) = \mathsf{P}(\alpha \circ y_{t-1} + \varepsilon_t = y_t). \quad (2.2)$$

Poisson INAR(1)

In (2.1), let ε_t be a Poisson distributed r.v., $\varepsilon_t \sim \mathrm{Poi}(\lambda)$, then $\{Y_t\}_{t\in\mathbb{Z}}$ is called a Poisson INAR(1) process. Al-Osh & Alzaid (1987) have shown that $Y_t \sim \mathrm{Poi}(\lambda/(1-\alpha))$ for all $t \in \mathbb{Z}$ when $Y_0 \sim \mathrm{Poi}(\lambda/(1-\alpha))$. Based on (2.2), the conditional probability distribution of the INAR(1) process $\{Y_t\}_{t\in\mathbb{Z}}$ is given by

$$p(y_t|y_{t-1}) = \mathsf{P}(Y_t = y_t | Y_{t-1} = y_{t-1}) = [\mathrm{Bin}(y_{t-1}, \alpha) * \mathrm{Poi}(\lambda)] =$$
$$\sum_{i=0}^{y_{t-1}\wedge y_t} \binom{y_{t-1}}{i} (\alpha)^i (1-\alpha)^{y_{t-1}-i} \exp(-\lambda) \frac{\lambda^{y_t-i}}{(y_t-i)!}, \quad (2.3)$$

where $*$ denotes convolution and $a \wedge b = \min\{a, b\}$.

Estimation methods and asymptotic properties

As pointed out by Al-Osh & Alzaid (1987), the estimation of the parameters of the INAR(1) process is more complicated than that of the AR(1) process because the conditional distribution of Y_t, given Y_{t-1}, is the convolution of the distribution of ε_t and a binomial with scale α and index Y_{t-1}. In what follows, we assume that $\varepsilon_t \sim \mathrm{Poi}(\lambda)$.

Yule-Walker (YW) estimators

The estimates of α and λ are obtained by replacing μ and $\gamma(h)$ with the sample mean $\overline{y}$ and sample autocovariance function, respectively, into the well known YW equations, obtained by multiplying (2.1) by Y_{t-i}, $i = 0, 1$, and taking expectations of both sides, which leads to

$$\begin{aligned}\widehat{\alpha}^{\mathrm{YW}} =& \frac{\sum_{t=1}^{n}(y_t - \overline{y})(y_{t-1} - \overline{y})}{\sum_{t=1}^{n}(y_t - \overline{y})^2},\\ \widehat{\lambda}^{\mathrm{YW}} =& n^{-1}\sum_{t=1}^{n}(y_t - \widehat{\alpha}^{\mathrm{YW}} y_{t-1}),\end{aligned}$$

where

$$\overline{y} = n^{-1}\sum_{k=1}^{n} y_k.$$

Conditional least squares (CLS) estimators

The CLS estimators $\widehat{\boldsymbol{\vartheta}}_n^{\mathrm{CLS}}$, $n \in \mathbb{N}$, of $\boldsymbol{\vartheta} = (\alpha, \lambda)^\top$ are obtained by minimizing the expression

$$Q_n(\boldsymbol{\vartheta}) = \sum_{t=1}^{n}(Y_t - \mathsf{E}(Y_t|Y_{t-1}))^2. \tag{2.4}$$

with respect to $\boldsymbol{\vartheta}$. From the derivatives of (2.4) with respect to α and λ, we obtain

$$\begin{aligned}\widehat{\alpha}^{\mathrm{CLS}} =& \frac{\sum_{t=1}^{n} y_t y_{t-1} - (\sum_{t=1}^{n} y_t)(\sum_{t=1}^{n} y_{t-1})/n}{\sum_{t=1}^{n} y_{t-1}^2 - (\sum_{t=1}^{n} y_{t-1})^2/n},\\ \widehat{\lambda}^{\mathrm{CLS}} =& n^{-1}\sum_{t=1}^{n}(y_t - \widehat{\alpha}^{\mathrm{CLS}} y_{t-1}).\end{aligned}$$

One can see that the real-valued penalty function $Q_n(\cdot)$ satisfies the assumptions of Theorem 3.1 in Klimko & Nelson (1978). The CLS estimator $\widehat{\boldsymbol{\vartheta}}^{\mathrm{CLS}} = (\widehat{\alpha}^{\mathrm{CLS}}, \widehat{\lambda}^{\mathrm{CLS}})$ of the parameter vector $\boldsymbol{\vartheta}$ is strongly consistent. In addition, if $\mathsf{E}(|\varepsilon_t|^3) < \infty$, which is verified for Poisson distribution, then by Theorem 3.2 of Klimko & Nelson (1978), $(\widehat{\alpha}^{\mathrm{CLS}}, \widehat{\lambda}^{\mathrm{CLS}})$ are asymptotically normally distributed as

$$n^{1/2}(\widehat{\boldsymbol{\vartheta}}_n^{\mathrm{CLS}} - \boldsymbol{\vartheta}_0) \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0}, V^{-1}WV^{-1}),$$

as $n \to \infty$, where $\boldsymbol{\vartheta}_0 = (\alpha_0, \lambda_0)$ denotes the 'true' value of the parameters, and the matrices V and W of dimension 2×2 with elements defined by

$$V_{i,j} = \mathsf{E}\left(\frac{\partial \mathsf{E}(Y_t|Y_{t-1})}{\partial \vartheta_i}\frac{\partial \mathsf{E}(Y_t|Y_{t-1})}{\partial \vartheta_j}\right)$$

and

$$W_{i,j} = \mathsf{E}\left((Y_t - \mathsf{E}(Y_t|Y_{t-1}))^2\frac{\partial \mathsf{E}(Y_t|Y_{t-1})}{\partial \vartheta_i}\frac{\partial \mathsf{E}(Y_t|Y_{t-1})}{\partial \vartheta_j}\right),$$

where $i, j = 1, 2$, and $(Y_t - \mathsf{E}(Y_t|Y_{t-1})$ and $\mathsf{E}(Y_t|Y_{t-1})$ are evaluated over $\boldsymbol{\vartheta}_0$.

Conditional maximum likelihood (CML) estimators

Based on the probability function defined in (2.3), the conditional likelihood function of the PINAR(1) model (conditioned on the first observation, which presents negligible influence when the sample size is large) can be written as

$$L_n(\boldsymbol{\vartheta}) = \prod_{t=1}^{n} p(y_t|y_{t-1}).$$

The conditional log-likelihood function $l_n(\boldsymbol{\vartheta}) = \log L_n(\boldsymbol{\vartheta}) = \sum_{t=1}^{n} \log p(y_t|y_{t-1})$ is maximized in order to obtain the CML estimator $\widehat{\boldsymbol{\vartheta}}_n^{\mathsf{CML}}$ of the parameter vector $\boldsymbol{\vartheta}$.

After a large set of simulation experiments using these three estimation methods, the conclusion stated by Al-Osh & Alzaid (1987) is that the CML estimators are the best, followed by the CLS and the YW estimators, respectively. Regarding the bias and MSE, in the CML estimates, when compared to the other two methods, it is worth the extra calculations.

1 .4 The INAR(p) model

The INAR(p) model, introduced independently by Alzaid & Al-Osh (1990) and Du & Li (1991), is an extension of the INAR(1) model that accounts the p-th order autoregressive structure. The two different approaches imply different second-order structure for the processes: the Du & Li (1991) formulation implies that the autocorrelation function of the process is the same as that of an AR(p) model, whereas the Alzaid & Al-Osh (1990) formulation gives an autocorrelation function of an ARMA$(p, p-1)$ process. Here, based on the approach of Du & Li (1991), the model and some of its statistical properties will be discussed.

A $\mathbb{Z}_+$-valued process $\{Y_t\}_{t\in\mathbb{Z}}$ is said to be an INAR(p) process if it satisfies the following equation

$$Y_t = \sum_{i=1}^{p} \alpha_i \circ Y_{t-i} + \varepsilon_t, \tag{2.5}$$

where $\alpha_i \in [0, 1]$ for all i and $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ is a sequence of $\mathbb{Z}_+$-valued i.i.d.r.v.'s, with finite mean $\mathsf{E}(\varepsilon_t) = \lambda$ and finite variance $\mathrm{Var}(\varepsilon_t) = \sigma^2 > 0$. Moreover, all counting r.v.'s are mutually independent and are independent of the sequence $\{\varepsilon_t\}_{t\in\mathbb{Z}}$.

In the following, we present the conditions on $\alpha_1, \ldots, \alpha_p$ obtained by Du & Li (1991) which guarantee the existence of a unique second-order stationary solution Y_t to (2.5). For this, we introduce the matricial Steuel and Van Harn operator.

Let $A\circ = (a_{i,j}\circ)$, $1 \leq i, j \leq p$, be a $p \times p$ *matricial binomial thinning operator*, also called the matricial Steuel and Van Harn operator, where $a_{i,j} \in [0, 1]$ for all $1 \leq i, j \leq p$. The action of $A\circ$ on $\boldsymbol{Y} = (Y_1, \ldots, Y_p)^\top$ is

$$A \circ \boldsymbol{Y} = \begin{pmatrix} \sum_{j=1}^{p} a_{1,j} \circ Y_j \\ \vdots \\ \sum_{j=1}^{p} a_{p,j} \circ Y_j \end{pmatrix}. \tag{2.6}$$

The operator $a_{i,j}\circ$, $1 \leq i, j \leq p$, is based on a sequence $\{U_l(a_{i,j})\}_{l\in\mathbb{Z}_+}$ of i.i.d.r.v.'s with a Bernoulli distribution. Based on Lemma 2.1 in Latour (1997), $\mathsf{E}(A \circ \boldsymbol{Y}) = A\mathsf{E}(\boldsymbol{Y})$, where $A = (a_{i,j})$, $1 \leq i, j \leq p$.

The matrix A related to model (2.5) is defined by

$$A\circ = \begin{bmatrix} \alpha_1\circ & \alpha_2\circ & \alpha_3\circ & \cdots & \alpha_{p-1}\circ & \alpha_p\circ \\ 1\circ & 0\circ & 0\circ & \cdots & 0\circ & 0\circ \\ 0\circ & 1\circ & 0\circ & \cdots & 0\circ & 0\circ \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0\circ & 0\circ & 0\circ & \cdots & 1\circ & 0\circ \end{bmatrix}. \tag{2.7}$$

Let $\boldsymbol{Y}_t = (Y_t, Y_{t-1}, \ldots, Y_{t-p+1})^\top$ and $\boldsymbol{\varepsilon}_t = (\varepsilon_t, 0, \ldots, 0)^\top$. The state-space representation of (2.5) is

$$\begin{aligned} \boldsymbol{Y}_t &= A \circ \boldsymbol{Y}_{t-1} + \boldsymbol{\varepsilon}_t, \\ Y_t &= B\boldsymbol{Y}_t, \end{aligned}$$

where $A\circ$ is defined by (2.7) and $B = (1, 0, \ldots, 0)$.

Let $\rho(A)$ be the spectral radius of matrix A, i.e., the maximum eigenvalue in modulus of A. Du & Li (1991) have shown that $\rho(A) < 1$ is a necessary and sufficient condition for the existence of a unique second-order stationary solution Y_t to (2.5). Moreover, this solution is causal in the sense that ε_t is independent of the past of Y_t, i.e. the σ-algebra generated by the random variables Y_{t-h} for $h > 0$.

2 Periodically autocorrelated series

In this section, the time index t is written as $t = kS + \nu$, where $\nu = 1, \ldots, S$ and $k \in \mathbb{Z}$, when emphasis on seasonality S is important. For example, in the case of daily data and weekly seasonality, $S = 7$, ν is the day of the week and k is the index of the week.

Let $\{Y_t\}$, $t \in \mathbb{Z}$, be a integer-valued stochastic process satisfying $\mathsf{E}(Y_t^2) < \infty$ for all $t \in \mathbb{Z}$. Denote the mean function $\mu(t) = \mathsf{E}(Y_t)$, and the covariance function

$$\gamma_{k,\nu}(h) = \mathrm{Cov}(Y_{kS+\nu}, Y_{kS+\nu-h}), \qquad h \in \mathbb{Z}.$$

The process $\{Y_t\}_{t\in\mathbb{Z}}$ is said to be PC with period S, $S \in \mathbb{N}$, if for $\nu = 1, \ldots, S$ and all $k \in \mathbb{Z}$,

(i) $\mu(kS + \nu) = \mu_\nu$;

(ii) $\gamma_{k,\nu}(h) = \gamma_\nu(h)$.

That is, if mean and covariances do not depend on k. This definition implies that the mean and covariance are periodic functions with period S. If $S = 1$, $\{Y_t\}_{t\in\mathbb{Z}}$ is weakly stationary in the usual sense.

Vecchia (1985) introduced the periodic autoregressive moving average (PARMA) model for real-valued time series. This model provides some tools for modeling series with properties that vary periodically within some basic time unit. A multivariate representation of the PARMA model was used to derive parameter space restrictions and difference equations for the periodic autocorrelations. The PARMA model was applied to model hydrologic time series.

Lund & Basawa (2000) explored the recursive prediction and likelihood evaluation techniques for PARMA models. Basawa & Lund (2001) studied the asymptotic properties of parameter estimates for causal and invertible PARMA models.

The models for counting time series, which contributed with the development of this thesis are briefly presented below.

2 .1 The PINAR$(1)_S$ model

The integer-valued autoregressive process with periodic structure, PINAR$(1)_S$ model was introduced by Monteiro et al. (2010) by extending the conventional periodic autoregressive model. $\{Y_t\}_{t\in\mathbb{Z}}$ is said to be a PINAR$(1)_S$ process with seasonal period $S \in \mathbb{N}$, if it satisfies the recursive equation

$$Y_t = \phi_t \circ Y_{t-1} + \varepsilon_t,$$

where $\phi_t = \alpha_\nu$, $t = kS + \nu$, $k \in \mathbb{Z}$ and $\nu = 1, \ldots, S$. The innovation process $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ is a sequence of $\mathbb{Z}_+$-valued r.v.'s such that for each $\nu \in \{1, \ldots, S\}$, the sequence $\{\varepsilon_{kS+\nu}\}_{k\in\mathbb{Z}}$ consists of independent Poisson-distributed r.v.'s with mean λ_ν, $\varepsilon_{kS+\nu} \sim \mathrm{Poi}(\lambda_\nu)$. The stationarity and ergodicity of the process are established. The moments-based (Yule-Walker), conditional least squares-type and conditional maximum likelihood parameters estimation methods are presented and the asymptotic distribution of the estimators is discussed. Their performances are compared through a simulation study. The paper does not present any application of the model to a real data set.

2 .2 An INAR$(2)_S$ model

Moriña et al. (2011) presented an integer-valued autoregressive time series model of order 2 to analyze the number of hospital emergency service arrivals caused by diseases that present seasonal behavior. We also introduce a method to describe this seasonality, on the basis of Poisson innovations with monthly means. Their model is given by

$$Y_t = p_1 \circ Y_{t-1} + p_2 \circ Y_{t-2} + \eta_t,$$

where the coefficients p_1 and p_2 involved in the thinning operations are fixed parameters in $]0;1[$, $\eta_t \sim \mathrm{Poi}(\lambda_t)$ with $\lambda_t = \lambda_{t+S}$ and $S \in \mathbb{N}$ is the period of the periodic process $\{Y_t\}$. This model is a particular case of the first model proposed in this thesis. The maximum likelihood parameter estimation method is discussed, and some methods for forecasting, on the basis of long-time means, and short-time and long-time prediction regions are presented. The proposed model was applied to model the number of hospital admissions per week caused by influenza.

2 .3 The MGINAR$(1)_S$ model

Latour (1997) introduces the multivariate GINAR(p), (MGINAR(1)) process. The matricial representation of the MGINAR(1) is based on the Steuel and van Harn matricial operator defined in (2.6), and some rules for the computation of the expected value of basic expressions involving this operator are presented at Section 2 of this paper. Lemma 2.1 of Latour (1997) presents basic properties of the matricial operators, some of then are given below. Let $A\circ$ and $M\circ$ be $p\times p$-matricial generalized Steuel and van Harn operators. Consider that $A = (\alpha_{i,j})$, $1 \leq i,j \leq p$ and $B = (\beta_{i,j})$, $1 \leq i,j \leq p$ denotes the mean and variance, respectively, of the operator $A\circ$, and M is the mean of the operator $M\circ$. Let $\mathbf{X}$ and $\mathbf{Y}$ be non-negative integer-valued random p-vectors. Then

(i) $\mathsf{E}[A \circ \mathbf{X}] = A\mathsf{E}[\mathbf{X}]$;

(ii) $\mathsf{E}[A \circ \mathbf{X}(A \circ \mathbf{X})^\top] = \mathrm{diag}(B)\mathsf{E}[\mathbf{X}] + A\mathsf{E}[\mathbf{X}\mathbf{X}^\top]A^\top$;

(iii) $\mathsf{E}[(A \circ \mathbf{X})^\top A \circ \mathbf{X}] = \mathbf{1}^\top B\mathsf{E}[\mathbf{X}] + \mathrm{trace}(A\mathsf{E}[\mathbf{X}\mathbf{X}^\top]A^\top)$;

(iv) if the counting series involved in $A\circ$ are independent of the counting series involved in $M\circ$, then $\mathsf{E}[A \circ \mathbf{X}(M \circ \mathbf{Y})^\top] = A\mathsf{E}[\mathbf{X}\mathbf{Y}^\top]M^\top$;

Proof. (i) The *i*th element of $A \circ \mathbf{X}$ being $\sum_{j=1}^{p} \alpha_{ij} \circ \mathbf{X}_j$, the result follows by talking the expectation.

(ii) In a similar way, the element (i,j) of $A \circ \mathbf{X}(A \circ \mathbf{X})^\top$ is $\sum_{k=1}^{p}\sum_{l=1}^{p} \alpha_{i,k} \circ \mathbf{X}_k\alpha_{j,l} \circ \mathbf{X}_l$. Taking the expectation gives

$$\sum_{k=1}^{p}\sum_{l=1}^{p} \mathsf{E}[\alpha_{i,k} \circ \mathbf{X}_k\alpha_{j,l} \circ \mathbf{X}_l] = \begin{cases} \sum_{k=1}^{p} \beta_{i,k}\mathsf{E}[\mathbf{X}_k] + \sum_{k=1}^{p}\sum_{l=1}^{p} \alpha_{i,k}\alpha_{i,l}\mathsf{E}[\mathbf{X}_k\mathbf{X}_l], & i = j, \\ \sum_{k=1}^{p}\sum_{l=1}^{S} \alpha_{i,k}\alpha_{j,l}\mathsf{E}[\mathbf{X}_k\mathbf{X}_l], & i \neq j. \end{cases}$$

So, $\mathsf{E}[A \circ \mathbf{X}(A \circ \mathbf{X})^\top] = \mathrm{diag}(B)\mathsf{E}[\mathbf{X}]) + A\mathsf{E}[\mathbf{X}\mathbf{X}^\top]A^\top$.

(iii) Since $\mathsf{E}[A \circ \mathbf{X})^\top A \circ \mathbf{X}] = \mathrm{trace}[\mathsf{E}[A \circ \mathbf{X}(A \circ \mathbf{X})^\top]]$, using (ii) one sees that this expression is equal to

$$\mathbf{1}^\top B\mathsf{E}[\mathbf{X}] + \mathrm{trace}(A\mathsf{E}[\mathbf{X}\mathbf{X}^\top]A^\top) \ .$$

(iv) The fourth assertion is proved in a similar way.

Based on an extension of Theorem 2.1 of Du & Li (1991), Latour (1997) presents a criterion for the existence of a stationary and causal multivariate integer-valued autoregressive process, MGINAR(p). The autocovariance function of this process is proved to be identical to the autocovariance function of a standard Gaussian multivariate AR(p).

3 The zero inflated Poisson (ZIP) model

Count time series presenting a large frequency of zeros are easily encountered in several areas of science. For example, in biomedical and public health domains, some types of rare diseases with low infection rates can be related to the daily number of hospital admittance's of affected patients. Usually in this type of series the large number of zeros, called zeros inflation, contrasts with high values, which can lead to bad statistical inference and offers spurious relations. Dealing with this kind of data, Yang et al. (2013) extended the classical Zero Inflated Poisson (ZIP) introduced by Lambert (1992) to accommodate count data series with an excess of zeros, autoregressive (AR) autocorrelation and time-dependent covariates regression framework. According to Dietz & Bohning (1997), one can see a ZIP distribution as a mixed distribution of a Poisson (λ) and a degenerate component with all its mass at zero. Thus, it is verified that ZIP is based on two distribution parameters: λ referring to the Poisson part of the distribution and ρ referring to the inflation parameter of zeros. To clarify, consider the count time series of random variables ε_t, $t \in \mathbb{Z}$, that follows the ZIP distribution with non-negative real parameters ρ and λ, $\varepsilon_t \sim \mathsf{ZIP}(\rho, \lambda)$. The probability mass function (p.m.f.) of ε_t is given by $P_{\varepsilon_t}(\varepsilon_t = m) = \rho\mathbb{I}_{m=0} + (1-\rho)\ \mathsf{exp}(-\lambda)\lambda^m/m!$, $m \in \mathbb{Z}_+$, where $\mathbb{I}_{m=0} = 1$, if $m = 0$ or $\mathbb{I}_{m=0} = 0$, if $m \neq 0$. The parameters ρ and λ connect the variable ε_t to the vectors of covariates $\boldsymbol{X}$ and $\boldsymbol{Z}$ of the model through equations $\log(\lambda) = \boldsymbol{X}^\top\boldsymbol{\beta}$ and $\log[\rho/(1-\rho)] = \boldsymbol{Z}^\top\gamma$, where $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ represent the vectors of coefficients.

4 INAR(1) process with zero inflated Poisson innovations

Jazi et al. (2012) introduced an INAR(1) process with zero inflated Poisson innovations (denoted by ZINAR(1)). The motivation for such a process comes from its potential in the modeling and analysis of non-negative integer values time series with excess of zeros. The ZINAR(1) could also be useful in other cases when the innovation process indicates over-dispersion of any value, not only zeros. Some mathematical and structural properties of the corresponding

marginal distribution of the process, such as the mean, variance, autocovariance and conditional maximum likelihood functions are derived. The ZINAR(1) model is an INAR(1) process

$$Y_t = \alpha \circ Y_{t-1} + \varepsilon_t,$$

where $\alpha \in [0, 1]$ and $\varepsilon_t \sim$ ZIP(ρ, λ). The marginal distributed r.v. Y_t of the ZINAR(1) process can be represented as an infinite sum of ZIP r.v.'s with the same ρ but geometrically decaying λ's. It follows that for small λ and ρ the marginal distribution will itself approximate a ZIP.

The ZINAR(1) model well fitted a real data given by the numbers of submissions to animal health laboratories, monthly 2003–2009, from a region in New Zealand.

Chapter 3

The periodic INAR$(1, 1_S)$ model

In this Chapter, we present the Periodic INAR$(1, 1_S)$ model, which is introduced in the first paper "A periodic and seasonal statistical model for dispensed medications in respiratory diseases" and studied in the second paper "The PINAR$(1, 1_S)$ model". These papers are presented below.

A A periodic and seasonal statistical model for dispensed medications in respiratory diseases

The first original contribution of this thesis is the introduction of a new class of models for counting time series which presents periodic and seasonal autoregressive structure. We discuss some statistical properties of the proposed model. We presented a likelihood-type method to estimate the parameters of the model. We found that our estimator is consistent and asymptotic normal distributed. A simulation study evaluates the performance of the estimator for small sample size. Finally, the proposed model was applied to model the daily number of antibiotics dispensed for the treatment of respiratory diseases.

This paper will be submitted to publication to the Journal of the Royal Statistical Society, Series C (Applied Statistics).

A periodic and seasonal statistical model for dispensed medications in respiratory diseases

Abstract

We introduce a new class of models for non-negative integer-valued time series with a periodic and seasonal autoregressive structure. Some properties of the model are discussed and the quasi-maximum likelihood method is used to estimate the parameters. The consistency and asymptotic normality of the estimator are also discussed. The performance of the estimator is investigated for small sample size and the empirical results indicate that the method gives accurate estimates. We analyze an application to model the daily number of antibiotic dispensing for the treatment of respiratory diseases.

Keywords: Count time series, INAR model, periodic stationarity, seasonality, quasi-maximum likelihood estimation.

A .1 Introduction

The study of medicine dispensing has become an important research topic since it can be very useful for public health issues such as to control and detect epidemic diseases, to promote public health education campaign, to reduce cost, to improve the quality of care, to propose intervention strategies, among others. See, for example, Organization et al. (1993). The papers McDowell et al. (2018), Caillaud et al. (2018), Oudin et al. (2017), Youngster et al. (2017), Holstiege & Garbe (2013) are some recent publications related to this theme.

The model proposed in this paper is mainly motivated by the analysis of the count time series of the daily number of people who received antibiotics for the treatment of respiratory diseases from the public health care system in the emergency service of the region of Vitória-ES (Brazil). Since the respiratory disease is strongly correlated to the air pollution levels and weather conditions, the correlation structure of the daily number of people who received antibiotics presents, among other phenomena, the periodicity and seasonality.

A count time series may be represented by the INteger-valued AutoRegressive (INAR) class of models which was initially introduced by Al-Osh & Alzaid (1987), that is, the INAR(1) process. These models are based on the thinning operator which is usually represented by "$\circ$", see Steutel & Van Harn (1979).

Let Y be a non negative integer-valued random variable (r.v.) and $\alpha \in [0, 1]$. The *binomial thinning operator* $\circ$ is defined as

$$\alpha \circ Y = \sum_{i=1}^{Y} U_i(\alpha), \tag{3.1}$$

where $\{U_i(\alpha)\}_{i\in\mathbb{N}}$ is a sequence of independent identically distributed (i.i.d.) r.v.'s which are Bernoulli distributed with parameter α. It is assumed that the sequence $\{U_i(\alpha)\}_{i\in\mathbb{N}}$ is mutually independent of Y. Note that the empty sum is set to 0 if $Y = 0$. The sequence $\{U_i(\alpha)\}_{i\in\mathbb{N}}$ is called a counting process. Observe that the probability of success in the thinning is $\mathrm{P}(U_i(\alpha) = 1) = \alpha$ and, conditionally on Y, $\alpha \circ Y \sim \mathrm{Bin}(Y, \alpha)$. For more details on thinning based count time series models see, e.g., Scotto et al. (2015) in the univariate and Latour (1997) in the multivariate case, respectively.

An extension of the INAR(1) model that takes into account the p-th order autoregressive structure is the INAR(p), introduced by Alzaid & Al-Osh (1990) and, independently by Du & Li (1991). The authors in Alzaid & Al-Osh (1990) introduced a model for count time series that has a correlation structure similar to the correlation structure of a conventional ARMA$(p, p-1)$ for continuous data. Du & Li (1991) suggested a model based on a process with a correlation structure identical to the correlation structure of a standard AR(p).

In Du & Li (1991), despite its flexibility in dealing with higher order autoregressive processes, the INAR(p) model does not account for the periodic phenomenon, which is a quite common time series characteristic in many areas of application, specially, in the air quality and health area.

Stochastic processes with periodically varying mean, variance and covariance were introduced by Gladyshev (1961) and are usually called periodically correlated processes (PC).

The occurrence of PC processes in time series is corroborated by real applications in many practical situations, see, e.g., Gardner et al. (2006). Basawa & Lund (2001) studied the asymptotic properties of parameter estimates for specific periodic autoregressive moving-average (PARMA) models among others. Recently, Sarnaglia et al. (2010) and Solci et al. (2018) presented robust estimation methods for periodic autoregressive (PAR) models applied to air pollution data.

Even though there are in the literature many studies that focus on periodically correlated processes, the vast majority is dedicated to the analysis and applications for discrete time processes with continuous marginal distributions (see Priestley (1981), Definition 3.2), for example, the PARMA model. However, not much attention has been paid to the analysis of periodically correlated count time series, that is, discrete parameter processes with discrete marginal distributions. See, for example, the ones discussed in Monteiro et al. (2010) and Moriña et al. (2011).

In the first example, Monteiro et al. (2010) introduced the Periodic INAR(1) (PINAR(1)) model and addressed some statistical properties of the parameter estimators together with some finite sample size investigation. However, the paper does not explore the model in a practical problem. In the second example, Moriña et al. (2011) presents a model based on two-order integer-valued autoregressive time series to analyze the number of hospital emergency service

arrivals caused by diseases that present seasonal behavior.

The first-order seasonal structure INAR was introduced by Bourguignon et al. (2016) and the class of subset INAR models is investigated in the forthcoming paper Bondon et al. (2018).

In the remainder of this paper, let $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Z}_+$, $\mathbb{R}$ and $\mathbb{R}_+$ denote the set of positive integers, integers, non-negative integers, real and non-negative numbers, respectively and I denotes identity matrix.

In the organization of the paper, Section 2 introduces the proposed model, presents the mean and the autocorrelation of the process and some probabilistic properties of the model. Section 3 discusses the estimation method of the parameters, namely the quasi-maximum likelihood framework. Section 4 presents the simulation and its results, real data application is presented in Section 5, conclusions and final comments are presented in the last section.

A .2 The periodic INAR$(1, 1_S)$ (PINAR$(1, 1_S)$) model

Let $\{Y_t\}_{t\in\mathbb{Z}}$ be a stochastic count process with seasonal characteristics of period S, $S \in \mathbb{N}$, defined on a probability space $(\Omega, \mathcal{A}, \mathsf{P})$, which depends on an unknown parameter $3S \times 1$-vector $\boldsymbol{\vartheta} = (\alpha_1, \beta_1, \lambda_1, \ldots, \alpha_S, \beta_S, \lambda_S)^\top$ lying in an open set Θ of Euclidean $3S-$space. $M^\top$ means transpose of a matrix M. Let $\mathsf{E}(\cdot)$ and $\mathsf{E}(\cdot|\cdot)$ denote the expectation and conditional expectation, respectively, under P and the true vector parameter value $\boldsymbol{\vartheta}_0$. In addition, let $\{\mathcal{F}_t\}_{t=0,1,\ldots}$ denote the sequence of sub-sigma fields with $\mathcal{F}_t$, $t \geq 1$, generated by an arbitrary subset of $Y_1, ..., Y_t$ and $\mathcal{F}_0 = \{\emptyset, \Omega\}$ is the trivial sigma field. The time index t may be written, by Euclidean division, as $t = kS + \nu$, where $\nu = 1, \ldots, S$ and $k \in \mathbb{Z}$. For example, in the case of daily data studied here, $S = 7$, ν and k represent the day of the week and the week, respectively.

Definition 1. $\{Y_t\}_{t\in\mathbb{Z}}$ is said to be a *periodic non-negative integer-valued process* of autoregressive order 2 with seasonal period S, for some $S \in \{2, 3, \ldots\}$, and is denoted by PINAR$(1, 1_S)$, if it satisfies the following stochastic recursion

$$Y_{kS+\nu} = \alpha_\nu \circ Y_{kS+\nu-1} + \beta_\nu \circ Y_{kS+\nu-S} + \varepsilon_{kS+\nu}, \tag{3.2}$$

where $k \in \mathbb{Z}$, and $\nu = 1, \ldots, S$, $\alpha_\nu, \beta_\nu \in [0, 1]$ are the autoregressive coefficients during the season ν. The immigration process $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ is a periodic sequence of $\mathbb{Z}_+$-valued r.v.'s such that for each ν, the sequence $\{\varepsilon_{kS+\nu}\}_{k\in\mathbb{Z}}$ consists of i.i.d.r.v.'s with finite mean $\mathsf{E}(\varepsilon_{kS+\nu}) = \lambda_\nu$, $\lambda_\nu \in \mathbb{R}_+$, and finite variance $\mathrm{Var}(\varepsilon_{kS+\nu}) = \sigma^2_\nu > 0$ for all $k \in \mathbb{Z}$. In addition, it is assumed that ε_t is independent of Y_{t-1}, $\alpha_\nu \circ Y_{t-1}$, Y_{t-S} and $\beta_\nu \circ Y_{t-S}$ and all counting processes are mutually independent.

As can be seen, for each seasonal period ν, the r.v. Y_t, in (3.2), has three random components; the immigration of the immediate past Y_{t-1} with survival probability α_ν, the immigration at $t - S$ with probability β_ν and the elements which entered in the system in the interval (t-1, t], which define the innovation term ε_t. Moreover, the autoregressive parameters α_ν, β_ν and immigration means λ_ν, $\nu = 1\ldots, S$, change periodically according to the seasonal period S. Note that

the above model becomes an extension of the models introduced in Moriña et al. (2011) and Monteiro et al. (2010). For example, in the model by Moriña et al. (2011), the autoregressive coefficients are fixed in time and only the immigration mean varies within a period. On the other hand, the PINAR$(1, 1_S)$ model, in addition to the periodic mean value, the autoregressive coefficients also vary periodically. In this context, the PINAR$(1, 1_S)$ model (3.2) also accommodates the periodicity in the autoregressive coefficients, that is, it can be considered as a kind of cyclostationary models introduced in Gladyshev (1961) for standard linear time series.

The mean of the process $Y_{kS+\nu}$, $k \in \mathbb{Z}$, in (3.2), is given by

$$\begin{aligned} \mu(kS+\nu) &= \mathsf{E}(Y_{kS+\nu}) = \alpha_\nu \mathsf{E}(Y_{kS+\nu-1}) + \beta_\nu \mathsf{E}(Y_{kS+\nu-S}) + \mathsf{E}(\varepsilon_{kS+\nu}), \\ \mu(\nu) &= \alpha_\nu \mu(\nu-1) + \beta_\nu \mu(\nu) + \lambda_\nu. \end{aligned} \tag{3.3}$$

In the above equation, $\mathsf{E}(\alpha \circ Y) = \alpha \mathsf{E}(Y)$. For more details of the thinning operator properties see, for example, Lemma 1 in da Silva & Oliveira (2004). It is worth noting that, the mean of arrivals at season ν, $\mu(t)$, corresponds to the proportion α_ν of the mean arrivals at $t-1$ plus the proportion β_ν of the mean arrivals at time $t-S$ and the mean λ_ν of the arrivals at t .

The analysis of the existence and uniqueness of a periodically stationary and causal PINAR$(1, 1_S)$ process, defined in (3.2), can be obtained analogously as the standard periodically ARMA processes introduced by Basawa & Lund (2001). In addition, these properties are well established for multivariate according to integer-valued autoregressive process Latour (1997), which were the basis for the model properties discussed in Monteiro et al. (2010) and and the PINAR(p) process in Filho et al. (n.d.). Following the same lines of the matrix representation properties of the PARMA process in Basawa & Lund (2001), some properties of the model (3.2) are now discussed.

Define the matrices $A = (a_{ij})$ and $B = (b_{ij})$ of dimension $S \times S$ as

$$a_{ij} = \begin{cases} 1 & \text{if } i = j, \\ -\alpha_i & \text{if } i = j+1, \\ 0 & \text{otherwise,} \end{cases} \qquad b_{ij} = \begin{cases} \beta_i & \text{if } i = j, \\ \alpha_1 & \text{if } i = 1, j = S, \\ 0 & \text{otherwise.} \end{cases} \tag{3.4}$$

Let $\boldsymbol{Y}_k = (Y_{kS+1}, \ldots, Y_{kS+S})^\top$ and $\boldsymbol{\varepsilon}_k = (\varepsilon_{kS+1}, \ldots, \varepsilon_{kS+S})^\top$, $k \in \mathbb{Z}$, and consider the non-negative integer stochastic processes $\{\boldsymbol{Y}_k\}_{k\in\mathbb{Z}}$ and $\{\boldsymbol{\varepsilon}_k\}_{k\in\mathbb{Z}}$ with finite mean, that is, $\mathsf{E}[\boldsymbol{\varepsilon}_k] = \boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_S)^\top$. Then, by (3.2) and the properties of matricial thinning operator presented by Latour (1997) in Lemma 2.1, one can easily see that the following stochastic equation holds

$$A \circ \boldsymbol{Y}_k = B \circ \boldsymbol{Y}_{k-1} + \boldsymbol{\varepsilon}_k, \tag{3.5}$$

where the matrices A and B are defined by (3.4).

Suppose that the process $\{\boldsymbol{Y}_k\}_{k\in\mathbb{Z}}$ has a constant mean vector $\boldsymbol{\mu}$. Then, $\mathsf{E}[A \circ \boldsymbol{Y}_k]$ is

$$A\boldsymbol{\mu} = B\boldsymbol{\mu} + \boldsymbol{\lambda}. \tag{3.6}$$

Note that A is a lower triangular non-singular matrix and its inverse $A^{-1} = (a_{ij}^{-1})$ is given by

$$a_{ij}^{-1} = \begin{cases} 1 & \text{if } i = j, \\ \prod_{k=j+1}^{i} \alpha_k & \text{if } i > j, \\ 0 & \text{if } i < j. \end{cases} \tag{3.7}$$

Thus A^{-1} and B are non-negative matrices, hence $A^{-1}B$ and $A^{-1}\boldsymbol{\lambda}$ are also non-negative matrix and vector, respectively. By multiplying A^{-1} in both sides of (3.6), it can be seen that

$$\boldsymbol{\mu} = A^{-1}B\boldsymbol{\mu} + A^{-1}\boldsymbol{\lambda} \tag{3.8}$$

or

$$(I - A^{-1}B)\boldsymbol{\mu} = A^{-1}\boldsymbol{\lambda}. \tag{3.9}$$

From Theorem 2.1 in Seneta (2006), since $A^{-1}B$ is a Perron-Frobenius matrix, a necessary and sufficient condition for a solution of $\boldsymbol{\mu}$ ($\boldsymbol{\mu} \geq \mathbf{0}, \neq \mathbf{0}$), where $\mathbf{0}$ is a S-dimensional vector of zeros, to (3.9) to exist for any $\boldsymbol{\lambda}^* = A^{-1}\boldsymbol{\lambda}$ ($\boldsymbol{\lambda}^* \geq \mathbf{0}, \neq \mathbf{0}$) is that the spectral radius $\rho(A^{-1}B) < 1$, which is the maximum eigenvalue in modulus of the matrix $A^{-1}B$. Note that, since $S \geq 2$, from the Perron-Frobenius Theorem in Horn & Johnson (2012) page 534, $\rho(A^{-1}B) > 0$. Therefore, $0 < \rho(A^{-1}B) < 1$.

Based on Graybill (1983), page 100, if $|\varphi| < 1$ for every characteristic root φ of $A^{-1}B$ and none of the sums of absolute values of row or column elements exceed unity, then $\sum_{i=1}^{\infty}(A^{-1}B)^i$ converges to $(I - A^{-1}B)^{-1}$. This condition assures the invertibility of $(I - A^{-1}B)$ and the positivity of its inverse. For the expected value of model 3.5, this condition may be stated as *the roots of the determinant equation* $\det(zI_S - A^{-1}B) = 0$*, for all complex* z*, are all less than 1 in absolute value*, which is also equivalent to *the roots of the characteristic polynomial* $P(z) = \prod_{j=1}^{S}(1 - \beta_j z) - z\prod_{j=1}^{S}\alpha_j$*, for all complex* z*, lie outside the complex unit circle*.

In this context, model PINAR$(1, 1_S)$ in (3.5) will be completely specified, if the

$$det(zI - A^{-1}B) \neq 0,$$

$z \in \mathbb{C}$, i.e., the characteristic roots will be inside the unit circle, and, then, the process in 3.5 will be strictly and second order periodic stationary process (Brockwell & Davis (2013), Latour (1997)). In addition, if all the eigenvalues of $A^{-1}B$ are inside the unit circle, $I - A^{-1}B$ is non-singular and $\boldsymbol{\mu} = (I - A^{-1}B)^{-1}\boldsymbol{\lambda}^*$ is the unique solution to (3.6). Some examples are now given.

Example 1. Consider the case when $\beta_j = 0$ for all $j = 1, \ldots, S$. Then the PINAR$(1, 1_S)$ model is reduced to a PINAR$(1)_S$ model introduced in Monteiro et al. (2010). The characteristic polynomial of this model is simplified to $P(z) = 1 - z\prod_{j=1}^{S}\alpha_j$ and a necessary and sufficient condition for the periodically stationarity of the process $\{Y_t\}$ is $\prod_{j=1}^{S}\alpha_j < 1$. Note that $\prod_{j=1}^{S}\alpha_j$ is the spectral radius of the matrix A defined on page 1531 in Monteiro et al. (2010).

Example 2. Consider the case $S = 2$, i.e., the PINAR$(1, 2)$ model. Then

$$A = \begin{bmatrix} 1 & 0 \\ -\alpha_2 & 1 \end{bmatrix}, \qquad A^{-1} = \begin{bmatrix} 1 & 0 \\ \alpha_2 & 1 \end{bmatrix}, \qquad B = \begin{bmatrix} \beta_1 & \alpha_1 \\ 0 & \beta_2 \end{bmatrix}. \tag{3.10}$$

The characteristic polynomial is given by $P(z) = (1 - \beta_1 z)(1 - \beta_2 z) - \alpha_1\alpha_2 z$. By solving the characteristic equation, it can be seen that $\beta_1+\beta_2-\beta_1\beta_2+\alpha_1\alpha_2 < 1$ is a necessary and sufficient stationarity condition. Note that this condition can be rewritten as $\alpha_1\alpha_2 < (1 - \beta_1)(1 - \beta_2)$.

The marginal distribution of $\{Y_t\}_{t\in\mathbb{Z}}$ process defined in 3.2 is given by

$$P(Y_{kS+\nu} = m) = \sum_{b_1,b_2=0}^{\infty} p_\nu(m|b_1, b_2)P(Y_{kS+\nu-1} = b_1, Y_{kS+\nu-S} = b_2), \tag{3.11}$$

where $m \in \mathbb{Z}_+$, $t = kS+\nu > S$ and $\nu = 1, \ldots, S$ and $p_\nu(m|b_1, b_2) = P(Y_t = m|Y_{t-1} = b_1, Y_{t-S} = b_2)$ for each ν.

Given starting values $Y_1, \ldots, Y_S$, by the definition of the conditional probability and the S-step Markov property of the PINAR$(1, 1_S)$ process, the conditional joint probability is given by

$$\begin{aligned} P(Y_t = y_t, \ldots, &Y_{S+1} = y_{S+1}|Y_s = y_s, \ldots, Y_1 = y_1) = \\ &= \frac{P(Y_t = y_t, \ldots, Y_1 = y_1)}{P(Y_{t-1} = y_{t-1}, \ldots, Y_1 = y_1)} \cdot \frac{P(Y_{t-1} = y_{t-1}, \ldots, Y_1 = y_1)}{P(Y_S = y_S, \ldots, Y_1 = y_1)}, \\ &= P(Y_t = y_t|Y_{t-1} = y_{t-1}, \ldots, Y_1 = y_1)\times \\ &\quad P(Y_{t-1} = y_{t-1}, \ldots, Y_{S+1} = y_{S+1}|Y_S = y_S, \ldots, Y_1 = y_1), \\ &= p_\nu(y_t|y_{t-1}, y_{t-s})P(Y_{t-1} = y_{t-1}, \ldots, Y_{S+1} = y_{S+1}|Y_S = y_S, \ldots, Y_1 = y_1), \end{aligned} \tag{3.12}$$

where $t = kS + \nu$, $t > S$ and $y_1, \ldots, y_t \in \mathbb{Z}_+$. Thus, by induction, if $T = nS$ where $n \in \mathbb{N}$, the conditional probability can be calculated as

$$\begin{aligned} P(Y_T = y_T, \ldots, &Y_{S+1} = y_{S+1}|Y_S = y_S, \ldots, Y_1 = y_1) = \\ &\prod_{\nu=1}^{S}\prod_{k=1}^{n-1} p_\nu(y_{kS+\nu}|y_{kS+\nu-1}, y_{kS+\nu-S}), \end{aligned} \tag{3.13}$$

where $y_1, \ldots, y_T \in \mathbb{Z}_+$.

Now, let the innovation process in (3.2) be an i.i.d. Poisson process with unconditional mean $\mathsf{E}(\varepsilon_{kS+\nu}) = \lambda_\nu$, $\lambda_\nu \in \mathbb{R}_+$. When $S = 1$, the model (3.2) becomes a variable with Poisson marginal distribution. See, for example, Bu et al. (2008).

When $S > 1$, it can be shown that the unconditional mean and variance of Y_t are generally not equal so that the marginal stationary distribution of Y_t is no longer Poisson even though the innovations are. However, an approximation to a Poisson distribution can be achieved if $\alpha_\nu.\beta_\nu \approx 0$ and Y_t becomes large due to the well-known Law of Small Numbers. See, also, Chen & Liu (1997).

In the case that the immigration $\varepsilon_{kS+\nu}$ follows a Poisson distribution, the conditional probability $p_\nu(\cdot|\cdot,\cdot)$ in 3.13 becomes

$$p_\nu(y_t|y_{t-1},y_{t-s}) = [\mathrm{Bin}(y_{t-1},\alpha_\nu) * \mathrm{Bin}(y_{t-s},\beta_\nu) * \mathrm{Poi}(\lambda_\nu)](y_t), \tag{3.14}$$
$$= \sum_{(c_1,c_2)\in\mathcal{J}} \binom{y_{t-1}}{c_1}\alpha_\nu^{c_1}(1-\alpha_\nu)^{y_{t-1}-c_1}\binom{y_{t-s}}{c_2}\beta_\nu^{c_2}(1-\beta_\nu)^{y_{t-s}-c_2}\frac{\lambda_\nu^{y_t-c_1-c_2}}{(y_t-c_2-c_1)!}\mathrm{e}^{-\lambda_\nu},$$

where $*$ denotes the convolution, and the index set $\mathcal{J}$ is defined by $\mathcal{J} = \{(c_1,c_2)\in\mathbb{Z}_+^2 | c_1 \le y_{t-1}, c_2 \le y_{t-s}, c_1+c_2 \le y_t\}$. Note that the definition of $\mathcal{J}$ depends on the values y_t, y_{t-1}, y_{t-s}.

Quasi-maximum likelihood (QML)

Before, lets establish the format of the vector of parameters $\vartheta_\nu = (\alpha_\nu,\beta_\nu,\lambda_\nu)^\top$, $\alpha_\nu,\beta_\nu\in(0;1)$ and $0<\lambda_\nu<\infty$, for $\nu = 1,\dots,S$ (S is fixed), and let $\boldsymbol{\vartheta} = (\vartheta_1^\top,\dots,\vartheta_S^\top)^\top$ represent the $3S$-dimensional unknown parameter vector of the PINAR$(1,1_S)$ model defined by (3.2). The parameter vector is assumed to be lying in the open set $\Theta = ([0,1]\times[0,1]\times(0,\infty))^S$, which contains the true parameter vector, denoted by $\boldsymbol{\vartheta}_0 = ((\vartheta_1^0)^\top,\dots,(\vartheta_S^0)^\top)^\top$. We assumed that here $Y_1,\dots,Y_T$ has n complete periods of observations, that is, consider a sample $Y_1,\dots,Y_T$ of size $T=nS$ from $\{Y_t\}$, the PINAR$(1,1_S)$ process. Our QML estimation approach is based on Taniguchi & Kakizawa (2000). Let the likelihood type penalty function of the PINAR$(1,1_S)$ model, conditioned on the first S observations, be

$$L_n(\boldsymbol{\vartheta}) = \sum_{k=1}^{n-1}\sum_{\nu=1}^{S}[\log\{f_{\vartheta_\nu}(t,t-1)\} + (Y_t - m_{\vartheta_\nu}(t,t-1))^2 f_{\vartheta_\nu}^{-1}(t,t-1)],$$

where

$$f_{\vartheta_\nu}(t,t-1) = \mathsf{E}\left[\{Y_t - m_{\vartheta_\nu}(t,t-1)\}^2\middle|\mathcal{F}_{t-1}\right]$$

and

$$m_{\vartheta_\nu}(t,t-1) = \mathsf{E}(Y_{kS+\nu}|\mathcal{F}_{kS+\nu-1}) = \alpha_\nu Y_{kS+\nu-1} + \beta_\nu Y_{kS+\nu-S} + \lambda_\nu. \tag{3.15}$$

The likelihood function $L_n(\boldsymbol{\vartheta}) = \sum_{\nu=1}^S l_{n,\nu}(\vartheta_\nu)$, where

$$l_{n,\nu}(\vartheta_\nu) = \sum_{k=1}^{n-1}[\log\{f_{\vartheta_\nu}(t,t-1)\} + (Y_t - m_{\vartheta_\nu}(t,t-1))^2 f_{\vartheta_\nu}^{-1}(t,t-1)] l_{n,\nu}(\vartheta_\nu) = \sum_{k=1}^{n-1}\phi_t(\vartheta_\nu)$$

is minimized in order to obtain the QML-estimator $\widehat{\boldsymbol{\vartheta}}_n^{\mathrm{QML}}$ of the parameter vector $\boldsymbol{\vartheta}$.
The function $f_{\vartheta_\nu}(t,t-1)$ is given by

$$f_{\vartheta_\nu}(t,t-1) = \alpha_\nu(1-\alpha_\nu)Y_{t-1} + \beta_\nu(1-\beta_\nu)Y_{t-S} + \lambda_\nu, \tag{3.16}$$

for Poisson distributed innovations. The function $l_{n,\nu}(\vartheta_\nu)$ can be obtained directly by replacing the results of (3.16) and (3.15) in (3.16). $\widehat{\boldsymbol{\vartheta}}_n^{\mathrm{QML}} = ((\widehat{\vartheta}_{n,1}^{\mathrm{QML}})^\top,\dots,(\widehat{\vartheta}_{n,S}^{\mathrm{QML}})^\top)^\top$ is a sequence of

estimators, such that $\widehat{\boldsymbol{\vartheta}}_n^{\mathrm{QML}} \to \boldsymbol{\vartheta}_0$ almost surely as $n \to \infty$, is the solution of

$$\frac{\partial}{\partial \boldsymbol{\vartheta}} L_n(\boldsymbol{\vartheta}) = 0, \tag{3.17}$$

which attains a relative minimum of the likelihood function $L_n(\boldsymbol{\vartheta})$.

The minimization of $L_n(\boldsymbol{\vartheta})$ can be done separately by minimizing the partial log-likelihood $l_{n,\nu}(\vartheta_\nu)$ for each season $\nu \in \{1, \ldots, S\}$. Similarly, one can solve the likelihood equation (3.17) by solving the partial likelihood equations

$$\frac{\partial}{\partial \vartheta_\nu} l_{n,\nu}(\vartheta_\nu) = 0, \qquad \nu = 1, \ldots, S,$$

separately.

Define IF_ν the matrix of dimension 3×3 for each season $\nu \in \{1, \ldots, S\}$ as

$$IF_\nu = U_{\vartheta_\nu}^{-1} V_{\vartheta_\nu} U_{\vartheta_\nu}^{-1}, \tag{3.18}$$

where

$$V_{\vartheta_\nu} = \mathsf{E}\left\{ \frac{\partial}{\partial \vartheta_\nu} \phi_t(\vartheta_\nu) \frac{\partial}{\partial \vartheta_\nu^\top} \phi_t(\vartheta_\nu) \right\} \tag{3.19}$$

and

$$U_{\vartheta_\nu} = \mathsf{E}\left\{ \frac{\partial^2}{\partial \vartheta_\nu \partial \vartheta_\nu^\top} \phi_t(\vartheta_\nu) \right\}. \tag{3.20}$$

Note that $\frac{\partial}{\partial \vartheta_\nu} \phi_t(\vartheta_\nu) = (\frac{\partial}{\partial \alpha_\nu} \phi_t(\vartheta_\nu), \frac{\partial}{\partial \beta_\nu} \phi_t(\vartheta_\nu), \frac{\partial}{\partial \lambda_\nu} \phi_t(\vartheta_\nu))$ is a 3-dimensional row vector. Then, the matrix IF of the PINAR$(1, 1_S)$ process is defined as the block diagonal matrix

$$IF = \mathrm{diag}\{IF_1, \ldots, IF_S\}. \tag{3.21}$$

The following theorem on the asymptotic normality of the QML-estimator $\widehat{\boldsymbol{\vartheta}}^{\mathrm{QML}}$ is given below.

Theorem 1. *Assume that* $\{\boldsymbol{Y}_t\}$ *in* (3.5)*, that is, the a PINAR*$(1, 1_S)$ *process, is a strictly stationary ergodic process with* $\mathsf{E}\,\|\boldsymbol{\varepsilon}_k\|^6 < \infty$ ($\mathsf{E}\,\|\varepsilon_t\|^6 < \infty$ *in model* (3.2))*, and* $m_{\vartheta_\nu}(t, t-1)$ *and* $f_{\vartheta_\nu}(t, t-1)$ *are almost surely three times continuously differentiable in the open set* Θ *containing the true parameter value* $\boldsymbol{\vartheta}_0$*. Then, the* QML *estimators* $\widehat{\boldsymbol{\vartheta}}_n^{\mathrm{QML}}$ *are asymptotically normal distributed as*

$$n^{1/2}(\widehat{\boldsymbol{\vartheta}}_n^{\mathrm{QML}} - \boldsymbol{\vartheta}_0) \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0}, IF), \tag{3.22}$$

as $n \to \infty$*, where* IF *is the matrix of dimension* $3S \times 3S$ *defined by* (3.21)*.*

The proof follows a straight generalization of the Theorem 3.2.26 in Taniguchi & Kakizawa (2000). The calculations are omitted here but available from authors upon request.

A .3 Monte Carlo simulations

In this section, the performance of the QML method to estimate the parameters of 3.2 is evaluated through simulations of time series with finite sample sizes T. $\{\varepsilon_t\}_{t=1,..,T}$ are generated from a periodic sequence of Poisson distributed r.v.'s such that for each ν, the mean $\mathrm{E}(\varepsilon_{kS+\nu}) = \lambda_\nu$, $\nu = 1, ..., S$. The parameters of the models are displayed in tables, as well as the sample sizes, and the period for each model was fixed for $S = 4, 7$. The results (the empirical bias and mean square error (MSE)) correspond to the mean of 1000 replications. All simulations were carried out using the R software.

Tables 3.1 and 3.2 display the results for $S = 4$ and $S = 7$, respectively. As was expected, in general, the performance of QML estimator presents estimates quite accurate even for a small sample size. By increasing n, the quantities bias and MSE of the estimates decrease, which corroborates the theoretical results described previously. Since the parameters α_ν and β_ν, for each ν, correspond to the coefficients of linear relation between the variable Y_t at time $t-1$ and $t-S$, respectively, their estimates perform nearly identical, that is, they present similar MSE. On the other hand, although the estimates of λ_ν also present accurate results, these are not precisely in terms of MSE as the ones of α_ν and β_ν. This fact may be mainly due to the minimization algorithm to estimate λ_ν, which is not linearly related to the observations Y_t. In practice, however, it may not be a big concern. Other parameter values were also considered in the simulation study and, in general, the conclusions were quite similar to those reported here. These results are available upon request.

A .4 Real data application

The time series of counts refers to the daily number of people who got antibiotics for the treatment of respiratory problems from the public health care system in the emergency service of the region of Vitória-ES, Brazil. This real data set was obtained from the network records system welfare of the municipality and corresponds to the period of May 26, 2013 to September 07, 2015, resulting in 834 daily observations. The series is displayed in Figure 3.1 contains persistence oscillation, that is, the mean changes periodically. This is clearly evidenced in the plots of the sample autocorrelation function (ACF), as discussed below.

Figure 3.2 shows the sample periodic mean and variance of the series over seasons $\nu = 1, \ldots, 7$, with $S = 7$, the sample ACF and the periodogram of the series. The analysis of the sample ACF suggests that this series has seasonal autocorrelation of period $S = 7$ which is an expected results by the fact that the series corresponds to daily data. The periodogram provides high peak at frequency 0.14, which corresponds to the period=$1/0.14 = 7$. The AR order identification per season $\nu = 1, \ldots, 7$ is identified by finding the lowest lag for which the sample Periodic Partial Autocorrelation (PePACF) function cuts off (McLeod (1994)). These are displayed in Table 3.4.

Tables 3.3 and 3.4 present the sample Periodic Autocorrelation (PeACF) and PePACF functions. In these tables, the values in bold are the sample ACFs that exceeded the confidence intervals

Table 3.1: Results of the simulation to estimate the parameters of the PINAR$(1, 1_4)$ model with sample size T=200, 800 and 2000 values. The real parameter values are: $\boldsymbol{\alpha} = \{0.1, 0.42, 0.23, 0.39\}$, $\boldsymbol{\beta} = \{0.47, 0.25, 0.36, 0.3\}$ and $\boldsymbol{\lambda} = \{4, 3, 2, 1\}$. Inside parenthesis is the MSE of each estimator.

	n=50, T=200	n=200, T=800	n=500, T=2000
	Bias$_{QML}$	Bias$_{QML}$	Bias$_{QML}$
α_1	0.025	-0.002	-0.004
	(0.018)	(0.005)	(0.003)
α_2	0.021	0.007	-0.004
	(0.014)	(0.004)	(0.002)
α_3	0.009	0.002	0.001
	(0.013)	(0.003)	(0.001)
α_4	0.004	0.006	0.000
	(0.010)	(0.002)	(0.001)
β_1	-0.028	-0.008	0.002
	(0.015)	(0.003)	(0.001)
β_2	-0.024	-0.007	-0.005
	(0.017)	(0.004)	(0.002)
β_3	-0.035	-0.006	-0.003
	(0.017)	(0.004)	(0.002)
β_4	-0.011	-0.005	-0.003
	(0.015)	(0.004)	(0.002)
λ_1	0.081	0.085	-0.008
	(1.324)	(0.278)	(0.151)
λ_2	0.003	0.017	0.068
	(1.427)	(0.342)	(0.157)
λ_3	0.11	0.005	0.015
	(1.16)	(0.208)	(0.091)
λ_4	0.058	-0.02	0.01
	(0.455)	(0.096)	(0.042)

Table 3.2: Results of the simulation to estimate the parameters of the PINAR$(1, 1_7)$ model with sample size T=350, 700 and 1400 values. The real parameter values are: $\boldsymbol{\alpha} = \{031, 0.35, 0.29, 0.29, 0.37, 0.29, 0.28\}$, $\boldsymbol{\beta} = \{0.27, 0.25, 0.26, 0.39, 0.27, 0.22, 0.33\}$ and $\boldsymbol{\lambda} = \{4.0, 3.3, 2.1, 2.5, 3.1, 2.6, 3.5\}$. Inside the parenthesis is the MSE of each estimate.

	T=350 n=50	T=700 n=100	T=1400, n=200
Pars	Bias$_{QML}$	Bias$_{QML}$	Bias$_{QML}$
α_1	0.017	0.005	0.003
	(0.021)	(0.009)	(0.002)
α_2	0.013	0.011	0.004
	(0.017)	(0.007)	(0.001)
α_3	0.003	0.006	0.000
	(0.011)	(0.006)	(0.001)
α_4	0.005	0.007	0.003
	(0.019)	(0.009)	(0.001)
α_5	0.012	-0.001	0.002
	(0.016)	(0.007)	(0.002)
α_6	0.004	0.002	0.001
	(0.013)	(0.006)	(0.001)
α_7	0.012	0.010	0.004
	(0.019)	(0.010)	(0.002)
β_1	-0.032	-0.010	-0.003
	(0.019)	(0.008)	(0.001)
β_2	-0.017	-0.014	-0.006
	(0.016)	(0.009)	(0.002)
β_3	-0.038	-0.009	0.001
	(0.018)	(0.008)	(0.002)
β_4	-0.028	-0.012	-0.006
	(0.018)	(0.007)	(0.001)
β_5	-0.034	-0.008	-0.001
	(0.018)	(0.008)	(0.002)
β_6	-0.016	-0.010	0.000
	(0.016)	(0.009)	(0.002)
β_7	-0.019	-0.011	0.003
	(0.017)	(0.008)	(0.001)
λ_1	0.139	0.052	0.002
	(2.096)	(0.920)	(0.166)
λ_2	0.036	0.014	0.016
	(1.997)	(0.849)	(0.177)
λ_3	0.188	0.012	0.013
	(1.194)	(0.581)	(0.096)
λ_4	0.170	0.046	0.024
	(1.110)	(0.522)	(0.090)
λ_5	0.177	0.083	-0.024
	(1.269)	(0.662)	(0.120)
λ_6	0.069	0.044	-0.001
	(1.048)	(0.594)	(0.108)
λ_7	0.047	0.045	-0.046
	(1.431)	(0.639)	(0.121)

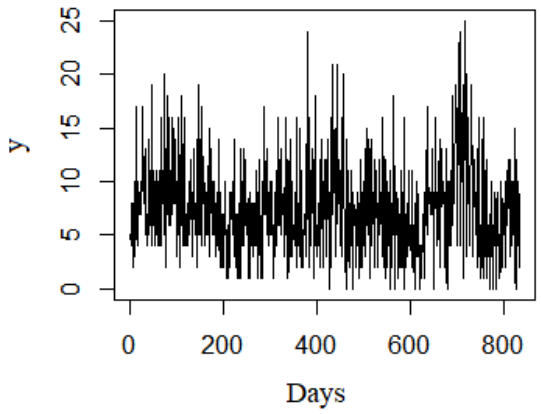


Figure 3.1: Daily number of people who received antibiotics for the treatment of respiratory problems from the public health care system in the emergency service of the region of Vitória-ES.

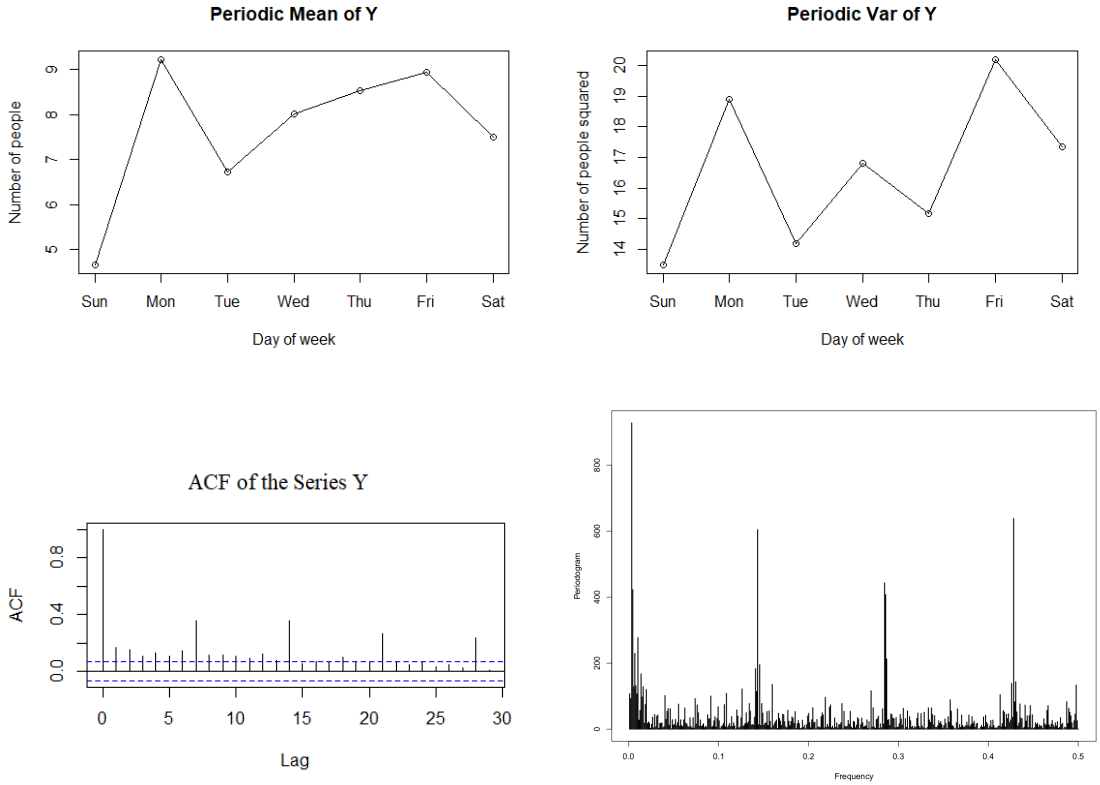


Figure 3.2: The periodic mean and periodic variance over the seasons $\nu = 1, \ldots, 7$, the ACF and the periodogram of $\{Y_t\}$.

given below. The approximate limits of the confidence intervals used in ACF and PACF tables were constructed for a significance level of 0.5%. This preliminary model identification step reinforce that a periodic INAR model could be adequate to capture the dynamic of the series.

Based on the previous and above discussion, the PINAR$(1, 1_7)$ model was used to fit the data. The estimates of the parameters are displayed in Table 3.5. The standard errors (the values

Table 3.3: Periodic ACF of the real data set.

	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=7$	$h=8$	$h=9$	$h=10$
$\nu=1$	0.01	**0.26**	**0.18**	**0.24**	**0.28**	0.11	0.15	0.02	0.07	**0.29**
$\nu=2$	**0.38**	-0.12	0.14	**0.23**	**0.18**	**0.19**	**0.29**	0.13	-0.11	0.04
$\nu=3$	**0.33**	**0.34**	-0.02	0.10	**0.23**	**0.39**	**0.42**	**0.18**	**0.37**	0.03
$\nu=4$	**0.27**	0.05	**0.17**	0.10	0.16	**0.33**	**0.29**	**0.23**	0.14	0.14
$\nu=5$	**0.18**	**0.36**	**0.23**	**0.31**	0.01	**0.18**	**0.29**	**0.22**	**0.25**	0.11
$\nu=6$	**0.25**	0.16	**0.20**	0.14	0.16	**0.17**	**0.18**	**0.30**	**0.23**	0.13
$\nu=7$	**0.20**	0.10	-0.03	-0.05	**−0.18**	0.03	**0.30**	0.10	-0.07	0.16

Table 3.4: Periodic PACF of the real data set.

	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=7$	$h=8$	$h=9$	$h=10$
$\nu=1$	0.01	**0.26**	0.12	**0.20**	**0.18**	0.00	0.03	-0.02	-0.01	0.16
$\nu=2$	**0.38**	-0.14	0.08	**0.18**	0.07	0.01	**0.22**	-0.06	-0.08	-0.02
$\nu=3$	**0.33**	**0.24**	0.00	-0.00	0.15	**0.32**	**0.29**	0.01	**0.26**	0.04
$\nu=4$	**0.27**	-0.04	0.10	0.10	0.11	**0.27**	**0.18**	0.03	0.09	-0.07
$\nu=5$	**0.18**	**0.33**	0.13	**0.18**	0.01	0.10	**0.17**	0.04	0.02	-0.02
$\nu=6$	**0.25**	0.12	0.10	0.06	0.03	**0.18**	0.08	**0.18**	0.13	-0.05
$\nu=7$	**0.20**	0.05	-0.07	-0.11	**−0.21**	0.08	**0.26**	0.03	-0.13	**0.21**

in parenthesis) were calculated using the inverse of the corresponding Hessian matrix. The strategy of model adequacy is based on the values of the ACF and PACF of the residuals, which were computed using

$$r_t = y_t - \hat{Y}_t = y_t - \widehat{\alpha}_\nu y_{t-1} - \widehat{\beta}_\nu y_{t-7} - \widehat{\lambda}_\nu, \tag{3.23}$$

where $t = 7k + \nu$, from $t > 7$, $k = 2, \ldots, n$ and $\nu = 1, \ldots, 7$.

Tables 3.6 and 3.7 display the values of PeACF and PePACF functions for the residuals, respectively. From these tables the expressive periodic correlations at lags 1 and 7 were removed and no systematic patterns is clearly observed. The fitted model seems to well capture the main dynamics of the data. Therefore, the estimated model can be useful in providing reliable forecast.

Table 3.5: Application of PINAR$(1, 1_7)$ model to the real data. The parameters were estimated by QML method. Inside parenthesis are the standard errors of the estimates.

Fitted model	ν=1	ν=2	ν=3	ν=4	ν=5	ν=6	ν=7
PINAR$(1, 1_7)$-Poisson Innovation:							
α_ν	0.095(0.039)	0.012(0.074)	0.209(0.045)	0.211(0.061)	0.133(0.060)	0.083(0.056)	0.126(0.045)
β_ν	0.192(0.047)	0.108(0.054)	0.217(0.055)	0.280(0.056)	0.150(0.061)	0.169(0.053)	0.097(0.051)
λ_ν	3.031(0.360)	8.209(0.654)	3.364(0.551)	4.361(0.562)	6.182(0.616)	6.739(0.640)	5.649(0.562)

Table 3.6: Periodic ACF of residuals after fitting the PINAR$(1, 1_7)$ model with Poisson distribution of innovations to the real data. The parameters were estimated by QML estimation method.

	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=7$	$h=8$	$h=9$	$h=10$
$\nu=1$	-0.02	0.11	-0.11	-0.02	-0.16	0.02	-0.06	0.06	-0.20	0.15
$\nu=2$	-0.04	0.21	0.04	0.13	0.13	0.04	-0.04	0.02	-0.05	0.21
$\nu=3$	0.00	-0.09	-0.00	0.15	0.06	-0.03	0.01	0.01	-0.03	0.01
$\nu=4$	-0.07	0.13	0.01	-0.05	0.01	0.26	0.05	-0.13	0.26	0.05
$\nu=5$	-0.02	-0.10	0.05	0.08	0.04	0.27	-0.06	0.07	0.11	0.04
$\nu=6$	-0.09	0.23	0.08	0.27	-0.04	0.11	-0.07	0.11	0.08	0.03
$\nu=7$	-0.07	0.05	0.09	0.04	0.03	0.16	-0.07	0.21	0.13	0.02

Table 3.7: Periodic PACF of residuals after fitting PINAR$(1, 1_7)$ model with Poisson distributed innovations to the real data set. Parameters estimates by QML estimation method.

	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=7$	$h=8$	$h=9$	$h=10$
$\nu=1$	-0.02	0.11	-0.10	-0.04	-0.19	-0.00	-0.06	0.05	-0.13	0.14
$\nu=2$	-0.04	0.21	0.06	0.13	0.10	0.05	-0.08	-0.02	-0.03	0.15
$\nu=3$	0.00	-0.09	-0.00	0.17	0.07	-0.07	-0.02	-0.04	-0.04	-0.01
$\nu=4$	-0.07	0.13	0.01	-0.08	0.00	0.25	0.04	-0.10	0.27	0.05
$\nu=5$	-0.02	-0.10	0.05	0.07	0.03	0.28	-0.03	-0.01	0.09	-0.04
$\nu=6$	-0.09	0.23	0.09	0.26	-0.02	0.08	-0.07	0.01	0.06	0.07
$\nu=7$	-0.07	0.04	0.11	0.06	0.04	0.16	-0.07	0.17	0.17	-0.01

A .5 Discussion

The PINAR$(1, 1_S)$ model with application in the health area was introduced in this paper. The main properties of this model are presented, such as the mean, matricial representation and transition probabilities function. The Conditional Maximum Likelihood method for estimating the parameters of the model was proposed and a simulation study was carried out to investigate its finite sample properties. The QML method presented a good performance in simulations.

The time series of counts of the daily number of people who received antibiotics for the treatment of respiratory problems from the public health care system in the emergency service of the region of Vitória-ES (Brazil) was used to illustrate the usefulness of the proposed model. This data set presents periodic, seasonal and serial correlation structure. The PINAR$(1, 1_7)$ model, under Poisson innovations (Poisson-PINAR$(1, 1_7)$ model), was fitted to this real data set.

Based on residual analysis, the Poisson-PINAR$(1, 1_7)$ model was able to capture the main dynamic the real data series, that is, periodicity in the data of dispensation of antibiotics for the treatment of respiratory infections.

Respiratory infections are among the leading causes of emergency service visits and the dispensation of antibiotics to treat these conditions can be used as an indicator of the effects of air pollution on human health for Europe et al. (2013). The first study on the relationship between drug dispensing and air pollution was conducted in France, published in 1998, and since then several authors have attempted to establish the relationship between dispensing drugs for

diseases and air pollution Zeghnoun et al. (1999).

Studies that evaluate the dispensation of medicines for the treatment of respiratory diseases are relevant from different points of view. According to the World Health Organization, drug use studies serve important purposes, such as: describing current treatment patterns; comparing the performance of individual facilities; observation of variations in therapeutic profiles over time; periodic monitoring and supervision of specific drug use behaviors; evaluation of the effects of educational activities; informational and regulatory measures; estimation of the most prevalent clinical conditions in the population, among others Zeghnoun et al. (1999), Organization et al. (1993).

Concerning the medicines used to treat respiratory diseases, the study of the dispensation of antibiotics is relevant because they are only dispensed with a medical prescription for a certain period, which provides more precise data. In fact, the drugs used in these situations, such as drugs used to treat asthma attacks (one of the diseases also influenced by the levels of air pollutants) come in multi-dose devices and are used on demand at times of crisis Zeghnoun et al. (1999).

In view of the increase in the use of public health services in the last decade, with a consequent increase in the consumption of medicines, good management is essential to meet the needs of the population Viacava et al. (2018), Gadelha et al. (2016).

Thus, the analysis of the behavior of drug consumption by a historical time series makes it possible to estimate the needs of the municipality and to identify potential risk factors associated with drug consumption. Therefore, the construction of models that allow the study of the trend of the use of drugs over time for subsequent correlation with socioeconomic and environmental data becomes relevant. These models can also serve as a tool to plan with greater precision and to avoid disorders caused by the lack or excess of medicines We used data from dispensing only patients treated in the public service, those who acquired antibiotics for respiratory diseases in private pharmacies were not considered. However, considering that the majority of the population of the municipality is public health system dependent, the sample studied is representative of the majority of residents in the municipality Viacava et al. (2018).

The model PINAR$(1, 1_S)$ can be very useful to make reliable predictions or forecast in the sense, for example, to better supply medications for the population in case of any emergency or not, and to determine the behavior of drug dispensing over time and its relation to local risk factors responsible for the worsening or triggering of diseases. In the future, it would be desirable to consider procedures which produce coherent (integer) forecasting.

B The PINAR$(1, 1_S)$ model

In this section we present a rigorous and comprehensive mathematical study of existence, uniqueness and stationary conditions of PINAR$(1, 1_S)$ model. We introduce a new class of models based on count time series with Poisson and Geometric innovations which have a periodic and seasonal second-order autoregressive structure. Statistical properties of the model, such as mean, variance, marginal and joint distributions, are discussed. We discussed the Moments-based (Yule-Walker equations), the conditional least squares and quasi-maximum likelihood method of estimation of the parameters. Their performances are investigated through Monte Carlo simulations, and we present a proof of consistency and asymptotically normality of the estimators. The usefulness of the PINAR$(1, 1_S)$ model is verified in an application to a real data referring to the daily number of visits of children with respiratory problems (International Classification of Diseases ICD-10) to the emergency service of the public health care system of the region of Vitória-ES. A section is focused on the forecast purposes.

This paper will be submitted to publication to the Journal of Time Series Analysis.

The PINAR$(1, 1_S)$ model

Abstract

This paper introduces a new class of models based on count time series with Poisson and Geometric immigrations which has a periodic and seasonal second-order autoregressive structure. Statistical properties of the model such as mean, variance, marginal and joint distributions are discussed. Moments-based, conditional least squares and quasi-maximum likelihood methods of estimation of the parameters are proposed, and their performances are investigated through Monte Carlo simulations. A real data application illustrates the use of the methodology in a practical situation, and a section is devoted to forecasting methods.

Keywords: INAR models, Binomial thinning, periodic stationarity, moment-based estimator, conditional least squares, quasi-maximum likelihood, asymptotic distribution.

B .1 Introduction

We are interested in count time series that presents at the same time periodic and seasonal serial correlation structures of the autoregressive (AR) type. The count time series analog of the standard AR model is the integer-valued AR (INAR) process. This model appears as an alternative to the well-known Poisson model family for modeling count time series, see, e.g., Fokianos et al. (2009). McKenzie (1985) and Al-Osh & Alzaid (1987) introduced independently the first-order INAR (INAR(1)) model. The pth-order (INAR(p)) extension of this process proposed by Alzaid & Al-Osh (1990) has a correlation structure similar to the correlation structure of an autoregressive moving average (ARMA) process with orders $(p, p-1)$ (ARMA$(p, p-1)$), while the INAR(p) model proposed by Du & Li (1991) has the same correlation structure as an AR model with order p (AR(p)).

Models that take into account the seasonal autocorrelation structure for INAR can be seen in the first-order seasonal structure introduced by Bourguignon et al. (2016) or on its extension, the subset INAR(p) process, which account both the first-order serial and seasonal correlations. The class of subset INAR models is investigated in the forthcoming paper Bondon et al. (2018).

These models cannot reproduce periodic correlations which are often present in many fields such as medicine, hydrology, climatology, air pollution, among others. Gladyshev (1961) introduced processes with periodically varying means and covariances that are denominated periodically correlated (PC) processes. For recent reviews on PC processes, see e.g. Gardner et al. (2006) and Hurd & Miamee (2007). A natural way to build models for PC processes

is to allow the parameters of stationary models to vary periodically with time. Thus, the periodic ARMA (PARMA) model extends the ARMA model to PC processes. Basawa & Lund (2001) investigated the asymptotic properties of weighted least squares parameter estimates for PARMA models, and Sarnaglia et al. (2010) proposed robust parameter estimates for PAR models. Although the literature is abundant about PC processes, its vast majority is dedicated to the analysis and the applications of PARMA models for continuous-valued data. Very little attention has been paid to modeling PC count data. To the best knowledge of the authors, only Monteiro et al. (2010) and Moriña et al. (2011) have proposed periodic models for correlated series of counts. Monteiro et al. (2010) have introduced periodicity with period T in the INAR(1) model, resulting in the PINAR$(1)_T$ model. Moriña et al. (2011) have considered a classical INAR(2) model where the innovation process follows a Poisson distribution whose intensity is a periodic function with period T.

The class of INAR models are based on the thinning operator, see Steutel & Van Harn (1979). In what follows, the thinning operator will be defined based on the Binomial distribution (for alternative thinning concepts see, for example, Weiß (2008)).

The *binomial thinning operator* $\alpha\circ$ for a random variable (r.v.) Y is defined as

$$\alpha \circ Y = \sum_{i=1}^{Y} U_i(\alpha), \tag{3.24}$$

where Y is a $\mathbb{Z}_+$-valued r.v., $\alpha \in [0,1]$ and $\{U_i(\alpha)\}_{i\in\mathbb{Z}_+}$ is a sequence of independent identically distributed (i.i.d.) r.v.'s which are Bernoulli distributed with parameter α. It is assumed that the sequence $\{U_i(\alpha)\}_{i\in\mathbb{Z}_+}$ is mutually independent of Y. Note that the empty sum is set to 0 if $Y = 0$. The sequence $\{U_i(\alpha)\}_{i\in\mathbb{Z}_+}$ is called a counting sequence. Remark that the probability of success in the thinning is $\mathsf{P}(U_i(\alpha) = 1) = \alpha$ and, conditionally on Y, $\alpha \circ Y \sim \mathrm{Bin}(Y, \alpha)$. For more details on thinning based count time series models see, e.g., Scotto et al. (2015) in the univariate and Latour (1997) in the multivariate case, respectively.

In the remainder of this paper, let $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Z}_+$, $\mathbb{R}$, $\mathbb{R}_+$ and $\mathbb{C}$ denote the set of positive integers, integers, non-negative integers, real numbers, non-negative real numbers and complex numbers, respectively.

Let $\{Y_t\}_{t\in\mathbb{Z}}$ be a stochastic process with seasonal characteristics of period S, $S \in \mathbb{N}$, defined on a probability space $(\Omega, \mathcal{A}, \mathsf{P})$, whose depends on an unknown parameter vector $\boldsymbol{\vartheta} = (\vartheta_1^\top, \ldots, \vartheta_p^\top)^\top$ lying in a some open set Θ of Euclidean $p-$space. $M^\top$ means transpose of a matrix M. Let $\mathsf{E}(\cdot)$ and $\mathsf{E}(\cdot|\cdot)$ be denoted as expectation and conditional expectation, respectively, under P and the true parameter value as $\boldsymbol{\vartheta}_0 = ((\vartheta_1^0)^\top, \ldots, (\vartheta_p^0)^\top)^\top$. In addition, let $\{\mathcal{F}_t\}_{t=0,1,\ldots}$ denote a sequence of sub-sigma fields with $\mathcal{F}_t$, $t \geq 1$, generated by an arbitrary subset of $Y_1, \ldots, Y_t$ and $\mathcal{F}_0 = \{\emptyset, \Omega\}$ is the trivial sigma field.

Denote by I_d the $d \times d$ identity matrix. If it is clear from the context, then the subscript d will be omitted. $\mathrm{Bin}(n, \alpha)$ denotes a binomial distribution with parameters $n \in \mathbb{N}$ and $\alpha \in [0,1]$; $\mathrm{Poi}(\lambda)$ denotes a Poisson distribution with mean parameter $\lambda \in \mathbb{R}_+$; $\mathrm{Geo}(p)$ denotes a Geometric distribution over $\mathbb{Z}_+$ with parameter $p \in (0,1]$ and mean $(1-p)/p$. Random variables are all

defined on a common probability space $(\Omega, \mathcal{A}, \mathrm{P})$. In addition, $\mathbf{0}$ is a S-dimensional column vector of zeros and, for a non-negative $S \times S$ matrix $M = m_{i,j}, i, j = 1, \ldots, S,$, i.e., $m_{i,j} \geq 0$ for all i, j, consider the notation $M \geq 0$, where *0* is a square S-dimensional matrix of zeros. If $m_{i,j} > 0$ for all i, j, then $M > 0$. These definition and notation should be extended naturally to column vectors.

The organization of the paper is as follows. Section 2 introduces the proposed model, presents the mean and the autocorrelation of the process and some probabilistic properties of the model. Section 3 discusses the estimation methods of the parameters, namely the conditional least squares, the Yule-Walker (moment-based) estimator, and the quasi-maximum likelihood framework. Section 4 presents the simulation and its results. A real data application is presented in Section 5. Section 6 focuses on forecasting purposes. Conclusions and final comments are presented in the last section. In the appendix we present some proofs.

B .2 The PINAR$(1, 1_S)$ model

In the following, the time index t is written as $t = kS + \nu$, where $\nu = 1, \ldots, S$ and $k \in \mathbb{Z}$, when emphasis on seasonality S is important. For example, in the case of daily data and weekly seasonality, $S = 7$, ν is the day of the week and k is the index of the week.

Let $\{Y_t\}$, $t \in \mathbb{Z}$, be a integer-valued stochastic process satisfying $\mathrm{E}(Y_t^2) < \infty$ for all $t \in \mathbb{Z}$. Denote the mean and autocovariance functions of (Y_t) by $\mu_t = \mathrm{E}(Y_t)$ and $\gamma_t(h) = \mathrm{Cov}(Y_t, Y_{t-h})$, respectively. (Y_t) is said to be PC with period S if, for every pair $(t, h) \in \mathbb{Z}^2$,

$$\mu_{t+S} = \mu_t \quad \text{and} \quad \gamma_{t+S}(h) = \gamma_t(h), \tag{3.25}$$

and there are no smaller values of S for which (3.25) is satisfied. This definition implies that μ_t and $\gamma_t(h)$ are periodic functions in t and need to be known only for $t = 1, \ldots, S$. If $S = 1$, (X_t) is weakly stationary in the usual sense.

$\{Y_t\}$ is said to be a periodic integer-valued process with period $S \in \mathbb{N}$, $S > 1$, and autoregressive orders $(1, 1_S)$, and is denoted by PINAR$(1, 1_S)$, if it satisfies the difference equation

$$Y_{kS+\nu} = \alpha_\nu \circ Y_{kS+\nu-1} + \beta_\nu \circ Y_{kS+\nu-S} + \varepsilon_{kS+\nu}, \tag{3.26}$$

where $\alpha_\nu, \beta_\nu \in (0, 1)$ are the thinning coefficients during the season ν. The random variables (RVs) $\{\varepsilon_t\}$ are non-negative and mutually independent, have finite second order moments, and for each ν, the RVs $\{\varepsilon_{kS+\nu}\}_{k\in\mathbb{Z}}$ have the same distribution and we denote $\mathrm{E}(\varepsilon_{kS+\nu}) = \lambda_\nu$ and $\mathrm{Var}(\varepsilon_{kS+\nu}) = \sigma_\nu^2 > 0$.

In addition, it is assumed that ε_t is independent of Y_{t-1}, $\alpha_\nu \circ Y_{t-1}$, Y_{t-S} and $\beta_\nu \circ Y_{t-S}$ and all counting processes are mutually independent.

As can be seen, for each seasonal period ν, Y_t in (3.26) has three random components; the immigration of the immediate past Y_{t-1} with survival probability α_ν, the immigration at $t-S$ with probability β_ν and the elements which entered in the system in the interval (t-1, t], which define

the innovation term ε_t for all $t \in \mathbb{Z}$, where $t = kS + \nu$, $k \in \mathbb{Z}$ and $\nu = 1, \ldots, S$. Moreover, the autoregressive parameters α_ν, β_ν and immigration means λ_ν, $\nu = 1. \ldots, S$, change periodically according to the seasonal period S. Note that the above model becomes an extension of models introduced in Moriña et al. (2011) in which the autoregressive coefficients are fixed in time and only the immigration mean varies within a period. Contrarily, in our proposed model, in addition to the periodic mean value, the autoregressive coefficients also vary periodically. In this context, the PINAR$(1, 1_S)$ model (3.26) also accommodates the periodicity in the autoregressive coefficients, that is, it can be considered as a kind of cyclostationary models introduced in Gladyshev (1961) for standard linear time series.

The unconditional mean of the process $\{Y_t\}$, in (3.26) is given by

$$\mathsf{E}(Y_{kS+\nu}) = \alpha_\nu \mathsf{E}(Y_{kS+\nu-1}) + \beta_\nu \mathsf{E}(Y_{kS+\nu-S}) + \mathsf{E}(\varepsilon_{kS+\nu}), \mu_\nu = \alpha_\nu \mu_{\nu-1} + \beta_\nu \mu_\nu + \lambda_\nu.$$

In the above equation, $\mathsf{E}(\alpha \circ Y) = \alpha \mathsf{E}(Y)$. For more details of the thinning operator properties see, for example, Lemma 1 in da Silva & Oliveira (2004). It is worth to note that, the mean of arrivals at season ν, μ_ν, corresponds to the proportion α_ν of the mean arrivals at $t-1$ plus the proportion β_ν of the mean arrivals at time $t-S$ and the mean of the new events λ_ν.

Equivalently to the PINAR$(1)_S$ model in Monteiro et al. (2010), the analysis of the existence and uniqueness of a periodically stationary and causal PINAR$(1, 1_S)$ process, defined in (3.26), can be obtained analogously as the multivariate integer-valued autoregressive process introduced by Latour (1997). The PINAR$(1, 1_S)$ model can be algebraically rewritten as follows. Firstly, consider the following definitions (see Definition 2.1 in Latour (1997)).

Definition 2. Let $A\circ = (a_{i,\nu}\circ)$, $1 \leq i, \nu \leq S$, be a $S \times S$ *matricial binomial thinning operator*, also called the matricial Steuel and Van Harn operator, where $a_{i,\nu} \in [0, 1]$ for all $1 \leq i, \nu \leq S$. The action of $A\circ$ on $\boldsymbol{Y} = (Y_1, \ldots, Y_S)^\top$, denoted by $A \circ \boldsymbol{Y}$, is

$$A \circ \boldsymbol{Y} = A \circ \begin{pmatrix} Y_1 \\ \vdots \\ Y_S \end{pmatrix} = \begin{pmatrix} \sum_{\nu=1}^{S} a_{1,\nu} \circ Y_\nu \\ \vdots \\ \sum_{\nu=1}^{S} a_{S,\nu} \circ Y_\nu \end{pmatrix}. \tag{3.27}$$

According to 3.24, the operator $a_{i,\nu}\circ$, $1 \leq i, \nu \leq S$, is based on a sequence $\{U_l(a_{i,\nu})\}_{l\in\mathbb{Z}_+}$ of independent identically Bernoulli distributed random variables. Based on Lemma 2.1 in Latour (1997), $\mathsf{E}(A \circ \boldsymbol{Y}) = A^* \mathsf{E}(\boldsymbol{Y})$, where $A^* = (a_{i,\nu})$, $1 \leq i, \nu \leq S$.

Definition 3. Let Y_t, $t \in \mathbb{Z}$, be a non-negative integer-valued random variable and $Z_t = \alpha \circ Y_t + \beta \circ Y_t = (\alpha \oplus \beta) \circ Y_t$ where $0 \leq \alpha, \beta < 1$. Z_t is the process satisfying

$$Z_t = (\alpha \oplus \beta) \circ Y_t = \alpha \circ Y_t + \beta \circ Y_t = \sum_{l_1=1}^{Y_t} U_{l_1}(\alpha) + \sum_{l_2=1}^{Y_t} U_{l_2}(\beta), \tag{3.28}$$

where the counting processes $\{U_{l_1}(\alpha)\}_{l_1\in\mathbb{Z}_+}$ and $\{U_{l_2}(\beta)\}_{l_2\in\mathbb{Z}_+}$ are sequences of i.i.d. r.v.'s

which are Bernoulli distributed with parameter α and β, respectively. It is assumed that the sequence $\{U_l(.)\}_{l\in\mathbb{Z}_+}$ are mutually independent and independent of the process Y_t. For a given $Y_t = y_t$ and $\alpha \neq \beta$, Z_t is Poisson-Binomial distributed and $\vec{p} = (p_1, \ldots, p_{y_t}, p_{y_t+1}, \ldots, p_{2y_t})'$, is the vector of parameters, where $p_1 = \ldots = p_{Y_t} = \alpha$ and $p_{y_t+1} = \ldots = p_{2y_t} = \beta$. However, for large Y_t and α and β small, but not necessary equals, the distribution of Z_t is well approximated by a Poisson distribution due to the well-known law of small numbers (see Chen & Liu (1997)).

Lemma 1. *Let $\mathcal{F}_t$ denotes the sub-sigma fields generated by $\{Y_1, \ldots, Y_t\}$. The conditional expected value and variance of Z_t, defined in 3.28, are given by*

$$\begin{aligned} \mathsf{E}(Z_t|\mathcal{F}_{t-1}) &= (\alpha+\beta)Y_t \quad \textit{and} \\ \mathrm{Var}(Z_t|\mathcal{F}_{t-1}) &= (\alpha(1-\alpha)+\beta(1-\beta))Y_t, \end{aligned} \tag{3.29}$$

respectively. Equivalently, $\mathsf{E}(Z_t) = (\alpha+\beta)\mathsf{E}(Y_t)$ *and* $\mathrm{Var}(Z_t) = (\alpha(1-\alpha)+\beta(1-\beta))\mathsf{E}(Y_t) + (\alpha+\beta)^2\,\mathrm{Var}(Y_t)$.

Now, let $A\circ$ and $B\circ$ be a $S \times S$ independent operators and $\boldsymbol{Y}_k = (Y_{kS+1}, \ldots, Y_{kS+S})^\top$, $\boldsymbol{\varepsilon}_k = (\varepsilon_{kS+1}, \ldots, \varepsilon_{kS+S})^\top$, $k \in \mathbb{Z}$, where Y_{kS+1} and ε_{kS+1} are defined in model (3.26).

By rearranging the simple iteration of Eq. (3.26), the following stochastic equation is obtained

$$\boldsymbol{Y}_k = A \circ \boldsymbol{Y}_{k-1} + \boldsymbol{\zeta_k}, \tag{3.30}$$

$k \in \mathbb{Z}$, where $\boldsymbol{\zeta_k} = B \circ \boldsymbol{\varepsilon}_k$ and the operators $A\circ$ and $B\circ$ are defined by

$$A\circ = \begin{bmatrix} \beta_1\circ & 0\circ & 0\circ & \cdots & 0\circ & 0\circ & \alpha_1\circ \\ \alpha_2\beta_1\circ & \beta_2\circ & 0\circ & \cdots & 0\circ & 0\circ & \alpha_2\alpha_1\circ \\ \alpha_3\alpha_2\beta_1\circ & \alpha_3\beta_2\circ & \beta_3\circ & \cdots & 0\circ & 0\circ & \alpha_3\alpha_2\alpha_1\circ \\ \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ \prod_{i=2}^{S-1}\alpha_i\beta_1\circ & \prod_{i=3}^{S-1}\alpha_i\beta_2\circ & \prod_{i=4}^{S-1}\alpha_i\beta_3\circ & \cdots & \alpha_{S-1}\beta_{S-2}\circ & \beta_{S-1}\circ & \prod_{i=1}^{S-1}\alpha_i\circ \\ \prod_{i=2}^{S}\alpha_i\beta_1\circ & \prod_{i=3}^{S}\alpha_i\beta_2\circ & \prod_{i=4}^{S}\alpha_i\beta_3\circ & \cdots & \alpha_S\alpha_{S-1}\beta_{S-2}\circ & \alpha_S\beta_{S-1}\circ & (\prod_{i=1}^{S}\alpha_i \oplus \beta_S)\circ \end{bmatrix} \tag{3.31}$$

and

$$B\circ = \begin{bmatrix} 1\circ & 0\circ & 0\circ & \cdots & 0\circ & 0\circ & 0\circ \\ \alpha_2\circ & 1\circ & 0\circ & \cdots & 0\circ & 0\circ & 0\circ \\ \alpha_3\alpha_2\circ & \alpha_3\circ & 1\circ & \cdots & 0\circ & 0\circ & 0\circ \\ \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ \prod_{i=2}^{S-1}\alpha_i\circ & \prod_{i=3}^{S-1}\alpha_i\circ & \prod_{i=4}^{S-1}\alpha_i\circ & \cdots & \alpha_{S-1}\circ & 1\circ & 0\circ \\ \prod_{i=2}^{S}\alpha_i\circ & \prod_{i=3}^{S}\alpha_i\circ & \prod_{i=4}^{S}\alpha_i\circ & \cdots & \alpha_S\alpha_{S-1}\beta_3\circ & \alpha_S\circ & 1\circ \end{bmatrix}, \tag{3.32}$$

with α_i's and β_i's, $i = 1, \ldots, S$, defined in (3.26).

Note that $\{\boldsymbol{\zeta_k}\}_{k\in\mathbb{Z}}$ is a sequence of i.i.d. $\mathbb{Z}_+^S$-valued random vectors independent of the matricial thinning operator $A\circ$, with finite mean and variance given by $\boldsymbol{\mu}_\zeta = \mathsf{E}(\boldsymbol{\zeta_k}) = \mathsf{E}(B \circ \boldsymbol{\varepsilon}_k) = B^*\mathsf{E}(\boldsymbol{\varepsilon}_k) = B^*(\lambda_1, \ldots, \lambda_S)^\top$ and $\mathrm{Var}(\boldsymbol{\zeta_k}) = B^*\Sigma_{\boldsymbol{\varepsilon}_k}(B^*)^\top$, respectively, where $\Sigma_{\boldsymbol{\varepsilon}_k} =$

$\mathrm{diag}(\sigma_1, \ldots, \sigma_S)$, for all $k \in \mathbb{Z}$. In addition, using Lemma 1, the matrix A^*, is given by

$$A^* = \begin{bmatrix} \beta_1 & 0 & 0 & \cdots & 0 & 0 & \alpha_1 \\ \alpha_2\beta_1 & \beta_2 & 0 & \cdots & 0 & 0 & \alpha_2\alpha_1 \\ \alpha_3\alpha_2\beta_1 & \alpha_3\beta_2 & \beta_3 & \cdots & 0 & 0 & \alpha_3\alpha_2\alpha_1 \\ \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ \prod_{i=2}^{S-1}\alpha_i\beta_1 & \prod_{i=3}^{S-1}\alpha_i\beta_2 & \prod_{i=4}^{S-1}\alpha_i\beta_3 & \cdots & \alpha_{S-1}\beta_{S-2} & \beta_{S-1} & \prod_{i=1}^{S-1}\alpha_i \\ \prod_{i=2}^{S}\alpha_i\beta_1 & \prod_{i=3}^{S}\alpha_i\beta_2 & \prod_{i=4}^{S}\alpha_i\beta_3 & \cdots & \alpha_S\alpha_{S-1}\beta_{S-2} & \alpha_S\beta_{S-1} & (\prod_{i=1}^{S}\alpha_i + \beta_S) \end{bmatrix}. \tag{3.33}$$

Thus, (3.26) is well defined by (3.30) and this is called the matricial representation of the PINAR$(1, 1_S)$ model defined in (3.26). The next step is to show that there exist a unique stationary non-negative integer-valued solution of $\boldsymbol{Y}_t$ to 3.30.

Let $\boldsymbol{\mu}_Y = \mathsf{E}(\boldsymbol{Y}_k) = (\mu_1, \ldots, \mu_\nu)^\top$, for all $k \in \mathbb{Z}$ and taking expectation on both sides of (3.30) leads to

$$\boldsymbol{\mu}_Y = A^*\boldsymbol{\mu}_Y + \boldsymbol{\mu}_\zeta, \tag{3.34}$$

$$(\boldsymbol{I} - A^*)\boldsymbol{\mu}_Y = \boldsymbol{\mu}_\zeta, \tag{3.35}$$

where I is an $S \times S$ identity matrix. A^* is a non-negative matrix and therefore, A^* satisfies $(I + A^*)^{S-1} > 0$. According to Theorem 6.2.23 in Horn & Johnson (2012), A^* is an irreducible matrix. Based on Theorem 2.1 in Seneta (2006), a necessary and sufficient condition for a solution $\boldsymbol{\mu}_Y$ ($\boldsymbol{\mu}_Y \geq \mathbf{0}, \neq \mathbf{0}$, i.e.,) to the equations $(x\boldsymbol{I} - A^*)\boldsymbol{\mu}_Y = \boldsymbol{\mu}_\zeta$ to exist for any $\boldsymbol{\mu}_\zeta$ ($\boldsymbol{\mu}_\zeta \geq \mathbf{0}, \neq \mathbf{0}$) is that $x > \rho(A^*)$ where $\rho(A^*)$ is the spectral radius, which is the maximum eigenvalue in modulus of the matrix A^*. In this case, there is only one solution, which is strictly positive and given by $\boldsymbol{\mu}_Y = (x\boldsymbol{I} - A^*)^{-1}\boldsymbol{\mu}_\zeta$. From (3.35), $x = 1$. Therefore, $\rho(A^*) < 1$ and $\boldsymbol{\mu}_Y = (\boldsymbol{I} - A^*)^{-1}\boldsymbol{\mu}_\zeta$ is the unique and strictly positive solution for (3.35). Since $S > 1$ and based on the Perron-Frobenius Theorem 8.4.4 in Horn & Johnson (2012) page 534, $\rho(A^*)$ is a positive value, that is, $\rho(A^*) > 0$. In addition, since $0 < \rho(A^*) < 1$ and based on item (1) of Corollary 5.6.10.1 in Graybill (1983), $(I - A^*)^{-1}$ is the limit of $\sum_{i=1}^{\infty}(A^*)^i$. Item (2) in this corollary ensures that $(I - A^*)^{-1}$ is a non-singular non-negative matrix.

Proposition 1. The matrix A^* in (3.33) is a primitive matrix.

Proof. As referenced above, the matrix A^*, in (3.33), is a non-negative and irreducible matrix. In addition, A^* is also an aperiodic matrix (see Frank Ayres (1967), pag 11), that is, there is not a positive integer k such that $(A^*)^{k-1} = A^*$. Since A^* satisfies the conditions of Theorem 1.4 in Seneta (2006), then A^* is primitive. □

The model PINAR$(1, 1_S)$ in (3.30) will be completely specified, if the $det(zI - A^*) \neq 0$, $z \in \mathbb{C}$, i.e., the characteristic roots will be inside of unit circle, and, therefore, the process in 3.30 will be stationary (Brockwell & Davis (2013), Latour (1997)). The following Lemma is the bridge to establish the model properties of the PINAR process defined in (3.30).

Lemma 2. *Let the matrix A^* be defined by* (3.33)*. Then, the following statements are equivalent:*

(i) $\rho(A^*) < 1$*;*

(ii) the roots of the determinant equation $\det(zI - A^*) = 0$*, for all complex* z*, are all less than 1 in absolute value;*

(iii) the roots of the characteristic polynomial

$$P(z) = \prod_{\nu=1}^{S-1}(z-\beta_\nu)(z - \prod_{i=1}^{S}\alpha_i - \beta_S) - \prod_{\nu=1}^{S}\alpha_\nu[\prod_{m=1}^{S-1}(\beta_m) + \sum_{n=1}^{S-1}[(z-\beta_n)\prod_{p\geq 1, p\neq n}^{S-1}\alpha_p] + \sum_{q=1}^{S-1}[(z-\beta_q)\sum_{a=q+1}^{S-1}(z-\beta_a)\prod_{b\geq 1, b\neq a}^{S-1}\alpha_b] + \ldots + \sum_{c=1}^{S-1}\alpha_c \prod_{d\geq 1, d\neq c}^{S-1}(z-\beta_c),$$

for all complex z*, lie inside of the complex unit circle.*

Proof. of Lemma 2. A^* is an $S \times S$ non-negative primitive matrix (Proposition 1) and $\rho(A^*) > 0$. ($i \rightarrow$ ii). $\rho(A^*) < 1$. From item (c) of Theorem 1.1 in Seneta (2006), $\rho(A^*) > |z|$; for any eigenvalue $z \neq \rho(A^*)$. Since z represents the roots of the characteristic polynomial $P(z) = \det(zI - A^*) = 0$, and $\rho(A^*) < 1$, then $|z| < 1$. ($ii \rightarrow$ i). $|z| < 1$ where z represents the roots of the characteristic polynomial $P(z) = \det(zI - A^*) = 0$. This directly implies that $\rho(A^*) < 1$. The proof of item iii is the simple Laplace formula applied on $P(z) = \det(zI - A^*) = 0$. □

Proposition 2. The PINAR$(1, 1_S)$ model, defined in (3.26), is periodically second-order stationary process.

Proof. Based on Lemma 2, all the eigenvalues of A^* are inside the complex unit circle, then $(I - A^*)$ is a non-singular matrix and $\boldsymbol{\mu}_Y = (I - A^*)^{-1}\boldsymbol{\mu}_\zeta$, which is finite. Based on these and the assumptions of the model in (3.26), the PINAR$(1, 1_S)$ model satisfies the conditions 3.1 in Latour (1997). Then, second moment is also finite (see Lemma 3.2 in Latour (1997)). Besides, under conditions of Lemma 2, it follows from Proposition 3.1 in Latour (1997) that there exist an almost surely unique non-negative integer-valued stationary process satisfying (3.30), and, consequently, the model (3.26) is periodically second-order stationary. □

Following the same lines of Proposition 4.1 in Latour (1997), the model in (3.30) can also be seen as just a standard vector AR(1) process and it is formalized below.

Proposition 3. Let $\{\boldsymbol{Y}_t\}_{t\in\mathbb{Z}}$ be the second order stationary process defined in (3.30). Then, $\{\boldsymbol{Y}_t\}_{t\in\mathbb{Z}}$ can be seen as an Vector AR(1) model with covariance matrix

$$\Gamma(h) = \begin{cases} A^*\Gamma(1)^\top + \operatorname{diag}(\tilde{A}\boldsymbol{\mu}_Y) + \operatorname{Var}(\boldsymbol{\zeta_k}), & \text{if } h = 0 \\ (A^*)^i\Gamma(0), & \text{if } h \geq 1. \end{cases} \tag{3.36}$$

The representation $\operatorname{diag}\{\boldsymbol{v}\}$ denotes a diagonal matrix with vector $\boldsymbol{v}$ in its diagonal. The matrix $\tilde{A}$ is the variance matrix of the operator $A\circ$ defined in (3.31).

The marginal distribution of the PC process $\{Y_t\}_{t\in\mathbb{Z}}$ at $t = kS + \nu$ is given by

$$P(Y_{kS+\nu} = c) = \sum_{m,n=0}^{\infty} p_\nu(c|m,n)P(Y_{kS+\nu-1} = m, Y_{(k-1)S+\nu} = n), \tag{3.37}$$

where $c \in \mathbb{Z}_+$, $k \in \mathbb{Z}$, $\nu = 1,\ldots,S$ and $p_\nu(c|b_1,b_2) = P(Y_t = c|Y_{t-1} = b_1, Y_{t-S} = b_2)$ for each ν. Given starting values $Y_1,\ldots,Y_S$ of $\{\boldsymbol{Y}_t\}_{t\in\mathbb{Z}}$, the conditional joint probability is given by

$$\begin{aligned} P(Y_t = y_t,\ldots,Y_{S+1} = y_{S+1}|Y_S = y_S,\ldots,Y_1 = y_1) = \\ p_\nu(y_t|y_{t-1},y_{t-s})P(Y_{t-1} = y_{t-1},\ldots,Y_{S+1} = y_{S+1}|Y_S = y_S,\ldots,Y_1 = y_1), \end{aligned} \tag{3.38}$$

where $y_1,\ldots,y_t \in \mathbb{Z}_+$. Thus, by induction, if $T = nS$ where $n \in \mathbb{N}$, it can be calculated as

$$P(Y_T = y_T,\ldots,Y_{S+1} = y_{S+1}|Y_S = y_S,\ldots,Y_1 = y_1) = \prod_{\nu=1}^{S}\prod_{k=1}^{n-1} p_\nu(y_{kS+\nu}|y_{kS+\nu-1},y_{kS+\nu-S}), \tag{3.39}$$

where $y_1,\ldots,y_T \in \mathbb{Z}_+$.

PINAR$(1,1_S)$ with Poisson immigration

Let the innovation process in (3.26) be an i.i.d Poisson process with unconditional mean $\mathsf{E}(\varepsilon_{kS+\nu}) = \lambda_\nu \in \mathbb{R}_+$. When S = 1, the model (3.26) becomes Poisson. Additionally, when $S > 1$ it can be shown that the unconditional mean and variance of $\{\boldsymbol{Y}_t\}_{t\in\mathbb{Z}}$ are generally not equal so that the marginal stationary distribution of $\{\boldsymbol{Y}_t\}_{t\in\mathbb{Z}}$ is no longer Poisson even though the innovations are (Bu et al. (2008)). However, due to the well-known Law of Small Numbers, an approximation to a Poisson distribution can be achieved if $\alpha_\nu.\beta_\nu \approx 0$ and when $\{\boldsymbol{Y}_t\}_{t\in\mathbb{Z}}$ becomes large. See, also, Chen & Liu (1997). Under some conditions, the following theorem establishes the distribution of model (3.26).

Lemma 3. *Let $\{U_i(\alpha)\}_{i\in\mathbb{Z}_+}$ be a sequence of independent identically Bernoulli distributed variable with P$(U_i(\alpha) = 1) = 1 -$ P$(U_i(\alpha) = 0) = \alpha$. It is also assumed that the sequence $\{U_i(\alpha)\}_{i\in\mathbb{Z}_+}$ is mutually independent of the random variable Y which follows a Poisson distribution with parameter $\theta > 0$. If*

$$M_Y = \alpha \circ Y = \sum_{i=1}^{Y} U_i(\alpha),$$

then $M_Y \sim$ Poi$(\alpha\theta)$.

See the proof in the Appendix.

Theorem 2. *Let $\{\varepsilon_t\}_{t\in\mathbb{Z}}$, $t = kS + \nu$, for all $k \in \mathbb{Z}$ and $\nu = 1,\ldots,S$, be a periodic sequence of independent Poisson distributed r.v.'s of period S, i.e., $\varepsilon_{kS+\nu} \sim \mathrm{Poi}(\lambda_\nu)$, $\lambda_\nu \in \mathbb{R}_+$. Let the starting values $(Y_\nu)_{1\le\nu\le S}$ follow periodic independent Poisson processes with mean $\mathsf{E}(Y_\nu) = \mu_\nu$, i.e., $Y_\nu \sim \mathrm{Poi}(\mu_\nu)$. Then, the process $\{Y\}_{t\in\mathbb{Z}}$ in* (3.26) *has the following distributions:*

1. *For* $1 \leq t \leq 2S-1$, $\{Y_t\}_{t\in\mathbb{Z}}$ *is a periodic Poisson process with periodic mean* μ_ν.

2. *For* $t \geq 2S$*:*

   (a) $\{Y_t\}_{t\in\mathbb{Z}}$ *follows a Poisson-Binomial process with periodic mean* μ_ν.

   (b) *if* $\alpha_i\beta_i \approx 0$*, for all* $i = 1,\ldots,S$*, then the marginal distribution of a PINAR*$(1,1_S)$ *process* $\{Y_t\}_{t\in\mathbb{Z}}$ *defined in* (3.26) *is a periodic Poisson process with periodic means* μ_ν*, i.e.,* $Y_{kS+\nu} \sim \text{Poi}(\mu_\nu)$.

The proof is in the Appendix.

Remark 1. The assumption $\alpha_i\beta_i \approx 0$ in 2.b is the necessary condition to guarantee that the periodic process $\{Y_t\}_{t\geq 2S}$ becomes a periodic Poisson process (see more details in the proof of Theorem 2). In fact, based on Definition 3, for $t \geq 2S$, $\{Y_t\}$ is a periodic Poisson-Binomial process (2.a). For example, for $t = 2S$, given the initial values $\{Y_\nu\}_{1\leq\nu\leq S}$, Y_{2S} can be written recursively as $Y_{2S} = (\prod_{i=1}^{S}\alpha_i \oplus \beta_S)\circ Y_S + \alpha_S\beta_{S-1}\circ Y_{S-1} + \ldots + \varepsilon_{2S}$ (see more details in the proof of Theorem 2). Note that, in $(\prod_{i=1}^{S}\alpha_i \oplus \beta_S)\circ Y_S$ the quantities $\prod_{i=1}^{S}\alpha_i \circ Y_S$ and $\beta_S \circ Y_S$ depend on Y_S. Then the marginal distribution of Y_{2S} has a recursive equation which presents a sum that violates the independence assumption among the variables of the sum of Poisson distributed variables. Therefore, the assumption $\alpha_i\beta_i \approx 0$ preserves that the recursive formula of $(Y_t)_{t\geq 2S}$ has variables Y_{l_1} and ε_{l_2} $(1 \leq l_1, l_2 < t)$, for all t, related to unique thinning operators, respectively.

Remark 2. Although the assumption $\alpha_i\beta_i \approx 0$, for all $i = 1,\ldots,S$, seems to be too restrictive, the empirical results in Section B .4 show that the parameters estimation methods present good results even for $\alpha_i\beta_i \neq 0$, as can be seen in Tables 3.8 and 3.9.

The periodic Markov-kernel of PINAR$(1,1_S)$ model with Poisson immigration is given by the following way. For all $t = kS+\nu$, where $k \in \mathbb{Z}$ and $\nu = 1,\ldots,S$, if y_t, y_{t-1}, y_{t-s} denote the values of the process $\{Y_t\}_{t\in\mathbb{Z}}$ at time $t, t-1, t-S$, then

$$
\begin{aligned}
p_\nu(y_t|y_{t-1},y_{t-s}) &= [\text{Bin}(y_{t-1},\alpha_\nu) * \text{Bin}(y_{t-s},\beta_\nu) * \text{Poi}(\lambda_\nu)](y_t) \qquad (3.40)\\
&= \sum_{(m,n)\in\mathcal{J}} \binom{y_{t-1}}{m}\alpha_\nu^m(1-\alpha_\nu)^{y_{t-1}-m}\binom{y_{t-s}}{n}\beta_\nu^n(1-\beta_\nu)^{y_{t-s}-n}\frac{\lambda_\nu^{y_t-m-n}}{(y_t-m-n)!}\text{e}^{-\lambda_\nu},
\end{aligned}
$$

where $*$ denotes convolution and the index set $\mathcal{J}$ is defined by $\mathcal{J} = \{(m,n) \in \mathbb{Z}_+^2 | m \leq y_{t-1}, n \leq y_{t-s}, m+n \leq y_t\}$ (Note that the definition of $\mathcal{J}$ depends on the values y_t, y_{t-1}, y_{t-s}).

PINAR$(1,1_S)$ with Geometric immigration

Let $\{\varepsilon_t\}_{t>S}$ follow a periodic sequence having Geometric distribution with parameter $(1+\lambda_\nu)^{-1}$, $0 < \lambda_\nu < \infty$, $\nu = 1,\ldots,S$, i.e $\varepsilon_{kS+\nu} \sim \text{Geo}((1+\lambda_\nu)^{-1})$ for all $k \in \mathbb{Z}$ and $\nu = 1,\ldots,S$. Then $\text{E}(\varepsilon_{kS+\nu}) = \lambda_\nu$.

The periodic Markov-kernel of PINAR$(1,1_S)$ model with Geometric immigration is given by the

following way. For all $t = kS + \nu$, where $k \in \mathbb{Z}$, $\nu = 1, \ldots, S$, if y_t, y_{t-1}, y_{t-s} denote the values of the process $\{Y_t\}_{t\in\mathbb{Z}}$ at time $t, t-1, t-S$, then

$$p_\nu(y_t|y_{t-1}, y_{t-s}) = [\mathrm{Bin}(y_{t-1}, \alpha_\nu) * \mathrm{Bin}(y_{t-s}, \beta_\nu) * \mathrm{Geo}((1+\lambda_\nu)^{-1})](y_t) \tag{3.41}$$
$$\sum_{(m,n)\in\mathcal{J}} \binom{y_{t-1}}{m} \alpha_\nu^m (1-\alpha_\nu)^{y_{t-1}-m} \binom{y_{t-s}}{n} \beta_\nu^n (1-\beta_\nu)^{y_{t-s}-n} \frac{\lambda_\nu^{y_t-m-n}}{(1+\lambda_\nu)^{y_t-m-n+1}},$$

where $*$ denotes convolution and the index set $\mathcal{J}$ is defined by $\mathcal{J} = \{(m,n) \in \mathbb{Z}_+^2 | m \le y_{t-1}, n \le y_{t-s}, m+n \le y_t\}$.

The probability distribution of $\{Y_t\}_{t\in\mathbb{Z}}$ when $\{\varepsilon_t\}_{t>S}$ follow a periodic sequence having Geometric distribution is not discussed here. The use of such innovation distribution to estimate the parameter of the PINAR$(1, 1_S)$ model will be considered in the simulation study.

B .3 Parameter estimation methods

In this section, the estimation methods moment-based or Yule-Walker (YW), quasi-maximum likelihood (QML) and conditional least squares (CLS) are discussed for the proposed model in (3.26) under a general immigration distribution. As examples, the properties of these estimators are also discussed under Poisson innovation marginal distribution, i.e., $\varepsilon_{kS+\nu} \sim \mathrm{Poi}(\lambda_\nu)$ for all $k \in \mathbb{Z}$ and $\nu \in \{1, \ldots, S\}$.

Let $\vartheta_\nu = (\alpha_\nu, \beta_\nu, \lambda_\nu)^\top$, $\alpha_\nu, \beta_\nu \in (0;1)$ and $0 < \lambda_\nu < \infty$, for $\nu = 1, \ldots, S$ (S is fixed), and let $\boldsymbol{\vartheta} = (\vartheta_1^\top, \ldots, \vartheta_S^\top)^\top$ represent the $3S$-dimensional unknown parameter vector of the PINAR$(1, 1_S)$ model defined by (3.26). The parameter vector is assumed to be lying in the open set $\Theta = ([0,1] \times [0,1] \times (0,\infty))^S$, which contains the true parameter vector, denoted by $\boldsymbol{\vartheta}_0 = ((\vartheta_1^0)^\top, \ldots, (\vartheta_S^0)^\top)^\top$. Now, without lost of generality, It is assumed here that $Y_1, \ldots, Y_T$ has n complete periods of observations, that is, consider a sample $Y_1, \ldots, Y_T$ of size $T = nS$ from $\{Y_t\}_{t\in\mathbb{Z}}$, the PINAR$(1, 1_S)$ process in model (3.26), where $n \in \mathbb{N}$.

Conditional least squares estimation (CLS)

The CLS-estimators $\widehat{\boldsymbol{\vartheta}}_n^{\mathrm{CLS}}$, $n \in \mathbb{N}$, of $\boldsymbol{\vartheta}$ are obtained by minimizing the expression

$$Q_n(\boldsymbol{\vartheta}) = \sum_{\nu=1}^{S} Q_n(\vartheta_\nu) = \sum_{\nu=1}^{S} \sum_{k=1}^{n-1} (Y_{kS+\nu} - \mathsf{E}(Y_{kS+\nu}|\mathcal{F}_{kS+\nu-1}))^2, \tag{3.42}$$

where, by (3.26),

$$\mathsf{E}(Y_{kS+\nu}|\mathcal{F}_{kS+\nu-1}) = \alpha_\nu Y_{kS+\nu-1} + \beta_\nu Y_{kS+\nu-S} + \lambda_\nu, \tag{3.43}$$

$t > S$ (see, Eq. 3.2.58 in Taniguchi & Kakizawa (2000)).

To find the CLS estimators of $\boldsymbol{\vartheta}$ is equivalent to find the solution of

$$\frac{\partial}{\partial \vartheta_{\boldsymbol{\vartheta}}} Q_n(\boldsymbol{\vartheta}) = \mathbf{0}. \tag{3.44}$$

$Q_n(.)$ in (3.44) attains a relative minimum at $\widehat{\boldsymbol{\vartheta}}_n^{\mathrm{CLS}}$. The parameter estimators can be computed numerically by separating the parameters according to the seasons. Using the properties of differential calculus, to minimize 3.42 is equivalent to minimize individually each $Q_n(\vartheta_\nu)$ for each $\nu = 1, \ldots, S$ as follows

$$\frac{\partial}{\partial(\vartheta_{\boldsymbol{\vartheta}})} Q_n(\boldsymbol{\vartheta}) = \frac{\partial}{\partial(\vartheta_{\nu_1})} Q_n(\vartheta_1) + \ldots + \frac{\partial}{\partial(\vartheta_{\nu_S})} Q_n(\vartheta_S). \tag{3.45}$$

Define the random vectors $\boldsymbol{U}_t = (Y_t, Y_{t-S+1}, 1)^\top$, $t > S$, and introduce, for all $\nu = 1, \ldots, S$,

$$\boldsymbol{Z}_\nu = \begin{bmatrix} Y_{S+\nu} \\ \vdots \\ Y_{(n-1)S+\nu} \end{bmatrix}, \qquad \boldsymbol{C}_\nu = \begin{bmatrix} \boldsymbol{U}_{S+\nu-1}^\top \\ \vdots \\ \boldsymbol{U}_{(n-1)S+\nu-1}^\top \end{bmatrix} = \begin{bmatrix} Y_{S+\nu-1} & Y_\nu & 1 \\ \vdots & \vdots & \vdots \\ Y_{(n-1)S+\nu-1} & Y_{(n-1)S-S+\nu} & 1 \end{bmatrix}. \tag{3.46}$$

$\boldsymbol{Z}_\nu$ is a $(n-1)$-dimensional random vector and $\boldsymbol{C}_\nu$ is a random matrix of dimension $(n-1) \times 3$. By (3.42), (3.43) and (3.45),

$$Q_{n,}(\vartheta_\nu) = \sum_{k=1}^{n-1} (Y_{kS+\nu} - \alpha_\nu Y_{kS+\nu-1} - \beta_\nu Y_{kS+\nu-S} - \lambda_\nu)^2 = \|\boldsymbol{Z}_\nu - \boldsymbol{C}_\nu \vartheta_\nu\|^2, \tag{3.47}$$

for each $\nu = 1, \ldots, S$. Thus, the CLS-estimator $\vartheta_\nu^{\mathrm{CLS}}$ of the parameter vector ϑ_ν can be expressed as

$$\vartheta_\nu^{\mathrm{CLS}} = \left(\boldsymbol{C}_\nu^\top \boldsymbol{C}_\nu\right)^{-1} \boldsymbol{C}_\nu^\top Z_\nu, \tag{3.48}$$

for each season $\nu = 1, \ldots, S$ (see, Theorem 7.2.2, in Bickel & Doksum (1977)).

One can see that the real-valued penalty function $Q_n(\boldsymbol{\vartheta})$ is twice continuously differentiable with respect to $\boldsymbol{\vartheta}$ in some neighborhood of $\boldsymbol{\vartheta}_0$ (see, Taniguchi & Kakizawa (2000), page 96).

Following Taniguchi & Kakizawa (2000), page 99, let the matrices V_ν and R_ν, $\nu = 1, \ldots, S$ and $t = kS + \nu$, of dimension 3×3 defined as

$$V_\nu = \mathsf{E}\left(\boldsymbol{U}_{t-1}\boldsymbol{U}_{t-1}^\top\right) = \begin{bmatrix} \gamma_{\nu-1}(0) & \gamma_\nu(S-1) & 0 \\ \gamma_\nu(S-1) & \gamma_\nu(0) & 0 \\ 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} \mu_{\nu-1} \\ \mu_\nu \\ 1 \end{bmatrix} \begin{bmatrix} \mu_{\nu-1} & \mu_\nu & 1 \end{bmatrix}, \tag{3.49}$$

where, for $\nu = 1$, $\mu_0 = \mu_S$ and

$$R_\nu = \mathsf{E}\left[\boldsymbol{U}_{t-1}(Y_t - \boldsymbol{U}_{t-1}^\top \vartheta_\nu)^2 \boldsymbol{U}_{t-1}^\top\right]. \tag{3.50}$$

The block diagonal matrices V and R of dimension $3S \times 3S$ are given by

$$V = \mathrm{diag}\{V_1, \ldots, V_S\} \quad \text{and} \quad R = \mathrm{diag}\{R_1, \ldots, R_S\}, \tag{3.51}$$

respectively.

The above results are the basis of the following theorem which shows that the CLS-estimator $\widehat{\boldsymbol{\vartheta}}^{\mathrm{CLS}} = ((\widehat{\vartheta}_1^{\mathrm{CLS}})^\top, \ldots, (\widehat{\vartheta}_S^{\mathrm{CLS}})^\top)^\top$ of the parameter vector $\boldsymbol{\vartheta}$ is consistent.

Theorem 3. *Assume that* $\{\boldsymbol{Y}_t\}$ *in* (3.30)*, that is, the a PINAR*$(1, 1_S)$ *process, is a strictly stationary ergodic process with* $\mathsf{E}\,\|\boldsymbol{\zeta_k}\|^4 < \infty$ ($\mathsf{E}\,\|\varepsilon_k\|^4 < \infty$ *in model* (3.26)) *and* $\mathsf{E}(Y_{kS+\nu}|\mathcal{F}_{kS+\nu-1})$ *is almost surely three times continuously differentiable in the open set* Θ *containing the true parameter value* $\boldsymbol{\vartheta}_0$*. Then, for the* CLS*-estimators* $\widehat{\boldsymbol{\vartheta}}_n^{\mathrm{CLS}}$*,* $n \in \mathbb{N}$*,*

$$n^{1/2}(\widehat{\boldsymbol{\vartheta}}_n^{\mathrm{CLS}} - \boldsymbol{\vartheta}_0) \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0}, V^{-1}RV^{-1}), \tag{3.52}$$

as $n \to \infty$*, where the matrices* V *and* R *of dimension* $3S \times 3S$ *are defined in* (3.51)*.*

The proof is in the Appendix.

Note that $V^{-1}RV^{-1}$ is a $3S \times 3S$ block diagonal matrix $\mathrm{diag}(V_1^{-1}R_1V_1^{-1}, \ldots, V_S^{-1}R_SV_S^{-1})$ where $V_\nu^{-1}R_\nu V_\nu^{-1}$ is the corresponding 3×3 covariance matrix of $\widehat{\vartheta}_\nu^{\mathrm{CLS}}$. From this on can see that $\widehat{\vartheta}_{\nu_i}^{\mathrm{CLS}}$ and $\widehat{\vartheta}_{\nu_j}^{\mathrm{CLS}}$ are asymptotically independent for $\nu_i \neq \nu_j$, $1 \leq, \nu_i, \nu_j \leq S$. This an equivalent conclusion of the theorem discussed in Shao & Ni (2004) for PAR processes (see, also, Basawa & Lund (2001)). In addition, the condition $\mathsf{E}\,\|\boldsymbol{\zeta_k}\|^4 < \infty$ is imposed so that the Assumption B4 in Theorem 3.2.24 in Taniguchi & Kakizawa (2000) may be used (see, also, Assumption 3.4 in Klimko & Nelson (1978), page 635).

Remark 3. The Assumption $\mathsf{E}\,\|\boldsymbol{\zeta_k}\|^4 < \infty$ is necessary to guarantee the Condition C4 in the proof of Theorem 3 and it is not difficult to be achieved in practical problems. For example, if ε_t follows a Poisson or Geometric distribution. These probability distributions are widely used in practical situations of modeling counting time series. For instance, let $G_{\varepsilon_t}(r)$ be the moment-generating function of $\varepsilon_t \sim \mathrm{Poi}(\lambda_\nu)$, $t = kS + \nu$, $0 < \lambda_\nu < \infty$. Then,

$$G_{\varepsilon_t}(r) = \mathsf{E}(\mathrm{e}^{r\,\varepsilon_t}) = \mathrm{e}^{\lambda_\nu(\mathrm{e}^r - 1)}.$$

For $n \in \mathbb{N}$, the n-th moment about 0 of the distribution of ε_t can be calculated through the n-th derivative of $G_{\varepsilon_t}(r)$, evaluated at $r = 0$, given by $\frac{d^n}{dr^n}G_{\varepsilon_t}(r)$. The sixth moment of ε_t is equal to $\lambda_\nu^6 + 15\lambda_\nu^5 + 65\lambda_\nu^4 + 90\lambda_\nu^3 + 31\lambda_\nu^2 + \lambda_\nu$ and, because $0 < \lambda_\nu < \infty$, for all $\nu = 1, \ldots, S$, the sixth moment is finite. Similarly, the sixth moment of $\varepsilon_{kS+\nu} \sim \mathrm{Geo}((1 + \lambda_\nu)^{-1})$, given by the convergent series $\sum_{r=1}^{\infty} r^6(1-p)^{r-1}$, is also finite, where $p = (1+\lambda_\nu)^{-1}$ and $0 < (1+\lambda_\nu)^{-1} < 1$. These ensure that the fourth moment of the Poisson and Geometric distributions are finite. This remark is also valid for Theorem 4.

The above theorem leads directly to the following Corollary.

Corollary 1. *Let* $\{\varepsilon_{kS+\nu}\}_{k\in\mathbb{Z}}$ *be a sequence of periodic i.i.d Poisson distributed variables with finite mean* $\mathsf{E}(\varepsilon_{kS+\nu}) = \lambda_\nu$*,* $\lambda_\nu \in \mathbb{R}_+$ *in model* (3.26)*. Then, for the* CLS*-estimators* $\widehat{\boldsymbol{\vartheta}}_n^{\mathrm{CLS}} = (\hat{\alpha}_\nu, \hat{\beta}_\nu, \hat{\lambda}_\nu)^\top$*,* $n \in \mathbb{N}$*,*

$$n^{1/2}(\widehat{\boldsymbol{\vartheta}}_n^{\mathrm{CLS}} - \boldsymbol{\vartheta}_0) \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0}, V^{-1}RV^{-1}), \tag{3.53}$$

as $n \to \infty$*, where the matrices* V *and* R *of dimension* $3S \times 3S$ *are defined in* (3.51)*.*

Similar result of Corollary 1 can be derived when the process $\{\varepsilon_{kS+\nu}\}_{k\in\mathbb{Z}}$ has Geometric distribution.

Moment-based estimators (Yule-Walker)

The first and second-order moments of the second-order stationary process $\{\boldsymbol{Y}_k\}_{k\in\mathbb{Z}}$ can be estimated as follows (see Section 11.2 in Brockwell & Davis (2013)). The unbiased estimate of $\boldsymbol{\mu_Y} = \mathsf{E}(\boldsymbol{Y}_k) = (\mu_1, \ldots, \mu_\nu)^\top$, is given by the vector of sample means

$$\widehat{\boldsymbol{\mu}}_Y = \overline{\boldsymbol{Y}}_n = n^{-1}\sum_{k=0}^{n-1}\boldsymbol{Y}_k, \tag{3.54}$$

where $\boldsymbol{Y}_k = (Y_{kS+1}, \ldots, Y_{kS+S})^\top$.

A natural estimate of the variance-covariance matrix of $\{\boldsymbol{Y}_k\}$, $\Gamma(h)$, is given by

$$\widehat{\Gamma}_n(h) = n^{-1}\sum_{k=h}^{n-1}(\boldsymbol{Y}_{k-h} - \overline{\boldsymbol{Y}}_n)(\boldsymbol{Y}_k - \overline{\boldsymbol{Y}}_n)^\top = n^{-1}\sum_{k=h}^{n}\widetilde{\boldsymbol{Y}}_{k-h}\widetilde{\boldsymbol{Y}}_k^\top, \tag{3.55}$$

for $0 \leq h \leq n-1$, where $\widetilde{\boldsymbol{Y}}_k = \boldsymbol{Y}_k - \overline{\boldsymbol{Y}}_n$, $k = 1, \ldots, n$.

For the periodic mean μ_ν, $\nu = 1, \ldots, S$, the estimator is

$$\widehat{\mu}_{n,\nu} = \overline{Y}_{n,\nu} = n^{-1}\sum_{k=1}^{n} Y_{kS+\nu}, \tag{3.56}$$

for all seasons $\nu = 1, \ldots, S$. For the periodic covariance functions γ_ν, the sample estimator is given by

$$\widehat{\gamma}_{n,\nu}(h) = n^{-1}\sum_{k=\lceil(h-\nu)/S\rceil}^{n-1}\widetilde{Y}_{kS+\nu-h}\widetilde{Y}_{kS+\nu}, \tag{3.57}$$

where $0 \leq h \leq nS - \nu$ and $\widetilde{Y}_{kS+\nu} = Y_{kS+\nu} - \overline{Y}_{n,\nu}$. Note that $\lceil x\rceil$ denotes the upper integer part of $x \in \mathbb{R}$.

In the sequel, it is omitted the index n of sample size in the estimators, i.e., simply $\widehat{\mu}_\nu$ and $\widehat{\gamma}_\nu$, $\nu = 1, \ldots, S$.

Using the general Yule-Walker equations derived for the periodic ACFs γ_ν, $\nu = 1, \ldots, S$, of the PC process $\{Y_t\}$ (as can be seen in McLeod (1994)), the estimators of $\alpha_\nu, \beta_\nu, \lambda_\nu$ are

$$\begin{aligned}
\widehat{\alpha}_\nu^{\mathrm{YW}} =& \frac{\widehat{\gamma}_\nu(1)\widehat{\gamma}_\nu(0) - \widehat{\gamma}_{\nu-1}(S-1)\widehat{\gamma}_\nu(S)}{\widehat{\gamma}_{\nu-1}(0)\widehat{\gamma}_\nu(0) - \widehat{\gamma}_{\nu-1}^2(S-1)},\\
\widehat{\beta}_\nu^{\mathrm{YW}} =& \frac{\widehat{\gamma}_\nu(1)\widehat{\gamma}_{\nu-1}(S-1) - \widehat{\gamma}_{\nu-1}(0)\widehat{\gamma}_\nu(S)}{\widehat{\gamma}_{\nu-1}^2(s-1)\widehat{\gamma}_\nu(0) - \widehat{\gamma}_{\nu-1}(0)\widehat{\gamma}_\nu(0)}, \quad \text{and}\\
\widehat{\lambda}_\nu^{\mathrm{YW}} =& (1 - \widehat{\beta}_\nu^{\mathrm{YW}})\widehat{\mu}_\nu - \widehat{\alpha}_\nu^{\mathrm{YW}}\widehat{\mu}_{\nu-1}.
\end{aligned} \tag{3.58}$$

From Section 4.4 in Reinsel (2003), since the PINAR$(1, 1_S)$ process can be seen as a VAR(1)

process (Proposition 3), the CLS and YW estimators of $\boldsymbol{\vartheta}$ are asymptotically equivalent, i.e., these estimators are equivalent for large n (see, discussion in Du & Li (1991), Section 4). Therefore, from Theorem 3, the large sample distribution of the YW-estimator $\widehat{\boldsymbol{\vartheta}}^{\mathrm{YW}}$ is asymptotically normal. In addition, as can be seen in Du & Li (1991), page 133, the $\widehat{\boldsymbol{\vartheta}}^{\mathrm{YW}} = ((\widehat{\vartheta}_1^{\mathrm{YW}})^\top, \ldots, (\widehat{\vartheta}_S^{\mathrm{YW}})^\top)^\top$ is strongly consistent.

Quasi-maximum likelihood (QML)

The approach of QML is based on Taniguchi & Kakizawa (2000), Section 3, page 101. Let the likelihood type penalty function of the PINAR$(1, 1_S)$ model, conditioned on the first S observations, be

$$f_{\vartheta_\nu}(t, t-1) = \mathsf{E}\left[\{Y_t - m_{\vartheta_\nu}(t, t-1)\}^2 \middle| \mathcal{F}_{t-1}\right],$$

where $m_{\vartheta_\nu}(t, t-1)$ is defined in (3.74). Define

$$L_n(\boldsymbol{\vartheta}) = \sum_{k=1}^{n-1}\sum_{\nu=1}^{S}[\log\{f_{\vartheta_\nu}(t, t-1)\} + (Y_t - m_{\vartheta_\nu}(t, t-1))^2 f_{\vartheta_\nu}^{-1}(t, t-1)].$$

The likelihood function $L_n(\boldsymbol{\vartheta}) = \sum_{\nu=1}^{S} l_{n,\nu}(\vartheta_\nu)$, where

$$l_{n,\nu}(\vartheta_\nu) = \sum_{k=1}^{n-1}[\log\{f_{\vartheta_\nu}(t, t-1)\} + (Y_t - m_{\vartheta_\nu}(t, t-1))^2 f_{\vartheta_\nu}^{-1}(t, t-1)], l_{n,\nu}(\vartheta_\nu) = \sum_{k=1}^{n-1}\phi_t(\vartheta_\nu),$$

is minimized in order to obtain the QML-estimator $\widehat{\boldsymbol{\vartheta}}_n^{\mathrm{QML}}$ of the parameter vector $\boldsymbol{\vartheta}$.

Corollary 2. *The function $f_{\vartheta_\nu}(t, t-1)$ is given by*

$$f_{\vartheta_\nu}(t, t-1) = \alpha_\nu(1-\alpha_\nu)Y_{t-1} + \beta_\nu(1-\beta_\nu)Y_{t-S} + \lambda_\nu, \tag{3.59}$$

in the case of Poisson innovations, and by

$$f_{\vartheta_\nu}(t, t-1) = \alpha_\nu(1-\alpha_\nu)Y_{t-1} + \beta_\nu(1-\beta_\nu)Y_{t-S} + \lambda_\nu(1+\lambda_\nu), \tag{3.60}$$

in the case of Geometric innovations.

The function $l_{n,\nu}(\vartheta_\nu)$, for Poisson or and Geometric innovations, can be obtained directly by replacing the results of (3.59), (3.60), respectively, and (3.74) in (3.59). From the strictly stationarity of $\{\boldsymbol{Y}_t\}$ it follows that $\mathsf{E}\,\|\varepsilon_t\|^6 < \infty$ (Remark 3) implies $\mathsf{E}\,\|Y_t\|^6 < \infty$, then one can prove that the real-valued penalty function $L_n(\boldsymbol{\vartheta})$ satisfies the assumptions of Theorem 3.2.26 in Taniguchi & Kakizawa (2000). Thus, there exists a sequence of estimators $\widehat{\boldsymbol{\vartheta}}_n^{\mathrm{QML}} = ((\widehat{\vartheta}_{n,1}^{\mathrm{QML}})^\top, \ldots, (\widehat{\vartheta}_{n,S}^{\mathrm{QML}})^\top)^\top$ such that $\widehat{\boldsymbol{\vartheta}}_n^{\mathrm{QML}} \to \boldsymbol{\vartheta}_0$ almost surely as $n \to \infty$, and for any $\epsilon > 0$, there exists an event E with $\mathsf{P}(E) > 1 - \epsilon$ and an $n_0 \in \mathbb{N}$ such that on E, for $n > n_0$, $\widehat{\boldsymbol{\vartheta}}_n^{\mathrm{QML}}$ is the solution of

$$\frac{\partial}{\partial \boldsymbol{\vartheta}} L_n(\boldsymbol{\vartheta}) = 0, \tag{3.61}$$

which attains a relative minimum of the likelihood function $L_n(\boldsymbol{\vartheta})$.

The minimization of $L_n(\boldsymbol{\vartheta})$ can be done separately by minimizing the partial log-likelihood $l_{n,\nu}(\vartheta_\nu)$ for each season $\nu \in \{1,\ldots,S\}$. Similarly, one can solve the likelihood equation (3.61) by solving the partial likelihood equations

$$\frac{\partial}{\partial \vartheta_\nu} l_{n,\nu}(\vartheta_\nu) = 0, \qquad \nu = 1,\ldots,S,$$

separately.

Define IF_ν the matrix of dimension 3×3 for each season $\nu \in \{1,\ldots,S\}$ as

$$IF_\nu = U_{\vartheta_\nu}^{-1} V_{\vartheta_\nu} U_{\vartheta_\nu}^{-1}, \tag{3.62}$$

where

$$V_{\vartheta_\nu} = \mathsf{E}\left\{ \frac{\partial}{\partial \vartheta_\nu}\phi_t(\vartheta_\nu) \frac{\partial}{\partial \vartheta_\nu^\top}\phi_t(\vartheta_\nu) \right\}, \tag{3.63}$$

and

$$U_{\vartheta_\nu} = \mathsf{E}\left\{ \frac{\partial^2}{\partial \vartheta_\nu \partial \vartheta_\nu^\top}\phi_t(\vartheta_\nu) \right\}. \tag{3.64}$$

Note that $\frac{\partial}{\partial \vartheta_\nu}\phi_t(\vartheta_\nu) = (\frac{\partial}{\partial \alpha_\nu}\phi_t(\vartheta_\nu), \frac{\partial}{\partial \beta_\nu}\phi_t(\vartheta_\nu), \frac{\partial}{\partial \lambda_\nu}\phi_t(\vartheta_\nu))$ is a 3-dimensional row vector. Then, the matrix IF of the PINAR$(1,1_S)$ process is defined as the block diagonal matrix

$$IF = \mathrm{diag}\{IF_1,\ldots,IF_S\}. \tag{3.65}$$

The following theorem on the asymptotic normality of the QML-estimator $\widehat{\boldsymbol{\vartheta}}^{\mathrm{QML}}$ is given below.

Theorem 4. *Assume that* $\{\boldsymbol{Y}_t\}$ *in* (3.30)*, that is, the a PINAR*$(1,1_S)$ *process, is a strictly stationary ergodic process with* $\mathsf{E}\,\|\boldsymbol{\zeta_k}\|^6 < \infty$ ($\mathsf{E}\,\|\varepsilon_t\|^6 < \infty$ *in model* (3.26))*, and* $m_{\vartheta_\nu}(t,t-1)$ *and* $f_{\vartheta_\nu}(t,t-1)$ *are almost surely three times continuously differentiable in the open set* Θ *containing the true parameter value* $\boldsymbol{\vartheta}_0$*. Then, the* QML *estimators* $\widehat{\boldsymbol{\vartheta}}_n^{\mathrm{QML}}$ *are asymptotically normal distributed as*

$$n^{1/2}(\widehat{\boldsymbol{\vartheta}}_n^{\mathrm{QML}} - \boldsymbol{\vartheta}_0) \xrightarrow{\mathcal{D}} \mathcal{N}(\boldsymbol{0}, IF), \tag{3.66}$$

as $n \to \infty$*, where* IF *is the matrix of dimension* $3S \times 3S$ *defined by* (3.65)*.*

The proof is in the Appendix.

B .4 Monte Carlo simulations

In this section, it is evaluated the empirical behavior of the estimation methods YW, CLS and QML for the parameters in the model (3.26) with $S = 4, 7$ when the innovations $\{\varepsilon_t\}$ are a periodic sequence of Poisson or Geometric distribution. The results are shown in Tables 3.8,

3.9 and 3.10 in which the true parameters and the sample sizes are displayed. To obtain the QML estimates, the initial choice was the values given by the YW estimator, as recommended in the literature. Note that, in order to verify the model property discussed in Remarks 1 and 2, it was considered the parameters α and β such that $\alpha_\nu\beta_\nu \neq 0$ and $\alpha_\nu\beta_\nu \approx 0$ and the results are displayed in Tables 3.8 and 3.9, respectively.

The empirical quantities, the bias and the mean square error (the values are in parenthesis), correspond to the mean of over 1000 replications. All simulation studies were performed using the statistical software R (R Development Core Team, 2009). To compute the QML method, the general non-linear optimization procedure was implemented using the augmented Lagrange multiplier method with numerical derivatives available in the *solnp* function of the *Rsolnp* package.

The results are shown in Tables 3.8, 3.9 and 3.10.

As previously mentioned, Table 3.8 display the parameter estimates when $\alpha_\nu\beta_\nu \neq 0$ and it can be seen that all methods performed well in which the QML is more accurate for T=200. In general, for moderate sample sizes, the QML and CLS are more accurate. However, as increasing T all estimators perform similarly, which corroborates the theoretical results, except for the parameter λ_ν in which the QML maintains its superiority in presenting smaller mse. Note that, independently of the sample size and method used, the estimates of λ_ν are always larger than the estimates of the parameters α_ν and β_ν. This fact may be mainly due to the minimization algorithm to estimate λ_ν, which is not linearly related to the observations Y_t. In practice, however, it may not be a big concern. Other parameter values were also considered in the simulation study and, in general, the conclusions were quite similar to those reported here. These results are available upon request.

An interesting feature is observed when $\alpha_\nu\beta_\nu \approx 0$ as displayed in Table 3.9. In this context, the performance of the estimation methods are similar to the previous table, that is, QML and CLS methods presented, in general, more accurate estimates. However, although the estimates of λ_ν are less accurate than the other parameters, their MSE values are much more precise than the ones in Table 3.8 and this can be justified by the assumed property that $\alpha_\nu\beta_\nu \approx 0$, that is, the process follows an Poisson distribution as pointed out in Remarks 1 and 2. To end this, in Table 3.10 a more complex model is considered, that is, now $S = 7$. A set of Monte Carlo simulations with 400 independent replications for series with a sample size of T=1001 values, i.e, n=143 values per season. True parameters given by $\boldsymbol{\alpha} = \{031, 0.35, 0.29, 0.29, 0.37, 0.29, 0.28\}$, $\boldsymbol{\beta} = \{0.27, 0.25, 0.26, 0.39, 0.27, 0.22, 0.33\}$ and $\boldsymbol{\lambda} = \{4.0, 3.3, 2.1, 2.5, 3.1, 2.6, 3.5\}$, for the Poisson innovations simulated series and $\boldsymbol{\alpha} = \{0.41, 0.35, 0.59, 0.19, 0.37, 0.29, 0.22\}$, $\boldsymbol{\beta} = \{0.27, 0.20, 0.16, 0.39, 0.21, 0.52, 0.33\}$ and $\boldsymbol{\lambda} = \{0.50, 0.40, 0.80, 0.30, 0.70, 0.54, 0.60\}$, $(\boldsymbol{\lambda})_\nu = (1 + \lambda_\nu)^{-1}$, for Geometric. Inside parenthesis are the MSE of each estimate above.

The performance of methods are also investigated under Geometric distribution. However, the method performance preserve similar behavior compared to the more simple model $S = 4$ in Table 3.8.

Table 3.8: Results of the simulation of a PINAR$(1, 1_4)$ model with sample size of T=200, 800 and 2000 values. The true parameters given by $\boldsymbol{\alpha} = \{0.1, 0.42, 0.23, 0.39\}$, $\boldsymbol{\beta} = \{0.47, 0.25, 0.36, 0.3\}$ and $\boldsymbol{\lambda} = \{4, 3, 2, 1\}$. Inside parenthesis are the MSE of each estimate above.

	n=50, T=200			n=200, T=800			n=500, T=2000		
	Bias$_{QML}$	Bias$_{YW}$	Bias$_{CLS}$	Bias$_{QML}$	Bias$_{YW}$	Bias$_{CLS}$	Bias$_{QML}$	Bias$_{YW}$	Bias$_{CLS}$
α_1	0.025	-0.013	0.006	-0.002	-0.01	0.001	-0.004	-0.004	-0.002
	(0.018)	(0.031)	(0.034)	(0.005)	(0.007)	(0.006)	(0.003)	(0.003)	(0.002)
α_2	0.021	0.022	0.055	0.007	0.002	0.009	-0.004	-0.003	0.005
	(0.014)	(0.021)	(0.020)	(0.004)	(0.005)	(0.004)	(0.002)	(0.002)	(0.002)
α_3	0.009	0.001	-0.005	0.002	-0.001	0.000	0.001	-0.001	0.000
	(0.013)	(0.014)	(0.023)	(0.003)	(0.003)	(0.004)	(0.001)	(0.001)	(0.002)
α_4	0.004	-0.001	0.036	0.006	0.001	0.004	0.000	0.000	0.006
	(0.010)	(0.013)	(0.018)	(0.002)	(0.003)	(0.004)	(0.001)	(0.001)	(0.002)
β_1	-0.028	-0.062	0.000	-0.008	-0.017	-0.001	0.002	-0.004	-0.001
	(0.015)	(0.021)	(0.013)	(0.003)	(0.005)	(0.003)	(0.001)	(0.002)	(0.001)
β_2	-0.024	-0.04	0.038	-0.007	-0.009	0.014	-0.005	-0.006	0.005
	(0.017)	(0.018)	(0.020)	(0.004)	(0.004)	(0.004)	(0.002)	(0.002)	(0.002)
β_3	-0.035	-0.058	-0.005	-0.006	-0.017	0.003	-0.003	-0.007	0.001
	(0.017)	(0.021)	(0.013)	(0.004)	(0.005)	(0.003)	(0.002)	(0.002)	(0.001)
β_4	-0.011	-0.024	0.044	-0.005	-0.012	0.008	-0.003	-0.005	0.002
	(0.015)	(0.017)	(0.017)	(0.004)	(0.004)	(0.004)	(0.002)	(0.002)	(0.002)
λ_1	0.081	0.548	-0.499	0.085	0.197	-0.087	-0.008	0.040	-0.034
	(1.324)	(2.312)	(2.114)	(0.278)	(0.448)	(0.365)	(0.151)	(0.215)	(0.161)
λ_2	0.003	0.017	-0.267	0.017	-0.045	-0.042	0.068	-0.049	-0.051
	(1.427)	(2.341)	(2.518)	(0.342)	(0.512)	(0.487)	(0.157)	(0.218)	(0.188)
λ_3	0.11	0.885	-0.225	0.005	0.671	-0.074	0.015	0.617	-0.018
	(1.16)	(1.976)	(1.708)	(0.208)	(0.663)	(0.354)	(0.091)	(0.488)	(0.133)
λ_4	0.058	0.644	-0.174	-0.02	0.569	-0.045	0.01	0.545	-0.015
	(0.455)	(1.052)	(0.754)	(0.096)	(0.449)	(0.168)	(0.042)	(0.359)	(0.066)

Table 3.9: Results of the simulation of a PINAR$(1, 1_4)$ model with sample size of T=200, 800 and 2000 values. The true parameters given by $\boldsymbol{\alpha} = \{0.0, 0.42, 0.0, 0.39\}$, $\boldsymbol{\beta} = \{0.47, 0.0, 0.36, 0.0\}$ and $\boldsymbol{\lambda} = \{4, 3, 2, 1\}$. Inside parenthesis are the MSE of each estimate above.

	n=50, T=200			n=200, T=800			n=500, T=2000		
	Bias$_{QML}$	Bias$_{YW}$	Bias$_{CLS}$	Bias$_{QML}$	Bias$_{YW}$	Bias$_{CLS}$	Bias$_{QML}$	Bias$_{YW}$	Bias$_{CLS}$
α_1	0.104	-0.001	0.021	0.047	-0.006	-0.003	0.029	-0.004	0.002
	(0.032)	(0.063)	(0.063)	(0.007)	(0.016)	(0.013)	(0.003)	(0.006)	(0.006)
α_2	-0.018	-0.017	0.049	-0.004	-0.001	0.010	-0.004	0.000	0.004
	(0.010)	(0.014)	(0.020)	(0.002)	(0.003)	(0.004)	(0.001)	(0.001)	(0.002)
α_3	0.048	0.002	-0.007	0.022	0.007	-0.004	0.013	0.001	-0.003
	(0.006)	(0.012)	(0.016)	(0.001)	(0.002)	(0.004)	(0.001)	(0.001)	(0.002)
α_4	-0.008	-0.001	0.026	0.005	0.007	0.013	-0.001	-0.002	0.002
	(0.008)	(0.012)	(0.018)	(0.002)	(0.004)	(0.004)	(0.001)	(0.001)	(0.002)
β_1	-0.010	-0.045	0.007	-0.004	-0.015	0.001	-0.005	-0.009	0.001
	(0.013)	(0.019)	(0.009)	(0.003)	(0.004)	(0.002)	(0.001)	(0.002)	(0.001)
β_2	0.042	-0.023	0.057	0.025	-0.006	0.009	0.017	-0.001	0.007
	(0.006)	(0.015)	(0.021)	(0.003)	(0.005)	(0.004)	(0.001)	(0.002)	(0.002)
β_3	-0.008	-0.038	-0.004	-0.011	-0.017	-0.002	-0.001	-0.003	0.000
	(0.016)	(0.020)	(0.016)	(0.004)	(0.005)	(0.003)	(0.001)	(0.002)	(0.001)
β_4	0.045	-0.017	0.026	0.022	-0.010	0.007	0.015	-0.005	0.001
	(0.008)	(0.017)	(0.015)	(0.002)	(0.005)	(0.004)	(0.001)	(0.002)	(0.002)
λ_1	-0.107	0.385	-0.375	-0.053	0.147	-0.067	-0.016	0.084	-0.034
	(1.051)	(1.693)	(1.491)	(0.185)	(0.341)	(0.298)	(0.081)	(0.160)	(0.131)
λ_2	-0.123	0.844	-0.119	-0.114	0.630	-0.047	-0.073	0.589	0.019
	(0.676)	(1.634)	(1.279)	(0.179)	(0.647)	(0.305)	(0.088)	(0.459)	(0.129)
λ_3	-0.259	0.127	-0.231	-0.117	0.018	-0.018	-0.088	-0.002	-0.020
	(0.363)	(0.402)	(0.624)	(0.088)	(0.075)	(0.137)	(0.036)	(0.030)	(0.056)
λ_4	-0.054	0.411	-0.030	-0.066	0.353	-0.013	-0.034	0.363	-0.005
	(0.115)	(0.300)	(0.220)	(0.030)	(0.162)	(0.046)	(0.010)	(0.147)	(0.015)

Table 3.10: Results of the simulation of a PINAR$(1, 1_7)$ model with sample size of T=1001 values, i.e, n=143 values per season.

	Poisson Innovations			Geometric Innovations		
Pars	Bias$_{QML}$	Bias$_{YW}$	Bias$_{CLS}$	Bias$_{QML}$	Bias$_{YW}$	Bias$_{CLS}$
α_1	-0.006	0.007	0.002	0.014	0.031	-0.001
	(0.006)	(0.007)	(0.007)	(0.005)	(0.007)	(0.009)
α_2	0.000	0.004	0.012	0.010	0.006	0.011
	(0.004)	(0.005)	(0.006)	(0.003)	(0.005)	(0.007)
α_3	-0.004	-0.001	0.003	0.004	0.000	-0.005
	(0.003)	(0.004)	(0.006)	(0.000)	(0.001)	(0.010)
α_4	0.000	0.001	0.016	0.022	0.034	0.013
	(0.005)	(0.006)	(0.006)	(0.005)	(0.014)	(0.006)
α_5	-0.009	0.004	0.001	0.004	0.001	0.004
	(0.005)	(0.006)	(0.005)	(0.001)	(0.002)	(0.003)
α_6	0.000	0.005	0.015	0.012	0.012	0.003
	(0.004)	(0.005)	(0.006)	(0.004)	(0.005)	(0.003)
α_7	0.001	-0.001	-0.006	0.005	0.000	-0.008
	(0.006)	(0.008)	(0.006)	(0.002)	(0.004)	(0.017)
β_1	0.004	-0.007	0.015	-0.001	0.010	0.023
	(0.005)	(0.005)	(0.006)	(0.000)	(0.006)	(0.007)
β_2	0.000	-0.012	0.002	-0.012	-0.024	0.001
	(0.006)	(0.005)	(0.006)	(0.001)	(0.003)	(0.002)
β_3	0.013	-0.013	0.011	0.000	0.006	0.009
	(0.005)	(0.005)	(0.006)	(0.002)	(0.003)	(0.004)
β_4	0.008	-0.014	-0.005	0.004	-0.010	0.002
	(0.003)	(0.005)	(0.005)	(0.002)	(0.004)	(0.006)
β_5	0.013	-0.018	0.009	0.006	0.000	0.018
	(0.005)	(0.005)	(0.006)	(0.002)	(0.004)	(0.005)
β_6	0.008	-0.008	-0.002	-0.011	-0.029	-0.003
	(0.006)	(0.006)	(0.007)	(0.002)	(0.005)	(0.004)
β_7	0.006	-0.014	0.008	-0.014	-0.017	0.010
	(0.004)	(0.006)	(0.006)	(0.006)	(0.006)	(0.007)
λ_1	0.011	-0.012	-0.130	0.012	0.033	-0.005
	(0.607)	(0.754)	(0.926)	(0.001)	(0.006)	(0.005)
λ_2	0.016	0.149	-0.174	0.002	-0.002	-0.005
	(0.533)	(0.697)	(0.784)	(0.001)	(0.002)	(0.004)
λ_3	-0.046	0.741	-0.104	0.012	0.017	-0.006
	(0.325)	(0.897)	(0.514)	(0.005)	(0.006)	(0.014)
λ_4	-0.052	-0.178	-0.070	0.010	0.011	0.002
	(0.303)	(0.580)	(0.475)	(0.000)	(0.002)	(0.002)
λ_5	-0.066	-0.190	-0.082	0.022	0.011	-0.009
	(0.407)	(0.670)	(0.653)	(0.005)	(0.009)	(0.016)
λ_6	-0.044	0.466	-0.017	-0.005	-0.014	0.005
	(0.375)	(0.619)	(0.499)	(0.001)	(0.005)	(0.008)
λ_7	-0.047	-0.343	-0.053	0.010	0.009	-0.011
	(0.373)	(0.938)	(0.579)	(0.002)	(0.007)	(0.011)

Table 3.11: Periodic ACF of the real data set.

	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=7$	$h=8$	$h=9$	$h=10$
$\nu=1$	0.12	-0.03	0.06	-0.11	-0.06	-0.03	**0.27**	0.06	0.01	-0.01
$\nu=2$	**0.25**	0.07	-0.04	-0.11	0.03	-0.01	0.11	0.07	-0.09	0.06
$\nu=3$	**0.35**	**0.22**	0.05	0.04	-0.07	0.12	**0.17**	**0.17**	0.12	-0.04
$\nu=4$	**0.20**	0.14	-0.09	0.03	-0.07	0.08	**0.38**	0.05	0.04	-0.01
$\nu=5$	**0.31**	0.02	0.01	0.03	0.08	**−0.26**	**0.32**	0.07	-0.02	-0.01
$\nu=6$	-0.06	0.02	-0.10	0.01	-0.01	0.05	**0.35**	-0.02	-0.01	-0.08
$\nu=7$	**0.46**	-0.12	-0.10	-0.13	-0.01	-0.03	0.06	**0.17**	-0.08	**−0.17**

B .5 Real data application

This application is based on the time series of counts referring to the daily number of visits of children with respiratory problems (International Classification of Diseases ICD-10) in the emergency service of the public health care system of the region of Vitória-ES. This data set was obtained from the network records system Welfare (*Rede Bem-Estar*) of the municipality. The period of the study corresponds to June 26, 2013 to April 7, 2016, resulting in 1022 daily observations. Figure 3.3 displays the plot of the real data which clearly shows that the persistence oscillation feature, that is, the mean changes periodically. This phenomenon is also clearly evidenced in the plots of Figure 3.4. The series correspond to daily data, therefore, S= 7. All the data analysis procedure was carried out using the R software.

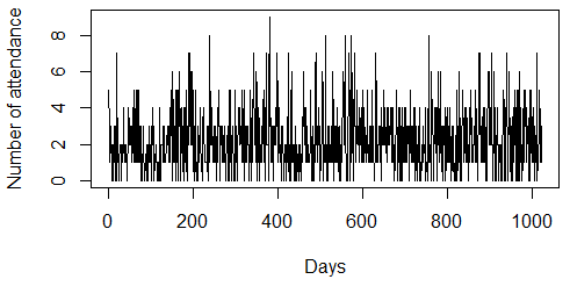


Figure 3.3: Daily number of visits of children with respiratory problems to emergency service service of the public health care system of the region of Vitória-ES.

Figure 3.4 shows the sample periodic mean (a), the variance (b), the sample ACF (c) and the periodogram (d). These clearly display the cycle-periodicity of the data with period S=7. For example, the periodogram has a high peak at the frequency f=0.14 which is related to the period=$1/0.14 = 7$.

Tables 3.11 and 3.12 show the PeACF ad PePACF sample functions. The elements in bold represent values that have exceeded the confidence interval. In addition, according to McLeod (1994) one can identify the AR order for each season by finding the lowest lag for which the sample PePACF cuts off. All these suggested the use of PINAR$(1, 1_7)$ to model the real data set.

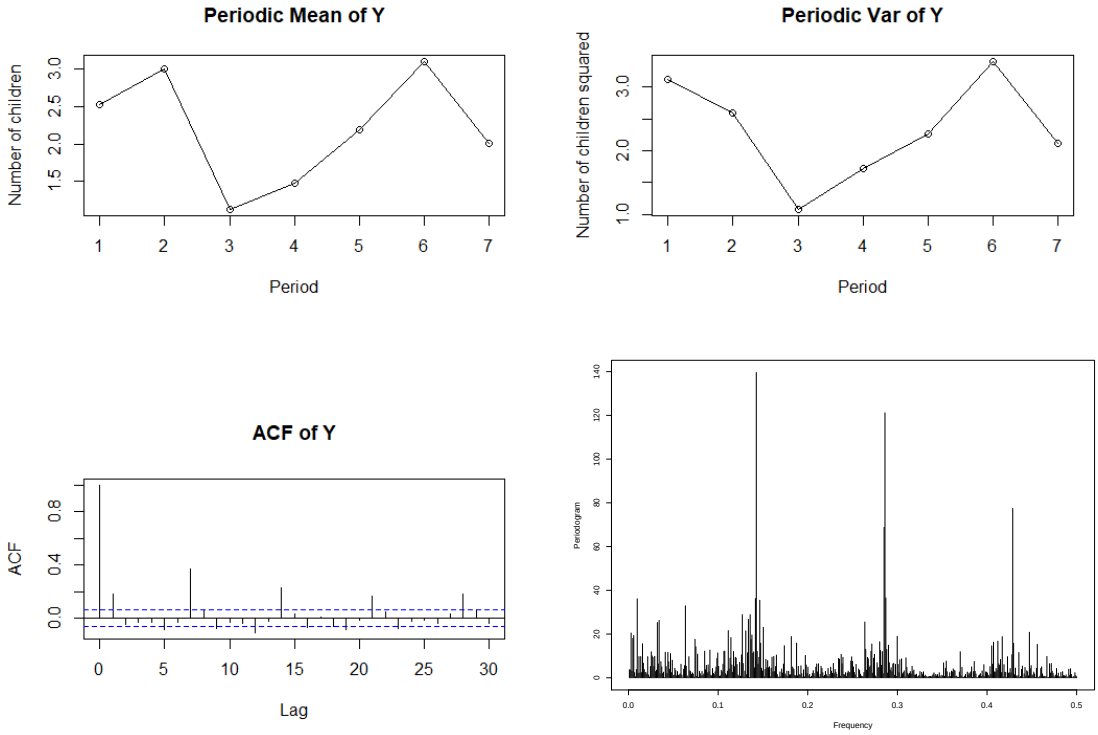


Figure 3.4: The sample periodic mean and variance over the seasons $\nu = 1, \cdots, 7$, the sample ACF and the periodogram of $\{Y_t\}$.

Table 3.12: Periodic PACF of the real data set.

	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=7$	$h=8$	$h=9$	$h=10$
$\nu=1$	0.12	-0.10	0.07	-0.12	-0.03	-0.00	**0.29**	0.03	0.06	-0.06
$\nu=2$	**0.25**	0.04	-0.06	-0.12	0.11	-0.02	0.13	-0.01	-0.10	0.13
$\nu=3$	**0.35**	0.14	0.01	0.07	-0.05	0.16	**0.18**	0.09	0.01	-0.05
$\nu=4$	**0.20**	0.07	-0.15	0.04	-0.12	0.13	**0.35**	-0.05	-0.04	0.07
$\nu=5$	**0.31**	-0.04	-0.02	0.08	0.07	**−0.32**	**0.34**	-0.11	-0.03	0.01
$\nu=6$	-0.06	0.04	-0.11	0.04	0.02	0.06	**0.39**	-0.05	-0.04	-0.02
$\nu=7$	**0.46**	-0.11	-0.09	-0.08	0.02	-0.02	0.05	-0.04	-0.05	-0.14

The parameters were estimated using QML method with the innovations distributed as Poisson and Geometric distributions. The results are displayed in Table 3.13. The standard errors of the parameter estimates were calculated from the inverse of the corresponding Hessian matrix. The adequacy of the adjusted model was evaluated by examining the residuals for serial dependency. The estimated residuals $\{r_t\}$ after fitting the PINAR$(1, 1_7)$ model were computed as

$$r_t = y_t - \hat{Y}_t = y_t - \widehat{\alpha}_\nu y_{t-1} - \widehat{\beta}_\nu y_{t-7} - \widehat{\lambda}_\nu, \tag{3.67}$$

where $\nu = \{t\}_7$. The Akaike (AIC) (see Bozdogan (1987)) and the Bayesian information criterion (BIC) (see Schwarz (1978)) were computed and the their values are also in the together with the parameter estimates in Table 3.13.

The PINAR$(1, 1_7)$ model with Poisson innovations presented to be more accurate than the Geometric models, i.e., both information criteria suggest that the Poisson-PINAR$(1, 1_7)$ has a

Table 3.13: Application of PINAR$(1, 1_7)$ model to the real data. The parameters were estimated by QML method. Inside parenthesis are the standard errors of the estimates.

Fitted model	ν=1	ν=2	ν=3	ν=4	ν=5	ν=6	ν=7	AIC	BIC
PINAR$(1, 1_7)$-Poisson Innovation:									
α_ν	0.173(0.093)	0.270(0.082)	0.217(0.044)	0.199(0.087)	0.318(0.089)	0.036(0.093)	0.323(0.052)		
β_ν	0.250(0.067)	0.160(0.093)	0.151(0.078)	0.327(0.069)	0.284(0.072)	0.332(0.069)	0.007(0.086)	7773.75	850.64
λ_ν	1.530(0.248)	1.848(0.353)	0.293(0.138)	0.774(0.137)	1.098(0.191)	1.987(0.341)	0.982(0.225)		
PINAR$(1, 1_7)$- Geometric innovation:									
α_ν	0.266(0.455)	0.393(0.026)	0.234(0.008)	0.260(0.307)	0.421(0.676)	0.236(0.070)	0.374(0.092)		
β_ν	0.347(0.185)	0.358(0.020)	0.166(0.120)	0.368(0.065)	0.352(0.335)	0.462(0.004)	0.115(0.702)	7800.20	877.10
$(1 + \lambda_\nu)^{-1}$	0.476(0.003)	0.514(0.003)	0.816(0.029)	0.608(0.009)	0.556(0.002)	0.466(0.006)	0.623(0.059)		

Table 3.14: Periodic ACF of residuals after fitting the PINAR$(1, 1_7)$ model with Poisson distribution of innovations to the real data. The parameters were estimated by QML estimation method.

	$h = 1$	$h = 2$	$h = 3$	$h = 4$	$h = 5$	$h = 6$	$h = 7$	$h = 8$	$h = 9$	$h = 10$
$\nu = 1$	0.04	-0.11	0.11	-0.09	-0.03	-0.11	0.04	0.03	0.03	-0.04
$\nu = 2$	-0.03	0.06	-0.05	-0.17	0.02	-0.04	-0.03	-0.04	-0.17	0.03
$\nu = 3$	-0.05	0.12	0.02	0.10	-0.05	0.05	0.01	0.05	0.09	0.00
$\nu = 4$	-0.01	0.12	-0.11	0.12	-0.06	-0.02	0.06	-0.02	-0.04	0.04
$\nu = 5$	0.05	-0.01	-0.01	0.03	$\mathbf{0.19}$	$\mathbf{-0.18}$	0.06	-0.08	-0.04	-0.02
$\nu = 6$	-0.01	0.07	-0.14	0.04	0.03	-0.14	0.03	-0.01	-0.03	-0.03
$\nu = 7$	0.06	-0.06	-0.05	-0.11	-0.01	-0.02	0.02	0.04	-0.00	$\mathbf{-0.18}$

better fit. This is corroborate with the residual PACFs displayed in Tables 3.14 and 3.15 for the marginal Poisson and Geometric distributions, respectively. It is clear that the model with Poisson distributed innovations was able to filter the autocorrelations of the data, especially at the lags 1 and 7, while the Geometric innovation does not. Therefore, using PINAR$(1, 1_7)$ with Poisson innovations no systematic pattern is clearly observed in the residuals, that is, the fitted model seems to well capture the main dynamics of the data, consequently, the estimated model can be very useful in providing reliable forecast.

All these empirical analyses, i.e., the values for the periodic ACF of the residuals (Table 3.14 and 3.15) and the results in Table 3.13 support the fact that the proposed model with Poisson distributed innovations is the best choice for modeling such data.

Table 3.15: Periodic ACF of residuals after fitting the PINAR$(1, 1_7)$ model with Geometric distribution of innovations to the real data. The parameters were estimated by QML estimation method.

	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=7$	$h=8$	$h=9$	$h=10$
$\nu=1$	-0.02	-0.15	0.11	-0.08	-0.03	-0.09	-0.05	0.01	0.03	-0.06
$\nu=2$	-0.13	0.06	-0.04	$\mathbf{-0.18}$	0.02	-0.03	$\mathbf{-0.18}$	-0.11	$\mathbf{-0.20}$	0.00
$\nu=3$	-0.10	0.11	0.02	0.11	-0.05	0.05	-0.01	0.01	0.08	0.01
$\nu=4$	-0.07	0.12	-0.12	0.14	-0.05	-0.01	0.02	-0.03	-0.04	0.05
$\nu=5$	-0.03	-0.02	-0.01	0.02	$\mathbf{0.26}$	$\mathbf{-0.25}$	0.02	-0.13	-0.04	-0.02
$\nu=6$	-0.13	0.04	-0.15	0.06	0.06	$\mathbf{-0.18}$	-0.04	-0.05	-0.06	-0.01
$\nu=7$	0.03	-0.05	-0.04	-0.10	-0.01	0.00	-0.10	-0.03	0.04	-0.17

B .6 Forecasting

The forecasting procedures here discussed were initially proposed by Du & Li (1991) and Freeland & McCabe (2004). The innovations are considered to have Poisson distribution. The interest is to forecast Y_{T+h}, $T+h = nS+\nu$, $n, h, S \in \mathbb{N}$ and $\nu = 1, \ldots, S$, given $\mathcal{F}_T$ the known information about $\{Y_t\}$ for $t = 1, \ldots, T$.
The first method is an extension of the approach presented in Section 5 of Du & Li (1991). The minimum variance predictor $Y_T(1)$ of Y_{T+1}, $h = 1$, based on the conditional mean, is given by

$$\hat{Y_T}(1) = \mathsf{E}(Y_{T+1}|\mathcal{F}_T) = \alpha_\nu Y_{nS+\nu} + \beta_\nu Y_{nS+\nu-S+1} + \lambda_\nu.$$

For $h > 1$, Y_{T+h} can be calculated as

$$\hat{Y_T}(h) = \mathsf{E}(Y_{T+h}|\mathcal{F}_T) = \mathsf{E}(\alpha_\nu \circ Y_{T+h-1}|\mathcal{F}_T) + \mathsf{E}(\beta_\nu \circ Y_{nS+\nu-S+1}|\mathcal{F}_T) + \lambda_{\nu+h},$$

were $\mathsf{E}(\alpha_\nu \circ Y_{T+h-1}|\mathcal{F}_T)$ can be obtained by

$$\mathsf{E}(\alpha_\nu \circ Y_{T+h-1}|\mathcal{F}_T) = \mathsf{E}(\mathsf{E}(\alpha_\nu \circ Y_{T+h-1}|Y_{T+h-1})|\mathcal{F}_T) = \alpha_\nu \mathsf{E}(Y_{T+h-1}|\mathcal{F}_T) = \alpha_\nu \hat{Y_T}(h-1).$$

Similarly $\mathsf{E}(\beta_\nu \circ Y_{nS+\nu-S+1}|\mathcal{F}_T) = \beta_\nu \hat{Y_T}(h-S)$. If $h - S \leq 0$, then $\hat{Y_T}(h-S) = Y_{T+S-h}$. So, the recursive formula of Y_{T+h} can be written as

$$\hat{Y_T}(h) = \mathsf{E}(Y_{T+h}|\mathcal{F}_T) = \alpha_\nu \hat{Y_T}(h-1) + \beta_\nu \hat{Y_T}(h-S) + \lambda_{\nu+h}. \tag{3.68}$$

The expression above is the most common procedure to obtain forecasts in time series models and, as pointed out by Du & Li (1991), offers minimum variance predictor and mean square forecast error. Besides, the use of conditional expectations is more easily implemented and less complicated than forecasting the distribution of Y_{T+h} to calculate its conditional median. However, the method in (3.68) does not produce integer-valued forecasts, usually signalized as the great incoherence of the count data models context. The following forecasting procedure is an approach based on extension of the method present in Moriña et al. (2011), which is based on the Section 4 in Freeland & McCabe (2004).

Given $Y_1, \ldots, Y_T$, the distribution of Y_{T+h}, $h \in \mathbb{N}$, is based on (3.26). For $h = 1$, it is given by (3.40). When $h > 1$, the distribution of Y_{T+h} can be written as a function of the S last known values of the process $\{Y_t\}_{t=1,\ldots,T}$, i.e., $Y_{t-S+1}, \ldots, Y_T$, for $T = nS$, $n, S \in \mathbb{N}$. For example, based on the properties of binomial thinning operator and the definition 3, for $h = 2$, Y_{T+2} is given by

$$Y_{T+2} = \alpha_2\alpha_1 \circ y_T + \beta_2 \circ y_{T-S+2} + \alpha_2\beta_1 \circ y_{T-S+1} + \alpha_2 \circ \varepsilon_{T+1} + \varepsilon_{T+2}.$$

Note that $(\alpha_2 \circ \varepsilon_{T+1}) \sim \mathrm{Poi}(\alpha_2\lambda_1)$, see Lemma 3. Then $\alpha_2 \circ \varepsilon_{T+1} + \varepsilon_{T+2} \sim \mathrm{Poi}(\lambda_{T+2})$, where $\lambda_{T+2} = \alpha_2\lambda_1 + \lambda_2$. The distribution of Y_{T+2} can be obtained analogously to (3.40), and is given

by

$$p_T(Y_{T+2}|Y_T = y_T, Y_{T-S+1} = y_{T-S+1}, Y_{T-S+2} = y_{T-S+2}) =$$
$$[\mathrm{Bin}(y_T, \alpha_2\alpha_1) * \mathrm{Bin}(y_{T-S+2}, \beta_2) * \mathrm{Bin}(y_{T-S+1}, \alpha_2\beta_1) * \mathrm{Poi}(\lambda_{T+2})](Y_{T+2})$$
$$= \sum_{(m,n,r)\in\mathcal{J}} \binom{y_T}{m}(\alpha_2\alpha_1)^m(1-(\alpha_2\alpha_1))^{y_T-m}\binom{y_{T-S+2}}{n}\beta_2^n(1-\beta_2)^{y_{T-S+2}-n}\binom{y_{T-S+1}}{r}(\alpha_2\beta_1)^r$$
$$(1-\alpha_2\beta_1)^{y_{T-S+1}-r}\frac{\lambda_{T+2}^{Y_{T+2}-m-n-r}}{(Y_{T+2}-m-n-r)!}\mathrm{e}^{-\lambda_{T+2}},$$

where $*$ denotes convolution and the index set $\mathcal{J}$ is defined by $\mathcal{J} = \{(m,n,r) \in \mathbb{Z}_+^3 | m \leq y_T, n \leq y_{T-S+2}, r \leq y_{T-S+1}, m+n+r \leq y_{T+2}\}$. Analogously, the distribution of Y_{T+h} for any h can be obtained. The approximated prediction regions with size $1-c$ can be calculated for each Y_{T+h} by replacing the estimates of the parameters in the recursive equations of the distribution of Y_{T+h}. The lower limit b_1 and upper limit b_2 can be obtained by

$$\sum_{i=0}^{b_1} p_T(Y_{T+h} = i|Y_T, \ldots, Y_{T-S+1}) \approx c/2, \sum_{i=b_1}^{b_2} p_T(Y_{T+h} = i|Y_T, \ldots, Y_{T-S+1}) \approx 1-c. \quad (3.69)$$

The approximated median for each Y_{T+h} is the value M such that

$$\sum_{i=0}^{M} p_T(Y_{T+h} = i|Y_T, \ldots, Y_{T-S+1}) \approx 0.5. \quad (3.70)$$

Figure 3.5 shows the prediction of the last week of the real data. The asymptotic stability of the model enables to obtain a prediction for any day, without using the last observations. For a certain confidence level, the results presented at Figure 3.5 show accuracy, since, in one week ahead prediction, only one predicted value is outside the prediction region.

B .7 Conclusions

The PINAR$(1, 1_S)$ model and its main properties were introduced in this paper. Three methods for estimating the parameters of the model, namely YW, CLS and QML, were proposed. The asymptotic properties of the estimators were fully provided. A simulation study was carried out to investigate their finite sample performances for standard sample sizes. The results corroborated the asymptotic theory. The QML outperformed the CLS and YW methods for a small sample size ($n = 50$). However, they becomes very competitive for $n \geq 200$. To illustrate the usefulness of the proposed model, it was analyzed the time series of counts of the daily number of visits of children with respiratory problems (International Classification of Diseases ICD-10) in the emergency service of the public health care system of the region of Vitória-ES. This data set displays seasonal and periodic serial correlation structures. The real data set was fitted using the PINAR$(1, 1_7)$ model under Poisson (Poisson-PINAR$(1, 1_7)$ model) and Geometric innovations. The adequacy of the fitted models were compared using goodness-of-fit statistics

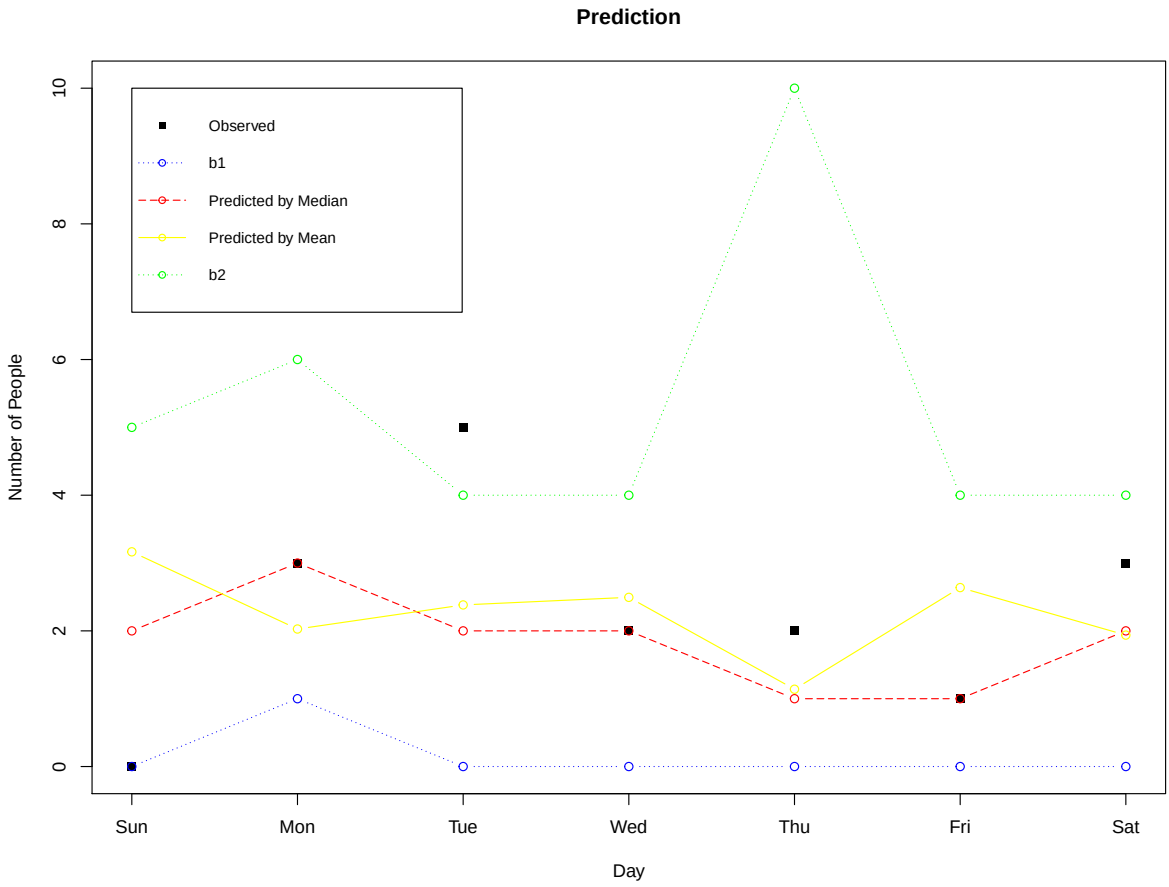


Figure 3.5: Prediction

AIC and BIC among other residual analyses. The two methods displayed good fit. However, the Poisson-PINAR$(1, 1_7)$ presented to be more accurate. In this context, the Poisson-PINAR$(1, 1_7)$ is the best choice to model this data set.

Appendix

Proof of (3.38)**.** Apply the definition of conditional probability and then the S-step Markov property:

$$\begin{aligned}
P(Y_t = y_t, \ldots, &Y_{S+1} = y_{S+1} | Y_s = y_s, \ldots, Y_1 = y_1) = \\
&= \frac{P(Y_t = y_t, \ldots, Y_1 = y_1)}{P(Y_{t-1} = y_{t-1}, \ldots, Y_1 = y_1)} \cdot \frac{P(Y_{t-1} = y_{t-1}, \ldots, Y_1 = y_1)}{P(Y_S = y_S, \ldots, Y_1 = y_1)} \\
&= P(Y_t = y_t | Y_{t-1} = y_{t-1}, \ldots, Y_1 = y_1) \times \\
&\quad P(Y_{t-1} = y_{t-1}, \ldots, Y_{S+1} = y_{S+1} | Y_S = y_S, \ldots, Y_1 = y_1) = \\
&= p_\nu(y_t | y_{t-1}, y_{t-s}) P(Y_{t-1} = y_{t-1}, \ldots, Y_{S+1} = y_{S+1} | Y_S = y_S, \ldots, Y_1 = y_1),
\end{aligned}$$

where $t = kS + \nu$, $\nu = 1, \ldots, S$ and $y_1, \ldots, y_t \in \mathbb{Z}_+$. □

Proof. of Lemma 3

Let M_Y be a positive discrete random variable such as

$$M_Y = \alpha \circ Y = \sum_{l=1}^{Y} U_l(\alpha),$$

where $U_k(\alpha)$, α and Y are defined in Lemma 3. Let $\phi_M(Z)$ be the characteristic function of M_Y. Then,

$$\phi_M(Z) = \mathsf{E}(e^{\iota Z M_Y}) = \mathsf{E}[e^{\iota Z(\sum_{l=1}^{Y} U_l(\alpha))}] = \prod_{l=1}^{Y} \mathsf{E}[e^{\iota Z(U_l(\alpha))}] = [\mathsf{E}(e^{\iota Z U_1(\alpha)})]^Y.$$

Then,

$$\phi_M(Z) = \mathsf{E}([\mathsf{E}(e^{\iota Z U_1(\alpha)})]^Y | Y).$$

Since $\mathsf{E}(e^{\iota Z U_1(\alpha)})$ =$P(U_1 = 1)e^{\iota Z 1} + P(U_1 = 0)e^{\iota Z 0}$ =$1 - \alpha + \alpha e^{\iota Z}$, then

$$\phi_M(Z) = \mathsf{E}[(1 - \alpha + \alpha e^{\iota Z})^Y].$$

Let $1 - \alpha + \alpha e^{\iota Z} = e^{\iota H}$, then:

$$\phi_M(Z) = \mathsf{E}[(e^{\iota H})^Y] = \phi_Y(H) = e^{\theta(H-1)} = e^{\alpha\theta(e^{\iota Z}-1)},$$

since Y $\sim$ Poi(θ). Based on Lemma 2.15 in Van der Vaart (2000), $M_Y \sim$ Poi$(\alpha\theta)$.

□

Proof. of Theorem 2

In (3.30), for $k = 1$, (3.30) becomes $\boldsymbol{Y}_1 = A \circ \boldsymbol{Y}_0 + \boldsymbol{\zeta_1}$ where $\boldsymbol{Y}_0 = \boldsymbol{\varepsilon_0}$ (see, Latour (1997) Eq. 3.2). The matricial equation of $\boldsymbol{Y}_1 = (Y_{S+1}. \ldots, Y_{2S})^\top$ results to the system

$$\begin{cases} Y_{S+1} = \alpha_1 \circ Y_S + \beta_1 \circ Y_1 + \varepsilon_{S+1} \\ Y_{S+2} = \alpha_2\alpha_1 \circ Y_S + \beta_2 \circ Y_2 + \alpha_2\beta_1 \circ Y_1 + \alpha_2 \circ \varepsilon_{S+1} + \varepsilon_{S+2} \\ \vdots \\ Y_{2S-1} = \prod_{i=1}^{S-1} \alpha_i \circ Y_S + \beta_{S-1} \circ Y_{S-1} + \ldots + \varepsilon_{S-1} \\ Y_{2S} = (\prod_{i=1}^{S} \alpha_i \oplus \beta_S) \circ Y_S + \alpha_S\beta_{S-1} \circ Y_{S-1} + \ldots + \varepsilon_{2S}, \end{cases} \tag{3.71}$$

where, by Definition 3,

$$(\prod_{i=1}^{S} \alpha_i \oplus \beta_S) \circ Y_S = (\prod_{i=1}^{S})\alpha_i \circ Y_S + \beta_S \circ Y_S. \tag{3.72}$$

For $t > 2S$, the general form of Y_t is algebraically written as

$$
\begin{aligned}
Y_t = &\left(\prod_{i=S+1}^{t} \alpha(i) \oplus \prod_{i=2S+1}^{t} \alpha(i)^{0(i=2S+1)}\beta_1 \oplus \ldots \oplus \prod_{i=2S+1}^{t} \alpha(i)\beta_S \oplus \prod_{i=3S+1}^{t} \alpha(i)(\beta_1)^2 \oplus \right. \\
&\left. \prod_{i=3S+1}^{t} \alpha(i)\beta_1\beta_2 \oplus \ldots \right) \circ Y_S + \left(\prod_{i=2S}^{t} \alpha(i)\beta_2 \oplus \prod_{i=3S}^{t} \alpha(i)(\beta_2)^2 \oplus \prod_{i=3S}^{t} \alpha(i)\beta_2\beta_3 \oplus \ldots \right) \circ Y_{S-1} \\
&+ \ldots + \left(\prod_{i=S+1}^{t} \alpha(i)^{0(i=S+1)}\beta_1 \oplus \prod_{i=2S+2}^{t} \alpha(i)(\beta_1)^2 \oplus \prod_{i=2S+2}^{t} \alpha(i)\beta_1\beta_2 \oplus \ldots \right) \circ Y_1 + \\
&\left(\prod_{i=S+2}^{t} \alpha(i) \oplus \prod_{i=2S+2}^{t} \alpha(i)\beta_1 \oplus \prod_{i=2S+2}^{t} \alpha(i)\beta_2 \oplus \ldots \right) \circ \varepsilon_{S+1} + \ldots + \alpha(t) \circ \varepsilon_{t-1} + \varepsilon_t, \qquad (3.73)
\end{aligned}
$$

where $\alpha(t) = \alpha_\nu$, for $t = kS + \nu$, $k \in \mathbb{Z}$, $\nu = 1, \ldots, S \in \mathbb{N}$, $0(i = x) = 0$ if $i = x$ and $0(i = x) = 1$ if $i \neq x$. (1) From Theorem 2, the starting values $Y_1, \ldots, Y_S$ follow a Poisson distribution with mean μ_ν ((3.25)), $\nu = 1, \ldots, S$, i.e., $Y_\nu \sim \text{Poi}(\mu_\nu)$ and $\varepsilon_{kS+\nu} \sim P(\lambda_\nu)$. From the assumptions of the model in (3.26), ε_t is independent of Y_{t-1}, $\alpha_\nu \circ Y_{t-1}$, Y_{t-S} and $\beta_\nu \circ Y_{t-S}$ and all counting processes are mutually independent, then, for each $\nu = 1, \ldots, s-1$, $Y_{S+\nu}$ corresponds to a sum of independent Poisson distributed variables. Therefore, based on Lemma 3, $Y_{S+\nu}$ is also a Poisson distributed variable with mean μ_ν and, for each $\nu = 1, \ldots, s-1$, μ_ν is obtained by computing the expected value of $Y_{S+\nu}$ in (3.71) using the property given in Lemma 1.

Let $t \geq 2S$. (2a) in (3.71), Y_t also involves the expansion of $(\prod_{i=1}^{S} \alpha_i \oplus \beta_S)$ in which the coefficients of Y_{l_1} and ε_{l_2} $(l_1, l_2 < t)$ may depend on the pair (α_i, β_i), $1 \leq i \leq S$. Using definition 3, $(\prod_{i=1}^{S} \alpha_i \oplus \beta_S) \circ Y_S$ is a Poisson-Binomial distributed variable. Therefore, $\{Y_t\}_{t\in\mathbb{Z}}$ follows a periodic Poisson-Binomial process with periodic mean μ_ν. (2b) Under the assumption $\alpha_i\beta_i \approx 0$, the assumptions in Lemma 3 are satisfied and, therefore, for all $\{Y_t\}_{t\in\mathbb{Z}}$, its recursive equation presents a sum of independent Poisson distributed random variables with periodic mean μ_ν. □

Proof. of Theorem 3.

This proof follows the lines in Theorem 3.2.24 in Taniguchi & Kakizawa (2000). Assume that $\{\boldsymbol{Y}_k\}_{k\in\mathbb{Z}}$, satisfies the conditions in Lemma 2, i.e., $\{\boldsymbol{Y}_k\}_{k\in\mathbb{Z}}$ is a strictly stationary and ergodic process with $\mathsf{E}||\boldsymbol{Y}_k||^2 < \infty$. (Lemma 3.3 in Latour (1997)).

Suppose that observations $(\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_n)$ are available.

By Theorem 1.3.3 in Taniguchi & Kakizawa (2000), the ergodicity of the process $\{\boldsymbol{Y}_k\}_{k\in\mathbb{Z}}$ implies the ergodicity of the process $\{Y_{kS+\nu}\}_{k\in\mathbb{Z}}$, for $\nu = 1, \ldots, S$. Since the first, second and third derivatives of $\mathsf{E}(Y_{kS+\nu}|\mathcal{F}_{kS+\nu-1})$ exist, it follows from (3.26) that it is easy evaluate the derivatives of the function $m_{\vartheta_\nu}(t, t-1)$, for $t = kS + \nu$, given by

$$
m_{\vartheta_\nu}(t, t-1) = \mathsf{E}(Y_{kS+\nu}|\mathcal{F}_{kS+\nu-1}) = \alpha_\nu Y_{kS+\nu-1} + \beta_\nu Y_{kS+\nu-S} + \lambda_\nu. \qquad (3.74)
$$

for $k = 1, \ldots, n-1$. Observe that $m_{\vartheta_\nu}(t, t-1) = \boldsymbol{U}_{t-1}^\top \vartheta_\nu$, where $\boldsymbol{U}_{t-1}$ is defined in (3.46). Note

that $m_{\vartheta_\nu}(t, t-1)$ is almost sure three times differentiable in the open set Θ which contains $\boldsymbol{\vartheta}_0$.

Let $\vartheta_\nu = (\alpha_\nu, \beta_\nu, \lambda_\nu)^\top$ and $(\vartheta_\nu)_i$ is the i-th element of the vector (ϑ_ν), i.e., $(\vartheta_\nu)_1 = \alpha_\nu, (\vartheta_\nu)_2 = \beta_\nu$ and $(\vartheta_\nu)_3 = \lambda_\nu$. To prove the Theorem 3, it has to be proved the following conditions:

(C1). For $1 \leq i, l \leq 3$, $\mathsf{E}\left\{\left|\frac{\partial}{\partial(\vartheta_\nu)_i} m_{\vartheta_\nu^0}(t, t-1)\right|^2\right\} < \infty$ and $\mathsf{E}\left\{\left|\frac{\partial^2}{\partial(\vartheta_\nu)_l \partial(\vartheta_\nu)_i} m_{\vartheta_\nu^0}(t, t-1)\right|^2\right\} < \infty$.

(C2). The functions $\frac{\partial}{\partial(\vartheta_\nu)_i} m_{\vartheta_\nu^0}(t, t-1)$, $i = 1, 2, 3$, are linearly independent in the sense that if $a_1(\nu), a_2(\nu), a_3(\nu)$ are arbitrary real numbers such that

$$\mathsf{E}\left\{\left|\sum_{i=1}^{3} a_i(\nu) \frac{\partial}{\partial(\vartheta_\nu)_i} m_{\vartheta_\nu^0}(t, t-1)\right|^2\right\} = 0,$$

then $a_1(\nu) = 0, a_2(\nu) = 0$, and $a_3(\nu) = 0$, for $\nu = 1, \ldots, S$.

(C3). For $\boldsymbol{\vartheta} \in \Theta$, there exists functions $G^{ijl}_{kS+\nu-1}(Y_{0S+\nu}, \ldots, Y_{kS+\nu-1})$ and $H^{ijl}_{kS+\nu}(Y_{0S+\nu}, \ldots, Y_{kS+\nu})$, $k \in \mathbb{Z}$, for each $\nu = 1, \ldots, S$, such that

$$\left|\frac{\partial}{\partial(\vartheta_\nu)_j} m_{\vartheta_\nu}(t, t-1) \frac{\partial^2}{\partial(\vartheta_\nu)_i \partial(\vartheta_\nu)_l} m_{\vartheta_\nu}(t, t-1)\right| \leq G^{ijl}_{kS+\nu-1}, \tag{3.75}$$

with $\mathsf{E}(G^{ijl}_{kS+\nu-1}) < \infty$ and

$$\left|\{Y_t - m_{\vartheta_\nu}(t, t-1)\} \frac{\partial^3}{\partial(\vartheta_\nu)_i \partial(\vartheta_\nu)_j \partial(\vartheta_\nu)_l} m_{\vartheta_\nu}(t, t-1)\right| \leq H^{ijl}_{kS+\nu}, \tag{3.76}$$

with $\mathsf{E}(H^{ijl}_{kS+\nu}) < \infty$ for $i, j, l = 1, 2, 3$.

(C4).

$$R_\nu = \mathsf{E}\left[\frac{\partial}{\partial \vartheta_\nu} m_{\vartheta_\nu^0}(t, t-1)^\top \{Y_t - m_{\vartheta_\nu^0}(t, t-1)\}\{Y_t - m_{\vartheta_\nu^0}(t, t-1)\} \frac{\partial}{\partial \vartheta_\nu} m_{\vartheta_\nu^0}(t, t-1)\right] < \infty.$$

Proof of the above conditions:

(C1)

For $t = kS + \nu$, the derivative $\frac{\partial}{\partial(\vartheta_\nu)_i} m_{\vartheta_\nu}(t, t-1)$ is given by

- for i=1: $\frac{\partial}{\partial(\alpha_\nu)} m_{\vartheta_\nu}(t, t-1) = Y_{kS+\nu-1}$;
- for i=2: $\frac{\partial}{\partial(\beta_\nu)} m_{\vartheta_\nu}(t, t-1) = Y_{kS+\nu-S}$;

- for i=3: $\frac{\partial}{\partial(\lambda_\nu)} m_{\vartheta_\nu}(t, t-1) = 1$;

- for any $j, l = 1, 2, 3$: $\frac{\partial^2}{\partial(\vartheta_\nu)_l \partial(\vartheta_\nu)_i} m_{\vartheta_\nu}(t, t-1) = 0$.

Then

$$\mathsf{E}\left\{\left|\frac{\partial}{\partial(\alpha_\nu)} m_{\vartheta_\nu}(t, t-1)\right|^2\right\} = \mathsf{E}\left\{Y_{kS+\nu-1}^2\right\} < \infty, \quad \mathsf{E}\left\{\left|\frac{\partial}{\partial(\beta_\nu)} m_{\vartheta_\nu}(t, t-1)\right|^2\right\} = \mathsf{E}\left\{Y_{kS+\nu-S}^2\right\} < \infty,$$

and

$$\mathsf{E}\left\{\left|\frac{\partial}{\partial(\lambda_\nu)} m_{\vartheta_\nu}(t, t-1)\right|^2\right\} = 1.$$

For any $j, l = 1, 2, 3$,

$$\mathsf{E}\left\{\left|\frac{\partial^2}{\partial(\vartheta_\nu)_l \partial(\vartheta_\nu)_i} m_{\vartheta_\nu}(t, t-1)\right|^2\right\} = 0.$$

Therefore the conditions in C1 are satisfied for the process defined in (3.26).

(C2).

$$\mathsf{E}\left\{\left|\sum_{i=1}^{3} a_i(\nu) \frac{\partial}{\partial(\vartheta_\nu)_i} m_{\vartheta_\nu}(t, t-1)\right|^2\right\} = a_1^2(\nu)\mathsf{E}(Y_{kS+\nu-1}^2) + a_2^2(\nu)\mathsf{E}(Y_{kS+\nu-S}^2) + a_3^2(\nu) +$$
$$a_1(\nu)a_2(\nu)\mathsf{E}(Y_{kS+\nu-1}Y_{kS+\nu-S}) + a_1(\nu)a_3(\nu)\mathsf{E}(Y_{kS+\nu-1}) + a_2(\nu)a_3(\nu)\mathsf{E}(Y_{kS+\nu-S}) = 0,$$

if and only if $a_1(\nu) = a_2(\nu) = a_3(\nu) = 0$ since $0 \leq \mathsf{E}(Y_t)^2 < \infty$, for all $t \in \mathbb{Z}$. Hence the conditions in C2 are satisfied.

(C3).

Since $\vartheta_\nu = (\alpha_\nu, \beta_\nu, \lambda_\nu)^\top$, it can be seen that

$$\frac{\partial^2}{\partial(\vartheta_\nu)_i \partial(\vartheta_\nu)_l} m_{\vartheta_\nu}(t, t-1) = 0 \quad \text{and} \quad \frac{\partial^3}{\partial(\vartheta_\nu)_i \partial(\vartheta_\nu)_j \partial(\vartheta_\nu)_l} m_{\vartheta_\nu}(t, t-1) = 0,$$

for any $i, j, l = 1, 2, 3$.

Then, it is clear that any non-negative function $G_{kS+\nu-1}^{ijl}$, with $\mathsf{E}(G_{kS+\nu-1}^{ijl}) < \infty$ satisfies (3.75) and any non-negative function $H_{kS+\nu}^{ijl}$, with $\mathsf{E}(H_{kS+\nu}^{ijl}) < \infty$, satisfies (3.76). For example, $G_{kS+\nu}^{ijl} = H_{kS+\nu}^{ijl} = \alpha_\nu \circ Y_{kS+\nu-1} + \beta_\nu \circ Y_{kS+\nu-S} + \varepsilon_{kS+\nu}$ and $\mathsf{E}(G_{kS+\nu}^{ijl}) = \mathsf{E}(H_{kS+\nu}^{ijl}) = \alpha_\nu \mu_{\nu-1} + \beta_\nu \mu_{\nu-S} + \lambda_\nu$.

Thus C3 is satisfied.

(C4).

Note that R_ν, in (3.77), may be re-written as

$$R_\nu = \mathsf{E}\left[\boldsymbol{U}_{t-1}(Y_t - \boldsymbol{U}_{t-1}^\top \vartheta_{\nu^0})^2 \boldsymbol{U}_{t-1}^\top\right], \tag{3.77}$$

where $t = kS + \nu$, see also (3.50). The expansion of $\mathsf{E}\left[\boldsymbol{U}_{t-1}(Y_t - \boldsymbol{U}_{t-1}^\top \vartheta_{\nu^0})^2 \boldsymbol{U}_{t-1}^\top\right]$ generates a 3×3 matrix R_ν with the forth moment $\mathsf{E}(\varepsilon_{t-1}^4)$ (the highest order) located in the diagonal, that is, in the elements $(R_\nu)_{1,1}$ and $(R_\nu)_{2,2}$. By assumption of the Theorem 3, the elements of R_ν are finite and, consequently, R_ν is finite for all $\nu = 1, \ldots, S$. Then conditions in C1, C2, C3 and C4 are satisfied and the Theorem 3 is proved. □

Proof. of Theorem 4.

This proof follows the same lines as in Theorem 3.2.26 in Taniguchi & Kakizawa (2000). Assume that $\{\boldsymbol{Y}_k\}_{k\in\mathbb{Z}}$ satisfies the conditions in Lemma 2, i.e., $\{\boldsymbol{Y}_k\}_{k\in\mathbb{Z}}$ is a strictly stationary (Lemma 3.3 in Latour (1997)). In addition, it is assumed that $\mathsf{E}\left\|\varepsilon_t\right\|^6 < \infty$ (Remark 3) which implies $\mathsf{E}\left\|Y_t\right\|^6 < \infty$. Let $(\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_n)$ a set of available observations. The ergodicity is also guaranteed using the same arguments in Latour (1997) page 243 according to Definition 1.3.2 in Taniguchi & Kakizawa (2000).

By Theorem 1.3.3 in Taniguchi & Kakizawa (2000), the ergodicity of the process $\{\boldsymbol{Y}_k\}_{k\in\mathbb{Z}}$ implies the ergodicity of the process $\{Y_{kS+\nu}\}_{k\in\mathbb{Z}}$, for $\nu = 1, \ldots, S$. The function $m_{\vartheta_\nu}(t, t-1)$, defined in (3.74), is almost sure three times differentiable in the open set Θ which contains $\boldsymbol{\vartheta}_0$. For $t = kS + \nu$, define $\phi_t(\vartheta_\nu)$ as

$$\phi_t(\vartheta_\nu) = [\log\{f_{\vartheta_\nu}(t, t-1)\} + (Y_t - m_{\vartheta_\nu}(t, t-1))^2 f_{\vartheta_\nu}^{-1}(t, t-1)], \tag{3.78}$$

where $f_{\vartheta_\nu}(t, t-1)$ is defined in Corollary 2. $f_{\vartheta_\nu}(t, t-1)$ is almost sure three times differentiable in the open set Θ which contains $\boldsymbol{\vartheta}_0$. Assume that $\phi_t(\vartheta_\nu)$ is almost sure three times differentiable in the open set Θ. The Theorem 4 is proved if the following conditions are satisfied:

C1. For $1 \le i, l \le 3$,

$$\mathsf{E}\left\{\left|\frac{\partial}{\partial(\vartheta_\nu)_i}\phi_t(\vartheta_\nu^0)\right|^2\right\} < \infty \quad \text{and} \quad \mathsf{E}\left\{\left|\frac{\partial^2}{\partial(\vartheta_\nu)_l \partial(\vartheta_\nu)_i}\phi_t(\vartheta_\nu^0)\right|^2\right\} < \infty,$$

where $(\vartheta_\nu)_i$ is the i-th element of the vector (ϑ_ν), i.e., $(\vartheta_\nu)_1 = \alpha_\nu, (\vartheta_\nu)_2 = \beta_\nu$ and $(\vartheta_\nu)_3 = \lambda_\nu$.

C2. The functions

$$f_{\vartheta_\nu^0}(t, t-1)^{-1/2}\frac{\partial}{\partial(\vartheta_\nu)_i} m_{\vartheta_\nu^0}(t, t-1),$$

for $i = 1, 2, 3$, are linearly independent.

C3. For $\boldsymbol{\vartheta} \in \Theta$, there exists a function $H_{kS+\nu}^{ijl}(Y_{0S+\nu}, \ldots, Y_{kS+\nu})$, $k \in \mathbb{Z}$, for each $\nu = 1, \ldots, S$, such that

$$\left|\frac{\partial^3}{\partial(\vartheta_\nu)_i \partial(\vartheta_\nu)_j \partial(\vartheta_\nu)_l}\phi_t(\vartheta_\nu)\right| \le H_{kS+\nu}^{ijl},$$

with $\mathsf{E}(H_{kS+\nu}^{ijl}) < \infty$ for $i, j, l = 1, 2, 3$.

C4.

$$V = \mathsf{E}\left\{\frac{\partial}{\partial(\vartheta_\nu)}\phi_t(\vartheta_\nu^0)\frac{\partial}{\partial(\vartheta_\nu)^\top}\phi_t(\vartheta_\nu^0)\right\} < \infty. \tag{3.79}$$

Proof of the above conditions for the PINAR$(1, 1_S)$ model.

C1: Let $g_{\vartheta_\nu}(t, t-1) = Y_t - m_{\vartheta_\nu}(t, t-1)$. Consider the followings derivatives

$$\frac{\partial}{\partial(\vartheta_\nu)_i}\phi_t(\vartheta_\nu) = \frac{1}{f_{\vartheta_\nu}(t,t-1)}\frac{\partial}{\partial(\vartheta_\nu)_i}f_{\vartheta_\nu}(t,t-1) + \frac{2g_{\vartheta_\nu}(t,t-1)}{f_{\vartheta_\nu}(t,t-1)}\frac{\partial}{\partial(\vartheta_\nu)_i}g_{\vartheta_\nu}(t,t-1) - \frac{g^2_{\vartheta_\nu}(t,t-1)}{f^2_{\vartheta_\nu}(t,t-1)}\frac{\partial}{\partial(\vartheta_\nu)_i}f_{\vartheta_\nu}(t,t-1), \tag{3.80}$$

and

$$\begin{aligned}\frac{\partial^2}{\partial(\vartheta_\nu)_i(\vartheta_\nu)_j}\phi_t(\vartheta_\nu) &= \frac{-1}{f^2_{\vartheta_\nu}(t,t-1)}\frac{\partial}{\partial(\vartheta_\nu)_i}f_{\vartheta_\nu}(t,t-1)\frac{\partial}{\partial(\vartheta_\nu)_j}f_{\vartheta_\nu}(t,t-1) + \frac{1}{f_{\vartheta_\nu}(t,t-1)}\frac{\partial^2}{\partial(\vartheta_\nu)_i\partial(\vartheta_\nu)_j}\\ &f_{\vartheta_\nu}(t,t-1) + 2\frac{\partial}{\partial(\vartheta_\nu)_i}\left(\frac{g_{\vartheta_\nu}(t,t-1)}{f_{\vartheta_\nu}(t,t-1)}\right)\frac{\partial}{\partial(\vartheta_\nu)_j}g_{\vartheta_\nu}(t,t-1) + \frac{2g_{\vartheta_\nu}(t,t-1)}{f_{\vartheta_\nu}(t,t-1)}\frac{\partial^2}{\partial(\vartheta_\nu)_i\partial(\vartheta_\nu)_j}g_{\vartheta_\nu}(t,t-1) -\\ &- \frac{\partial}{\partial(\vartheta_\nu)_i}\left(\frac{g^2_{\vartheta_\nu}(t,t-1)}{f^2_{\vartheta_\nu}(t,t-1)}\right)\frac{\partial}{\partial(\vartheta_\nu)_j}f_{\vartheta_\nu}(t,t-1) - \frac{g^2_{\vartheta_\nu}(t,t-1)}{f^2_{\vartheta_\nu}(t,t-1)}\frac{\partial^2}{\partial(\vartheta_\nu)_i\partial(\vartheta_\nu)_j}f_{\vartheta_\nu}(t,t-1),\end{aligned} \tag{3.81}$$

where $i, j = 1, 2, 3$, and $(\vartheta_\nu)_i$ is the i-th element of the vector (ϑ_ν), i.e., $(\vartheta_\nu)_1 = \alpha_\nu, (\vartheta_\nu)_2 = \beta_\nu$ and $(\vartheta_\nu)_3 = \lambda_\nu$.

From above calculations, the following inequality can be obtained

$$\mathsf{E}\left\{\left|\frac{\partial}{\partial(\vartheta_\nu)_i}\phi_t(\vartheta_\nu)\right|\right\} \le c_1\,\mathsf{E}(Y^6) < \infty \quad \text{and} \quad \mathsf{E}\left\{\left|\frac{\partial^2}{\partial(\vartheta_\nu)_i(\vartheta_\nu)_j}\phi_t(\vartheta_\nu)\right|\right\} \le c_2\,\mathsf{E}(Y^6) < \infty,$$

for any $i, j = 1, 2, 3$, where c_1 and c_2 are positive constants. Therefore the conditions in C1 are satisfied for the process defined in (3.26).

C2: Similarly to condition 2 in Theorem 2, the functions $f_{\vartheta_\nu^0}(t, t-1)^{-1/2}\frac{\partial}{\partial(\vartheta_\nu)_i}m_{\vartheta_\nu^0}(t, t-1)$, $i = 1, 2, 3$, are linearly independent in the sense that if $a_1(\nu), a_2(\nu), a_3(\nu)$ are arbitrary real numbers such that

$$\mathsf{E}\left\{\left|\sum_{i=1}^{3} a_i(\nu) f_{\vartheta_\nu^0}(t,t-1)^{-1/2}\frac{\partial}{\partial(\vartheta_\nu)_i}m_{\vartheta_\nu^0}(t,t-1)\right|^2\right\} = 0, \tag{3.82}$$

then $a_1(\nu) = 0, a_2(\nu) = 0$, and $a_3(\nu) = 0$, for $\nu = 1, \ldots, S$. The conditional variance of Y_t, $f_{\vartheta_\nu}(t, t-1)$, is a finite number, since $V[Y_t] = \mathsf{E}[Y_t^2] - \mathsf{E}[Y_t]^2$ and, by assumption, $\mathsf{E}[Y_t^2] < \infty$. Let $f_{\vartheta_\nu}(t, t-1) = {\sigma_Y}^2(\nu)$. As shown in Theorem 2, $\frac{\partial}{\partial\vartheta_\nu}m_{\vartheta_\nu^0}(t, t-1) = (Y_{t-1}, Y_{t-s}, 1)$, then Eq. (3.82) implies

$$\frac{a_1(\nu)}{{\sigma_Y}^2(\nu)}\mathsf{E}(Y_{kS+\nu-1}) + \frac{a_2(\nu)}{{\sigma_Y}^2(\nu)}\mathsf{E}(Y_{kS+\nu-S}) + \frac{a_3(\nu)}{{\sigma_Y}^2(\nu)} = 0,$$

which together with (3.26) results $a_1(\nu) = 0, a_2(\nu) = 0$ and $a_3(\nu) = 0$, for $\nu = 1, \ldots, S$. Hence C2 is satisfied.

C3: The proof of this conditions follows the same line of C3 in Theorem 3.

C4: From (3.80) and condition C1, $\frac{\partial}{\partial(\vartheta_\nu)_i}\phi_t(\vartheta_\nu)$ exists and is finite for all $i = 1, 2, 3$. The element of the i-th row and j-th column of the matrix V, in (3.79), is given by

$$(V)_{i,j} = \frac{\partial}{\partial(\vartheta_\nu)_i}\phi_t(\vartheta_\nu)\frac{\partial}{\partial(\vartheta_\nu)_j}\phi_t(\vartheta_\nu).$$

For all $i, j = 1, 2, 3$, there exists a defined positive constant $c_{i,j}$ such that $\mathsf{E}[(V)_{i,j}] < c_{i,j}\,\mathsf{E}[Y_t^6] < \infty$. Then the matrix V exists and is also finite. Hence C4 is satisfied.

□

Chapter 4

The S-periodic integer autoregressive model of order p (PINAR$(p)_S$)

A natural extension of the Periodic Integer-valued autoregressive model to periodically autocorrelated count time series with seasonal period S and autoregressive structure of order p is introduced in this paper. We present some statistical properties of the model and three parameter estimation methods. A simulation study is presented to investigate the performance of the estimators for some finite sample sizes. We show that the estimators are asymptotically Normal distributed with rate of convergence of $n^{1/2}$, where n is the sample size of each season. A section of application to real data series is included, referring to the daily number of people who got medicine based on salbutamol sulphate for the treatment of respiratory problems from the public health care system in the hospital emergency service of the region of Vitória-ES, Brazil.

This paper will be submitted to publication to the Journal of Multivariate Analysis.

The S-Periodic Integer Autoregressive model of order p (PINAR$(p)_S$)

Abstract

This paper introduces a new class of models for S-periodically autocorrelated count time series, which has autoregressive structure of order p. This model is an extension of the PINAR$(1)_S$ model. Statistical properties of the model such as mean, variance, marginal and joint distributions are discussed. Moments-based, conditional least squares and quasi-maximum likelihood estimation methods of the parameters are studied and their performances are investigated through Monte Carlo simulations. Under some assumptions, the estimators are asymptotically Normal distributed with rate of convergence of $n^{1/2}$, where n is the sample size of each season. The performance of the estimator was investigated for small sample size and the empirical results indicated that the method presented accurate estimates. The model well adjusted the series daily number of medicine dispensing for the treatment of respiratory disease.

Keywords: INAR, Periodic Stationarity, PINAR, Moment-based estimators, Conditional Least Squares, Conditional Maximum Likelihood Estimation.

1 Introduction

The integer autoregressive (INAR) models, initially introduced by the INAR(1) model in Al-Osh & Alzaid (1987), appears as an alternative to the well-known Poisson model family for modeling count time series, see, e.g., Fokianos et al. (2009). These models are based on the thinning operator, see Steutel & Van Harn (1979). In this article, the thinning operator is based on Bernoulli distribution, called *binomial thinning operator*. The binomial thinning operator $\circ$ applied on a random variable (r.v.) Y is defined as

$$\alpha \circ Y = \sum_{i=1}^{Y} U_i(\alpha), \tag{4.1}$$

where Y is a $\mathbb{Z}_+$-valued r.v., $\alpha \in [0,1]$ and $\{U_i(\alpha)\}_{i\in\mathbb{Z}_+}$ is a sequence of independent identically distributed (i.i.d.) r.v.'s which are Bernoulli distributed with parameter α. We assume that the sequence $\{U_i(\alpha)\}_{i\in\mathbb{Z}_+}$ is mutually independent of Y. Note that the empty sum is set to 0 if $Y = 0$. The sequence $\{U_i(\alpha)\}_{i\in\mathbb{Z}_+}$ is called a counting sequence. Observe that the probability of success in the thinning is $\mathrm{P}(U_i(\alpha) = 1) = \alpha$ and, conditionally on Y, $\alpha \circ Y \sim \mathrm{Bin}(Y, \alpha)$. Further details about thinning based count time series models are given by Scotto et al. (2015) for the univariate and Latour (1997) in the multivariate case, respectively.

An extension of the INAR(1) model that account the p-th order autoregressive structure is the INAR(p), introduced by Alzaid & Al-Osh (1990) and, independently by Du & Li (1991). A discrete time non-negative integer-valued stochastic process $\{Y_t\}_{t\in\mathbb{Z}}$, is said to be an INAR(p) process if it satisfies the following equation,

$$Y_t = \alpha_1 \circ Y_{t-1} + \cdots + \alpha_p \circ Y_{t-p} + \varepsilon_t,$$

where $0 \leq \alpha_i < 1$ for $i = 1, \ldots, p-1$ and $0 < \alpha_p < 1$, $\{\varepsilon_t\}$ is a sequence of independent and identically distributed (IID) non-negative integer-valued random variables with finite mean and variance. Alzaid & Al-Osh (1990) presented a model for count time series that has a correlation structure similar to the correlation structure of a conventional ARMA$(p, p-1)$ for continuous data. We introduce a model based on an extension of the INAR(p) presented by Du & Li (1991), which model is based on a process with a correlation structure identical to the correlation structure of a standard AR(p).

In spite of its flexibility in dealing with higher order autoregressive processes, the INAR(p) model do not account the periodic phenomenon which is quite common in many area of application. Time series with periodically varying mean, variance and covariance, were introduced by Gladyshev (1961) and are usually called periodically correlated processes (PC). The occurrence of PC processes in time series is corroborated by real applications in many practical situations, see, e.g., Gardner et al. (2006). Basawa & Lund (2001) studied the asymptotic properties of parameter estimates for specifics periodic autoregressive moving-average (PARMA) models among others, and, recently, Sarnaglia et al. (2010) and Solci et al. (2018) presented robust estimation methods for periodic autoregressive processes (PAR) with application in air pollution data. Even though there are in the literature many studies that focus on periodically correlated processes, the vast majority are dedicated to the analysis and the applications for discrete parameter processes (see Priestley (1981), Definition 3.2), with the application of the PARMA model. However, not much attention has been paid to the analysis of periodically correlated count series, for example, Monteiro et al. (2010) and Moriña et al. (2011). In the former paper, the authors introduced the PINAR(1) model and addressed some statistical properties of the parameter estimators together with some empirical investigation. However, the paper does explore the model in a practical problem. The later paper presents a model based on two-order integer-valued autoregressive time series to analyze the number of hospital emergency service arrivals caused by diseases that present seasonal behavior. The first-order seasonal structure INAR was introduced by Bourguignon et al. (2016) and the class of subset INAR models will be investigated in the forthcoming paper Bondon et al. (2018). The PINAR$(1, 1_S)$ model is a particular case of the PINAR$(p)_S$,

In the remainder of this paper, let $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Z}_+$, $\mathbb{R}$, $\mathbb{R}_+$ and $\mathbb{C}$ denote the set of positive integers, integers, non-negative integers, real numbers, non-negative real numbers and complex numbers, respectively. The integer part of $x \in \mathbb{R}$ is denoted by $\lfloor x \rfloor$ and the modulus of $n \in \mathbb{N}$ with respect to $S \in \mathbb{Z}_+$ is defined as n MOD $S = n - S\lfloor n/S \rfloor$. Let $\{n\}_S = S$, if n MOD $S = 0$ and $\{n\}_S = n$ MOD S, otherwise. Clearly, $\{n\}_S \in \{1, \ldots, S\}$ for all $n \in \mathbb{N}$. Let $\{\boldsymbol{e}_i\}_{i=1,\ldots,d}$ be the standard basis in $\mathbb{R}^d$, i.e., $(\boldsymbol{e}_i)_\nu = \delta_{i\nu}$ for all $i, \nu = 1, \ldots, d$ where δ denotes the Kronecker delta.

For all $d \in \mathbb{N}$ let $\mathbf{1}_d = (1, \ldots, 1)^\top \in \mathbb{N}^d$ and let us denote by I_d the $d \times d$ identity matrix. If it is clear from the context, then we omit the subscript d. $\mathrm{Bin}(n, \alpha)$ denotes a binomial distribution with parameters $n \in \mathbb{N}$ and $\alpha \in (0, 1)$; $\mathrm{Poi}(\lambda)$ denotes a Poisson distribution with mean parameter $\lambda \in \mathbb{R}_+$; $\mathrm{Geo}(q)$ denotes a Geometric distribution over $\mathbb{Z}_+$ with parameter $q \in (0, 1)$ and mean $(1 - q)/q$. Let $\mathsf{E}(\cdot)$ and $\mathsf{E}(\cdot|\cdot)$ represent the expectation and the conditional expectation, respectively. Random variables are all defined on a common probability space $(\Omega, \mathcal{A}, \mathrm{P})$.

The organization of the paper is as follows. Section 2 introduces the proposed model, presents the mean and the autocorrelation of the process and some probabilistic properties of the model. Section 3 discuss estimation methods of the parameters, namely the Yule-Walker (moment-based) estimator, the conditional least squares and the quasi-maximum likelihood framework and an alternative estimation procedure. Section 4 presents the simulation and its results, real data application is presented in the Section 5, finally conclusions and final comments are presented at the last section. The appendix shows some proofs and equations mentioned in this article.

2 The PINAR$(p)_S$ model

Let $\{Y_t\}_{t\in\mathbb{Z}}$, $Y_t \in \mathbb{Z}_+$, be a stochastic process with seasonal characteristics of period S, $S \in \mathbb{N}$. The time index t may be written as the Euclidean division between t and S, i.e., as $t = kS + \nu$, where $k \in \mathbb{Z}$ and $\nu = 1, \ldots, S$. For example, in the case of monthly data, $S = 12$, ν and k represent the month of the year and the year, respectively, or in the case of daily data, $S = 7$, ν and k represent the day of the week and the week, respectively.

Define the mean function, $\mu(t) = \mathsf{E}(Y_t)$ for all $t \in \mathbb{Z}$, and the covariance function, the scalar $\gamma_{k,\nu}$, $\nu = 1, \ldots, S$ and $k \in \mathbb{Z}$, on $\mathbb{Z}$ as

$$\gamma_{k,\nu}(h) = \mathrm{Cov}(Y_{kS+\nu}, Y_{kS+\nu-h}), \qquad h \in \mathbb{Z}. \tag{4.2}$$

Definition 4. The stochastic process $\{Y_{kS+\nu}\}_{k\in\mathbb{Z},\nu=1,\ldots,S}$ is said to be a *periodically correlated process* (PC) of period S, $S \in \mathbb{N}$, if, for $\nu = 1, \ldots, S$ and all integers k,

(i) $\mathsf{E}(Y^2_{kS+\nu}) < \infty$;

(ii) $\mu(kS + \nu) = \mu_\nu$;

(iii) $\gamma_{k,\nu}(h) = \gamma_\nu(h)$.

That is, if mean and variance are finite and if they do not depend on k.

Remark 4. Note that, if $\{Y_t\}_{t\in\mathbb{Z}}$ is a periodically correlated process then its mean and covariance are periodic functions with period S. If $S = 1$, then $\{Y_t\}$ represents a homogeneous stochastic process and the condition of periodic stationarity is equivalent to the non-periodic ones.

Definition 5. A $\mathbb{Z}_+$-valued process $\{Y_t\}_{t\in\mathbb{Z}}$ is said to be a *periodic non-negative integer-valued autoregression* (PINAR) with seasonal period S, for some $S \in \{2,3,\ldots\}$, and is denoted by PINAR$(p)_S$, where $p = \max(\vec{p})$ and $\vec{p}$ is the $1 \times S$ vector of autoregressive orders of $\{Y_t\}$, if it satisfies the following stochastic recursion

$$Y_{\nu+kS} = \sum_{i=1}^{p_\nu} \alpha_i(\nu) \circ Y_{\nu-i+kS} + \varepsilon_{\nu+kS}, \tag{4.3}$$

where $k \in \mathbb{Z}$ and $t = kS + \nu$. In this paper $p \leq S$. Because of the similarity of INAR models to the standard autoregressive (AR) model for continuous data, the $\alpha_i(\nu)$, $i = 1,\ldots,p_\nu$, $\nu = 1,2,\ldots,S$ and $p_\nu = 1,\ldots,S$, are called autoregressive coefficients. The vector of AR orders $\vec{p}$ has the form $(p_1,p_2,\ldots,p_S)_{1\times S}$, $S \in \mathcal{N}$, where p_ν represents the AR order of the $\nu-$th season. For each season $\nu = 1,2,\ldots,S$, the set of autoregressive coefficients has the form $\{\alpha_1(\nu),\ldots,\alpha_{p_\nu}(\nu)\} \subset [0;1]^{p_\nu}$. The immigration process $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ is a periodic sequence of $\mathbb{Z}_+$-valued r.v.'s such that for each $\nu \in \{1,\ldots,S\}$ the sequence $\{\varepsilon_{kS+\nu}\}_{k\in\mathbb{Z}}$ consists of i.i.d.r.v.'s. The r.v.'s $Y_1,\ldots,Y_S$ are known as the starting values for the recursion (4.3). Finally, in (4.3), we assume that all counting random variables are mutually independent and they are independent of the sequence $\{\varepsilon_t\}_{t\in\mathbb{Z}}$.

It is assumed that the immigration process and starting values have finite second moments, i.e., the mean function μ and the variance of $\{Y_t\}_{t\in\mathbb{Z}}$ exist and are finite. Moreover, let $\lambda_\nu = \mathrm{E}(\varepsilon_{kS+\nu})$, $\sigma_\nu^2 = \mathrm{Var}(\varepsilon_{kS+\nu})$ for all $k \in \mathbb{Z}$ and $\nu = 1,\ldots,S$.

As can be seen that in the seasonal period ν, Y_t in (4.3) has $p_\nu + 1$ random components; the immigration part of the past Y_{t-i}, $t = \nu + kS$ and $i = 1,\ldots,p_\nu$, with survival probability $\alpha_i(\nu)$ and the elements which entered in the system in the interval (t-1, t], which define the innovation term ε_t for all $t \in \mathbb{Z}$. Moreover, the autoregressive parameters $\alpha_i(\nu)$ and immigration means λ_ν, $\nu = 1.\ldots,S$, change periodically according to the seasonal period S.

It is worth to remark that INAR(p), PINAR$(1,1_S)$ and PINAR$(1)_S$ are particular cases of the model in 4.3.

The mean of the process $\{Y_t\}_{t\in\mathbb{Z}}$, in (4.3) is given by

$$\mu(kS+\nu) = \sum_{i=1}^{p_\nu} \alpha_i(\nu)\mu(kS+\nu-i) + \lambda_\nu. \tag{4.4}$$

Following the same lines of PAR model in Basawa & Lund (2001), the PINAR$(p)_S$ model can be algebraically rewritten as follows. Initially, consider the following definitions.

Definition 6. Let $\boldsymbol{A}\circ = (a_i(\nu)\circ)_{i,\nu}$, $1 \leq i,\nu \leq S$, be a $S \times S$ *matricial binomial thinning operator*, also called the matricial Steuel and Van Harn operator, where $a_i(\nu) \in [0,1]$ for all $1 \leq i,\nu \leq S$.

The action of $A\circ$ on $\boldsymbol{Y} = (Y_1, \ldots, Y_S)^\top$, denoted by $\boldsymbol{A} \circ \boldsymbol{Y}$, is

$$\boldsymbol{A} \circ \boldsymbol{Y} = \boldsymbol{A} \circ \begin{pmatrix} Y_1 \\ \vdots \\ Y_S \end{pmatrix} = \begin{pmatrix} \sum_{\nu=1}^{S} a_1(\nu) \circ Y_\nu \\ \vdots \\ \sum_{\nu=1}^{S} a_S(\nu) \circ Y_\nu \end{pmatrix}. \tag{4.5}$$

In the above definition, the operators $a_i(\nu)\circ$, $1 \leq i, \nu \leq S$, are supposed to be mutually independent, see Definition 2.1 in Latour (1997). Based on Lemma 2.1 in Latour (1997), $\mathsf{E}(\boldsymbol{A} \circ \boldsymbol{Y}) = \boldsymbol{A}\mathsf{E}(\boldsymbol{Y})$, where $\boldsymbol{A} = (a_i(\nu))_{i,\nu}$ for $1 \leq i, \nu \leq S$. $\boldsymbol{A}$ is the mean of the operator $\boldsymbol{A}\circ$.

Now, let $\boldsymbol{Y}_k = (Y_{kS+1}, \ldots, Y_{kS+S})^\top$, $\boldsymbol{\varepsilon}_k = (\varepsilon_{kS+1}, \ldots, \varepsilon_{kS+S})^\top$, $k \in \mathbb{Z}$, where Y_{kS+1} and $\boldsymbol{\varepsilon}_{kS+1}$ are defined in Definition 5. Then, by (4.3), for $k \in \mathbb{Z}$, one can see that the following stochastic equation holds

$$\boldsymbol{A} \circ \boldsymbol{Y}_k = \boldsymbol{B} \circ \boldsymbol{Y}_{k-1} + \boldsymbol{\varepsilon}_k, \tag{4.6}$$

where $\boldsymbol{A}\circ$ and $\boldsymbol{B}\circ$ are $S \times S$ independent matricial binomial thinning operators, defined by

$$(\boldsymbol{A}\circ)_{i,j} = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{if } i < j \\ -\alpha_{i-j}(i)\circ, & \text{if } i > j \end{cases} \quad \text{and} \quad (\boldsymbol{B}\circ)_{i,j} = \alpha_{S+i-j}(i)\circ, \tag{4.7}$$

with the convention $\alpha_m(n) = 0$ if $m > p_n$.

Following, are established conditions for the periodically distribution of the process $\{\boldsymbol{Y}_k\}_{k\in\mathbb{Z}}$ in (4.3).

Suppose that the process $\{\boldsymbol{Y}_k\}_{k\in\mathbb{Z}}$ has a constant mean vector $\boldsymbol{\mu}$. By definition of $\{\boldsymbol{\varepsilon}_k\}$, the mean vector $\mathsf{E}[\{\boldsymbol{\varepsilon}_k\}] = \boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_S)^\top$ is finite and do not depend on k. Then, $\mathsf{E}[A \circ \boldsymbol{Y}_k]$ is

$$A\boldsymbol{\mu} = B\boldsymbol{\mu} + \boldsymbol{\lambda}. \tag{4.8}$$

Note that A is a lower triangular non-singular matrix and its inverse has only non-negative elements. Thus A^{-1} and B are non-negative matrices, hence $A^{-1}B$ and $A^{-1}\boldsymbol{\lambda}$ are also non-negative matrix and vector, respectively. By multiplying A^{-1} in both sides of (4.8), it can be seen that

$$\boldsymbol{\mu} = A^{-1}B\boldsymbol{\mu} + A^{-1}\boldsymbol{\lambda}. \tag{4.9}$$

or

$$(I - A^{-1}B)\boldsymbol{\mu} = A^{-1}\boldsymbol{\lambda}. \tag{4.10}$$

Consider the following lemma.

Lemma 4. *Let the mean matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ of the two operators $\boldsymbol{A}\circ$ and $\boldsymbol{B}\circ$, respectively,*

defined by (4.7)*. Then, the following statements are equivalent:*

(i) $\rho(\boldsymbol{A}^{-1}\boldsymbol{B}) < 1$*;*

(ii) the roots of the determinant equation $\det(z\boldsymbol{I}_S - \boldsymbol{A}^{-1}\boldsymbol{B}) = 0$*, for all complex* z*, are all less than 1 in absolute value;*

(iii) the roots of the matricial autoregressive polynomial $\boldsymbol{P}(z) = \boldsymbol{A} - z\boldsymbol{B}$*, for all complex* z*, lie outside of the complex unit circle;*

Proof. $(i) \Leftrightarrow (ii)$ Since the complex eigenvalues of $A^{-1}B$ can be derived as the solutions to the characteristic equation $\det(zI_S - A^{-1}B) = 0$ the equivalence is clear. $(ii) \Leftrightarrow (iii)$ follows from the identity

$$\det(zI_S - A^{-1}B) = \det(zA^{-1}(A - z^{-1}B)) = z^S \det(A - z^{-1}B))$$

with $z \neq 0$ and we used that $\det(A) = 1$. □

Theorem 5. *Let* $\{Y_t\}_{t\in\mathbb{Z}}$ *be a PINAR*$(p)_S$ *process defined by* (4.3) *and let the matrices* $\boldsymbol{A}$ *and* $\boldsymbol{B}$ *be defined by* (4.7)*. If one of the statements of Lemma 4 holds for matrices* $\boldsymbol{A}$ *and* $\boldsymbol{B}$*, then there exists a second order periodically stationary solution to* $\{Y_t\}_{t\in\mathbb{Z}}$*.*

Proof. By Theorem 6.2.24 in Horn & Johnson (2012), the strict positivity of α's implies that $A^{-1}B$ is irreducible in the sense of Definitions 6.2.21 and 6.2.22 of Horn & Johnson (2012). From Theorem 2.1 in Seneta (2006) and since $A^{-1}B$ is a Perron-Frobenius matrix, a necessary and sufficient condition for a solution of $\boldsymbol{\mu}$ ($\boldsymbol{\mu} \geq \mathbf{0}, \neq \mathbf{0}$), where $\mathbf{0}$ is a S-dimensional vector of zeros, to (4.10) to exist for any $\boldsymbol{\lambda}^* = A^{-1}\boldsymbol{\lambda}$ ($\boldsymbol{\lambda}^* \geq \mathbf{0}, \neq \mathbf{0}$) is that the spectral radius $\rho(A^{-1}B) < 1$, which is the maximum eigenvalue in modulus of the matrix $A^{-1}B$. Note that, since $S \geq 2$, from the Perron-Frobenius Theorem in Horn & Johnson (2012) page 534, $\rho(A^{-1}B) > 0$. Therefore, $0 < \rho(A^{-1}B) < 1$.

Based on Graybill (1983), page 100, if $|\varphi| < 1$ for every characteristic root φ of $A^{-1}B$ and none of sums of absolute values of row or column elements exceed unity, then $\sum_{i=1}^{\infty}(A^{-1}B)^i$ converges to $(I - A^{-1}B)^{-1}$. This condition assures the invertibility of $(I - A^{-1}B)$ and the positivity of its inverse. In this context, model PINAR$(p)_S$ in (4.6) will be completely specified, if the $det(zI - A^{-1}B) \neq \mathbf{0}$, $z \in \mathbb{C}$, i.e., the characteristic roots will be inside of unit circle, and, therefore, the process in 4.6 will be second order periodic stationary process (Brockwell & Davis (2013)). In addition, if all the eigenvalues of $A^{-1}B$ are inside the unit circle, $I - A^{-1}B$ is non-singular and $\boldsymbol{\mu} = (I - A^{-1}B)^{-1}\boldsymbol{\lambda}^*$ has unique solution. □

Conversely, if a process $\{Y_t\}_{t\in\mathbb{Z}}$ which satisfies a proper PINAR$(p)_S$ model (its parameters $\alpha_i(\nu), \lambda_\nu$, $1 \leq i \leq p$ and $\nu = 1, \ldots, S$, are strictly positive) is a second order periodically stationary process, then all of the statements of Lemma 4 are true. Some examples are now given.

Example 3. Consider the case when $\alpha_i(\nu) = 0$ for all $i \neq 1$ and $\nu = 1, \ldots, S$. Then the PINAR$(p)_S$ model is reduced to a PINAR$(1)_S$ model. The characteristic polynomial of this

model is simplified to $P(z) = 1 - z\prod_{\nu=1}^{S}\alpha_1(\nu)$ and a necessary and sufficient condition of asymptotic stability in the mean is $\prod_{\nu=1}^{S}\alpha_1(\nu) < 1$. (Note that $\prod_{\nu=1}^{S}\alpha_\nu$ is the spectral radius of the matrix A defined on page 1531 in Monteiro et al. (2010))

Example 4. Consider the case $S = 3$, the PINAR$(2)_3$ model, with $\vec{p} = (1, 2, 2)$. Then

$$A = \begin{bmatrix} 1 & 0 & 0 \\ -\alpha_1(2) & 1 & 0 \\ -\alpha_2(3) & -\alpha_1(3) & 1 \end{bmatrix}, \qquad A^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ \alpha_1(2) & 1 & 0 \\ \alpha_2(3) + \alpha_1(3)\alpha_1(2) & \alpha_1(3) & 1 \end{bmatrix}, \qquad B = \begin{bmatrix} 0 & 0 & \alpha_1(1) \\ 0 & 0 & \alpha_2(2) \\ 0 & 0 & 0 \end{bmatrix}. \tag{4.11}$$

The characteristic polynomial is given by $P(z^{-1}) = 1-[\alpha_1(1)\alpha_1(2)\alpha_1(3)+\alpha_1(1)\alpha_2(3)+\alpha_1(3)\alpha_2(1)]z^{-1}$. By solving the characteristic equation, one can see that $1/(\alpha_1(1)\alpha_1(2)\alpha_1(3) + \alpha_1(1)\alpha_2(3) + \alpha_1(3)\alpha_2(1)) > 1$ is a necessary and sufficient condition for the second order periodically stationarity. (Note that this condition can be rewritten as $\alpha_1(1)\alpha_1(2)\alpha_1(3)+\alpha_1(1)\alpha_2(3)+\alpha_1(3)\alpha_2(1) < 1$)

A PINAR$(p)_S$ process defined by (4.3) has a PAR representation similarly to the AR representation of the INAR models. Namely, let define the random variables $X_t = Y_t - \mathsf{E}(Y_t)$ and $M_t = Y_t - \mathsf{E}(Y_t|\mathcal{F}_{t-1})$, where $\mathcal{F}_t$ denotes the σ-algebra generated by the random variables until time t, for all $t \in \mathbb{Z}$ and $\mathcal{F}_0 = \{\emptyset, \Omega\}$. Clearly, by (4.3), it can be seen that

$$\mathsf{E}(Y_{\nu+kS}|\mathcal{F}_{\nu+kS-1}) = \sum_{i=1}^{p_\nu}\alpha_i(\nu)Y_{\nu+kS-i} + \lambda_\nu \tag{4.12}$$

and thus

$$M_{\nu+kS} = \sum_{i=1}^{p_\nu}(\alpha_i(\nu) \circ Y_{\nu+kS-i} - \alpha_i(\nu)Y_{\nu+kS-i}) + (\varepsilon_{\nu+kS} - \lambda_\nu), \tag{4.13}$$

where $k \in \mathbb{Z}$ and $\nu = 1, \ldots, S$. Since the counting r.v.'s involved into the model and the immigrations are mutually independent, $\{M_t\}_{t\in\mathbb{Z}}$ is a sequence of martingale differences with respect to the filtration $\{\mathcal{F}_t\}_{t\in\mathbb{Z}}$. Moreover, for all $k \in \mathbb{Z}$ and $\nu = 1, \ldots, S$,

$$\mathsf{E}(M_{kS+\nu}^2|\mathcal{F}_{kS+\nu-1}) = \sum_{i=1}^{p_\nu}\alpha_i(\nu)(1 - \alpha_i(\nu))Y_{\nu+kS-i} + \sigma_\nu^2. \tag{4.14}$$

Hence, $\{M_t\}_{t\in\mathbb{Z}}$ is a heteroscedastic white noise process. The process $\{X_t\}_{t\in\mathbb{Z}}$ satisfies a *periodic autoregressive* (PAR) model defined by

$$X_{\nu+kS} = \sum_{i=1}^{p_\nu}\alpha_i(\nu)X_{\nu+kS-i} + M_{\nu+kS}, \tag{4.15}$$

where $k \in \mathbb{Z}$ and $\nu = 1, \ldots, S$, with autoregressive parameters $\alpha_i(\nu)$, $i = 1, \ldots, p_\nu$, and $\{M_t\}_{t\in\mathbb{Z}}$ is a periodic innovation process with zero mean and variance $\mathsf{E}(M_{kS+\nu}^2) = \sum_{i=1}^{p_\nu}\alpha_i(\nu)(1 - \alpha_i(\nu))\mu_{\nu-i} + \sigma_\nu^2$ for all $k \in \mathbb{Z}$ and $\nu = 1, \ldots, S$, where $\mu_0 = \mu_S$ and $\mu_{-a} = \mu_{S-a}$, for $a = 1, \ldots, S$. Let $\boldsymbol{X}_k = (X_{kS+1}, \ldots, X_{kS+S})^\top$, $k \in \mathbb{Z}$, and $\boldsymbol{M}_k = (M_{kS+1}, \ldots, M_{kS+S})^\top$, $k \in \mathbb{Z}$, and consider

the $\mathbb{R}^S$-valued stochastic processes $\{\boldsymbol{X}_k\}_{k\in\mathbb{Z}}$ and $\{\boldsymbol{M}_k\}_{k\in\mathbb{Z}}$. Then, we have the S-variate VAR representation of the PAR process $\{X_t\}_{t\in\mathbb{Z}}$ as

$$\boldsymbol{A}\boldsymbol{X}_k = \boldsymbol{B}\boldsymbol{X}_{k-1} + \boldsymbol{M}_k, \tag{4.16}$$

$k \in \mathbb{Z}$, and the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are defined in (4.7). The covariance matrix of the random vector $\boldsymbol{M}_k$ is diagonal and since $\{Y_t\}_{t\in\mathbb{Z}}$ is periodically stationary, then this covariance matrix does not depend on k and can be written in the following form. Define the S-dimensional vectors $\boldsymbol{\alpha}_i = (\alpha_i(1), \ldots, \alpha_i(S))^\top$, $\widetilde{\boldsymbol{\mu}}_i = (\mu_{S-i}, \mu_{S-i+1}, \ldots, \mu_S, \mu_1, \ldots, \mu_{S-i-1})^\top$, $i = 1, \ldots, S$, where $\mu_0 = \mu_S$ and $\mu_{-a} = \mu_{S-a}$, for $a = 1, \ldots, S$, with the convention $\alpha_m(n) = 0$ if $m > p_n$, and $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_S^2)^\top$. Then, the common covariance matrix $\Sigma_{\boldsymbol{M}}$ of $\boldsymbol{M}_k$, $k \in \mathbb{Z}$, is given by

$$\Sigma_{\boldsymbol{M}} = \operatorname{diag}\left(\sum_{i=1}^{p} \boldsymbol{\alpha}_i \odot (\mathbf{1} - \boldsymbol{\alpha}_i) \odot \widetilde{\boldsymbol{\mu}}_i + \boldsymbol{\sigma}^2\right).$$

where $\operatorname{diag}\{\boldsymbol{v}\}$ denotes a diagonal matrix with vector $\boldsymbol{v}$ in its diagonal, $p = \max(\vec{p})$ and $\odot$ is the element wise (Hadamard) product. Consider $\alpha_i(\nu) = 0$ for $i > p_\nu$, $\nu = 1, \ldots, S$.

It is noted that, $\boldsymbol{X}_k = \boldsymbol{Y}_k - \mathsf{E}(\boldsymbol{Y}_k)$ and $\boldsymbol{M}_k = (B \circ \boldsymbol{Y}_{k-1} - B\boldsymbol{Y}_{k-1}) - (A \circ \boldsymbol{Y}_k - A\boldsymbol{Y}_k) + (\boldsymbol{\varepsilon}_k - \boldsymbol{\lambda})$ for all $k \in \mathbb{Z}$. Thus, the autoregressive representation of the matricial INAR model (4.6) coincides with the vector autoregressive model (4.16).

Remark 5. The state vector $\boldsymbol{Y}_k$ in (4.6) is the state-space representation of the process in (4.3), which form is usual in the theory of real valued periodic processes, see, e.g., Franses & Paap (2004). The matricial polynomial $\boldsymbol{P}$ in Lemma 4 can be interpreted as the formal characteristic polynomial to the state-space representation (4.6). Moreover, $\boldsymbol{P}$ is the matricial autoregressive polynomial of the VAR process $\{\boldsymbol{X}_k\}_{k\in\mathbb{Z}}$ which satisfies (4.16), see (3.12) in Franses & Paap (2004). The determinant equation in (ii) of Lemma 4 is also well-known in the field of real-valued PC processes, see, e.g., Vecchia (1985, eqn. (4)), Ula & Smadi (1997, eqn. (12)) and Franses & Paap (2004, eqn. (3.26)).

The first and second order moments of the $\mathbb{Z}_+^S$-valued stationary process (4.6) can be described following the lines in Section $2.1.4$ in Lütkepohl (2005). Plainly, $\mathsf{E}(\boldsymbol{Y}_k) = \boldsymbol{\mu}$ for all $k \in \mathbb{Z}$. Define the $S \times S$-dimensional covariance matrices $\Gamma(k, \ell) = \operatorname{Cov}(\boldsymbol{Y}_k, \boldsymbol{Y}_\ell)$ for all $k, \ell \in \mathbb{Z}$. We have $\Gamma(k, \ell) = \Gamma(\ell, k)^\top$ for all $k, \ell \in \mathbb{Z}$. Clearly, $(\Gamma(k, \ell))_{i,\nu} = \operatorname{Cov}(Y_{kS+i}, Y_{\ell S+\nu})$ for all $k, \ell \in \mathbb{Z}$ and $i, \nu = 1, \ldots, S$ and $\Gamma(k, \ell) = \mathsf{E}(\boldsymbol{X}_k \boldsymbol{X}_\ell^\top)$ for all $k, \ell \in \mathbb{Z}$. Moreover, let $\Sigma = \Gamma(k, k)$, i.e., $\Sigma = \operatorname{Var}(\boldsymbol{Y}_k) = \operatorname{Var}(\boldsymbol{X}_k)$, for all $k \in \mathbb{Z}$. It can be seen, by recursion (4.16), that Σ is the unique solution to the matrix equation

$$A\Sigma A^\top = B\Sigma B^\top + \Sigma_{\boldsymbol{M}}. \tag{4.17}$$

The covariance matrix $\Gamma(k, \ell)$ depends only on $k - \ell$. Thus,it can be defined that the covariance matrix function $\Gamma(h)$, $h \in \mathbb{Z}$, of the second order stationary process $\{\boldsymbol{Y}_k\}_{k\in\mathbb{Z}}$ can be written as

$$\Gamma(h) = \begin{cases} \Gamma(k+h, k) = \left(A^{-1}B\right)^h \Sigma & \text{if } h \geq 0, \\ \Gamma(k, k-h) = \Sigma \left(\left(A^{-1}B\right)^\top\right)^{-h} & \text{if } h \leq 0, \end{cases} \tag{4.18}$$

where $k \in \mathbb{Z}$ is arbitrary. Clearly, $\Gamma(h) = \Gamma(-h)^\top$ for all $h \in \mathbb{Z}$. The covariance kernel R of a PC process $\{Y_t\}_{t\in\mathbb{Z}}$ can be extended onto $\mathbb{Z}^2$. Moreover, by this extension, we have

$$(\Gamma(h))_{i,\nu} = R(hS + i, \nu) \qquad \text{for all} \quad h \in \mathbb{Z}_+,\ i, \nu \in \{1, \dots, S\}. \tag{4.19}$$

Now, let introduce the Yule-Walker equations of the second order stationary process $\{\boldsymbol{Y}_k\}_{k\in\mathbb{Z}}$ as

$$A\Gamma(h) = B\Gamma(h-1). \tag{4.20}$$

By (4.17) we obtain the equation

$$A\Gamma(0) = B\Gamma(-1) + \Sigma_{\boldsymbol{M}}(A^{-1})^\top, \tag{4.21}$$

which can be considered as an extension of (4.20) for $h = 0$. Define the scalar functions γ_ν, $\nu = 1, \dots, S$, on $\mathbb{Z}$ as

$$\gamma_\nu(h) = R(\nu + h, \nu), \qquad h \in \mathbb{Z}. \tag{4.22}$$

If $\{Y_t\}_{t\in\mathbb{Z}}$ is a PC process of period S then the functions γ_ν, $\nu = 1, \dots, S$, determine the covariance kernel R and thus the covariance matrix function Γ of the state process $\{\boldsymbol{Y}_k\}_{k\in\mathbb{Z}}$. Namely, if $s = kS + i$ and $t = \ell S + \nu$ where $k, \ell \in \mathbb{Z}$ and $i, \nu \in \{1, \dots, S\}$, then

$$R(s,t) = R(kS + i, \ell S + \nu) = \gamma_\nu((k-\ell)S + i - \nu). \tag{4.23}$$

The functions γ_ν, $\nu = 1, \dots, S$, are called the periodic autocovariance functions (ACF) of the PC process $\{Y_t\}_{t\in\mathbb{Z}}$. The periodic autocovariance functions satisfy the symmetry property

$$\gamma_\nu(hS + i) = \begin{cases} \gamma_{i+\nu}(-hS - i) & \text{if } i + \nu \le S, \\ \gamma_{i+\nu-S}(-hS - i) & \text{if } i + \nu > S \end{cases} \tag{4.24}$$

for all $h \in \mathbb{Z}$ and $i, \nu \in \{1, \dots, S\}$. Especially, in the case of $h = 0$ we have

$$\gamma_\nu(i) = \begin{cases} \gamma_{i+\nu}(-i) & \text{if } i + \nu \le S \\ \gamma_{i+\nu-S}(-i) & \text{if } i + \nu > S, \end{cases} \tag{4.25}$$

for all $i, \nu \in \{1, \dots, S\}$. The relation between the covariance matrix function Γ of the state process $\{\boldsymbol{Y}_k\}_{k\in\mathbb{Z}}$ and the periodic autocovariance functions γ_ν, $\nu = 1, \dots, S$, can be expressed as

$$(\Gamma(h))_{i,\nu} = \gamma_\nu(hS + i - \nu), \tag{4.26}$$

for all $h \in \mathbb{Z}$ and $i, \nu \in \{1, \dots, S\}$. Thus, the covariance matrix $\Gamma(h)$ can be written as

$$\Gamma(h) = \begin{bmatrix} \gamma_1(hS) & \gamma_2(hS-1) & \cdots & \gamma_S(hS-S+1) \\ \gamma_1(hS+1) & \gamma_2(hS) & \cdots & \gamma_S(hS-S+2) \\ \vdots & \vdots & & \vdots \\ \gamma_1(hS+S-1) & \gamma_2(hS+S-2) & \cdots & \gamma_S(hS) \end{bmatrix}. \tag{4.27}$$

By property (4.25), in case of $h = 0$, the covariance matrix Σ can also be expressed as

$$\Sigma = \Gamma(0) = (\gamma_{i\wedge\nu}(|i-\nu|))_{i,\nu=1}^{S} = \begin{bmatrix} \gamma_1(0) & \gamma_1(1) & \cdots & \gamma_1(S-2) & \gamma_1(S-1) \\ \gamma_1(1) & \gamma_2(0) & \cdots & \gamma_2(S-3) & \gamma_1(S-2) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \gamma_1(S-2) & \gamma_2(S-3) & \cdots & \gamma_{S-1}(0) & \gamma_{S-1}(1) \\ \gamma_1(S-1) & \gamma_2(S-2) & \cdots & \gamma_{S-1}(1) & \gamma_S(0) \end{bmatrix}, \quad (4.28)$$

where $i \wedge \nu = \min\{i, \nu\}$.

Now, let establish the marginal distribution of the PINAR$(p)_S$ process, $\{Y_t\}$. Since the counting r.v.'s involved into the (4.3) and the immigration process are mutually independent, the process $\{Y_t\}_{t\in\mathbb{Z}}$, for $t = \nu + kS$, $k \in \mathbb{Z}$, is an in-homogeneous p-step Markov chain, i.e., for all $t > p$, with $p = \mathsf{max}(\vec{p})$, and $y_1, \ldots, y_t \in \mathbb{Z}_+$, the conditional probability function is given by

$$\begin{aligned} \mathsf{P}(Y_t = y_t | Y_s = y_s,\ s = 1, \ldots, t-1) =& \mathsf{P}(Y_t = y_t | Y_{t-1} = y_{t-1}, \ldots, Y_{t-p_\nu} = y_{t-p_\nu}) \\ =& P(\alpha_1(\nu) \circ y_{t-1} + \ldots + \alpha_{p_\nu}(\nu) \circ y_{t-p_\nu} + \varepsilon_t = y_t). \end{aligned} \quad (4.29)$$

Denote the periodic Markov kernel of the PC process $\{Y_t\}$ as

$$p_\nu(m|m_1, \ldots, m_{p_\nu}) = P(\alpha_1(\nu) \circ m_1 + \ldots + \alpha_{p_\nu}(\nu) \circ m_{p_\nu} + \varepsilon_\nu = m), \qquad m, m_1, \ldots, m_{p_\nu} \in \mathbb{Z}_+. \quad (4.30)$$

The marginal distribution of the PC process $\{Y_t\}$ is given by

$$P(Y_{kS+\nu} = m) = \sum_{m_1,\ldots,m_{p_\nu}=0}^{\infty} p_\nu(m|m_1, \ldots, m_{p_\nu}) P(Y_{kS+\nu-1} = m_1, \ldots, Y_{kS+\nu-p_\nu} = m_{p_\nu}), \quad (4.31)$$

where $l \in \mathbb{Z}$, for all $k \in \mathbb{Z}$ and $\nu = 1, \ldots, S$.

Given starting values $Y_1, \ldots, Y_S$, by the definition of the conditional probability and the p-step Markov property of the PINAR$(p)_S$ process, it can be seen that

$$\begin{aligned} P(Y_t = y_t, \ldots, &Y_{S+1} = y_{S+1} | Y_S = y_S, \ldots, Y_1 = y_1) = \\ & p_\nu(y_t | y_S, \ldots, y_{t-p_\nu}) P(Y_{t-1} = y_{t-1}, \ldots, Y_{S+1} = y_{S+1} | Y_S = y_S, \ldots, Y_1 = y_1), \end{aligned} \quad (4.32)$$

where $\nu = \{t\}_S$ and $y_1, \ldots, y_t \in \mathbb{Z}_+$. Thus, by induction, if $T = nS$ where $n \in \mathbb{N}$, we have

$$P(Y_T = y_T, \ldots, Y_{S+1} = y_{S+1} | Y_S = y_S, \ldots, Y_1 = y_1) = \prod_{\nu=1}^{S} \prod_{k=1}^{n-1} p_\nu(y_{kS+\nu} | y_{kS+\nu-1}, \ldots, y_{kS+\nu-p_\nu}), \quad (4.33)$$

where $y_1, \ldots, y_T \in \mathbb{Z}_+$.

2 .1 PINAR$(p)_S$ model with Poisson immigration

Let $\{\varepsilon_{kS+\nu}\}$ be a periodic sequence, where $\varepsilon_{kS+\nu} \sim \mathrm{Poi}(\lambda_\nu)$, for $k \in \mathbb{Z}$ and $\nu \in \{1,\ldots,S\}$. The periodic Markov-kernel of PINAR$(p)_S$ model with Poisson immigration, Poisson-PINAR$(p)_S$, is given by the following equation:

$$p_\nu(y_t|Y_{t-1}=y_{t-1},\ldots,Y_{t-p_\nu}=y_{t-p_\nu}) = [\mathrm{Bin}(y_{t-1},\alpha_1(\nu)) * \ldots * \mathrm{Bin}(y_{t-p_\nu},\alpha_{p_\nu}(\nu)) * \mathrm{Poi}(\lambda_\nu)](k)$$
$$= \sum_{i_1=0}^{y_{t-1}\wedge y_t} \binom{y_{t-1}}{i_1} (\alpha_1(\nu))^{i_1}(1-\alpha_1(\nu))^{y_{t-1}-i_1} \sum_{i_2=0}^{y_{t-2}\wedge(y_t-i_1)} \binom{y_{t-2}}{i_2} (\alpha_2(\nu))^{i_2}(1-\alpha_2(\nu))^{y_{t-2}-i_2}$$
$$\ldots \sum_{i_{p_\nu}=0}^{y_{t-p_\nu}\wedge(y_t-(i_1+i_2+\ldots+i_{p_\nu-1}))} \binom{y_{t-p_\nu}}{i_{p_\nu}} (\alpha_{p_\nu}(\nu))^{i_{p_\nu}}(1-\alpha_{p_\nu}(\nu))^{y_{t-p_\nu}-i_{p_\nu}} \exp(-\lambda_\nu)$$
$$\frac{\lambda_\nu^{y_t-(i_1+i_2+\ldots+i_{p_\nu})}}{(y_t-(i_1+i_2+\ldots+i_{p_\nu}))!},$$

where $*$ denotes convolution and $i \wedge \nu = \min\{i,\nu\}$.

3 Parameter estimation methods

In this section, the standard estimation methods such as moment-based or Yule-Walker (YW), conditional least squares (CLS) and quasi-maximum likelihood (QML) estimators are discussed for the proposed model. The closed forms of Yule-Walker and CLS estimators are derived for the PINAR$(p)_S$ model (4.3) under general immigration distribution. Moreover, the QML estimator is discussed when the immigration follows Poisson distribution, i.e., $\varepsilon_{kS+\nu} \sim \mathrm{Poi}(\lambda_\nu)$ for all $k \in \mathbb{Z}$ and $\nu \in \{1,\ldots,S\}$. The asymptotic properties of these estimators are also investigated.

Let $\vartheta_\nu = (\boldsymbol{\alpha}_\nu^\top, \lambda_\nu)$, where $\boldsymbol{\alpha}_\nu = (\alpha_1(\nu),\ldots,\alpha_{p_\nu}(\nu))^\top$, $\alpha_\nu \in (0;1)$ and $\lambda_\nu \in \mathbb{R}_+$, for all seasons $\nu = 1,\ldots,S$ and let $\boldsymbol{\vartheta} = (\vartheta_1,\ldots,\vartheta_S)$ represent the p^*-dimensional, where $p^* = p_1 + \ldots + p_S + S$, unknown parameter vector of the PINAR$(p)_S$ model defined by (4.3). In this paper, all the parameter vectors are column vector. The parameter vector is assumed to be lying in the open set Θ. Its true value is denoted by $\boldsymbol{\vartheta}_0$. Consider a sample $Y_1,\ldots,Y_T$ of size $T = nS$ where $n \in \mathbb{N}$ for the PINAR$(p)_S$ process $\{Y_t\}$. The notation used assumes n complete periods of observations. This implies that a sample $\boldsymbol{Y}_0,\ldots,\boldsymbol{Y}_{n-1}$ of size n is given for the state process $\{\boldsymbol{Y}_k\}$. In this section, it is supposed that $\{\boldsymbol{Y}_k\}$ is a strictly stationary ergodic process. By (4.3) and (4.33) we may conjecture that all estimators of the parameter ϑ_ν will depend on the sequence of data $\{Y_{kS+\nu},\ldots,Y_{kS+\nu-p_\nu}\}_{k=1,\ldots,n-1}$, respectively, for each season $\nu \in \{1,\ldots,S\}$.

3 .1 Moment-based estimation (Yule-Walker)

The first and second order moments of the periodically stationary process $\{Y_k\}$ can be estimated, following Vecchia (1985), page 724, as follows. For the periodic means μ_ν, $\nu = 1, \ldots, S$, we have the estimators

$$\widehat{\mu}_{n,\nu} = n^{-1} \sum_{k=0}^{n-1} Y_{kS+\nu}, \tag{4.34}$$

for all seasons $\nu = 1, \ldots, S$. For the periodic covariance functions γ_ν, $\nu = 1, \ldots, S$, we have the estimator

$$\widehat{\gamma}_{n,\nu}(h) = n^{-1} \sum_{k=0}^{n-\lceil (h+\nu)/S \rceil} \widetilde{Y}_{kS+h+\nu} \widetilde{Y}_{kS+\nu}, \tag{4.35}$$

where $0 \leq h \leq nS - \nu$ and $\widetilde{Y}_{kS+\nu} = Y_{kS+\nu} - \overline{Y}_{n,\nu}$, $k = 0, \ldots, n-1$ and $\nu = 1, \ldots, S$, are the periodically mean standardized observations. (Note that $\lceil x \rceil$ denotes the upper integer part of $x \in \mathbb{R}$.) Clearly, $\widetilde{Y}_{kS+\nu} = (\widetilde{\boldsymbol{Y}}_{k-1})_\nu$ for all $k = 1, \ldots, n$ and $\nu = 1, \ldots, S$. The estimators $\widehat{\mu}_{n,\nu}$ and $\widehat{\gamma}_{n,\nu}$, $\nu = 1, \ldots, S$, are the same as for real-valued PC processes, see, e.g., Vecchia (1985, eqn. (10)). The estimators $\widehat{\gamma}_{n,\nu}$, $\nu = 1, \ldots, S$, are called the *sample periodic autocovariance functions* (ACF) of the integer-valued PC process $\{Y_t\}$. In the sequel, if it is clear from the context, we omit the index n of sample size in the estimators and, e.g., we write simply $\widehat{\mu}_\nu$ and $\widehat{\gamma}_\nu$, $\nu = 1, \ldots, S$.

For each season $\nu \in \{1, \ldots, S\}$ the parameter vector ϑ_ν can be estimated through the Yule-Walker equations, as can be seen below, by replacing the means μ_ν's and the autocovariance functions γ_ν's by the corresponding sample means $\widehat{\mu}_\nu$'s and sample autocovariance functions $\widehat{\gamma}_\nu$'s, respectively. Multiplying (4.3) by $Y_{kS+\nu-i}$, $i = 0, 1, \ldots, p_\nu$, and taking expectations of both leads to

$$\Gamma_\nu \boldsymbol{\alpha}_\nu = \boldsymbol{\gamma}_\nu, \tag{4.36}$$

where $(\Gamma_\nu)_{i,\nu} = \gamma_{\nu-i}(\nu - i)$, for $i, \nu = 1, \ldots, p_\nu$, and $\boldsymbol{\gamma}_\nu = (\gamma_\nu(1), \ldots, \gamma_\nu(p_\nu))^\top$. From (4.4), the following equation is obtained

$$\lambda_\nu = \mu(\nu) - \sum_{i=1}^{p_\nu} \alpha_i(\nu) \mu(\nu - i). \tag{4.37}$$

Applying these equations, it is possible to estimate the coefficients separately for each season. Note that this involves an inversion of a $p_\nu \times p_\nu$ positive definite matrix for each season ν.

Since the PINAR$(p)_S$ model can be written as a PAR model, the asymptotic properties of the Yule-Walker estimates of its parameters behave similarly as in the case of autoregressive model discussed in (Brockwell & Davis 2013, chaps. 7& 8). The estimators $\widehat{\mu}_\nu$ and $\widehat{\gamma}_\nu$, $\nu = 1, \ldots, S$, are strongly consistent and asymptotically normal. Similarly, the Yule-Walker estimator $\widehat{\boldsymbol{\vartheta}}^{\mathrm{YW}} = (\widehat{\vartheta}_1^{\mathrm{YW}}, \ldots, \widehat{\vartheta}_S^{\mathrm{YW}})$, where $\widehat{\vartheta}_\nu^{\mathrm{YW}} = (\widehat{\alpha}_1^{\mathrm{YW}}(\nu), \ldots, \widehat{\alpha}_{p_\nu}^{\mathrm{YW}}(\nu), \widehat{\lambda}_\nu^{\mathrm{YW}})$ for each $\nu = 1, \ldots, S$, is strongly consistent and we obtain the following theorem. Note that these results are analogous to those derived for INAR(p) processes, see Theorem 3.1 and Section 4.1 in Du & Li (1991).

3 .2 Conditional least squares estimation

The CLS-estimators $\widehat{\boldsymbol{\vartheta}}_n^{\mathrm{CLS}}$, $n \in \mathbb{N}$, of $\boldsymbol{\vartheta}$ are obtained by minimizing the expression

$$Q_n(\boldsymbol{\vartheta}) = \sum_{\nu=1}^{S} Q_{n,\nu}(\vartheta_\nu) = \frac{1}{2}\sum_{\nu=1}^{S}\sum_{k=1}^{n-1}(Y_{kS+\nu} - \mathsf{E}(Y_{kS+\nu}|\mathcal{F}_{kS+\nu-1}))^2. \tag{4.38}$$

The CLS-estimators $\widehat{\boldsymbol{\vartheta}}_n^{\mathrm{CLS}}$, $n \in \mathbb{N}$, of $\boldsymbol{\vartheta}$ are obtained by minimizing the expression

$$Q_n(\boldsymbol{\vartheta}) = \sum_{\nu=1}^{S} Q_{n,\nu}(\vartheta_\nu) = \sum_{\nu=1}^{S}\sum_{k=1}^{n-1}(Y_{kS+\nu} - \mathsf{E}(Y_{kS+\nu}|\mathcal{F}_{kS+\nu-1}))^2, \tag{4.39}$$

where, by (4.3),

$$\mathsf{E}(Y_{kS+\nu}|\mathcal{F}_{kS+\nu}) = \sum_{i=1}^{p_\nu}\alpha_i(\nu)Y_{kS+\nu-i} + \lambda_\nu. \tag{4.40}$$

Since, for $\nu \neq \nu^*$, with $\nu, \nu^* = 1, \ldots, S$, $Q_{n,\nu}(\vartheta_\nu)$ and $Q_{n,\nu^*}(\vartheta_\nu^*)$ are uncorrelated, then minimizing $Q_n(\boldsymbol{\vartheta})$ means to minimize individually each $Q_{n,\nu}(\vartheta_\nu)$ for $\nu = 1, \ldots, S$. Define the random vectors $\boldsymbol{U}_t = (Y_t, Y_{t-1}, \ldots, Y_{t-p_\nu+1}, 1)^\top$, $t \geq S$, and introduce, for all $\nu = 1, \ldots, S$,

$$\boldsymbol{Z}_\nu = \begin{bmatrix} Y_{S+\nu} \\ \vdots \\ Y_{(n-1)S+\nu} \end{bmatrix}, \qquad \boldsymbol{C}_\nu = \begin{bmatrix} \boldsymbol{U}_{S+\nu-1}^\top \\ \vdots \\ \boldsymbol{U}_{(n-1)S+\nu-1}^\top \end{bmatrix} = \begin{bmatrix} Y_{S+\nu-1} & Y_{S+\nu-2} & \cdots & Y_S+\nu-p_\nu & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ Y_{(n-1)S+\nu-1} & Y_{(n-1)S+\nu-2} & \cdots & Y_{(n-1)S+\nu-p_\nu} & 1 \end{bmatrix}. \tag{4.41}$$

$\boldsymbol{Z}_\nu$ is a $(n-1)$-dimensional random vector and $\boldsymbol{C}_\nu$ is a random matrix of dimension $(n-1) \times (p_\nu + 1)$. By (4.39) and (4.40),

$$Q_{n,\nu}(\vartheta_\nu) = \sum_{k=1}^{n-1}(Y_{kS+\nu} - \sum_{i=1}^{p_\nu}\alpha_i(\nu)Y_{kS+\nu-i} - \lambda_\nu)^2 = \|\boldsymbol{Z}_\nu - \boldsymbol{C}_\nu\vartheta_\nu\|^2, \tag{4.42}$$

for each $\nu = 1, \ldots, S$. Thus, the CLS-estimator $\vartheta_\nu^{\mathrm{CLS}}$ of the parameter ϑ_ν can be expressed as

$$\vartheta_\nu^{\mathrm{CLS}} = \left(\boldsymbol{C}_\nu^\top\boldsymbol{C}_\nu\right)^{-1}\boldsymbol{C}_\nu^\top Z_\nu, \tag{4.43}$$

for each season $\nu = 1, \ldots, S$ (see, Theorem 7.2.2, in Bickel & Doksum (1977)).

One can see that the real-valued penalty function $Q_n(\nu)$ satisfies the assumptions of Theorem 3.2.24 in Taniguchi & Kakizawa (2000), see also Theorem 2.1 and Theorem 2.2 in Klimko & Nelson (1978). Define the matrices V_ν and R_ν, $\nu = 1, \ldots, S$ and $t = kS + \nu$, of dimension 3×3

as

$$V_\nu = \mathsf{E}\left(\boldsymbol{U}_{t-1}\boldsymbol{U}_{t-1}^\top\right),$$

$$V_\nu = \begin{bmatrix} \gamma_{\nu-1}(0) & \gamma_{\nu-1}(1) & \gamma_{\nu-1}(2) & \cdots & \gamma_{\nu-1}(p_\nu-1) & 0 \\ \gamma_{\nu-1}(1) & \gamma_{\nu-2}(0) & \gamma_{\nu-2}(1) & \cdots & \gamma_{\nu-2}(p_\nu-2) & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \gamma_{\nu-1}(p_\nu-1) & \gamma_{\nu-2}(p_\nu-2) & \gamma_{\nu-3}(p_\nu-3) & \cdots & \gamma_{\nu-p_\nu}(0) & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{bmatrix} + \mathsf{E}\left(\boldsymbol{U}_{t-1}\right)\mathsf{E}\left(\boldsymbol{U}_{t-1}\right)^\top, \tag{4.44}$$

where $\mathsf{E}\left(\boldsymbol{U}_{t-1}\right) = (\mu_{\nu-1}, \mu_{\nu-2}, \ldots, 1)^\top$ and $\mu_0 = \mu_S$, and

$$R_\nu = \mathsf{E}\left[\boldsymbol{U}_{t-1}(Y_t - \boldsymbol{U}_{t-1}^\top\vartheta_\nu)^2\boldsymbol{U}_{t-1}^\top\right], \tag{4.45}$$

Finally, define the block diagonal matrices V and W of dimension $p^* \times p^*$, where $p^* = p_1 + \ldots + p_S + S$, as

$$V = \mathrm{diag}\{V_1, \ldots, V_S\}, \qquad R = \mathrm{diag}\{R_1, \ldots, R_S\}. \tag{4.46}$$

The CLS-estimator $\widehat{\boldsymbol{\vartheta}}^{\mathrm{CLS}} = (\widehat{\vartheta}_1^{\mathrm{CLS}}, \ldots, \widehat{\vartheta}_S^{\mathrm{CLS}})$ of the parameter vector $\boldsymbol{\vartheta}$ is strongly consistent and the following theorem is obtained.

Theorem 6. *Assume that* $\{\boldsymbol{Y}_t\}_{t\in\mathbb{Z}}$ *in* (4.6)*, that is, the a PINAR*$(p)_S$ *process, is a strictly stationary ergodic process with* $\mathsf{E}\left\|\boldsymbol{Y}_t\right\|^2 < \infty$ *and* $\mathsf{E}(Y_{kS+\nu}|\mathcal{F}_{kS+\nu})$ *is almost surely three times continuously differentiable in the open set* Θ*. Then for the* CLS*-estimators* $\widehat{\boldsymbol{\vartheta}}_n^{\mathrm{CLS}}$*,* $n \in \mathbb{N}$*,*

$$n^{1/2}(\widehat{\boldsymbol{\vartheta}}_n^{\mathrm{CLS}} - \boldsymbol{\vartheta}_0) \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0}, V^{-1}RV^{-1}), \tag{4.47}$$

as $n \to \infty$*, where the matrices* V *and* R *of dimension* $p^* \times p^*$*,* $p^* = p_1 + \ldots + p_S + S$*, are defined in* (4.46)*.*

Proof. This proof follows the lines in Theorem 3.2.24 in Taniguchi & Kakizawa (2000), in straightforward generalization of the proof of Theorem 2 in PINAR$(1, 1_S)$ paper. □

3 .3 Quasi-maximum likelihood (QML)

We based our method on the approach of QML presented by Taniguchi & Kakizawa (2000), Section 3, page 101. First, let define

$$f_{\vartheta_\nu}(t, t-1) = \mathsf{E}\left[\{Y_t - m_{\vartheta_\nu}(t, t-1)\}^2\middle|\mathcal{F}_{t-1}\right],$$

where $t = kS + \nu$, and

$$m_{\vartheta_\nu}(t, t-1) = \mathsf{E}(Y_{kS+\nu}|\mathcal{F}_{kS+\nu-1}) = \mathsf{E}(Y_{kS+\nu}|\mathcal{F}_{kS+\nu}) = \sum_{i=1}^{p_\nu}\alpha_i(\nu)Y_{kS+\nu-i} + \lambda_\nu, \tag{4.48}$$

for $k = 1, \ldots, n-1$. The likelihood type penalty function of the PINAR$(p)_S$ model, conditioned on the first S observations, is given by

$$L_n(\boldsymbol{\vartheta}) = \sum_{k=1}^{n-1} \sum_{\nu=1}^{S} [\log\{f_{\vartheta_\nu}(t, t-1)\} + (Y_t - m_{\vartheta_\nu}(t, t-1))^2 f_{\vartheta_\nu}^{-1}(t, t-1)].$$

The likelihood function $L_n(\boldsymbol{\vartheta}) = \sum_{\nu=1}^{S} l_{n,\nu}(\vartheta_\nu)$, where

$$l_{n,\nu}(\vartheta_\nu) = \sum_{k=1}^{n-1} [\log\{f_{\vartheta_\nu}(t, t-1)\} + (Y_t - m_{\vartheta_\nu}(t, t-1))^2 f_{\vartheta_\nu}^{-1}(t, t-1)] l_{n,\nu}(\vartheta_\nu) = \sum_{k=1}^{n-1} \phi_t(\vartheta_\nu)$$

is minimized in order to obtain the QML-estimator $\widehat{\boldsymbol{\vartheta}}_n^{\mathrm{QML}}$ of the parameter vector $\boldsymbol{\vartheta}$.

Corollary 3. *The function $f_{\vartheta_\nu}(t, t-1)$ is given by*

$$f_{\vartheta_\nu}(t, t-1) = \sum_{i=1}^{p_\nu} \alpha_i(\nu)(1 - \alpha_i(\nu)) Y_{kS+\nu-i} + \lambda_\nu, \tag{4.49}$$

in the case of Poisson innovations.

The function $l_{n,\nu}(\vartheta_\nu)$ can be obtained directly by replacing the results of (4.49) and (4.48) in (4.49). From the second order stationarity of $\{\boldsymbol{Y}_t\}$ it follows that $\mathsf{E}\,\|\varepsilon_t\|^6 < \infty$ implies $\mathsf{E}\,\|Y_t\|^6 < \infty$, then one can prove that the real-valued penalty function $L_n(\boldsymbol{\vartheta})$ satisfies the assumptions of Theorem 3.2.26 in Taniguchi & Kakizawa (2000). Thus, there exists a sequence of estimators $\widehat{\boldsymbol{\vartheta}}_n^{\mathrm{QML}} = ((\widehat{\vartheta}_{n,1}^{\mathrm{QML}})^\top, \ldots, (\widehat{\vartheta}_{n,S}^{\mathrm{QML}})^\top)^\top$ such that $\widehat{\boldsymbol{\vartheta}}_n^{\mathrm{QML}} \to \boldsymbol{\vartheta}_0$ almost surely as $n \to \infty$, and for any $\epsilon > 0$, there exists an event E with $\mathsf{P}(E) > 1 - \epsilon$ and an $n_0 \in \mathbb{N}$ such that on E, for $n > n_0$, $\widehat{\boldsymbol{\vartheta}}_n^{\mathrm{QML}}$ is the solution of

$$\frac{\partial}{\partial \boldsymbol{\vartheta}} L_n(\boldsymbol{\vartheta}) = 0, \tag{4.50}$$

which attains a relative minimum of the likelihood function $L_n(\boldsymbol{\vartheta})$.

The minimization of $L_n(\boldsymbol{\vartheta})$ can be done separately by minimizing the partial log-likelihood $l_{n,\nu}(\vartheta_\nu)$ for each season $\nu \in \{1, \ldots, S\}$. Similarly, one can solve the likelihood equation (4.50) by solving the partial likelihood equations

$$\frac{\partial}{\partial \vartheta_\nu} l_{n,\nu}(\vartheta_\nu) = 0, \qquad \nu = 1, \ldots, S,$$

separately.

Define IF_ν the matrix of dimension $(p_\nu + 1) \times (p_\nu + 1)$ for each season $\nu \in \{1, \ldots, S\}$ as

$$IF_\nu = U_{\vartheta_\nu}^{-1} V_{\vartheta_\nu} U_{\vartheta_\nu}^{-1}, \tag{4.51}$$

where

$$V_{\vartheta_\nu} = \mathsf{E}\left\{ \frac{\partial}{\partial \vartheta_\nu} \phi_t(\vartheta_\nu) \frac{\partial}{\partial \vartheta_\nu^\top} \phi_t(\vartheta_\nu) \right\}, \tag{4.52}$$

and

$$U_{\vartheta_\nu} = \mathsf{E}\left\{\frac{\partial^2}{\partial\vartheta_\nu \partial\vartheta_\nu^\top}\phi_t(\vartheta_\nu)\right\}. \tag{4.53}$$

Note that $\frac{\partial}{\partial\vartheta_\nu}\phi_t(\vartheta_\nu) = (\frac{\partial}{\partial\alpha_1(\nu)}\phi_t(\vartheta_\nu), \ldots, \frac{\partial}{\partial\alpha_{p_\nu}(\nu)}\phi_t(\vartheta_\nu), \frac{\partial}{\partial\lambda_\nu}\phi_t(\vartheta_\nu))$ is a $(p_\nu + 1)$-dimensional vector. Then, the matrix IF of the PINAR$(p)_S$ process is defined as the block diagonal matrix

$$IF = \mathrm{diag}\{IF_1, \ldots, IF_S\}. \tag{4.54}$$

The following theorem on the asymptotic normality of the QML-estimator $\widehat{\boldsymbol{\vartheta}}^{\mathrm{QML}}$ is given below.

Theorem 7. *Assume that* $\{\boldsymbol{Y}_t\}_{t\in\mathbb{Z}}$ *in* (4.6)*, that is, the a PINAR*$(p)_S$ *process, is a strictly stationary ergodic process with* $\mathsf{E}\,\|\boldsymbol{\varepsilon}_k\|^6 < \infty$ ($\mathsf{E}\,\|\varepsilon_t\|^6 < \infty$ *in model* (4.3))*, and* $m_{\vartheta_\nu}(t, t-1)$ *and* $f_{\vartheta_\nu}(t, t-1)$ *are almost surely three times continuously differentiable in the open set* Θ *containing the true parameter value* $\boldsymbol{\vartheta}_0$*. Then, the* QML *estimators* $\widehat{\boldsymbol{\vartheta}}_n^{\mathrm{QML}}$ *are asymptotically normal distributed as*

$$n^{1/2}(\widehat{\boldsymbol{\vartheta}}_n^{\mathrm{QML}} - \boldsymbol{\vartheta}_0) \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0}, IF), \tag{4.55}$$

as $n \to \infty$*, where* IF *is the matrix of dimension* $p^* \times p^*$*,* $p^* = p_1 + \ldots + p_S + S$ *defined by* (4.54)*.*

Proof. This proof follows the lines in Theorem 3.2.26 in Taniguchi & Kakizawa (2000), in straightforward generalization of the proof of Theorem 3 in PINAR$(1, 1_S)$ paper. □

4 Monte Carlo simulations

In this section, we evaluate the behavior of the estimation methods discussed previously, that is, YW, CLS, and QML. The simulated model is a Poisson-PINAR$(3)_4$ model with $\vec{p} = (1, 2, 1, 3)$, $S = 4$, with sample size $n = 50, 200, 500$ per season. Each model was simulated 500 times. The sets of true parameters are given by $\alpha = \{0.49, 0.12, 0.27, 0.28, 0.30, 0.15, 0.22\}$ and $\lambda = \{1.50, 2.50, 5.25, 2.80\}$. The results are shown in table 4.1, where the bias and MSE (mean square error) are the mean of 500 estimates of each parameter. The initial values for the QML were obtained by the YW estimator. The simulated series were generated in the statistical software R (R Development Core Team, 2009). For the QML method, we applied a general non-linear optimization procedure using augmented Lagrange multiplier method with numerical derivatives as implemented in the *solnp* function of R.

In general, the results presented in tables suggest that the YW, CLS and QML estimators have good finite sample properties. These results show that the YW and CLS methods were outperformed by the QML procedure. When compared with the YW and CLS estimates, QML estimates present smaller biases and MSE.

Table 4.1: Poisson-PINAR$(3)_4$ model with $\vec{p} = (1, 2, 1, 3)$. 500 replications. The sets of true parameters are given by alphas=$\{$0.49, 0.12, 0.27, 0.28, 0.30, 0.15, 0.22$\}$ and lambdas=$\{$1.50, 2.50, 5.25, 2.80$\}$.

			n=50			
	BiasQML	MSEQML	BiasYW	MSEYW	BiasCLS	MSECLS
$\alpha_1(1)$	-0.004	0.011	0.005	0.010	0.007	0.010
$\alpha_2(1)$	-0.023	0.020	0.004	0.031	-0.011	0.031
$\alpha_2(2)$	0.011	0.021	-0.001	0.021	0.015	0.021
$\alpha_3(1)$	0.003	0.031	0.011	0.031	-0.002	0.030
$\alpha_4(1)$	0.010	0.021	0.010	0.021	0.006	0.020
$\alpha_4(2)$	-0.021	0.023	0.006	0.032	0.002	0.030
$\alpha_4(3)$	-0.008	0.021	0.002	0.030	0.014	0.030
λ_1	0.023	0.260	-0.031	0.500	-0.039	0.431
λ_2	0.035	0.591	-0.007	0.820	-0.053	0.712
λ_3	-0.046	0.721	-0.085	0.780	0.018	0.821
λ_4	0.097	1.000	-0.075	1.300	-0.119	1.601

			n=200			
	BiasQML	MSEQML	BiasYW	MSEYW	BiasCLS	MSECLS
$\alpha_1(1)$	0.001	0.000	0.004	0.000	0.002	0.000
$\alpha_2(1)$	0.001	0.010	0.004	0.010	0.005	0.010
$\alpha_2(2)$	0.002	0.000	0.000	0.010	-0.003	0.010
$\alpha_3(1)$	-0.003	0.010	-0.001	0.010	0.000	0.010
$\alpha_4(1)$	-0.001	0.000	-0.002	0.001	0.005	0.010
$\alpha_4(2)$	0.003	0.011	0.004	0.010	0.002	0.010
$\alpha_4(3)$	0.006	0.010	0.006	0.011	0.004	0.010
λ_1	-0.003	0.050	-0.028	0.110	-0.017	0.100
λ_2	-0.017	0.130	-0.017	0.160	0.000	0.170
λ_3	0.005	0.151	-0.001	0.151	0.010	0.172
λ_4	-0.023	0.283	-0.024	0.350	-0.059	0.410

			n=500			
	BiasQML	MSEQML	BiasYW	MSEYW	BiasCLS	MSECLS
$\alpha_1(1)$	0.000	0.000	0.000	0.000	0.001	0.000
$\alpha_2(1)$	-0.003	0.000	-0.001	0.000	0.001	0.000
$\alpha_2(2)$	0.000	0.000	0.002	0.000	-0.002	0.000
$\alpha_3(1)$	-0.001	0.000	0.002	0.000	0.000	0.000
$\alpha_4(1)$	-0.002	0.000	0.002	0.000	0.001	0.000
$\alpha_4(2)$	0.004	0.000	0.000	0.000	0.002	0.000
$\alpha_4(3)$	-0.002	0.000	0.004	0.000	0.003	0.000
λ_1	-0.001	0.020	-0.003	0.050	-0.001	0.040
λ_2	0.021	0.060	0.000	0.080	0.001	0.060
λ_3	0.000	0.060	0.000	0.070	0.002	0.070
λ_4	-0.004	0.110	-0.027	0.130	-0.033	0.130

5 Real data application

5 .1 The data

This application is based on the time series of counts referring to the daily number of people who got medicine based on salbutamol sulphate for the treatment of respiratory problems from the public health care system in the hospital emergency service of the region of Vitória-ES, Brazil. These data were obtained from the network records system Welfare of the municipality. The series corresponds from January 03, 2013 to July 18, 2017, resulting in 1659 daily observations. Figure 4.1 displays the plot of the real data, from this it is observed a persistence oscillation aspect, i.e., the mean changes periodically. This phenomenon is also clearly evidenced in the plots of Figure 4.2. In addition, the series correspond to daily data, which corroborates for S= 7 to be our choice for the period length.

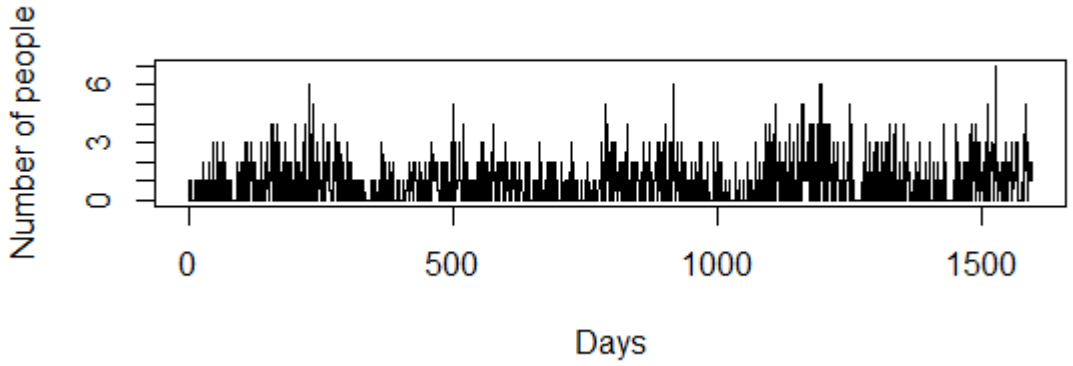


Figure 4.1: Daily number of people who received medicine based on salbutamol sulphate indicated for the relief of bronchial spasm associated with asthma attacks, chronic bronchitis and emphysema from the public health care system in the hospital emergency service of the region of Vitória-ES.

5 .2 Data analysis and discussion

Figure 4.2 shows the sample periodic mean and variance of the series varying over the seasons $j = 1, \ldots, 7$ with $S = 7$. According to McLeod (1994) one can identify the AR order of each season by finding the lowest lag for which the sample PePACF cuts off. Tables 4.2, 4.3, 4.5 and 4.6 were obtained through the R's Pear package based on McLeod (1994). The elements in bold represent values that have exceeded the confidence interval. We analyze the sample PePACF and PeACF of the real data in Tables 4.2 and 4.3, respectively, regarding to a parsimonious choice of the AR orders of the model, and the use of the Poisson-PINAR$(7)_7$ model with $\vec{p} = (1, 4, 4, 1, 3, 7, 6)$ is suggested as a candidate to fit the real data.

The parameters were estimated by quasi-maximum likelihood method and the results are displayed in Table 4.4. The estimated residuals $\hat{e}_t$ after fitting the Poisson-PINAR$(7)_7$ model were defined by

$$\hat{e}_t = Y_t - \sum_{i=1}^{p_\nu} \widehat{\alpha}_\nu Y_{t-i} - \widehat{\lambda}_\nu,$$

where $\nu = \{t\}_7$.

Following Bu et al. (2008), one can assess the adequacy of the fitted model examining the residuals for serial dependence. The sample PeACFs and PePACFs of residuals are shown at the Tables 4.5 and 4.6. By the analysis of these tables and the ACF and PACF of the residuals shown in Table 4.3, we conclude that the suggested model was able to filter the expressive autocorrelations, especially in the lowers lags (up to lag 14), which indicate that the residuals are not correlated and that there is no particular point causing a expressive impact in the model.

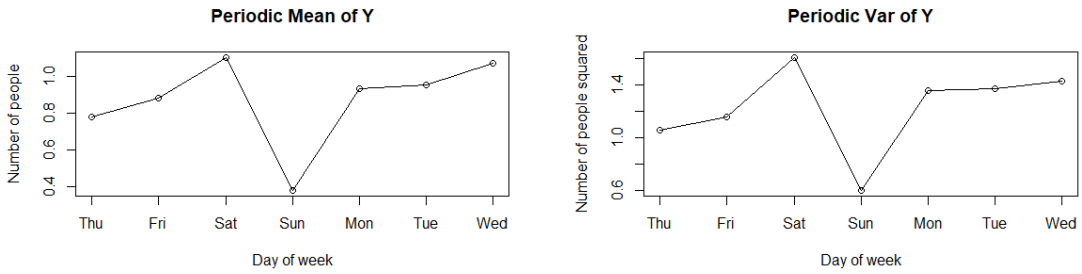


Figure 4.2: Plots of the means and variances of the seasons of the real data.

Table 4.2: Sample periodic ACF of the real data.

	Sample periodic ACF of the real data														
	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=7$	$h=8$	$h=9$	$h=10$	$h=11$	$h=12$	$h=13$	$h=14$	$h=15$
$\nu=1$	**0.16**	**0.14**	**0.20**	0.07	0.12	**0.16**	0.05	0.12	0.13	**0.17**	0.08	0.11	0.05	0.12	-0.01
$\nu=2$	0.07	0.09	**0.20**	**0.24**	0.02	**0.23**	**0.14**	**0.18**	0.10	**0.17**	0.02	0.04	**0.20**	0.00	0.06
$\nu=3$	**0.19**	**0.14**	**0.24**	**0.27**	**0.14**	-0.04	**0.27**	**0.14**	0.04	**0.14**	**0.23**	0.13	0.07	**0.17**	**0.14**
$\nu=4$	0.06	-0.05	0.05	0.02	0.06	0.06	0.06	0.07	-0.01	0.10	-0.02	-0.12	-0.04	0.07	0.08
$\nu=5$	0.06	**0.16**	**0.27**	0.09	0.10	0.12	**0.14**	0.07	**0.20**	0.08	**0.15**	0.10	**0.18**	0.13	0.05
$\nu=6$	**0.19**	-0.03	**0.28**	**0.22**	**0.15**	**0.14**	**0.30**	0.09	-0.07	**0.23**	**0.21**	0.09	0.09	**0.29**	0.13
$\nu=7$	**0.30**	**0.25**	-0.08	**0.23**	**0.14**	**0.20**	**0.14**	**0.17**	**0.17**	-0.05	0.06	0.04	0.01	0.12	**0.17**

Table 4.3: Sample periodic PACF of the real data.

Sample periodic PACF of the real data															
	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=7$	$h=8$	$h=9$	$h=10$	$h=11$	$h=12$	$h=13$	$h=14$	$h=15$
$\nu=1$	**0.16**	0.09	**0.16**	0.07	0.05	0.10	0.00	0.08	0.04	0.09	0.08	0.00	-0.02	0.05	-0.08
$\nu=2$	0.07	0.08	**0.18**	**0.20**	0.02	**0.16**	0.03	0.13	0.02	0.07	-0.08	0.04	0.08	-0.09	-0.01
$\nu=3$	**0.19**	0.13	**0.21**	**0.18**	0.02	-0.03	**0.16**	0.03	-0.06	0.05	0.10	0.05	0.10	0.03	0.06
$\nu=4$	0.06	-0.06	0.04	0.01	0.06	0.06	0.06	0.05	-0.05	0.10	-0.05	**-0.17**	-0.07	0.06	0.08
$\nu=5$	0.06	**0.16**	**0.25**	0.05	0.05	0.02	0.04	0.07	0.11	0.00	0.08	0.03	0.10	0.07	0.02
$\nu=6$	**0.19**	-0.04	**0.26**	**0.15**	0.11	0.06	**0.20**	-0.04	-0.07	0.09	0.11	0.02	-0.02	**0.15**	0.05
$\nu=7$	**0.30**	**0.21**	-0.10	**0.14**	0.01	**0.15**	0.05	0.04	0.09	-0.06	-0.10	-0.09	-0.05	0.07	0.02

Table 4.4: The estimated parameters of a Poisson-PINAR$(7)_7$ model with $\vec{p}=(1,4,4,1,3,7,6)$ using QML estimation. The standard error of is below each estimate, inside parenthesis. Values are rounded to three decimal places.

	$\nu=1$	$\nu=2$	$\nu=3$	$\nu=4$	$\nu=5$	$\nu=6$	$\nu=7$
$\alpha_1(\nu)$	0.099	0.023	0.147	0.042	0.095	0.117	0.169
	(0.046)	(0.066)	(0.060)	(0.036)	(0.072)	(0.058)	(0.081)
$\alpha_2(\nu)$		0.000	0.092		0.097	0.000	0.169
		(0.092)	(0.064)		(0.049)	(0.916)	(0.062)
$\alpha_3(\nu)$		0.156	0.176		0.240	0.126	0.000
		(0.061)	(0.058)		(0.063)	(0.051)	(1.111)
$\alpha_4(\nu)$		0.161	0.172			0.136	0.107
		(0.058)	(0.058)			(0.079)	(0.065)
$\alpha_5(\nu)$						0.124	0.029
						(0.059)	(0.060)
$\alpha_6(\nu)$						0.019	0.184
						(0.063)	(0.123)
$\alpha_7(\nu)$						0.147	
						(0.059)	
λ_ν	0.674	0.573	0.555	0.335	0.575	0.335	0.469
	(0.077)	(0.099)	(0.092)	(0.054)	(0.084)	(0.218)	(0.160)

Table 4.5: The sample PeACF of the residuals after fitting the Poisson-PINAR$(7)_7$ model with $\vec{p}=(1,4,4,1,3,7,6)$.

Periodic ACF of the residuals															
	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=7$	$h=8$	$h=9$	$h=10$	$h=11$	$h=12$	$h=13$	$h=14$	$h=15$
$\nu=1$	-0.03	0.03	0.13	0.08	0.04	0.11	0.02	0.03	0.04	0.13	0.08	0.08	-0.01	0.13	-0.11
$\nu=2$	-0.01	-0.06	-0.07	0.00	0.01	0.12	0.03	0.13	0.03	0.09	-0.03	0.03	**0.14**	-0.06	0.02
$\nu=3$	0.00	0.01	-0.02	-0.01	0.00	-0.04	0.13	0.02	-0.06	0.03	0.09	0.03	0.08	0.07	0.07
$\nu=4$	-0.01	-0.09	0.04	-0.04	0.06	0.05	0.06	0.08	0.00	0.09	-0.02	**-0.17**	-0.08	0.06	0.06
$\nu=5$	0.00	-0.01	0.01	0.05	0.00	-0.02	0.04	0.06	0.08	-0.02	0.09	0.02	0.12	0.11	0.03
$\nu=6$	-0.02	-0.06	0.04	0.00	-0.02	-0.01	0.01	-0.07	-0.09	0.05	0.09	0.02	-0.06	0.13	0.05
$\nu=7$	0.04	0.02	-0.12	-0.01	-0.03	-0.03	0.05	0.07	0.11	-0.06	-0.09	-0.06	-0.07	0.07	0.07

Table 4.6: The sample PePACF of the residuals after fitting the Poisson-PINAR$(7)_7$ model with $\vec{p}=(1,4,4,1,3,7,6)$.

Periodic PACF of the residuals															
	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=7$	$h=8$	$h=9$	$h=10$	$h=11$	$h=12$	$h=13$	$h=14$	$h=15$
$\nu=1$	-0.03	0.03	**0.14**	0.08	0.04	0.11	0.01	0.05	0.05	0.13	0.08	0.04	-0.02	0.09	-0.11
$\nu=2$	-0.01	-0.06	-0.07	0.00	0.00	0.12	0.03	0.13	0.04	0.10	-0.04	0.03	0.12	-0.07	0.03
$\nu=3$	0.00	0.01	-0.02	-0.01	0.00	-0.04	0.13	0.01	-0.06	0.03	0.10	0.05	0.11	0.05	0.09
$\nu=4$	-0.01	-0.09	0.04	-0.04	0.05	0.05	0.06	0.09	0.01	0.10	-0.01	**-0.16**	-0.11	0.05	0.05
$\nu=5$	0.00	-0.01	0.01	0.05	0.00	-0.02	0.04	0.06	0.09	-0.02	0.08	0.02	0.12	0.09	0.03
$\nu=6$	-0.02	-0.06	0.04	-0.01	-0.02	-0.02	0.01	-0.06	-0.09	0.05	0.08	0.04	-0.06	0.14	0.06
$\nu=7$	0.04	0.02	-0.11	-0.01	-0.04	-0.02	0.05	0.07	0.13	-0.04	-0.08	-0.07	-0.06	0.07	0.05

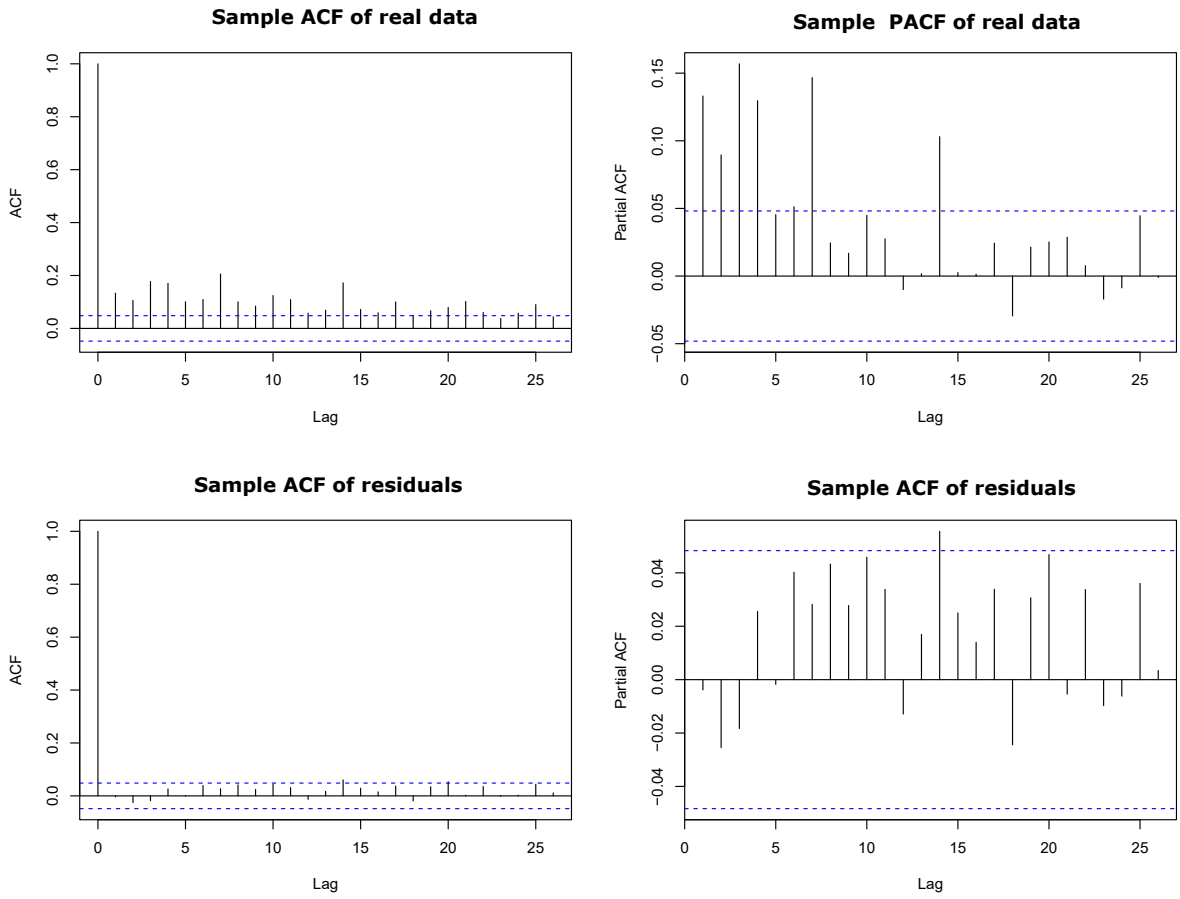


Figure 4.3: The sample ACF and PACF of the real data and of the residuals after fitting Poisson-PINAR$(7)_7$ with $\vec{p} = (1, 4, 4, 1, 3, 7, 6)$.

All these empirical analyses, i.e. the plots of PeACF and PePACF of the residuals, support the fact that the Poisson-PINAR$(1, 4, 4, 1, 3, 7, 6)$ seems to be well fitted to the real data.

6 Conclusions

The PINAR$(p)_S$ model was introduced in this paper. The main properties of this model are presented, such as the mean, autocorrelation function and transition probabilities. Several methods for estimating the model parameters were considered and a simulation study was performed to investigate the sample properties of YW, CLS and QML estimates. As expected the QML procedure outperforms the YW and CLS procedures, and therefore, this is the preferred method for model fitting.

To illustrate the proposed model, we considered a time series of counts on daily number of people who received medicine for the treatment of respiratory problems from the public health care system in the emergency service of the region of Vitória-ES (Brazil). This medicine is a medication used for opens up the airways in the lungs. It is mainly used to treat asthma and chronic obstructive pulmonary disease. It is a medication only dispensed with a medical prescription for a certain period, which provides more precise data. In fact, drugs used to treat asthma attacks are used on demand at times of crisis Zeghnoun et al. (1999).

Based on residual analysis, the Poisson-PINAR$(7)_7$ model was able to capture the main dynamic the real data series, that is, periodicity in the data.

The results presented in this paper will hopefully stimulate further research on this theme. For future work, it would be desirable to consider procedures that produce coherent forecasting. In addition, an extension of periodic models for integer-valued time series, taking into account different marginal distributions for innovations, such as Poisson inflated of zeros or negative binomial. Additionally, to extend the PINAR$(p)_S$ model to incorporate explanatory covariates.

Appendix

Proof of (4.19)**.** If $h \geq 0$ then we have

$$(\Gamma(h))_{i,\nu} = (\Gamma(h,0))_{i,\nu} = \mathrm{Cov}(Y_{hS+i}, Y_\nu) = R(hS+i,\nu).$$

On the other hand, if $h > 0$ then we have

$$\begin{aligned}(\Gamma(-h))_{i,\nu} =&(\Gamma(h))_{\nu,i} = (\Gamma(h,0))_{\nu,i} = \mathrm{Cov}(Y_{hS+\nu}, Y_i)\\ =&R(hS+\nu,i) = R(i,hS+\nu) = R(-hS+i,\nu).\end{aligned}$$

□

Proof of (4.21)**.** Since A is non-singular we have that

$$A\Sigma = B\Sigma B^\top (A^\top)^{-1} + \Sigma_{\boldsymbol{M}}(A^\top)^{-1}.$$

Then the statement follows from the fact that $\Gamma(0) = \Sigma$ and $\Gamma(-1) = \Sigma(A^{-1}B)^\top$. □

Proof of (4.24)**.** By the definition of functions γ_ν, $\nu = 1, \ldots, S$, we obtain

$$\gamma_\nu(hS+i) = \gamma_\nu(hS+i+\nu, \nu) = \gamma_\nu(\nu, hS+i+\nu) = \gamma_\nu(-hS-i+(i+\nu), i+\nu).$$

Since $\gamma_\nu(-hS-i+(i+\nu), i+\nu) = \gamma_{i+\nu}(-hS-i)$ if $i+\nu \le S$ and $\gamma_\nu(-hS-i+(i+\nu), i+\nu) = \gamma_\nu(-hS-i+(i+\nu)-S, i+\nu-S) = \gamma_{i+\nu-S}(-hS-i)$ if $i+\nu > S$ we have the equation. □

Proof of (4.29)**.** Apply the definition of conditional probability and the independence of the thinning operators and immigration at time t from the past of the process $\{Y_t\}$:

$$\begin{aligned}
\mathsf{P}(Y_t = y_t | Y_{t-1} = y_{t-1}, \ldots, Y_1 = y_1) &= \frac{\mathsf{P}(Y_t = y_t, Y_{t-1} = y_{t-1}, \ldots, Y_1 = y_1)}{\mathsf{P}(Y_{t-1} = y_{t-1}, \ldots, Y_1 = y_1)} \\
&= \frac{\mathsf{P}(\sum_{i=1}^{p_\nu} \alpha_i(\nu) \circ y_{t-i} + \varepsilon_t = y_t, Y_{t-1} = y_{t-1}, \ldots, Y_1 = y_1)}{\mathsf{P}(Y_{t-1} = y_{t-1}, \ldots, Y_1 = y_1)} \\
&= \frac{\mathsf{P}(\sum_{i=1}^{p_\nu} \alpha_i(\nu) \circ y_{t-i} + \varepsilon_t = y_t)\mathsf{P}(Y_{t-1} = y_{t-1}, \ldots, Y_1 = y_1)}{\mathsf{P}(Y_{t-1} = y_{t-1}, \ldots, Y_1 = y_1)} \\
&= \mathsf{P}\left(\sum_{i=1}^{p_\nu} \alpha_i(\nu) \circ y_{t-i} + \varepsilon_t = y_t\right)
\end{aligned}$$

and similarly we have

$$\mathsf{P}(Y_t = y_t | Y_{t-1} = y_{t-1}, \ldots, Y_{t-p_\nu} = y_{t-p_\nu}) = \mathsf{P}(\sum_{i=1}^{p_\nu} \alpha_i(\nu) \circ y_{t-i} + \varepsilon_t = y_t)$$

where $\nu = \{t\}_S$ and $y_1, \ldots, y_t \in \mathbb{Z}_+$. □

Proof of (4.32)**.** Apply the definition of conditional probability and then the S-step Markov property:

$$\begin{aligned}
P(Y_t = y_t, \ldots, &Y_{S+1} = y_{S+1} | Y_s = y_s, \ldots, Y_1 = y_1) = \\
&= \frac{P(Y_t = y_t, \ldots, Y_1 = y_1)}{P(Y_{t-1} = y_{t-1}, \ldots, Y_1 = y_1)} \cdot \frac{P(Y_{t-1} = y_{t-1}, \ldots, Y_1 = y_1)}{P(Y_S = y_S, \ldots, Y_1 = y_1)} \\
&= P(Y_t = y_t | Y_{t-1} = y_{t-1}, \ldots, Y_1 = y_1) \times \\
&\quad P(Y_{t-1} = y_{t-1}, \ldots, Y_{S+1} = y_{S+1} | Y_S = y_S, \ldots, Y_1 = y_1) = \\
&= p_\nu(y_t | y_{t-1}, \ldots, y_{t-p_\nu}) P(Y_{t-1} = y_{t-1}, \ldots, Y_{S+1} = y_{S+1} | Y_S = y_S, \ldots, Y_1 = y_1),
\end{aligned}$$

where $\nu = \{t\}_S$ and $y_1, \ldots, y_t \in \mathbb{Z}_+$. □

Chapter 5

The regression ZIP-PINAR$(p)_S$ model

An extension of the PINAR$(p)_S$ model for count time series with inflation of zeros is introduced in this paper. The use of explanatory covariates is proposed in order to extend the applicability of the model. Statistical properties such as conditional mean and variance, marginal and joint probability distributions are presented. The quasi-maximum likelihood estimation of the parameters of the model was considered. The finite sample behavior of the estimators will be illustrated via computational simulations. The proposed methodology will be applied to model the relation between the count time series of the number of visits to the hospital emergency service for people with respiratory diseases (asthma and rhinitis) and the series of air pollutant concentrations.

This paper will be submitted to publication to the Journal of Statistical Planning and Inference.

The Regression ZIP-PINAR$(p)_S$ model

Abstract

This paper introduces the PINAR$(p)_S$ model for periodic count time series with inflation of zeros and covariates, denoted ZIP-PINAR$(p)_S$ model. The model properties such as stationarity and ergodicity as well as the asymptotic properties of the conditional quasi-maximum likelihood parameter estimators are fully established. Simulations were carried out in order to verify the estimation method performance for finite sample sizes. The count time series of the number of visits to the hospital emergency service for people with respiratory diseases (asthma and rhinitis) was analyzed using the proposed model with the covariate air pollutant concentrations.

Keywords: INAR, periodic stationarity, PINAR, zero inflated Poisson, quasi-maximum likelihood.

1 Introduction

In the literature of time series, a series with a excessive number of zeros is usually defined as time series with zero-inflation and this phenomenon is quite common in many area of application. For example, in the biomedical and public health domains, some types of rare diseases with low infection rates can lead to a count time series with a large number of zeros. Ignoring zeros in the data may conduct to a wrong model choice and inference and a spurious association between the count time series with covariates.

Yang et al. (2013) extended the classical regression based on the Zero Inflated Poisson (ZIP) distribution, introduced by Lambert (1992), for counting time series with an excess of zeros, autoregressive (AR) structure and time-dependent covariates regression framework. The ZIP distribution can be seen as a mixed distribution of a Poisson with parameter λ, and a degenerate component with all its mass at zero, parameter ρ, called the inflation parameter of zeros. The regression ZIP is referred here as ZIP model.

To illustrate the mechanism of this model, consider the process $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ of $\mathbb{Z}_+$-valued independent ZIP distributed random variables (r.v.'s) ε_t, with inflation of zeros parameter $\rho \in [0;1]$ and Poisson parameter $\lambda \in \mathbb{R}_+$, $\varepsilon_t \sim$ ZIP(ρ, λ). The probability mass function (p.m.f.) of ε_t is given by $P_{\varepsilon_t}(\varepsilon_t = m) = \rho\mathbb{I}_{m=0} + (1-\rho)\exp(-\lambda)\lambda^m/m!$, $m \in \mathbb{Z}_+$, where $\mathbb{I}_{m=0} = 1$, if $m = 0$ or $\mathbb{I}_{m=0} = 0$, if $m \neq 0$. The parameters ρ and λ relate the variable ε to the vectors of covariates $\boldsymbol{X}$ and $\boldsymbol{Z}$ through equations $\log(\lambda) = \boldsymbol{X}^\top\boldsymbol{\beta}$ and $\log[\rho/(1-\rho)] = \boldsymbol{Z}^\top\gamma$, where $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ represent the vectors of coefficients of the model. $X^\top$ is the transpose matrix of X.

A simple model for a stationary sequence of integer-valued random variables with lag-one dependence referred to as the integer-valued autoregressive of order one 1 (INAR1) model was introduced by Al-Osh & Alzaid (1987). This model has a special advantage over the ZIP model due its similarity to the Box and Jenkins ARMA models for continuous data. INAR model has the same additive structure of ARMA models instead of the multiplicative structure presented in ZIP. This additive characteristic and the discreteness of the modeled process are proportioned by the Thinning Operator $\circ$. INAR model can be represented as the process $Y_t = \alpha \circ Y_{t-1} + \varepsilon_t$, where $\alpha \circ Y_{t-1} = \sum_{i=1}^{Y_{t-1}} B_i(\alpha)$. $B_i(\alpha)$ represents a sequence of independent random variables (R.V.s) with Bernoulli distribution and probability of success P$(B_{i,t}(\alpha) = 1) = \alpha$, $0 \leq \alpha \leq 1$. In this case $\circ$ is called binomial thinning operator. For the Poisson INAR model, $\varepsilon_t \sim Poisson(\lambda)$ represents a sequence of independent R.V.s, assumed independent of Y_{t-1} and $\alpha \circ Y_{t-1}$.

A natural extension of Poisson INAR model to provide a consistent fit to count times series with over-dispersion proportioned by the excess of zeros is the first-order integer valued AR processes with zero inflated Poisson innovations (ZINAR) developed by Jazi et al. (2012). The ZINAR model has similar equation of INAR models, in which the innovations $\varepsilon_t \sim$ ZIP(ρ, λ).

The inclusion of explanatory variables to extend the applicability of INAR models was briefly introduced by Brännäs (1993) and studied by Enciso-Mora et al. (2009). The former authors introduced a model based on the INAR(p) model and developed an efficient Markov Chain Monte Carlo algorithm which analyze both explanatory variable an model order selection. The methodology was applied to the analysis of monthly polio incidences in the USA 1970-1983 and claims from the logging industry to the British Columbia Workers' Compensation Board 1985-1994.

The models discussed in the previous paragraphs are based on the assumption of stationarity in the mean and variance, that is, standard count time series models. However, it is quite common in many area of application to have time series that varies periodically in the mean, the variance and the autocovariance. This type of time series was introduced by Gladyshev (1961) as Periodically Correlated process (PC).

There is a lot of research related to PC processes for continuous time series. For a review, from theoretical and applied point of view, of the periodic autoregressive moving-average (PARMA) models, one can mention Gardner et al. (2006), Sarnaglia et al. (2010), Basawa & Lund (2001) among others. However, no much attention has been paid to the analysis of periodically correlated count series. For example, Monteiro et al. (2010) introduced the periodic integer-valued autoregressive model of order 1, with Poisson distributed data (PINAR). The stationarity and ergodicity properties of the process were established following the same lines in Latour (1997). The Yule-Walker based equations, least squares-type and quasi-maximum likelihood estimators of the model parameters were the estimation methods discussed.

Filho et al. (n.d.) extended the PINAR$(1)_S$ model to the PINAR$(p)_S$. Statistical properties of the model such as mean, variance, marginal and joint distributions are presented in the paper. The Moments-based, conditional least squares and quasi-maximum likelihood estimation methods

of the parameters were considered. An application to medication dispensing is given to show the usefulness of the proposed model.

Here, it is introduced an extension of PINAR$(p)_S$ model, denoted as ZIP-PINAR$(p)_S$ model, which takes into account periodic cont time series with zero-inflates and covariates as explanatory variables of the model.

In the remainder of this paper, let $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Z}_+$, $\mathbb{R}$ and $\mathbb{R}_+$ denote the set of positive integers, integers, non-negative integers, real numbers, and non-negative real numbers, respectively. Denote by I_d the $d \times d$ identity matrix. If it is clear from the context, then we omit the subscript d. $\mathrm{Bin}(n, \alpha)$ denotes a binomial distribution with parameters $n \in \mathbb{N}$ and $\alpha \in [0, 1]$; $\mathrm{Poi}(\lambda)$ denotes a Poisson distribution with mean parameter $\lambda \in \mathbb{R}_+$. Let $\mathsf{E}(\cdot)$ and $\mathsf{E}(\cdot|\cdot)$ represent the expectation and the conditional expectation, respectively. Random variables are all defined on a common probability space $(\omega, \mathcal{A}, \mathsf{P})$ and $\mathcal{F}_t$ denotes the σ-algebra generated by the random variables until time t, for all $t \in \mathbb{Z}$, $t > S$ and $\mathcal{F}_0 = \{\emptyset, \omega\}$.

The organization of the paper is as follows: Section 2 presents the PINAR$(p)_S$ model; Section 3 introduces the regression ZIP-PINAR$(p)_S$ model, and some of its statistical and probabilistic properties; The transition probability function of the process established on the ZIP model is introduced in Section 4; the Section 5 discusses the Quasi-Maximum Likelihood(QML) estimation method of the parameters of the model; Section 6 presents a set of simulations; a real data application is the motivation of the Section 7, and, finally, conclusions and final comments are presented at the last section. Some proofs are in the Appendix.

2 The PINAR$(p)_S$ model

Let $\{Y_t\}_{t\in\mathbb{Z}}$, $Y_t \in \mathbb{Z}_+$, be a stochastic process with seasonal characteristics of period S, $S \in \mathbb{N}$. The time index t may be written as the Euclidean division between t and S, i.e., as $t = kS + \nu$, where $k \in \mathbb{Z}$ and $\nu = 1, \ldots, S$. For example, in the case of daily data, $S = 7$, ν and k represent the day of the week and the week, respectively.

Define the mean function, $\mu(t) = \mathsf{E}(Y_t)$ for all $t \in \mathbb{Z}$, and the covariance function, the scalar $\gamma_{k,\nu}$, $\nu = 1, \ldots, S$ and $k \in \mathbb{Z}$, on $\mathbb{Z}$ as

$$\gamma_{k,\nu}(h) = \mathrm{Cov}(Y_{kS+\nu}, Y_{kS+\nu-h}), \qquad h \in \mathbb{Z}. \tag{5.1}$$

The stochastic process $\{Y_{kS+\nu}\}_{k\in\mathbb{Z},\nu=1,\ldots,S}$, satisfying $\mathsf{E}(Y^2_{kS+\nu}) < \infty$, is said to be a *periodically correlated process* (PC) of period S, $S \in \mathbb{N}$, if, for $\nu = 1, \ldots, S$ and all integers k,

$$\mu(kS + \nu) = \mu_\nu \quad \text{and} \quad \gamma_{k,\nu}(h) = \gamma_\nu(h).$$

That is, if mean and variance are finite and if they do not depend on k. Note that, if $\{Y_t\}_{t\in\mathbb{Z}}$ is a periodically correlated process then its mean and covariance are periodic functions with period S. If $S = 1$, then $\{Y_t\}$ represents a homogeneous stochastic process and the condition

of periodic stationarity is equivalent to the non-periodic ones.

A $\mathbb{Z}_+$-valued process $\{Y_t\}$ is said to be a *periodic non-negative integer-valued autoregression* (PINAR) with seasonal period S, for some $S \in \{2, 3, \ldots\}$, and is denoted by PINAR$(p)_S$, where $p = \max(\vec{p})$ and $\vec{p}$ is the $1 \times S$ vector of autoregressive orders of $\{Y_t\}$, if it satisfies the following stochastic recursion

$$Y_{\nu+kS} = \sum_{i=1}^{p_\nu} \alpha_i(\nu) \circ Y_{\nu-i+kS} + \varepsilon_{\nu+kS}, \tag{5.2}$$

where $k \in \mathbb{Z}$ and $t = \nu + kS$. In this paper assume that $p \leq S$. Because of the similarity of INAR models to the standard autoregressive (AR) model for continuous data, the $\alpha_i(\nu)$, $i = 1, \ldots, p_\nu$, $\nu = 1, 2, \ldots, S$ and $p_\nu = 1, \ldots, S$, are called autoregressive coefficients. The vector of AR orders $\vec{p}$ has the form $(p_1, p_2, \ldots, p_S)_{1\times S}$, $S \in \mathcal{N}$, where p_ν represents the AR order of the $\nu-$th season. For each season $\nu = 1, 2, \ldots, S$, the set of autoregressive coefficients has the form $\{\alpha_1(\nu), \ldots, \alpha_{p_\nu}(\nu)\} \subset [0; 1]^{p_\nu}$. The immigration process $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ is a periodic sequence of $\mathbb{Z}_+$-valued r.v.'s such that for each $\nu \in \{1, \ldots, S\}$ the sequence $\{\varepsilon_{kS+\nu}\}_{k\in\mathbb{Z}}$ consists of i.i.d.r.v.'s. The r.v.'s $Y_1, \ldots, Y_S$ are known as the starting values for the recursion (5.2). Finally, in (5.2), we assume that all counting random variables are mutually independent and they are independent of the sequence $\{\varepsilon_t\}_{t\in\mathbb{Z}}$.

It is assumed that the immigration process and starting values have finite second moments, i.e., the mean function μ and the variance of $\{Y_t\}_{t\in\mathbb{Z}}$ exist and are finite. Moreover, let $\mu_\varepsilon(\nu) = \mathrm{E}(\varepsilon_{kS+\nu})$, $\sigma^2_\nu = \mathrm{Var}(\varepsilon_{kS+\nu})$ for all $k \in \mathbb{Z}$ and $\nu = 1, \ldots, S$. As can be seen that in the seasonal period ν, Y_t in (5.2) has $p_\nu + 1$ random components; the immigration part of the past Y_{t-i}, $t = \nu + kS$ and $i = 1, \ldots, p_\nu$, with survival probability $\alpha_i(\nu)$ and the elements which entered in the system in the interval (t-1, t], which define the innovation term ε_t for all $t \in \mathbb{Z}$. Moreover, the autoregressive parameters $\alpha_i(\nu)$ and immigration means $\mu_\varepsilon(\nu)$, $\nu = 1. \ldots, S$, change periodically according to the seasonal period S.

The mean of the process $\{Y_t\}_{t\in\mathbb{Z}}$, $t > p$, where $p = \max(\vec{p})$, in (5.2) is given by:

$$\mu(kS + \nu) = \sum_{i=1}^{p_\nu} \alpha_i(\nu)\mu(kS + \nu - i) + \mu_\varepsilon(\nu). \tag{5.3}$$

A PINAR$(p)_S$ model is called proper if its parameters $\alpha_i(\nu), \mu_\varepsilon(\nu)$, $1 \leq i \leq p$ and $\nu = 1, \ldots, S$ are positive values. Consider that $\{Y_t\}_{t\in\mathbb{Z}}$ is a proper PINAR$(p)_S$ model and satisfies all of the statements of Lemma 1 in PINAR$(p)_S$ paper. Then, from the Theorem 1 in PINAR$(p)_S$ paper, the process $\{Y_t\}_{t\in\mathbb{Z}}$ is a second order periodically stationary process.

3 The regression ZIP-PINAR$(p)_S$ model

Let $\{\varepsilon_t\}_{t\in\mathbb{Z}}$, in (5.2), be a sequence of non-negative integer-valued R.V.s, where $\varepsilon_t \sim$ ZIP(ρ_ν, λ_t), with $t = \nu + kS$, $S \in \mathbb{N}$, $k \in \mathbb{Z}$, $\nu = 1, \ldots, S$, for $0 \leq \rho_\nu \leq 1$ and $\lambda_t \in \mathbb{R}_+$. The PINAR$(p)_S$ model in (5.2) is extended to the Periodic INteger of AutoRegressive order $(\vec{p})$ with Zero Inflated

Poisson distributed innovations. Suppose that for each t there are q explanatory variables. Let $X_{i,t}$, $i = 1, \ldots, q$, denote the value of the i-th explanatory variable at time t. The ZIP-PINAR$(p)_S$ model is introduced by the set of equations

$$\begin{cases} Y_{\nu+kS} = \sum_{i=1}^{p_\nu} \alpha_i(\nu) \circ Y_{\nu-i+kS} + \varepsilon_{\nu+kS} \\ \log \lambda_{\nu+kS} = \boldsymbol{X}_{\nu+kS-d}^\top \boldsymbol{\beta}_\nu, \end{cases} \tag{5.4}$$

where $\alpha_i(\nu) \in (0;1)$, $\nu = 1, \ldots, S$ and $i = 1, \ldots, p_\nu$. p_ν represents the AR order of the ν-th season of the period $S \in \mathbb{N}$. $\boldsymbol{\beta}_\nu = (\beta_{1,\nu}, ..., \beta_{q,\nu})^\top$, $q \in \mathcal{N}$,$\nu = 1, \ldots, S$, is a vector of parameters related to $\boldsymbol{X}_{t-d} = (X_{1,t-d}, \ldots, X_{q,t-d})^\top$, which is the vector of explanatory covariates at time $t - d$. $d \in \mathbb{Z}_+$ is a predefined constant which represents a time delay in the relation that expresses the influence of $\boldsymbol{X}_{t-d}$ on the parameter λ_t. For $d = 0$ this relation is parallel in time.

One may say that write $\alpha_i(t)$ instead of $\alpha_i(\nu)$ is more appropriate because this notation indicates that for each time t there is a sequence of variables $\{Y_t\}_{t\in\mathbb{Z}}$, nevertheless the notation $\alpha_i(\nu)$ will be utilized due to its closer similarity to the standard AR model coefficients.

The conditional probability mass function (p.m.f.) of $\{\varepsilon_t\}_{t\in\mathbb{Z}}$, $t = \nu + kS$, is given by

$$P_{\varepsilon_t}(\varepsilon_t = m|\rho_\nu, \lambda_t) = \rho_\nu \mathbb{I}_{m=0} + (1 - \rho_\nu) \text{ exp}(-\lambda_t)\frac{\lambda_t^m}{m!},\ m \in \mathbb{Z}_+, \tag{5.5}$$

where

$$\mathbb{I}_{m=0} = \begin{cases} 1, & \text{if } m = 0 \\ 0, & \text{if } m \neq 0 \end{cases}. \tag{5.6}$$

As pointed by Enciso-Mora et al. (2009), a set of explanatory variables can be used to model a linear trend or periodicity. Assume that the variables $\{X_{i,t}\}_{i\in\{1,...,q\},t\in\mathbb{Z}}$ are linearly independents.

where $\mu_{X_i}(\nu)$ do not depend on the K value.

The conditional mean of $\{\varepsilon_t\}_{t\in\mathbb{Z}}$, for $t = kS + \nu$, $k \in \mathbb{Z}$ and $\nu = 1, ..., S$, is given by

$$\mu_\varepsilon(\nu) = E(\varepsilon_t|\rho_\nu, \boldsymbol{\beta}_\nu, \boldsymbol{X}_{t-d}) = (1 - \rho_\nu)\lambda_t = (1 - \rho_\nu)\exp(\boldsymbol{X}_{t-d}^\top \boldsymbol{\beta}_\nu).$$

The conditional variance of ε_t is given by

$$\sigma_\varepsilon^2(\nu) = \text{VAR}[\varepsilon_t|\rho_\nu, \boldsymbol{\beta}_\nu, \boldsymbol{X}_{t-d}] = \mu_\varepsilon(\nu)(1 + \rho_\nu\lambda_t) = \mu_\varepsilon(\nu)(1 + \rho_\nu \exp(\boldsymbol{X}_{t-d}^\top \boldsymbol{\beta}_\nu)).$$

Now it is possible to obtain the marginal mean and variance of the process $\{y_t\}_{t\in\mathbb{Z}}$ in (5.4),

conditioned to the known covariates $\boldsymbol{X}_{t-d}$ and parameters $\alpha_i(\nu), \rho_\nu, \boldsymbol{\beta}_\nu,\ i = 1, \ldots, p_\nu$. For $t = kS + \nu$, the mean and the variance are given by, respectively,

$$\mu_\nu = \mathsf{E}(Y_t) = \sum_{i=1}^{p_\nu} \alpha_i(\nu)\mu_{\nu-i} + \mu_\varepsilon(\nu), \tag{5.7}$$

and

$$\gamma_\nu(0) = \mathrm{Var}(Y_t) = \sum_{i=1}^{p_\nu} \alpha_i^2(\nu - i)\,\mathrm{Var}(Y_{t-i}) + \alpha_i(\nu - i)(1 - \alpha_i(\nu - i))\mu_{\nu-i} + \sigma_\varepsilon^2(\nu). \tag{5.8}$$

4 The transition probability

The variable Y_t, $t = \nu + kS$ for $t > p$, is assumed to be generated according to (5.4). In the same sense of Bu et al. (2008), we propose a recursive representation of the transition probability of the model. For the periodic autoregressive process of order p, where $p = \max\{p_\nu\}_{\nu=1}^{S}$, established on Y_t, defined in (5.4), one can characterize this sequence, in terms of probability, as a state machine where each observation, from the $(p+1)-$th, represents a state partially dependent on the preceding states. The transition probability from a state where $Y_{t-p} = y_{t-p}, Y_{t-p+1} = y_{t-p+1}, ..., Y_{t-1} = y_{t-1}$ to state $Y_t = y_t$, in a single step, conditionally at the $\nu-$th section of S, can be written considering y_t as the convolution between $\alpha_1(\nu) \circ y_{t-1}$ and $\alpha_2(\nu) \circ y_{t-2} + ... + \alpha_{p_\nu}(\nu) \circ y_{t-p_\nu} + \varepsilon_t$, which are mutually independent, given the p past observed lags. In turn, $\alpha_2(\nu) \circ y_{t-2} + ... + \alpha_{p_\nu}(\nu) \circ y_{t-p_\nu} + \varepsilon_t$ can be seen as the convolution between $\alpha_2(\nu) \circ y_{t-2}$ and $\alpha_3(\nu) \circ y_{t-3} + ... + \alpha_{p_\nu}(\nu) \circ y_{t-p_\nu} + \varepsilon_t$ leading us to a clear recursion. Thus, for $t = kS + \nu$, the recursive form for the transition probability is given by

$$\begin{aligned}
&P(Y_t = y_t | Y_{t-1} = y_{t-1}, \cdots, Y_{t-p_\nu+1} = y_{t-p_\nu+1}, Y_{t-p_\nu} = y_{t-p_\nu}) = p_\nu(y_{kS+\nu} | y_{kS+\nu-1}, \ldots, y_{kS+\nu-p_\nu}) \\
&= \sum_{i_1=0}^{min(y_{t-1}, y_t)} \binom{y_{t-1}}{i_1} (\alpha_1(\nu))^{i_1} (1 - \alpha_1(\nu))^{y_{t-1}-i_1} \sum_{i_2=0}^{min(y_{t-2}, y_t - i_1)} \binom{y_{t-2}}{i_2} (\alpha_2(\nu))^{i_2} (1 - \alpha_2(\nu))^{y_{t-2}-i_2} \\
&\quad \sum_{i_{p_\nu}=0}^{min(y_{t-p_\nu}, y_t - (i_1+i_2+...+i_{p_\nu-1}))} \binom{y_{t-p_\nu}}{i_{p_\nu}} (\alpha_{p_\nu}(\nu))^{i_{p_\nu}} (1 - \alpha_{p_\nu}(\nu))^{y_{t-p_\nu}-i_{p_\nu}} \{\rho_\nu \mathbb{I}_{y_t-(i_1+i_2+...+i_{p_\nu})=0} + \\
&\qquad (1 - \rho_\nu)\, \exp(-\lambda_t) \frac{\lambda_t^{y_t-(i_1+i_2+...+i_{p_\nu})}}{(y_t - (i_1 + i_2 + ... + i_{p_\nu}))!}\}.
\end{aligned} \tag{5.9}$$

With $\lambda_t = \exp(\boldsymbol{X}_{t-d}^\top \boldsymbol{\beta}_\nu)$.

5 The quasi-maximum likelihood (QML) method

In this section, the quasi-maximum likelihood (QML) estimation method is discussed for the model in (5.4). The asymptotic properties of these estimators are also investigated.

Let $\vartheta_\nu = (\boldsymbol{\alpha}_\nu^\top, \boldsymbol{\beta}_\nu^\top, \rho_\nu)^\top$, where $\boldsymbol{\alpha}_\nu = (\alpha_1(\nu), \ldots, \alpha_{p_\nu}(\nu))^\top$, $\alpha_i(\nu) \in (0;1)$, $i = 1, \ldots, p_\nu$, $\rho_\nu \in [0;1]$ and $\boldsymbol{\beta}_\nu = (\beta_{1,\nu}, ..., \beta_{q,\nu})^\top$, $q \in \mathcal{N}$, $\beta_{i,\nu} \in \mathbb{R}$, for $i = 1, \ldots, q$ and $\nu = 1, \ldots, S$. Let $\boldsymbol{\vartheta} = (\vartheta_1, \ldots, \vartheta_S)$ represent the p^*-dimensional, where $p^* = p_1 + \ldots + p_S + S(q+1)$, unknown parameter vector of the ZIP-PINAR$(p)_S$ model defined by (5.4). In this paper, all the parameter vectors are column vector. The parameter vector is assumed to be lying in the open set Θ. Its true value is denoted by $\boldsymbol{\vartheta}_0$. Consider a sample $y_1, \ldots, y_T$ of size $T = nS$ where $n \in \mathbb{N}$ for the ZIP-PINAR$(p)_S$ process $\{Y_t\}_{t\in\mathbb{Z}}$. By (5.4) and (5.9) it is possible to conjecture that all estimators of the parameter ϑ_ν will depend on the sequence of data $\{y_{kS+\nu-p_\nu}, \ldots, y_{kS+\nu}\}_{k=1,\ldots,n-1}$, respectively, for each season $\nu \in \{1, \ldots, S\}$.

We present a likelihood type penalty function of the PINAR$(p)_S$ model, conditioned on the first S observations, based on the approach of QML introduced by Taniguchi & Kakizawa (2000). The function to be minimized is given by

$$L_n(\boldsymbol{\vartheta}) = \sum_{k=1}^{n-1}\sum_{\nu=1}^{S}[\log\{f_{\vartheta_\nu}(t,t-1)\} + (Y_t - m_{\vartheta_\nu}(t,t-1))^2 f_{\vartheta_\nu}^{-1}(t,t-1)],$$

where

$$f_{\vartheta_\nu}(t,t-1) = \mathsf{E}\left[\{Y_t - m_{\vartheta_\nu}(t,t-1)\}^2 \middle| \mathcal{F}_{t-1}\right],$$

and

$$m_{\vartheta_\nu}(t,t-1) = \mathsf{E}(Y_{kS+\nu}|\mathcal{F}_{kS+\nu-1}) = \mathsf{E}(Y_{kS+\nu}|\mathcal{F}_{kS+\nu}) = \sum_{i=1}^{p_\nu}\alpha_i(\nu)Y_{kS+\nu-i} + \mu_\varepsilon(\nu), \qquad (5.10)$$

for $k = 1, \ldots, n-1$ and $t = kS + \nu$.

The conditional likelihood function $L_n(\boldsymbol{\vartheta}) = \sum_{\nu=1}^{S} l_{n,\nu}(\vartheta_\nu)$, where

$$l_{n,\nu}(\vartheta_\nu) = \sum_{k=1}^{n-1}[\log\{f_{\vartheta_\nu}(t,t-1)\} + (Y_t - m_{\vartheta_\nu}(t,t-1))^2 f_{\vartheta_\nu}^{-1}(t,t-1)] l_{n,\nu}(\vartheta_\nu) = \sum_{k=1}^{n-1}\phi_t(\vartheta_\nu)$$

is maximized in order to obtain the QML-estimator $\widehat{\boldsymbol{\vartheta}}_n^{\mathrm{QML}}$ of the parameter vector $\boldsymbol{\vartheta}$.

Corollary 4. *The function $f_{\vartheta_\nu}(t,t-1)$ is given by*

$$f_{\vartheta_\nu}(t,t-1) = \sum_{i=1}^{p_\nu}\alpha_i(\nu)(1-\alpha_i(\nu))Y_{kS+\nu-i} + \sigma_\varepsilon^2(\nu), \qquad (5.11)$$

in the case of ZIP distributed innovations.

The function $l_{n,\nu}(\vartheta_\nu)$ can be obtained directly by replacing the results of (5.11) and (5.10)

in (5.11). From the second order periodic stationarity of $\{Y_t\}_{t\in\mathbb{Z}}$ it follows that $\mathsf{E}\,\|\varepsilon_t\|^6 < \infty$ implies $\mathsf{E}\,\|Y_t\|^6 < \infty$, then one can prove that the real-valued penalty function $L_n(\boldsymbol{\vartheta})$ satisfies the assumptions of Theorem 3.2.26 in Taniguchi & Kakizawa (2000). Thus, there exists a sequence of estimators $\widehat{\boldsymbol{\vartheta}}_n^{\mathrm{QML}} = ((\widehat{\vartheta}_{n,1}^{\mathrm{QML}})^\top, \ldots, (\widehat{\vartheta}_{n,S}^{\mathrm{QML}})^\top)^\top$ such that $\widehat{\boldsymbol{\vartheta}}_n^{\mathrm{QML}} \to \boldsymbol{\vartheta}_0$ almost surely as $n \to \infty$, and for any $\epsilon > 0$, there exists an event E with $\mathsf{P}(E) > 1 - \epsilon$ and an $n_0 \in \mathbb{N}$ such that on E, for $n > n_0$, $\widehat{\boldsymbol{\vartheta}}_n^{\mathrm{QML}}$ is the solution of

$$\frac{\partial}{\partial \boldsymbol{\vartheta}} L_n(\boldsymbol{\vartheta}) = 0, \tag{5.12}$$

which attains a relative minimum of the likelihood function $L_n(\boldsymbol{\vartheta})$.

The minimization of $L_n(\boldsymbol{\vartheta})$ can be done separately by minimizing the partial log-likelihood $l_{n,\nu}(\vartheta_\nu)$ for each season $\nu \in \{1, \ldots, S\}$. Similarly, one can solve the likelihood equation (5.12) by solving the partial likelihood equations

$$\frac{\partial}{\partial \vartheta_\nu} l_{n,\nu}(\vartheta_\nu) = 0, \qquad \nu = 1, \ldots, S,$$

separately.

Define IF_ν the matrix of dimension $(p_\nu + q + 1) \times (p_\nu + q + 1)$ for each season $\nu \in \{1, \ldots, S\}$ as

$$IF_\nu = U_{\vartheta_\nu}^{-1} V_{\vartheta_\nu} U_{\vartheta_\nu}^{-1}, \tag{5.13}$$

where

$$V_{\vartheta_\nu} = \mathsf{E}\left\{ \frac{\partial}{\partial \vartheta_\nu} \phi_t(\vartheta_\nu) \frac{\partial}{\partial \vartheta_\nu^\top} \phi_t(\vartheta_\nu) \right\} \tag{5.14}$$

and

$$U_{\vartheta_\nu} = \mathsf{E}\left\{ \frac{\partial^2}{\partial \vartheta_\nu \partial \vartheta_\nu^\top} \phi_t(\vartheta_\nu) \right\}. \tag{5.15}$$

Note that $\frac{\partial}{\partial \vartheta_\nu}\phi_t(\vartheta_\nu) = (\frac{\partial}{\partial \alpha_1(\nu)}\phi_t(\vartheta_\nu), \ldots, \frac{\partial}{\partial \alpha_{p_\nu}(\nu)}\phi_t(\vartheta_\nu), \frac{\partial}{\partial \beta_1(\nu)}\phi_t(\vartheta_\nu), \ldots, \frac{\partial}{\partial \beta_q(\nu)}\phi_t(\vartheta_\nu), \frac{\partial}{\partial \rho_\nu}\phi_t(\vartheta_\nu))$ is a $(p_\nu + q + 1)$-dimensional vector. Then, the matrix IF of the ZIP-PINAR$(p)_S$ process is defined as the $p^* \times p^*$ block diagonal matrix, $p^* = p_1 + \ldots + p_S + S(q+1)$, given by

$$IF = \mathrm{diag}\{IF_1, \ldots, IF_S\}. \tag{5.16}$$

The following theorem on the asymptotic normality of the QML-estimator $\widehat{\boldsymbol{\vartheta}}^{\mathrm{QML}}$ is given below.

Theorem 8. *Assume that $\{Y_t\}_{t\in\mathbb{Z}}$ in* (5.4)*, that is, the a ZIP-PINAR$(p)_S$ process, is a second order periodic stationary process with $\mathsf{E}\,\|\varepsilon_t\|^6 < \infty$, and $m_{\vartheta_\nu}(t, t-1)$ and $f_{\vartheta_\nu}(t, t-1)$ are almost surely three times continuously differentiable in the open set Θ containing the true parameter value $\boldsymbol{\vartheta}_0$. Then, the QML estimators $\widehat{\boldsymbol{\vartheta}}_n^{\mathrm{QML}}$ are asymptotically normal distributed as*

$$n^{1/2}(\widehat{\boldsymbol{\vartheta}}_n^{\mathrm{QML}} - \boldsymbol{\vartheta}_0) \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0}, IF), \tag{5.17}$$

as $n \to \infty$, where IF is the matrix of dimension $p^* \times p^*$, $p^* = p_1 + \ldots + p_S + S(q+1)$ defined by (5.16).

Proof. This proof follows the lines in Theorem 3.2.26 in Taniguchi & Kakizawa (2000), in straightforward generalization of the proof of Theorem 3 in PINAR$(1, 1_S)$ paper. □

6 Monte Carlo simulations

This section is developed in order to evaluate the behavior of the QML estimation for different sample sizes, this way illustrating the theoretical findings presented in this article. We present two models: model 01, summarized in the Table 5.1, which displays the QML estimation results of a Monte Carlo simulation for a ZIP-PINAR$(1)_7$ with $\vec{p} = (1,1,1,1,1,1,1)$ and model 02, summarized in table 5.2, which presents the QML estimation results of a Monte Carlo simulation for a ZIP-PINAR$(2)_7$ with $\vec{p} = (1,2,1,1,1,1,2)$, both of then assessed for the lengths 140, 980 and 1960, multiples of $S = 7$, with 500 independent simulated series. For each simulation, the time series was generated by

$$\begin{cases} Y_t = \sum_{i=1}^{p_\nu} \alpha_{i,\nu} \circ Y_{t-i} + \varepsilon_t \\ \log \lambda_t = \beta_\nu X_{t-1}, \end{cases} \tag{5.18}$$

where $t = \nu + kS$ for $\nu, k, S \in \mathcal{N}_0$, $\nu, t \neq 0$. It is assumed that X_t follows several Normal distributions with different means and variance: ζ_ν and υ_ν, respectively, such that $\zeta_\nu = \zeta_{\nu+S}$ and $\upsilon_\nu = \upsilon_{\nu+S}$, where ζ_ν and υ_ν, for each $\nu = 1, \ldots, S$, were randomly obtained following uniform distributions over the interval $[1;2]$ and $(0;0.1]$, respectively.
The simulated series were generated in the statistical software R (R Development Core Team, 2009). For the QML method, we applied a general non-linear optimization procedure using augmented Lagrange multiplier method with numerical derivatives as implemented in the *solnp* function of R.

In general, the results presented in Tables 5.1 and 5.2 suggest that the QML estimators have good finite sample properties. These results show that the estimates present smaller biases and MSE when the sample size increases what is expected and agree with the asymptotic properties of the estimators, i.e., consistency and unbiasedness. In addition, it is observed that the number of estimates which tend to be biased to the left also reduces when the sample size increases.

7 Real data application

This section is dedicated to the application of the model to a real data set referring to the daily number of hospital emergency service visits for people with respiratory airway diseases,

Table 5.1: Results of 500 simulated ZIP-PINAR$(1)_7$ with $\vec{p} = (1, 1, 1, 1, 1, 1, 1)$ processes, where Mean and MSE represent the mean of the estimated parameters and the mean square errors of the Real parameters, respectively. The quasi-maximum likelihood method of parameters estimation was applied.

	Real	Mean	Bias	MSE		Real	Mean	Bias	MSE		Real	Mean	Bias	MSE
		n=140					n=980					n=1960		
$\alpha_1(1)$	0,480	0,480	0,000	0,009	$\alpha_1(1)$	0,480	0,480	0,000	0,002	$\alpha_1(1)$	0,480	0,480	0,000	0,001
$\alpha_1(2)$	0,670	0,663	0,007	0,024	$\alpha_1(2)$	0,670	0,668	0,002	0,004	$\alpha_1(2)$	0,670	0,671	-0,001	0,002
$\alpha_1(3)$	0,240	0,235	0,005	0,020	$\alpha_1(3)$	0,240	0,238	0,002	0,004	$\alpha_1(3)$	0,240	0,237	0,003	0,002
$\alpha_1(4)$	0,400	0,395	0,005	0,028	$\alpha_1(4)$	0,400	0,401	-0,001	0,004	$\alpha_1(4)$	0,400	0,400	0,000	0,002
$\alpha_1(5)$	0,750	0,748	0,002	0,022	$\alpha_1(5)$	0,750	0,750	0,000	0,002	$\alpha_1(5)$	0,750	0,750	0,000	0,001
$\alpha_1(6)$	0,860	0,853	0,007	0,007	$\alpha_1(6)$	0,860	0,860	0,000	0,001	$\alpha_1(6)$	0,860	0,860	0,000	0,000
$\alpha_1(7)$	0,200	0,197	0,003	0,008	$\alpha_1(7)$	0,200	0,200	0,000	0,001	$\alpha_1(7)$	0,200	0,200	0,000	0,001
ρ_1	0,810	0,790	0,020	0,022	ρ_1	0,810	0,805	0,005	0,003	ρ_1	0,810	0,806	0,004	0,002
ρ_2	0,310	0,307	0,003	0,034	ρ_2	0,310	0,305	0,005	0,008	ρ_2	0,310	0,307	0,003	0,004
ρ_3	0,450	0,444	0,006	0,049	ρ_3	0,450	0,447	0,003	0,007	ρ_3	0,450	0,447	0,003	0,004
ρ_4	0,550	0,530	0,020	0,029	ρ_4	0,550	0,545	0,005	0,003	ρ_4	0,550	0,549	0,001	0,002
ρ_5	0,320	0,318	0,002	0,032	ρ_5	0,320	0,312	0,008	0,008	ρ_5	0,320	0,315	0,005	0,003
ρ_6	0,440	0,420	0,020	0,026	ρ_6	0,440	0,435	0,005	0,003	ρ_6	0,440	0,437	0,003	0,002
ρ_7	0,640	0,621	0,019	0,019	ρ_7	0,640	0,636	0,004	0,004	ρ_7	0,640	0,639	0,001	0,002
β_1	0,300	0,267	0,033	0,042	β_1	0,300	0,282	0,018	0,023	β_1	0,300	0,291	0,009	0,013
β_2	0,150	0,161	-0,011	0,017	β_2	0,150	0,149	0,001	0,006	β_2	0,150	0,147	0,003	0,004
β_3	0,250	0,274	-0,024	0,053	β_3	0,250	0,243	0,007	0,005	β_3	0,250	0,247	0,003	0,003
β_4	0,500	0,479	0,021	0,049	β_4	0,500	0,497	0,003	0,004	β_4	0,500	0,498	0,002	0,004
β_5	0,320	0,320	0,000	0,023	β_5	0,320	0,318	0,002	0,017	β_5	0,320	0,315	0,005	0,003
β_6	0,440	0,423	0,017	0,032	β_6	0,440	0,435	0,005	0,003	β_6	0,440	0,440	0,000	0,003
β_7	0,740	0,716	0,024	0,019	β_7	0,740	0,738	0,002	0,011	β_7	0,740	0,738	0,002	0,002

Table 5.2: Simulation of ZIP-PINAR$(2)_7$ with $\vec{p} = (1, 2, 1, 1, 1, 1, 2)$ process. 500 repetitions. The quasi-maximum likelihood method of parameters estimation was applied.

	Real	Mean	Bias	MSE		Real	Mean	Bias	MSE		Real	Mean	Bias	MSE
		n=140					n=980					n=1960		
$\alpha_1(1)$	0,590	0,583	0,007	0,021	$\alpha_1(1)$	0,590	0,590	0,000	0,002	$\alpha_1(1)$	0,590	0,590	0,000	0,001
$\alpha_2(1)$	0,120	0,142	-0,022	0,025	$\alpha_2(1)$	0,120	0,122	-0,002	0,005	$\alpha_2(1)$	0,120	0,121	-0,001	0,002
$\alpha_2(2)$	0,270	0,258	0,012	0,033	$\alpha_2(2)$	0,270	0,268	0,002	0,005	$\alpha_2(2)$	0,270	0,269	0,001	0,002
$\alpha_3(1)$	0,680	0,661	0,019	0,030	$\alpha_3(1)$	0,680	0,684	-0,004	0,003	$\alpha_3(1)$	0,680	0,680	0,000	0,001
$\alpha_4(1)$	0,360	0,369	-0,009	0,062	$\alpha_4(1)$	0,360	0,358	0,002	0,007	$\alpha_4(1)$	0,360	0,362	-0,002	0,003
$\alpha_5(1)$	0,190	0,194	-0,004	0,020	$\alpha_5(1)$	0,190	0,189	0,001	0,002	$\alpha_5(1)$	0,190	0,191	-0,001	0,002
$\alpha_6(1)$	0,250	0,257	-0,007	0,033	$\alpha_6(1)$	0,250	0,250	0,000	0,004	$\alpha_6(1)$	0,250	0,248	0,002	0,002
$\alpha_7(1)$	0,170	0,174	-0,004	0,027	$\alpha_7(1)$	0,170	0,170	0,000	0,005	$\alpha_7(1)$	0,170	0,167	0,003	0,002
$\alpha_7(2)$	0,390	0,381	0,009	0,033	$\alpha_7(2)$	0,390	0,390	0,000	0,003	$\alpha_7(2)$	0,390	0,389	0,001	0,001
ρ_1	0,710	0,678	0,032	0,028	ρ_1	0,710	0,704	0,006	0,003	ρ_1	0,710	0,709	0,001	0,001
ρ_2	0,610	0,631	-0,021	0,071	ρ_2	0,610	0,603	0,007	0,012	ρ_2	0,610	0,602	0,008	0,007
ρ_3	0,450	0,447	0,003	0,044	ρ_3	0,450	0,449	0,001	0,006	ρ_3	0,450	0,447	0,003	0,003
ρ_4	0,250	0,247	0,003	0,021	ρ_4	0,250	0,250	0,000	0,002	ρ_4	0,250	0,249	0,001	0,002
ρ_5	0,360	0,350	0,010	0,052	ρ_5	0,360	0,357	0,003	0,007	ρ_5	0,360	0,359	0,001	0,002
ρ_6	0,290	0,290	0,000	0,043	ρ_6	0,290	0,291	-0,001	0,009	ρ_6	0,290	0,288	0,002	0,003
ρ_7	0,520	0,545	-0,025	0,100	ρ_7	0,520	0,524	-0,004	0,019	ρ_7	0,520	0,513	0,007	0,012
β_1	0,600	0,549	0,051	0,072	β_1	0,600	0,586	0,014	0,022	β_1	0,600	0,596	0,004	0,002
β_2	0,150	0,204	-0,054	0,063	β_2	0,150	0,146	0,004	0,014	β_2	0,150	0,147	0,003	0,000
β_3	0,250	0,278	-0,028	0,072	β_3	0,250	0,248	0,002	0,007	β_3	0,250	0,250	0,000	0,003
β_4	0,800	0,789	0,011	0,024	β_4	0,800	0,799	0,001	0,001	β_4	0,800	0,796	0,004	0,003
β_5	0,660	0,615	0,045	0,087	β_5	0,660	0,656	0,004	0,008	β_5	0,660	0,659	0,001	0,001
β_6	0,530	0,518	0,012	0,075	β_6	0,530	0,532	-0,002	0,011	β_6	0,530	0,529	0,001	0,003
β_7	0,170	0,263	-0,093	0,107	β_7	0,170	0,187	-0,017	0,020	β_7	0,170	0,174	-0,004	0,015

classified according to International Classification of Diseases (ICD-10, j31 and j45).

The data selected were the people of any age group who visited the hospital emergency service in Vitória-ES city, specifically those living in the neighborhoods of Praia do Suá, Enseada do Suá, Bento Ferreira and Ilha do Boi. The choice of these neighborhoods was due to the close proximity to the air quality monitoring station located at the Enseada do Suá. These data were obtained from the network records system Welfare of the municipality. This network, the Health Management System Welfare Network.

The Welfare Network computer system is implemented at the central level of the Secretary (SEMUS) and in all the basic health units, in such as: reference centers, emergency services, among others. The data records and information of the service network of the municipal health system have digital certification in accordance with Municipal Decree 15.913, of February 13, 2014.

The period of study was from June 26, 2013 to April 7, 2016, resulting in 1022 daily observations. Figure 5.2 shows the line graphs, histogram, periodic mean and variance graphs from the series of attendances for respiratory diseases. In this application, the daily values of the concentrations of the pollution variable will be the covariate used in the modeling of the health variable, which represents the main variable of the model.

As can be seen at the Figures 5.2(c) and 5.2(d), the mean and the variance, respectively, vary over the seasons, this is a characteristic of periodically correlated time series. Information on daily levels of atmospheric pollutant Particulate matter particles with a diameter of 10 micrometers or less (PM_{10}) was obtained from the State Institute of Environment and Water Resources (IEMA), with data collected at the Suá Station, in Vitória, belonging to the Automatic Air Quality Monitoring Network (RAMQAr).

The data collection comprised a period of 24 hours, which began in the first half hour of the day. The mean of 24 hours for the pollutant PM_{10} was considered. Analyzing the graph of Cross-correlation function, Figure 5.5, the value $d = 0$ is suggested, since the strong peak is given for the lag 0. For the computational optimization procedure, we divided each value of X by 100, in order to obtain a value between 0 and 1.

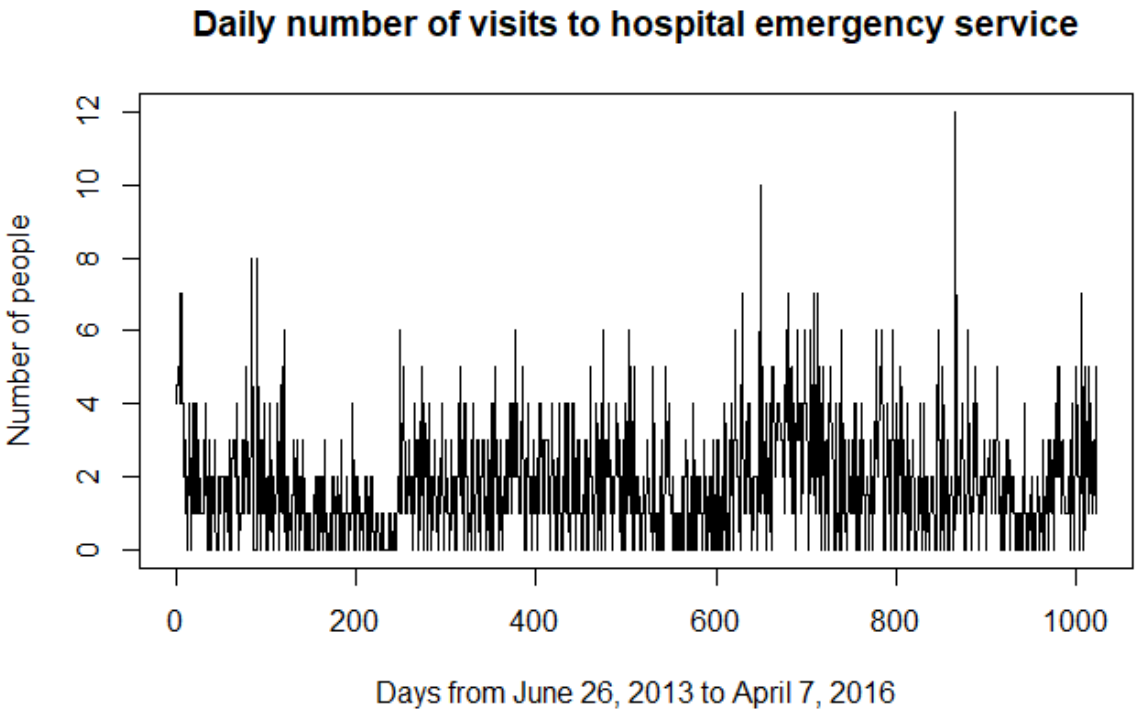


Figure 5.1: Number of hospital visits to people with respiratory airway diseases of the region of Vitória, (ES, Brazil), from June 26,2013 to April 7, 2016, resulting in 1022 daily observations.

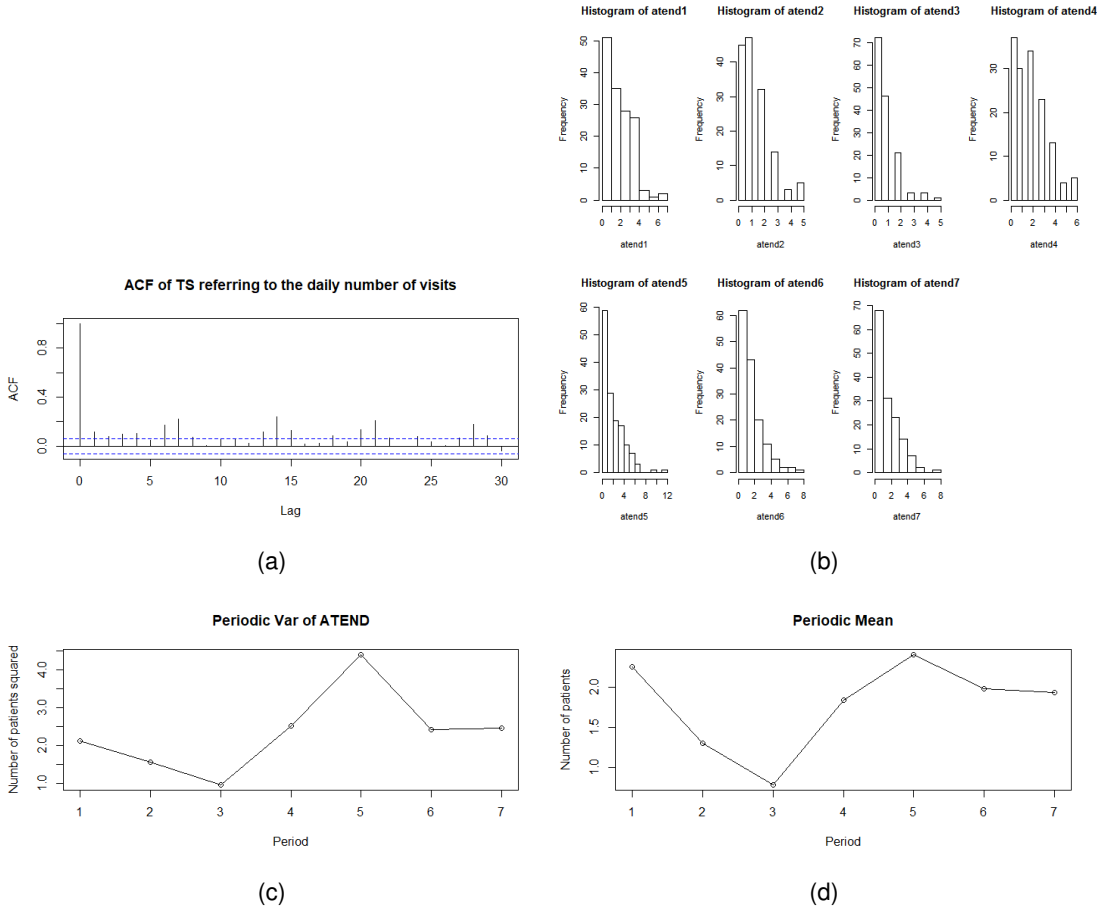


Figure 5.2: ACF(a), histograms by seasons (b), the sample variances (c) and means (d) for each season.

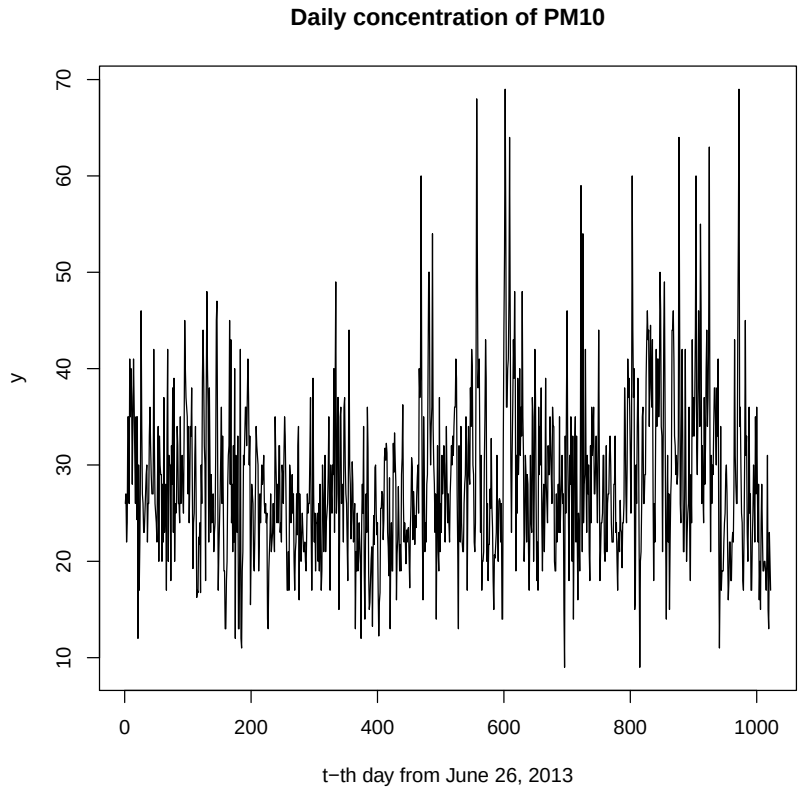


Figure 5.3: Daily concentrations of the pollution covariate, 1022 observations.

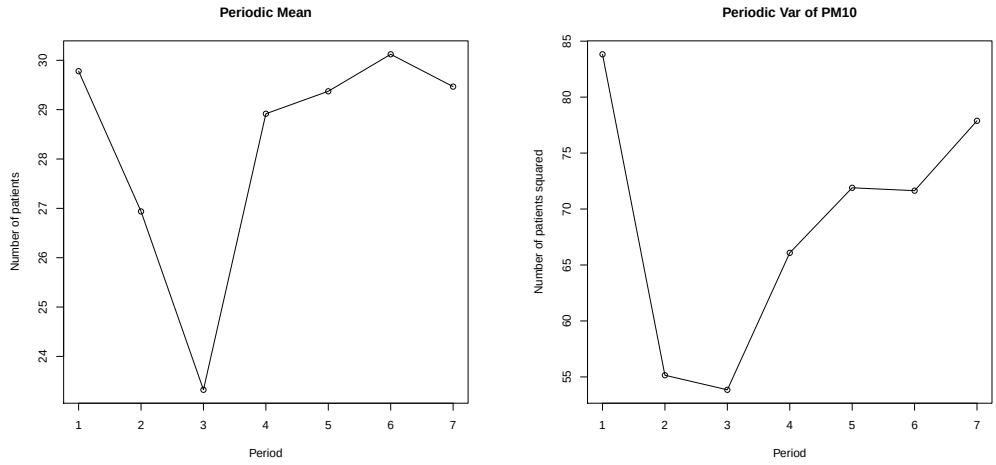


Figure 5.4: The sample variances and means of the seasons (left and right, respectively) of the daily values of the concentrations of the pollution covariate variable.

Tables 5.3 and 5.4 show the PeACF and PePACF sample functions of the health data. The elements in bold represent values that have exceeded the confidence interval. In addition, according to McLeod (1994) one can identify the AR order for each season by finding the lowest lag for which the sample PePACF cuts off. All these suggested the use of ZIP-PINAR$(7)_7$ model with $\vec{p} = (3, 4, 4, 2, 7, 6, 6)$ to model the real data set.

The parameters were estimated using QML method with ZIP distributed innovations. The re-

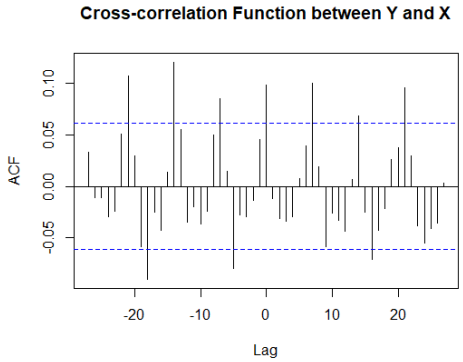


Figure 5.5: The sample cross correlation function between the daily number of visits to emergency service and the values of the concentrations of PM_{10}.

Table 5.3: Sample PeACF function of the daily number of hospital visits to people with respiratory airway diseases (health data).

Sample PeACF function of the health data														
Season	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=7$	$h=8$	$h=9$	$h=10$	$h=11$	$h=12$	$h=13$	$h=14$
$\nu=1$	**0.03615743**	**0.23453908**	**0.17623970**	**0.15191931**	0.11969754	0.14124429	0.07167236	0.126107594	0.045404826	0.05324367	0.070631229	0.13351077	0.09798248	0.07554284
$\nu=2$	**0.02477357**	0.03293201	0.08230342	**0.25894745**	**0.17944770**	0.08855555	0.08007309	**0.178249162**	0.109950767	0.05996379	**0.167380529**	0.14084816	0.06141951	**0.17011477**
$\nu=3$	**0.02038011**	**0.28945655**	0.14283592	0.13791804	**0.16975064**	**0.16335145**	0.06427256	**0.249987471**	-0.038438806	-0.04905219	0.035516699	0.02381582	-0.03205135	0.12310399
$\nu=4$	0.08793825	**0.23601359**	0.02380250	**0.14362365**	0.07198004	0.09602562	0.09819808	-0.116819686	0.057062353	**0.15567545**	0.074686223	0.11706798	**0.22280860**	**0.16698385**
$\nu=5$	**0.15013911**	**0.17397431**	**0.22338219**	**0.15882988**	0.06830728	0.11685891	**0.31416136**	-0.005926477	0.007155263	**0.14918424**	0.145326921	0.11242536	0.06747111	**0.30106483**
$\nu=6$	0.11152528	0.08277729	0.12372636	0.09062921	0.12911949	**0.35621800**	0.03101838	0.029310409	0.105633386	0.04696810	0.007433698	0.03582651	0.10881090	0.04249913
$\nu=7$	**0.22769719**	0.14163874	**0.20608829**	0.06239895	**0.15333057**	**0.22069450**	0.03942914	0.056054172	0.107844041	**0.16080872**	0.118217587	0.10267796	0.08652887	0.02233563

sults are displayed in Table 5.5. As can be seen, the obtained estimates corroborate the results of the graphical analysis presented in Figure 5.2(b), the estimated values for ρ_3, ρ_4 and ρ_5 are more expressive. In relation to the estimates of β, we are struck by the fact that pollutant levels did not influence the number of visits on Sunday, Monday and Thursday, and the related coefficient of Tuesday was the largest one.

The adequacy of the adjusted model was evaluated by examining the residuals for serial dependency. The estimated residuals $\{r_t\}$ after fitting the ZIP-PINAR$(7)_7$ model with $\vec{p} = (3,4,4,2,7,6,6)$ model were computed as

$$r_t = y_t - \hat{Y}_t = y_t - \sum_{i=1}^{p_\nu} \widehat{\alpha}_i(\nu) \circ y_{\nu-i+kS} + \widehat{\lambda}_\nu, \tag{5.19}$$

where $t = 7k + \nu$.

The residual PACFs displayed in Table 5.6 show that the fitted model was able to filter the autocorrelations of the data. So, using ZIP-PINAR$(7)_7$ with $\vec{p} = (3,4,4,2,7,6,6)$ we do not found clearly systematic pattern observed in the residuals, it means that the fitted model seems

Table 5.4: Sample PePACF function of the daily number of hospital visits to people with respiratory airway diseases.

Sample PePACF function of the Health data														
Season	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=7$	$h=8$	$h=9$	$h=10$	$h=11$	$h=12$	$h=13$	$h=14$
$\nu=1$	0.03615743	**0.23256319**	**0.158722640**	0.12612913	0.06492867	0.079628171	0.019069296	0.0209107467	-0.002644775	-0.03361488	0.02400030	0.14727568	0.069057844	-0.012349114
$\nu=2$	0.02477357	0.03206707	0.073406103	**0.25458203**	**0.15520358**	0.036350531	-0.0007933577	**0.1502159844**	0.065909280	-0.02531138	0.06568170	0.11286327	0.070332383	0.105922946
$\nu=3$	0.02038011	**0.28910042**	0.138021901	0.04506810	0.10932210	0.095757407	0.0008223576	**0.1750181010**	-0.106652269	-0.13736176	0.04747391	-0.05280139	-0.079718617	0.109055669
$\nu=4$	0.08793825	**0.23518116**	-0.006555898	0.12945598	0.01799826	0.008571614	0.0225815584	**-0.1625314828**	-0.009289720	0.14782016	0.07072478	0.07794849	**0.198648806**	0.130563082
$\nu=5$	**0.15013911**	**0.16324703**	**0.198352738**	0.11461482	0.02833369	0.049192647	**0.2350460795**	-0.1180930425	-0.058935749	0.06771839	0.08922126	0.09442689	-0.013366452	**0.204392203**
$\nu=6$	0.11152528	0.06720933	0.102776274	0.05885116	0.09016581	**0.341131358**	-0.0989328623	-0.0892845392	0.009527345	0.02895840	-0.08180822	-0.07527491	0.114449871	0.006591626
$\nu=7$	**0.22769719**	0.12013006	**0.178964881**	0.00618928	0.08080766	**0.194609296**	-0.0651045775	-0.0009858756	0.034095234	0.11055805	0.10889168	0.06091423	0.007841388	-0.089447812

Table 5.5: Estimated parameters of ZIP-PINAR$(7)_7$ with $\vec{p} = (3, 4, 4, 2, 7, 6, 6)$

	Estimated parameters						
	ν =1	ν =2	ν =3	ν =4	ν =5	ν =6	ν =7
$\alpha_1(\nu)$	0.048	0.000	0.000	0.272	0.116	0.023	0.158
$\alpha_2(\nu)$	0.332	0.000	0.147	0.360	0.097	0.000	0.071
$\alpha_3(\nu)$	0.160	0.053	0.066	–	0.318	0.000	0.155
$\alpha_4(\nu)$	–	0.136	0.071	–	0.140	0.055	0.000
$\alpha_5(\nu)$	–	–	–	–	0.045	0.109	0.139
$\alpha_6(\nu)$	–	–	–	–	0.108	0.301	0.165
$\alpha_7(\nu)$	–	–	–	–	0.284	–	–
ro	0.000	0.144	0.839	0.042	0.762	0.000	0.400
beta	0.145	0.206	0.000	0.000	1	0.015	0.000

to well adjust the data and capture its main dynamics. This estimated model can be very useful in providing reliable forecast.

All these empirical analyses, i.e., the values for the periodic ACF and PACF of the residuals (Tables 5.6 and 5.7) and the sample ACF (Table 5.6) support the fact that the proposed model with ZIP distributed innovations is a good choice for modeling such data.

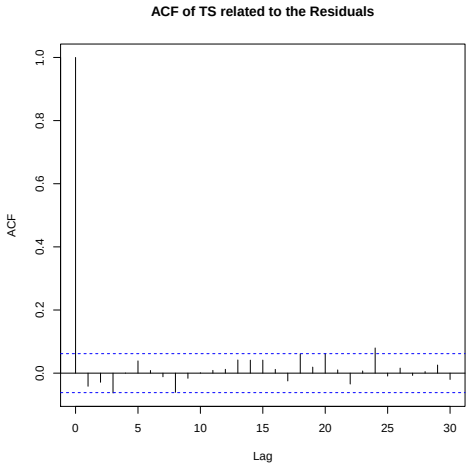


Figure 5.6: The sample ACF of the residuals after fitted the ZIP-PINAR$(7)_7$ model with $\vec{p} = (3, 4, 4, 2, 7, 6, 6)$ to the daily number of visits to emergency service with the values of the concentrations of PM_{10} as covariates.

8 Conclusions

The ZIP-PINAR$(p)_S$ model and some of its main properties were introduced in this paper. The quasi-maximum likelihood method for estimating the parameters of the model, namely QML, was proposed. The asymptotic properties of the estimators were presented. A simulation study was carried out to investigate their finite sample performances for standard sample sizes. The results corroborated the asymptotic theory. The QML methods even for a small sample size

Table 5.6: PeACF of residuals after fitted the ZIP-PINAR$(7)_7$ model with $\vec{p} = (3, 4, 4, 2, 7, 6, 6)$ to the series of the number of visits to emergency service.

	PeACF of residuals													
Season	h =1	h =2	h =3	h =4	h =5	h =6	h =7	h =8	h =9	h =10	h =11	h =12	h =13	h =14
ν =1	-0,13872	-0,14264	-0,10767	0,067104	0,069014	0,103165	-0,00765	-0,03151	-0,02032	-0,10681	0,002484	0,072219	-0,00405	-0,0721
ν =2	-0,05459	-0,06207	-0,06802	-0,00629	0,12804	0,024868	0,005328	0,096429	0,003884	0,017753	-0,05418	0,09729	0,093409	0,135216
ν =3	-0,10345	0,020297	-0,02546	-0,05584	0,033175	0,052987	0,006162	**0,189072**	-0,10731	-0,12561	0,032689	-0,02284	-0,1034	0,113723
ν =4	-0,09099	-0,06911	-0,07765	0,079055	0,045033	-0,08591	0,062069	**-0,21126**	-0,03547	0,059288	0,003747	0,06726	**0,195332**	**0,171328**
ν =5	-0,03818	0,073125	-0,13608	-0,04929	0,04889	-0,024	-0,02628	**-0,22059**	0,020241	0,051603	0,059536	-0,03556	-0,01482	0,05926
ν =6	0,000795	-0,00361	-0,02095	-0,00833	0,021793	0,085469	-0,04997	-0,01727	0,047925	0,069993	-0,04227	-0,04529	0,071884	-0,06013
ν =7	0,06592	-0,02829	0,020397	-0,05313	-0,05074	0,037545	-0,07353	-0,05516	-0,05037	0,056968	0,083343	0,008234	-0,0151	-0,05566

Table 5.7: PePACF of the residuals after fitted the ZIP-PINAR$(7)_7$ model with $\vec{p} = (3, 4, 4, 2, 7, 6, 6)$ to the series of the number of visits to emergency service.

	Sample periodic PACF of the residuals													
Season	h =1	h =2	h =3	h =4	h =5	h =6	h =7	h =8	h =9	h =10	h =11	h =12	h =13	h =14
ν =1	-0,13872	-0,13509	-0,11341	0,066821	0,075625	0,097262	0,004862	-0,0234	-0,02853	-0,11995	-0,03287	0,121542	-0,02403	-0,09575
ν =2	-0,05459	-0,07043	-0,07365	-0,01639	0,13501	0,039571	0,022706	0,113405	0,009563	0,022045	-0,0466	0,084773	**0,156237**	0,149119
ν =3	-0,10345	0,014751	-0,03033	-0,06069	0,032407	0,06952	0,010327	**0,2054**	-0,07636	-0,13557	0,040595	-0,0399	-0,1099	**0,153048**
ν =4	-0,09099	-0,07927	-0,08069	0,061655	0,020594	-0,09074	0,080279	**-0,19156**	-0,03544	0,060926	-0,00735	0,068651	**0,214626**	**0,184618**
ν =5	-0,03818	0,069993	-0,13256	-0,06202	0,038086	-0,03948	-0,03972	**-0,20753**	0,00458	0,026771	0,064771	0,00769	0,006481	0,051483
ν =6	0,000795	-0,00358	-0,02146	-0,01084	0,02134	0,088929	-0,05323	-0,0122	0,048017	0,077818	-0,02621	-0,05001	0,074446	-0,06989
ν =7	0,06592	-0,02841	0,019618	-0,04841	-0,05871	0,033649	-0,08085	-0,05338	-0,04844	0,059947	0,091004	0,029828	-0,00597	-0,07398

($n = 50$), showed good performance when analyzed the bias and the MSE for each estimate. The usefulness of the proposed model was verified by an application to the time series of counts of the daily number of visits of people with respiratory problems (International Classification of Diseases ICD-10) in the hospital emergency service of the public health care system of the region of Vitória-ES. The ZIP-PINAR$(7)_7$ model with $\vec{p} = (3, 4, 4, 2, 7, 6, 6)$ was fitted to the real data. The adequacy of the fitted model was assessed through residual analyses. In this context, the ZIP-PINAR$(7)_7$ with $\vec{p} = (3, 4, 4, 2, 7, 6, 6)$ represents a good choice to model this data set.

Chapter 6

Conclusions and perspectives

In this work, we were interested in the extension of periodic models for counting time series. The contributions of this thesis were presented in four articles.

First, we proposed the PINAR$(1, 1_S)$ model, mainly motivated by the analysis of the time series of the number of people who received antibiotics for the treatment of respiratory diseases. As the respiratory disease is strongly correlated with air pollution levels and climatic conditions (Oudin et al. (2017), Caillaud et al. (2018)), the correlation structure of the daily number series of people who received antibiotics showed, among other phenomena, periodicity, and seasonality, what is often observed in series of daily average concentrations of atmospheric pollutants (Hies et al. (2000)). We briefly presented some properties of the PINAR$(1, 1_S)$ model, such as existence and uniqueness, stationarity conditions, and the mean of the process. The quasi-maximum likelihood method was used to estimate the parameters of the model. A section was devoted to a simulation study. For different sample sizes, the performance of the estimator was investigated, and the empirical results indicated that the method provided accurate estimates regarding the bias and the mean square error of the estimates of the parameters. We obtained a good fit with the application of the model to the real data.

Next, we studied the mathematical properties of the process described by the PINAR$(1, 1_S)$ model. We presented the scalar and matrix representation of the process and, with this, we were able to demonstrate its conditions of existence, uniqueness, and stationarity. For Poisson distributed innovations, we discussed the marginal distribution of the process and, under certain conditions, we found that it is also Poisson distributed. The marginal and joint conditional probability distributions of the process were presented for the cases of Poisson and Geometric distributed innovations. To estimate the parameters of the model, we proposed the methods of conditional least squares (CLS), moment-based method (Yule-Walker) and quasi-maximum likelihood (QML). We demonstrated that the estimators were consistent and had asymptotic normality. Then a comprehensive simulation study with finite samples was conducted, and we were able to prove the consistency of our results. We presented a section for forecasting purposes. Finally, an application on health data was conducted. The PINAR$(1, 1_S)$ proved to be very useful for making reliable forecasts.

We then proposed the PINAR$(p)_S$ model for count time series with periodic autocorrelation structure with order larger than or equal to 1. The statistical properties of the model, such as mean, variance, marginal and joint distributions were also discussed. We proposed the CLS, Yule-Walker and QML methods to estimate the model parameters. Under some hypotheses, we showed that the estimators were consistent and had asymptotic normality. Their performances were investigated through Monte Carlo simulations, whose empirical results indicated that estimates were accurate. The model effectively adjusted the daily number of dispensing medications for the treatment of respiratory disease (asthma).

Thus, a final aspect of this work was to focus on the extension of PINAR$(p)_S$ model with ZIP distributed innovations, and also on the use of explanatory covariates. Some statistical properties and the probability of transition between the states of the process were calculated, and then the QML method was proposed for the estimation of the parameters of the model. Finally, we applied the model to the series of health data with air pollutants as explanatory covariates.

This research together with the real data set used lead to several promising research lines that can be pursued in the future, such as, in the time domain, to extend the PINAR$(p)_S$ model by proposing new distributions of probability for the innovations; to explore the statistical and mathematical properties of the ZIP-PINAR$(p)_S$ model and develop a general model that considers the inflation or deflation of zeros or any other specific value; to investigate innovative and coherent forecasting methods to periodically correlated integer-valued series; to fit the PINAR model to incomplete time series, under the effect or not of aberrant observations, through the development of imputation method to cyclostationary count time series with missing data and the development of robust estimators for the parameters of the model, respectively; to propose an extension of PINAR$(p)_S$ for multivariate and spacial-time count series modeling among others. All these proposed models and their parameter estimation methods can be studied in the frequency domain approach, which is also an interesting research line from the theoretical and applied point of view.

References

Al-Osh, M. & Alzaid, A. A. (1987), 'First-order integer-valued autoregressive (INAR(1)) process', *Journal of Time Series Analysis* **8**(3), 261–275.

Alzaid, A. A. & Al-Osh, M. (1990), 'An integer-valued pth-order autoregressive structure (INAR(p)) process', *Journal of Applied Probability* **27**(2), 314–324.

Baldacci, S., Maio, S., Cerrai, S., Sarno, G., Baïz, N., Simoni, M., Annesi-Maesano, I., Viegi, G. & Study, H. (2015), 'Allergy and asthma: Effects of the exposure to particulate matter and biological allergens', *Respiratory medicine* **109**(9), 1089–1104.

Basawa, I. & Lund, R. (2001), 'Large sample properties of parameter estimates for periodic ARMA models', *Journal of Time Series Analysis* **22**(6), 651–663.

Bickel, P. J. & Doksum, K. A. (1977), *Mathematical statistics: basic ideas and selected topics*, Vol. 1, Prentice Hall.

Bondon, P., Ispány, M., Paraiba, C. C. M. & Reisen, V. A. (2018), On subset integer-valued autoregressions, Technical report, CentraleSupélec, UFES, University of Debrecen.

Bourguignon, M., Vasconcellos, K. L. P., Reisen, V. A. & Ispány, M. (2016), 'A Poisson INAR(1) process with a seasonal structure', *Journal of Statistical Computation and Simulation* **86**(2), 373–387.

Bozdogan, H. (1987), 'Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions', *Psychometrika* **52**(3), 345–370.

Brännäs, K. (1993), *Estimation and testing in integer-valued AR (1) models*, University of Umeå.

Brockwell, P. J. & Davis, R. A. (2013), *Time series: theory and methods*, Springer Science & Business Media.

Bu, R., McCabe, B. & Hadri, K. (2008), 'Maximum likelihood estimation of higher-order integer-valued autoregressive processes', *Journal of Time Series Analysis* **29**(6), 973–994.

Caillaud, D., Cheriaux, M., Martin, S., Ségala, C., Dupuy, N., Evrard, B. & Thibaudon, M. (2018), 'Short-term effect of outdoor mold spore exposure on prescribed allergy medication sales in Central France', *Clinical & Experimental Allergy* **48**(7), 837–845.

Chen, S. X. & Liu, J. S. (1997), 'Statistical applications of the poisson-binomial and conditional bernoulli distributions', *Statistica Sinica* pp. 875–892.

da Silva, M. E. & Oliveira, V. L. (2004), 'Difference equations for the higher-order moments and cumulants of the inar(1) model', *Journal of Time Series Analysis* **25**(3), 317–333.

Dietz, E. & Bohning, D. (1997), 'The use of two-component mixture models with one completely or partly known component', *Computational Statistics* **12**(2).

Du, J.-G. & Li, Y. (1991), 'The integer-valued autoregressive (INAR(p)) model', *Journal of Time Series Analysis* **12**(2), 129–142.

Enciso-Mora, V., Neal, P. & Rao, T. S. (2009), 'Integer valued AR processes with explanatory variables', *Sankhyā: The Indian Journal of Statistics, Series B (2008-)* pp. 248–263.

Filho, P. R. P., Reisen, V. A., Bondon, P. & Ispány, M. (n.d.), The s-periodic integer autoregressive model of order p (PINAR$(p)_s$). Manuscript in preparation.

Fokianos, K., Rahbek, A. & Tjostheim, D. (2009), 'Poisson autoregression', *Journal of the American Statistical Association* **104**(488), 1430–1439.

for Europe, W. R. O. et al. (2013), 'Review of evidence on health aspects of air pollution–revihaap project'.

Frank Ayres, J. (1967), *Schaum's outline of theory and problems of matrices*, McGraw-Hill.

Franke, J. & Seligmann, T. (1993), Conditional maximum likelihood estimates for INAR(1) processes and their application to modelling epileptic seizure counts, *in* T. S. Rao, ed., 'Developments in Time Series', Chapman and Hall, London, chapter 22, pp. 310–330.

Franses, P. H. & Paap, R. (2004), *Periodic Time Series Models*, Advanced Texts in Econometrics, Oxford University Press.

Freeland, R. K. & McCabe, B. P. (2004), 'Forecasting discrete valued low count time series', *International Journal of Forecasting* **20**(3), 427–434.

Gadelha, C. A. G., Costa, K. S., Nascimento, J. M. d., Soeiro, O. M., Mengue, S. S., Motta, M. L. d. & Carvalho, A. C. C. d. (2016), 'Pnaum: abordagem integradora da assistência farmacêutica, ciência, tecnologia e inovação', *Rev. Saúde Pública* **50**(suppl 2).

Gardner, W. A., Napolitano, A. & Paura, L. (2006), 'Cyclostationarity: Half a century of research', *Signal Processing* **86**(4), 639–697.

Gladyshev, E. G. (1961), 'Periodically correlated random sequences', *Sov. Math.* **2**, 385–388.

Graybill, F. A. (1983), *Matrices with applications in statistics*, Wadsworth Inc.

Hies, T., Treffeisen, R., Sebald, L. & Reimer, E. (2000), 'Spectral analysis of air pollutants. part 1: elemental carbon time series', *Atmospheric Environment* **34**(21), 3495–3502.

Holstiege, J. & Garbe, E. (2013), 'Systemic antibiotic use among children and adolescents in Germany: a population-based study', *European Journal of Pediatrics* **172**(6), 787–795.

Horn, R. A. & Johnson, C. R. (2012), *Matrix analysis*, 2 edn, Cambridge University Press.

Hurd, H. L. & Miamee, A. (2007), *Periodically correlated random sequences: Spectral theory and practice*, Wiley Series in Probability and Statistics, John Wiley & Sons, Hoboken, NJ.

IBGE (2010), 'Demographic census, Brazilian Institute of Geography and Statistics , Brazil. Retrieved on 31 December 2018 from: www.ibge.gov.br'.

IEMA (2013), 'Annual air quality report for Greater Vitória, Estate Institute of Environment and Water Resources of ES, Brazil. Retrieved on 31 December 2018 from: www.iema.gov.br'.

IEMA (2015), 'Annual air quality report for Greater Vitória, Estate Institute of Environment and Water Resources of ES, Brazil. Retrieved on 31 December 2018 from: www.iema.gov.br'.

Jazi, M. A., Jones, G. & Lai, C.-D. (2012), 'First-order integer valued AR processes with zero inflated poisson innovations', *Journal of Time Series Analysis* **33**(6), 954–963.

Klimko, L. A. & Nelson, P. I. (1978), 'On conditional least squares estimation for stochastic processes', *The Annals of Statistics* **6**(3), 629–642.

Lambert, D. (1992), 'Zero-inflated poisson regression, with an application to defects in manufacturing', *Technometrics* **34**(1), 1–14.

Latour, A. (1997), 'The multivariate GINAR(p) process', *Advances in Applied Probability* **29**(1), 228–248.

Lund, R. & Basawa, I. (2000), 'Recursive prediction and likelihood evaluation for periodic ARMA models', *Journal of Time Series Analysis* **21**(1), 75–93.

Lütkepohl, H. (2005), *New Introduction to Multiple Time Series Analysis*, Springer, Berlin Heidelberg New York.

McDowell, R., Bennett, K., Moriarty, F., Clarke, S., Barry, M. & Fahey, T. (2018), 'An evaluation of prescribing trends and patterns of claims within the Preferred Drugs Initiative in Ireland (2011–2016): an interrupted time-series study', *BMJ Open* p. 8:e019315. doi:10.1136/bmjopen-2017-0193.

McKenzie, E. (1985), 'Some simple models for discrete variate time series', *J. Amer. Water. Res. Assoc.* **21**, 645–650.

McLeod, A. I. (1994), 'Diagnostic checking of periodic autoregression models with application', *Journal of Time Series Analysis* **15**(2), 221–233.

Monteiro, M., Scotto, M. G. & Pereira, I. (2010), 'Integer-valued autoregressive processes with periodic structure', *Journal of Statistical Planning and Inference* **140**(6), 1529–1541.

Moriña, D., Puig, P., Ríos, J., Vilella, A. & Trilla, A. (2011), 'A statistical model for hospital admissions caused by seasonal diseases', *Statistics in Medicine* **30**(26), 3125–3136.

Organization, W. H. et al. (1993), 'How to investigate drug use in health facilities: selected drug use indicators'.

Oudin, A., Bråbäck, L., Oudin Åström, D. & Forsberg, B. (2017), 'Air pollution and dispensed medications for asthma, and possible effect modifiers related to mental health and socio-economy: A longitudinal cohort study of swedish children and adolescents', *International Journal of Environmental Research and Public Health* **14**(11), 1392.

Priestley, M. B. (1981), *Spectral analysis and time series*, Academic press.

Reinsel, G. C. (2003), *Elements of multivariate time series analysis*, Springer Science & Business Media.

Sarnaglia, A. J. Q., Reisen, V. A. & Lévy-Leduc, C. (2010), 'Robust estimation of periodic autoregressive processes in the presence of additive outliers', *Journal of Multivariate Analysis* **101**(9), 2168–2183.

Schwarz, G. (1978), 'Estimating the dimension of a model', *The Annals of Statistics* **6**(2), 461–464.

Scotto, M. G., Weiss, C. H. & Gouveia, S. (2015), 'Thinning-based models in the analysis of integer-valued time series: a review', *Statistical Modelling* **15**(6), 590–618.

Seneta, E. (2006), *Non-negative matrices and Markov chains*, Springer Science & Business Media.

Shao, Q. & Ni, P. (2004), 'Least-squares estimation and anova for periodic autoregressive time series', *Statistics & probability letters* **69**(3), 287–297.

Solci, C. C., Reisen, V. A., Sarnaglia, A. J. Q. & Bondon, P. (2018), 'Empirical study of robust estimation methods for PAR models with application to the air quality area', *Communications in Statistics - Theory and Methods* .

Steutel, F. & Van Harn, K. (1979), 'Discrete analogues of self-decomposability and stability', *The Annals of Probability* pp. 893–899.

Taniguchi, M. & Kakizawa, Y. (2000), *Asymptotic theory of statistical inference for time series*, Springer, New York Berlin Heidelberg.

Ula, T. A. & Smadi, A. A. (1997), 'Periodic stationarity conditions for periodic autoregressive moving average processes as eigenvalue problems', *Water Resources Research* **33**(8), 1929–1934.

Van der Vaart, A. W. (2000), *Asymptotic statistics (Cambridge series in statistical and probabilistic mathematics)*, Cambridge University Press.

Vecchia, A. (1985), 'Periodic autoregressive-moving average (PARMA) modeling with applications to water resources', *Journal of the American Water Resources Association* **21**(5), 721–730.

Viacava, F., Oliveira, R. A. D. d., Carvalho, C. d. C., Laguardia, J. & Bellido, J. G. (2018), 'Sus: oferta, acesso e utilização de serviços de saúde nos últimos 30 anos', *Ciência & Saúde Coletiva* **23**, 1751–1762.

Weiß, C. H. (2008), 'Thinning operations for modeling time series of counts-a survey', *AStA Advances in Statistical Analysis* **92**(3), 319.

Yang, M., Zamba, G. K. & Cavanaugh, J. E. (2013), 'Markov regression models for count time series with excess zeros: A partial likelihood approach', *Statistical Methodology* **14**, 26–38.

Youngster, I., Avorn, J., Belleudi, V., Cantarutti, A., Díez-Domingo, J., Kirchmayer, U., Park, B.-J., Peiró, S., Sanfélix-Gimeno, G., Schröder, H. et al. (2017), 'Antibiotic use in children–a cross-national analysis of 6 countries', *The Journal of Pediatrics* **182**, 239–244.

Zeghnoun, A., Beaudeau, P., Carrat, F., Delmas, V., Boudhabhay, O., Gayon, F., Guincètre, D. & Czernichow, P. (1999), 'Air pollution and respiratory drug sales in the city of le havre, france, 1993–1996', *Environmental Research* **81**(3), 224–230.

Appendix A

Co-authored papers

This Section presents additional papers that I have also worked on during my PhD studies. In addition to my PhD supervisors, other collaborators in Brazil and elsewhere were involved in these papers. My contributions in these works are mainly related to the simulation and applied exercises. Apart from my contributions in the papers below, the experience that I acquired greatly contributed to the development of my thesis because it allowed me to work with a multidisciplinary team. In addition, this experience developed my skills in the analysis of real data, which are also related to air quality control variables, observed in the Greater Vitória region.

These papers are listed below.

A.1 On generalized additive models with dependent time series covariates;

A.2 Management of air quality monitoring networks using robust principal component analysis;

A.3 Parameters influencing population annoyance due to air pollution;

A.4 Deconstruction of annoyance due to air pollution by multiple correspondence analyses;

A.5 Spatial and temporal analysis of the effect of air pollution on children's health.

1 On generalized additive models with dependent time series covariates

Some properties of the GAM-PCA-VAR model are discussed theoretically and verified by simulation in this paper. A real data set is also analyzed with the aim to describe the association between respiratory disease and air pollution concentrations. A hybrid called GAM-PCA-VAR model composed by three statistical tools, the VAR model, PCA and the GAM, with Poisson marginal distribution, was developed. A three-stage estimation method was proposed and studied by simulation for some examples. A real data application was conducted to describe the dependence between the number of hospital admissions for respiratory diseases and air pollutant covariates.

This is a published book chapter of Time Series Analysis and Forecasting, Springer, DOI 10.1007/978-3-319-96944-2. ISBN 9783319969442 (on-line) and 9783319969435 (print).

On generalized additive models with dependent time series covariates

Márton Ispány[1], Valdério A. Reisen[24], Glaura C. Franco[3], Pascal Bondon[4], Higor H. A. Cotta[4], Paulo R. P. Filho[24], and Faradiba S. Serpa[2]

[1] University of Debrecen, Debrecen, Hungary,
ispany.marton@inf.unideb.hu,
WWW home page: https://www.inf.unideb.hu/en/ispanymarton
[2] Federal University of Espírito Santo, Vitória, Brazil
[3] Federal University of Minas Gerais, Belo Horizonte, Brazil
[4] Laboratoire des Signaux et Systèmes (L2S), CNRS-CentraleSupélec-Université Paris-Sud, Gif sur Yvette, France

Abstract. The generalized additive model (GAM) is a standard statistical methodology and is frequently used in various fields of applied data analysis where the response variable is non-normal, e.g., integer valued, and the explanatory variables are continuous, typically normally distributed. Standard assumptions of this model, among others, are that the explanatory variables are independent and identically distributed vectors which are not multicollinear. To handle the multicollinearity and serial dependence together a new hybrid model, called GAM-PCA-VAR model, was proposed in [17] which is the combination of GAM with the principal component analysis (PCA) and the vector autoregressive (VAR) model. In this paper, some properties of the GAM-PCA-VAR model are discussed theoretically and verified by simulation. A real data set is also analysed with the aim to describe the association between respiratory disease and air pollution concentrations.

Keywords: air pollution, generalized additive model, multicollinearity, principal component analysis, time series, vector autoregressive model

1 Introduction

In the recent literature of time series, there has been an outstanding growth in models proposed for data that do not satisfy the Gaussian assumption. This is mainly the case when the response variable under study is a count series or an integer valued series. Procedures developed to analyse this kind of data comprises, for example, observation driven models, see [3] and [6], integer valued autoregressive (INAR) processes, see [1] and [2], or non-Gaussian state space models, see [8] and [10].

This paper is based on the talk "An application of the GAM-PCA-VAR model to respiratory disease and air pollution data" given by the first author.

Particularly in health and environmental studies, where the response variable is typically a count time series, the generalized additive model (GAM) has been widely used to associate the dependent series, such as the number of respiratory or cardiovascular diseases to some pollutant or climate variables, see, for example, [5], [13], [14], [16], [17] and [18] among others. Therefore, in general, the researches related to the study of the association between pollution and adverse health effects usually consider only one pollutant. This simple model choice may be due to the fact that the pollutants are linearly time correlated variables, see the discussion and references in the recent paper [17].

Recently, it has become common practice to use principal component analysis (PCA) in regression models to reduce the dimensionality of an independent set of data, especially the pollutants, which in some instances can include a large number of variables. The PCA is highly indicated to this purpose, as it can handle the multicollinearity problem that can cause biased regression estimates, see, for example, [21].

Nevertheless, use of PCA in the time series context can bring some misspecifications in the fit of the GAM model, as this technique requires that the data should be independent. This problem arises due to the fact that the principal components are linear combinations of the variables. In this context, as the covariates are time series, the autocorrelation present in the observations are promptly transferred to the principal components, see [20].

One solution to this issue was recently proposed by [17], see, also, [18], who introduced a model which combines GAM, PCA and the vector autoregressive (VAR) process. The authors suggest to apply the VAR model to the covariates, in order to eliminate the serial correlation and produce white noise processes, which in turn will be used to build the principal components in the PCA. The new variables obtained in the PCA are finally used as covariates in the GAM model, originating the so called GAM-PCA-VAR model. In their work, the authors have focused on presenting the model and showing its superiority compared to the sole use of GAM or the GAM-PCA procedures, but have not deepened on the theoretical properties of the model.

Thus, to cover this gap, this work aims to state and prove some properties of the GAM-PCA-VAR model, as well as to perform some simulation study to check the results for small samples.

The paper is organized as follows. Section 2 presents the main statistical model, GAM-PCA-VAR, addressed here and its related models as GAM, PCA and VAR, in some detail. In Section 3 the theoretical results are proved for the main model. Section 4 discusses the simulation results and Section 5 is devoted to the analysis of a real data set. Section 6 concludes the work.

2 The GAM-PCA-VAR model

The generalized additive model (GAM), see [11] and [19], with a Poisson marginal distribution is typically used to relate a non-negative integer valued response variable Y with a set of covariates or explanatory variables $X_1, \ldots, X_p$. In GAM

the expected value $\mu = \mathsf{E}(Y)$ of the response variable depends on the covariates via the formula

$$g(\mu) = \beta_0 + \sum_{i=1}^{p} f_i(X_i),$$

where g denotes the link function, β_0 is the intercept parameter and f_i's are functions with a specified parametric form, e.g., they are linear functions $f_i(x) = \beta_i x$, $\beta_i \in \mathbb{R}$, $i = 1, \ldots, p$, or non-parametric, e.g., they are simple smoothing functions like splines or moving averages. The unknown parameters β_0 and f_i, $i = 1, \ldots, p$ can be estimated by various algorithms, e.g., backfitting or restricted maximum likelihood (REML) method. However, if the data observed for variables Y and X_i, $i = 1, \ldots, p$, form a time series the observations cannot be considered as a result of independent experiments and the covariates present strong interdependence, e.g., multicollinearity or concurvity, the standard fitting methods result in remarkable bias, see, e.g., [7] and [17].

Let $\{Y_t\} \equiv \{Y_t\}_{t\in\mathbb{Z}}$ be a count time series, i.e., it is composed of non-negative integer valued random variables. We suppose that the explanatory variables form a zero-mean stationary vector time series $\{\boldsymbol{X}_t\} \equiv \{\boldsymbol{X}_t\}_{t\in\mathbb{Z}}$ of dimension p, i.e., $\boldsymbol{X}_t = (X_{1t}, \ldots, X_{pt})^\top$ where $\top$ denotes the transpose, with the covariance matrix $\Sigma_{\boldsymbol{X}} = \mathsf{E}(\boldsymbol{X}_t \boldsymbol{X}_t^\top)$. Let $\mathcal{F}_t$ denote the σ-algebra which contains the available information up to time t for all $t \in \mathbb{Z}$ from the point of view of the response variable, e.g., $\boldsymbol{X}_t$ is $\mathcal{F}_{t-1}$-measurable. The GAM-PCA-VAR model is introduced in [17] as a probabilistic latent variable model. In this paper, we define this model in a more general form as

$$Y_t \,|\, \mathcal{F}_{t-1} \sim \mathrm{Poi}(\mu_t), \tag{1}$$

$$\boldsymbol{X}_t = \Phi \boldsymbol{X}_{t-1} + A\boldsymbol{Z}_t \tag{2}$$

with link

$$g(\mu_t) = \beta_0 + \sum_{i=1}^{p} \sum_{j=0}^{\infty} f_{ij}(Z_{i(t-j)}), \tag{3}$$

where $\mathrm{Poi}(\cdot)$ denotes the Poisson distribution, the latent variables $\{\boldsymbol{Z}_t\}$, $\boldsymbol{Z}_t = (Z_{1t}, \ldots, Z_{pt})^\top$, form a Gaussian vector white noise process of dimension p with diagonal variance matrix $\Lambda = \mathrm{diag}\{\lambda_1, \ldots, \lambda_p\}$, where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$, A is an orthogonal matrix of dimension $p \times p$, Φ is a matrix of dimension $p \times p$, g is a known link function, β_0 denotes the intercept, and f_{ij}'s are unknown functions. For a zero-mean Gaussian vector white noise process $\{\boldsymbol{Z}_t\}$ with covariance matrix Σ we shall use the notation $\{\boldsymbol{Z}_t\} \sim \mathrm{GWN}(\Sigma)$. See also [4, Definition 11.1.2]. Clearly, for all i, the univariate time series $\{Z_{it}\} \sim \mathrm{GWN}(\lambda_i)$, and $\{Z_{it}\}$ is mutually independent from $\{Z_{jt}\}$ for all $j \neq i$. We assume that all the eigenvalues of Φ are less than 1 in modulus which implies that equation (2) has a unique stationary causal solution. In the case of a Poisson distributed response variable the two widely used link functions are the identity link, $g(z) = z$, and the canonical logarithmic link, $g(z) = \log z$. The set $(\beta_0, \{f_{ij}\}, A, \Lambda, \Phi)$ forms the parameters of the GAM-PCA-VAR model to be estimated. We remark that

in the case of canonical logarithmic link function no additional assumption is needed for the parameters, while in the case of identity link function all the parameters in equation (3), i.e., β_0 and f_{ij}'s, have to be non-negative. It should be also emphasized that the underlying intensity process $\{\mu_t\}$ of $\{Y_t\}$ is also a time series with a complex dependence structure, and μ_t is $\mathcal{F}_{t-1}$-measurable for all $t \in \mathbb{Z}$. One can see that the time series $\{\boldsymbol{X}_t\}$ of covariates depends on $\{\boldsymbol{Z}_t\}$ by formula $\boldsymbol{X}_t = \sum_{k=0}^{\infty} \Phi^k A \boldsymbol{Z}_{t-k}$ for all t, see [4, Example 11.3.1].

The dependence of the response time series $\{Y_t\}$ from the explanatory vector time series $\{\boldsymbol{X}_t\}$ in the GAM-PCA-VAR model can be described by three transformation steps. Clearly, by equation (2), the latent variable can be expressed as $\boldsymbol{Z}_t = A^\top \boldsymbol{U}_t$, where $\boldsymbol{U}_t := \boldsymbol{X}_t - \Phi \boldsymbol{X}_{t-1}$ for all t. Thus, as the first step, the intermediate vector times series $\{\boldsymbol{U}_t\}$ is derived from filtering $\{\boldsymbol{X}_t\}$ by a VAR(1) filter. One can see that $\{\boldsymbol{U}_t\} \sim \mathrm{GWN}(\Sigma_{\boldsymbol{U}})$ where $\Sigma_{\boldsymbol{U}} := A \Lambda A^\top$. Then, as the second step, the latent vector time series $\{\boldsymbol{Z}_t\}$ as principal component (PC) vector is derived by principal component transformation of the intermediate vector white noise $\{\boldsymbol{U}_t\}$. The transformation matrix of the PCA is given by the spectral decomposition of $\Sigma_{\boldsymbol{U}}$. Finally, as the third step, the standard GAM with link (3) is fitting for the response time series $\{Y_t\}$ using the latent vector time series $\{\boldsymbol{Z}_t\}$. The impact of the VAR(1) filter in the first step is to eliminate the serial correlation present in the original covariates. On the other hand, the impact of the PCA in the second step is to eliminate the correlation in the state space of the original covariates. Hence, the result of these two consecutive transformations is the latent vector time series $\{\boldsymbol{Z}_t\}$ whose components, Z_{it}, $i = 1, \ldots, p$, $t \in \mathbb{Z}$, are independent Gaussian variables both in space and time. In the case of logarithmic link function, large positive values in a coordinate of the latent variable indicate locally high influence according to this latent factor. On the contrary, large negative values indicate negligible influence on the response, see, for example, [20]. The order of models in the acronym GAM-PCA-VAR corresponds to these steps starting with the third one and finishing with the first one.

The GAM-PCA-VAR model contains several submodels with particular dependence structure. If $\Phi = 0$ then the VAR equation (2) is simplified to a principal component transformation. In this case, we suppose that there is no serial correlation and we only have to handle the correlation in the state space of covariates. We have two transformation steps: PCA and GAM. This kind of models is called GAM-PCA model that is intensively studied nowadays, see, e.g., [15] and [22]. Beside the full PCA when all PCs are involved into the GAM, we can fit a restricted PCA model by defining $f_{ij} = 0$ for all $i > r$ and $j \geq 0$ where $r < p$. In this case, the first rth PCs are applied as covariates in the GAM step. If the matrices in VAR(1) model (2) have the following block structures

$$\Phi = \begin{bmatrix} \Phi_q & 0 \\ 0 & 0 \end{bmatrix}, \qquad A = \begin{bmatrix} A_q & 0 \\ 0 & I_{p-q} \end{bmatrix},$$

where the eigenvalues of the $q \times q$ matrix Φ_q are less than one in modulus, A_q is an orthogonal matrix of dimension $q \times q$ ($q \leq p$), and $f_{i1}(z) = \beta_i z$ with $\beta_i \in \mathbb{R}$ for

$i = 1, \ldots, r$ ($r \leq q$), f_{i1} is a general smoothing function for $i = q+1, \ldots, p$, $f_{ij} = 0$ otherwise, then we obtain the model that was studied in [17] and applied in the data analysis of Section 5. In this model it is supposed that the set of covariates can be partitioned into two sets: $(X_1, \ldots, X_q)$ are normal covariates, e.g., the pollutant variables in the terminology of Section 5, while $(X_{q+1}, \ldots, X_p)$ are so-called confounding variables as trend, seasonality, etc. The normal covariates satisfy a q-dimensional VAR(1) model, however, instead of the all coordinates of the innovation, only its first rth PCs are involved into the GAM taking into consideration that the covariates present strong inter-correlation. Finally, we note that our model can be further generalized by replacing equation (2) by the more general VARMA or VARIMA or their seasonal variants (SVARMA or SVARIMA) models.

Since the latent variables $\{\boldsymbol{Z}_t\}$ form a Gaussian vector time series, given a sample $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)$, the log-likelihood can be expressed in an explicit form, see [17] for a particular case. Because this log-likelihood is rather complicated a three-stage estimation method is proposed. Firstly, VAR(1) model is fitted to the original covariates by applying standard time series techniques. Secondly, PCA is applied for the residuals defined by $\widehat{\boldsymbol{Z}}_t = \boldsymbol{X}_t - \widehat{\Phi}\boldsymbol{X}_{t-1}$, $t = 2, \ldots, n$, where $\widehat{\Phi}$ denotes the estimated autoregressive coefficient matrix in the fitted VAR(1) model. Thirdly, GAM model is fitted using the PCs. The approach discussed above is similar to the principal component regression, see, e.g., [12, Chapter 8], and it can be considered as a three-stage non-linear regression method.

The first two steps of the above proposed parameter estimation method for GAM-PCA-VAR model can be interpreted as consecutive orthogonalizations, firstly in time and then in the state space of covariates. In [17, Remark] we argued that the order of VAR filter and PCA can not be interchanged because the orthogonalization in the state space does not eliminate the serial correlation and, as the necessary next step, the orthogonalization in time by VAR filter bring back the inter-correlation between the covariates. In what follows, we demonstrate this phenomena by giving a simple example. Let $\{\boldsymbol{X}_t\}$ be a zero-mean causal VAR(1) process defined by

$$\boldsymbol{X}_t = \Psi \boldsymbol{X}_{t-1} + \boldsymbol{W}_t,$$

where $\{\boldsymbol{W}_t\}$ is a zero-mean vector white noise process with variance matrix Σ_W. Suppose that the variance matrix Σ_X of $\{\boldsymbol{X}_t\}$ is diagonal, i.e., the coordinates of $\{\boldsymbol{X}_t\}$ can be interpreted as PCs after PCA. Then Σ_W is not necessarily a diagonal matrix, which implies that a VAR(1) filter may result in an inter-correlated white noise. Namely, consider the following parameters $\Sigma_W = A\Lambda A^\top$ and $\Psi = ASA^\top$, where Λ and S are diagonal matrices and A is an orthogonal matrix. In other words, we suppose that the orthogonal matrix A in the spectral decomposition of Σ_W diagonalizes the autoregressive coefficient matrix as well. Then, we have, by formula (11.1.13) in [4], that

$$\Sigma_X = \sum_{j=0}^{\infty} \Psi^j \Sigma_W (\Psi^\top)^j = \sum_{j=0}^{\infty} AS^j \Lambda S^j A^\top = A\mathrm{diag}\left\{\frac{\lambda_i}{1 - s_i^2}\right\} A^\top.$$

Let $\sigma^2 > \max_i\{\lambda_i\}$ arbitrary and define $s_i := \sqrt{1-\lambda_i/\sigma^2}$ for all i. Clearly, Ψ is a causal matrix since all its eigenvalues are less than 1 in modulus and $\Sigma_X = \sigma^2 I$, i.e., the coordinates of $\{\boldsymbol{X}_t\}$ are uncorrelated. However, the innovation variance matrix Σ_W can be arbitrary proving that the application of VAR filter for a non-intercorrelated vector time series can give inter-correlated vector white noise in its coordinates.

Now, we present some particular examples of GAM-PCA-VAR models.

Example 1. One of the simplest GAM-PCA-VAR models is the model with dimension $p = 1$ and log-linear link function. In this case, there is only one covariate $\{X_t\}$, and the VAR equation (2) is an AR(1) model

$$X_t = \phi X_{t-1} + Z_t, \tag{4}$$

where $|\phi| < 1$ which guarantees the existence of a unique stationary causal solution, $\{Z_t\} \sim \mathrm{GWN}(\lambda)$, $\lambda > 0$. We remark that $A = 1$ in equation (2) in order for the model to be identifiable. The link is log-linear expressed as

$$\log \mu_t = \beta_0 + \beta_1 Z_t. \tag{5}$$

The parameter set of this model is $(\beta_0, \beta_1, \lambda, \phi)$ with parameter space $\mathbb{R}^2 \times \mathbb{R}_+ \times (-1, 1)$. In this model, there is no dimension reduction. Clearly, $Z_t = X_t - \phi X_{t-1}$, thus the response depends on the covariate through the link

$$\log \mu_t = \gamma_0 + \gamma_1 X_t + \gamma_2 X_{t-1}, \tag{6}$$

where there is a one-to-one correspondence between the parameter sets (β_0, β_1, ϕ) and $(\gamma_0, \gamma_1, \gamma_2)$ defined by the equations $\gamma_0 = \beta_0$, $\gamma_1 = \beta_1$ and $\gamma_2 = -\phi\beta_1$ provided $\phi \neq 0$. However, if we fit the standard GAM by using the link (6) with covariates X_t and X_{t-1} at time t, we take no count of the interdependence in time series $\{X_t\}$ which can result in biased and inconsistent estimators of the GAM parameters.

Example 2. Define a particular two-dimensional ($p = 2$) GAM-PCA-VAR model with logarithmic link function in the following way. The two-dimensional covariate vector process $\{\boldsymbol{X}_t\}$, $\boldsymbol{X}_t = (X_{1t}, X_{2t})^\top$, satisfies the VAR(1) model

$$\begin{bmatrix} X_{1t} \\ X_{2t} \end{bmatrix} = \begin{bmatrix} \phi_1 & 0 \\ 0 & \phi_2 \end{bmatrix} \begin{bmatrix} X_{1(t-1)} \\ X_{2(t-1)} \end{bmatrix} + \begin{bmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{bmatrix} \begin{bmatrix} Z_{1t} \\ Z_{2t} \end{bmatrix},$$

where $|\phi_1| < 1$, $|\phi_2| < 1$ and $\{Z_{it}\} \sim \mathrm{GWN}(\lambda_i)$ with $\lambda_i > 0$, $i = 1, 2$, which are independent from each other. Note that the set of two-dimensional orthogonal matrices, A, can be parametrized by an angle parameter $\varphi \in [0, 2\pi)$. We assume that the link is

$$\log \mu_t = \beta_0 + \beta_1 Z_{1t}.$$

The parameter set of this model is $(\beta_0, \beta_1, \varphi, \lambda_1, \lambda_2, \phi_1, \phi_2)$ and the parameter space is $\mathbb{R}^2 \times [0, 2\pi) \times \mathbb{R}_+^2 \times (-1, 1)^2$. Note that, in this model, there is a PCA step as a dimension reduction since only the first coordinate $\{Z_{1t}\}$ of the vector

innovation is involved into the GAM as covariate. One can see that the response depends on the covariates through the link

$$\log \mu_t = \gamma_0 + \gamma_1 X_{1t} + \gamma_2 X_{2t} + \gamma_3 X_{1(t-1)} + \gamma_4 X_{2(t-1)},$$

where $\gamma_0 = \beta_0$, $\gamma_1 = \beta_1 \cos\varphi$, $\gamma_2 = \beta_1 \sin\varphi$, $\gamma_3 = -\beta_1\phi_1 \cos\varphi$ and $\gamma_4 = -\beta_1\phi_2 \sin\varphi$. Thus, the intensity process $\{\mu_t\}$ depends on all coordinates of $\boldsymbol{X}_t$ and $\boldsymbol{X}_{t-1}$. Clearly, there is a one-to-one correspondence between the two parameter sets $(\beta_0, \beta_1, \varphi, \phi_1, \phi_2)$ and $(\gamma_0, \gamma_1, \gamma_2, \gamma_3, \gamma_4)$.

Example 3. A seasonal one-dimensional GAM-PCA-VAR model with linear link function can be defined in the following way. Suppose that the one-dimensional covariate process $\{X_t\}$ satisfies the $\mathrm{SAR}_s(1)$ model:

$$X_t = \phi X_{t-s} + Z_t,$$

where $|\phi| < 1$, $\{Z_t\} \sim \mathrm{GWN}(\lambda)$ with $\lambda > 0$ and $s \in \mathbb{Z}_+$ denotes the seasonal period. The link is linear and is given by

$$\mu_t = \beta_0 + \beta_1 f(Z_t),$$

where $f : \mathbb{R} \to \mathbb{R}_+$ is a known function and $\beta_0, \beta_1 \in \mathbb{R}_+$ are parameters. The parameter set of this model is $(\beta_0, \beta_1, \lambda, \phi)$ with parameter space $\mathbb{R}_+^3 \times (-1, 1)$. The response variable depends on the original covariates through the link

$$\mu_t = \beta_0 + \beta_1 f(X_t - \phi X_{t-s}).$$

If the function f is sufficiently smooth we have by approximation $f(X_t - \phi X_{t-s}) \approx f(X_t) - \phi f'(X_t) X_{t-s}$, and then

$$\mu_t = \gamma_0 + \gamma_1 f_1(X_t) + \gamma_2 f_2(X_t, X_{t-s}), \tag{7}$$

where f_1, f_2 are known functions and $\gamma_0 = \beta_0$, $\gamma_1 = \beta_1$ and $\gamma_2 = -\beta_1\phi$. Thus, the response depends on the original covariate and its s-step lagged series through the standard GAM. However, the covariates in equation (7) are clearly dependent.

3 Theoretical results

In this section, we prove some theoretical results for particular classes of GAM-PCA-VAR models. Consider the log-linear model defined by the link

$$\log \mu_t = \beta_0 + \sum_{i=1}^{p} \sum_{j=0}^{\infty} \beta_{ij} Z_{i(t-j)}, \tag{8}$$

where $\beta_0, \beta_{ij} \in \mathbb{R}$, $i = 1, \ldots, p$, $j \in \mathbb{Z}_+$. The first proposition is about the existence of log-linear GAM-PCA-VAR models.

Proposition 1. *Suppose that* $\sigma^2 := \sum_{i=1}^{p} \lambda_i \sum_{j=0}^{\infty} \beta_{ij}^2$ *is finite. Then the GAM-PCA-VAR model with log-linear link* (8) *has solution* $\{(Y_t, \boldsymbol{X}_t)\}$ *which is a strictly stationary process and* $\mathsf{E}(Y_t) = \mathsf{E}(\mu_t) = \exp(\beta_0 + \sigma^2/2)$ *for all* $t \in \mathbb{Z}$.

Proof. By conditioning we have that

$$\mathsf{E}(Y_t) = \mathsf{E}(\mathsf{E}(Y_t \,|\, \mathcal{F}_{t-1})) = \mathsf{E}(\mu_t) = \mathsf{E}(\exp(\log \mu_t)) = \exp(\beta_0 + \sigma^2/2) \tag{9}$$

is finite since, by equation (8), $\log \mu_t \sim \mathcal{N}(\beta_0, \sigma^2)$, i.e., μ_t has a lognormal distribution, and the moment generating function of $\xi \sim \mathcal{N}(\beta_0, \sigma^2)$ is given by $M_\xi(t) := \mathsf{E}(\exp(t\xi)) = \exp(\beta_0 t + (\sigma t)^2/2)$. Thus, the non-negative integer valued random variable Y_t is finite with probability one for all $t \in \mathbb{Z}$. The vector time series $\{\boldsymbol{Z}_t\}$ forms a Gaussian white noise. Hence it is strictly stationary process with backshift operator $B(\boldsymbol{Z}_t) = \boldsymbol{Z}_{t-1}$ for all $t \in \mathbb{Z}$. Since both stochastic processes $\{Y_t\}$ and $\{\boldsymbol{X}_t\}$ depend on $\{\boldsymbol{Z}_t\}$ through time-invariant functionals, we have the strict stationarity of $\{(Y_t, \boldsymbol{X}_t)\}$ and $B(\boldsymbol{X}_t) = \boldsymbol{X}_{t-1}$, $B(Y_t) = Y_{t-1}$ for all $t \in \mathbb{Z}$. □

In the next proposition, we prove that all moments of the log-linear GAM-PCA-VAR model are finite.

Proposition 2. *Suppose that* σ^2 *defined in Proposition 1 is finite. Then all moments of the stochastic process* $\{(Y_t, \boldsymbol{X}_t)\}$ *are finite. In particular, we have, for all* $t \in \mathbb{Z}$,

$$\begin{aligned}\mathrm{Var}(Y_t) &= \exp(2\beta_0 + \sigma^2)(\exp(\sigma^2) - 1 + \exp(-\beta_0 - \sigma^2/2)),\\ \mathrm{Var}(\mu_t) &= \exp(2\beta_0 + \sigma^2)(\exp(\sigma^2) - 1).\end{aligned}$$

Proof. Let $r \in \mathbb{N}$. Define the rth factorial of a non-negative integer k as $k^{[r]} := k(k-1)\cdots(k-r+1)$ and let $k^{[0]} := 1$. For the rth factorial moment of Y_t we have by conditioning that

$$\begin{aligned}\mathsf{E}(Y_t^{[r]}) =& \sum_{k=0}^{\infty} k^{[r]} \mathsf{P}(Y_t = k) = \mathsf{E} \sum_{k=0}^{\infty} k^{[r]} \mathsf{P}(Y_t = k \,|\, \mathcal{F}_{t-1})\\ =& \mathsf{E} \sum_{k=r}^{\infty} \frac{\mu_t^k}{(k-r)!} e^{-\mu_t} = \mathsf{E}(\mu_t^r)\end{aligned}$$

for all $t \in \mathbb{Z}$. Similarly to (9), we have that the factorial moments are finite, since

$$\mathsf{E}(Y_t^{[r]}) = \mathsf{E}(\mu_t^r) = \mathsf{E}(\exp(r \log \mu_t)) = \exp\{\beta_0 r + (\sigma r)^2/2\}. \tag{10}$$

Since the higher order moments can be expressed by the factorial moment via the formula

$$\mathsf{E}(Y^r) = \sum_{j=0}^{r} S(r, j) \mathsf{E}(Y^{[j]}),$$

where $S(r, j)$'s denotes Stirling numbers of the second kind, the finiteness of all higher order moments follows easily. Since $\{\boldsymbol{X}_t\}$ is a Gaussian process all

its moments are finite. Finally, the existence of mixed moments follows by the Cauchy-Schwarz inequality.

From Equation (10), we have

$$\begin{aligned}\operatorname{Var}(\mu_t) =& \mathsf{E}(\mu_t^2) - \mathsf{E}^2(\mu_t) = \exp(2\beta_0 + (2\sigma)^2/2) - \exp(2\beta_0 + \sigma^2) \\ =& \exp(2\beta_0 + \sigma^2)(\exp(\sigma^2) - 1).\end{aligned}$$

Finally, the formula for $\operatorname{Var}(Y_t)$ can be derived by

$$\operatorname{Var}(Y_t) = \mathsf{E}(\operatorname{Var}(Y_t \,|\, \mathcal{F}_{t-1})) + \operatorname{Var}(\mathsf{E}(Y_t \,|\, \mathcal{F}_{t-1})) = \mathsf{E}(\mu_t) + \operatorname{Var}(\mu_t). \qquad \square$$

The existence of all moments for the log-linear GAM-PCA-VAR process is to be compared with the same result for the integer valued GARCH, so-called INGARCH, process, see [9, Proposition 6]. This implies that the log-linear GAM-PCA-VAR process possesses second and higher order structures, e.g., the autocorrelation function, the spectral density function, the cumulants and the higher order spectra exist. Let ρ_Y denotes the autocorrelation function of the time series $\{Y_t\}$.

Proposition 3. *For the auto- and cross-correlation functions of the GAM-PCA-VAR process $\{(Y_t, \boldsymbol{X}_t)\}$ with intensity process $\{\mu_t\}$, we have $\rho_Y(h) = c_Y \rho(h)$, $\rho_\mu(h) = c_\mu \rho(h)$ and $\rho_{Y\mu}(h) = c_{Y\mu}\rho(h)$ where*

$$\rho(h) := \exp\left(\sum_{i=1}^{p} \lambda_i \sum_{j=0}^{\infty} \beta_{i(j+|h|)}\beta_{ij}\right) - 1, \qquad h \in \mathbb{Z} \setminus \{0\},$$

and the constants $c_Y, c_\mu, c_{Y\mu}$ are defined by

$$c_Y := (\exp(\sigma^2) - 1 + \exp(-\beta_0 - \sigma^2/2))^{-1}, \quad c_\mu := (\exp(\sigma^2) - 1)^{-1}, \quad c_{Y\mu} := \sqrt{c_Y c_\mu}.$$

Moreover, $\operatorname{Cov}(Y_{t+h}, \boldsymbol{X}_t) = \operatorname{Cov}(\mu_{t+h}, \boldsymbol{X}_t) = \mathsf{E}(Y_{t+h}\boldsymbol{X}_t) = \mathsf{E}(\mu_{t+h}\boldsymbol{X}_t) = C(h)$ *with*

$$C(h) := \exp(\beta_0 + \sigma^2/2) \times \begin{cases} \sum_{k=0}^{\infty} \boldsymbol{\Phi}^k A(\boldsymbol{\lambda} \circ \boldsymbol{\beta}_{h+k}) & \text{if} \quad h \geq 0, \\ \sum_{k=0}^{\infty} \boldsymbol{\Phi}^{k-h} A(\boldsymbol{\lambda} \circ \boldsymbol{\beta}_k) & \text{if} \quad h \leq 0, \end{cases} \tag{11}$$

where $\boldsymbol{\lambda} := (\lambda_1, \ldots, \lambda_p)^\top$, $\boldsymbol{\beta}_j := (\beta_{1j}, \ldots, \beta_{pj})^\top$, $j \in \mathbb{Z}_+$, and $\circ$ denotes the entrywise (Hadamard) product.

Proof. Let $h \in \mathbb{N}$. One can see that for the intensity process we have $\mu_{t+h} = \mu_{th}^{(1)} \mu_{th}^{(2)}$ where

$$\log \mu_{th}^{(1)} := \beta_0 + \sum_{i=1}^{p} \sum_{j=1}^{h} \beta_{i(h-j)} Z_{i(t+j)}, \qquad \log \mu_{th}^{(2)} := \sum_{i=1}^{p} \sum_{j=0}^{\infty} \beta_{i(j+h)} Z_{i(t-j)}.$$

Clearly, $\mu_{th}^{(1)}$ is independent of $\mathcal{F}_{t-1}$ and Y_t, while $\mu_{th}^{(2)}$ is $\mathcal{F}_{t-1}$-measurable. Hence, we have by conditioning that

$$\begin{aligned}\mathsf{E}(Y_{t+h}Y_t) =& \mathsf{E}(Y_t\mathsf{E}(Y_{t+h}\,|\,\mathcal{F}_{t+h-1})) = \mathsf{E}(\mu_{t+h}Y_t) = \mathsf{E}(\mu_{th}^{(1)}\mu_{th}^{(2)}Y_t)\\ =& \mathsf{E}(\mu_{th}^{(1)})\mathsf{E}(\mu_{th}^{(2)}\mathsf{E}(Y_t\,|\,\mathcal{F}_{t-1})) = \mathsf{E}(\mu_{th}^{(1)})\mathsf{E}(\mu_{th}^{(2)}\mu_t) = \mathsf{E}(\mu_{t+h}\mu_t)\end{aligned}$$

since μ_t is independent of $\mu_{th}^{(1)}$. This gives the result for $h > 0$. On the other hand, for all $h > 0$, again by conditioning, $\mathsf{E}(Y_{t+h}\mu_t) = \mathsf{E}(\mu_{t+h}\mu_t)$. Thus

$$\mathrm{Cov}(Y_{t+h}, Y_t) = \mathrm{Cov}(\mu_{t+h}, \mu_t) = \mathrm{Cov}(Y_{t+h}, \mu_t), \quad h \in \mathbb{Z}\setminus\{0\}.$$

Since

$$\mathsf{E}(\mu_{t+h}\mu_t) = \mathsf{E}(\mu_{th}^{(1)}\mu_{th}^{(2)}\mu_t) = \mathsf{E}(\mu_{th}^{(1)})\mathsf{E}(\mu_{th}^{(2)}\mu_t)$$

similarly to equation (9) we have

$$\begin{aligned}\mathsf{E}(\mu_{t+h}\mu_t) =& \exp\left(2\beta_0 + \frac{1}{2}\sum_{i=1}^{p}\lambda_i\left(\sum_{j=0}^{h-1}\beta_{ij}^2 + \sum_{j=0}^{\infty}(\beta_{i(j+h)}+\beta_{ij})^2\right)\right)\\ =& \exp\left(\sum_{i=1}^{p}\lambda_i\sum_{j=0}^{\infty}\beta_{i(j+h)}\beta_{ij}\right)\mathsf{E}(\mu_{t+h})\mathsf{E}(\mu_t).\end{aligned}$$

Thus, the first part of the proposition follows by Proposition 2.

Next we prove the formula (11) for the cross-correlations of response and covariate variables. Clearly, by conditioning, $\mathsf{E}(Y_{t+h}\boldsymbol{X}_t) = \mathsf{E}(\mu_{t+h}\boldsymbol{X}_t)$ for all $h \in \mathbb{Z}_+$. On the other hand, for all $t \in \mathbb{Z}$, $h \in \mathbb{Z}_+$, we have $\boldsymbol{X}_{t+h} = \boldsymbol{X}_{th}^{(1)} + \boldsymbol{X}_{th}^{(2)}$ where

$$\boldsymbol{X}_{th}^{(1)} := \sum_{k=1}^{h}\Phi^{h-k}A\boldsymbol{Z}_{t+k}, \qquad \boldsymbol{X}_{th}^{(2)} := \sum_{k=0}^{\infty}\Phi^{h+k}A\boldsymbol{Z}_{t-k}.$$

One can see that $\boldsymbol{X}_{th}^{(1)}$ is independent of $\mathcal{F}_{t-1}$ and Y_t, while $\boldsymbol{X}_{th}^{(2)}$ is $\mathcal{F}_{t-1}$-measurable. Thus, we have that

$$\begin{aligned}\mathsf{E}(\boldsymbol{X}_{t+h}Y_t) =& \mathsf{E}((\boldsymbol{X}_{th}^{(1)} + \boldsymbol{X}_{th}^{(2)})Y_t) = \mathsf{E}(\boldsymbol{X}_{th}^{(1)})\mathsf{E}(Y_t) + \mathsf{E}(\boldsymbol{X}_{th}^{(2)}\mathsf{E}(Y_t\,|\,\mathcal{F}_{t-1}))\\ =& \mathsf{E}(\boldsymbol{X}_{th}^{(1)})\mathsf{E}(\mu_t) + \mathsf{E}(\boldsymbol{X}_{th}^{(2)}\mu_t) = \mathsf{E}(\boldsymbol{X}_{t+h}\mu_t).\end{aligned}$$

Hence $\mathsf{E}(Y_{t+h}\boldsymbol{X}_t) = \mathsf{E}(\mu_{t+h}\boldsymbol{X}_t)$ for all $h \in \mathbb{Z}$ and it is enough to compute the cross-correlation between $\{\boldsymbol{X}_t\}$ and $\{\mu_t\}$. Let $h \geq 0$. For all $\ell \in \{1,\dots,p\}$, $k \in \mathbb{Z}_+$ let $\mathcal{I}_{\ell k}^h := \{1,\dots,p\} \times \mathbb{Z}_+ \setminus (\ell, k+h)$ and define the random variables

$$\log\xi_{\ell k}^{th} := \beta_0 + \sum_{(i,j)\in\mathcal{I}_{\ell k}}\beta_{ij}Z_{i(t+h-j)}, \qquad \log\eta_{\ell k}^{th} := \beta_{\ell(k+h)}Z_{\ell(t-k)}.$$

Then $\mu_{t+h} = \xi_{\ell k}^{th}\eta_{\ell k}^{th}$, where the factors in this decompostion are independent. Since $\mathsf{E}(\mu_{t+h}\boldsymbol{X}_t) = \sum_{k=0}^{\infty}\Phi^k A\mathsf{E}(\mu_{t+h}\boldsymbol{Z}_{t-k})$ and, using the fact that, for $Z \sim \mathcal{N}(0,\lambda)$ and $\beta \in \mathbb{R}$, $\mathsf{E}(Z\exp(\beta Z)) = \beta\lambda\exp(\lambda\beta^2/2)$, we have

$$\mathsf{E}(\mu_{t+h}Z_{\ell(t-k)}) = \mathsf{E}(\xi_{\ell k}^{th}\eta_{\ell k}^{th}Z_{\ell(t-k)}) = \mathsf{E}(\xi_{\ell k}^{th})\mathsf{E}(\eta_{\ell k}^{th}Z_{\ell(t-k)}) = \mathsf{E}(\mu_{t+h})\beta_{\ell(k+h)}\lambda_\ell$$

we obtain the formula (11). The proof is similar for the case $h < 0$. □

Remark 1. It is easy to see that if $\beta_{ij} = \beta_i^j$ for all i, j, then the function ρ is given by $\rho(h) = \exp(\sum_{i=1}^p \lambda_i \beta_i^{|h|}/(1-\beta_i^2)) - 1$, $h \in \mathbb{Z}$. If β_i's are all positive then ρ is positive everywhere and we have autocorrelation functions which are similar to what is displayed in Figure 1. For the one-dimensional model in Example 1 we have the cross-correlation function (CCF) $C(h) = \exp(\beta_0 + \lambda\beta_1^2/2)\lambda\beta_1\phi^{-h}$ for $h \leq 0$ and $C(h) = 0$ for $h > 0$. If $\phi > 0$ then, according to positive or negative β_1, we obtain everywhere positive or negative CCFs. For example, see the CCFs in Figure 2 between the response (Admissions) and pollutants CO, NO_2 that are positive and the CCFs between the response (Admissions) and O_3, SO_2 that are negative at every lag, respectively.

Consider another widely used link function, the linear one, and define the linear GAM-PCA-VAR model by the link

$$\mu_t = \beta_0 + \sum_{i=1}^{p} \sum_{j=0}^{\infty} \beta_{ij} f(Z_{i(t-j)}), \tag{12}$$

where $\beta_0, \beta_{ij} \in \mathbb{R}_+$, $i = 1, \ldots, p$, $j \in \mathbb{Z}_+$ are parameters and $f : \mathbb{R} \to \mathbb{R}_+$ is a known function, e.g., $f(z) = \exp(z)$. Let $\varphi(x \,|\, \lambda)$ denote the probability density function of the normal distribution with mean 0 and variance λ.

Proposition 4. *Suppose that, for all* $i = 1, \ldots, p$, $\sum_{j=0}^{\infty} \beta_{ij} < \infty$ *and* $\tau_i := \int_{-\infty}^{\infty} f(x)\varphi(x \,|\, \lambda_i)\mathsf{d}x < \infty$. *Then the GAM-PCA-VAR model with linear link* (12) *has a strictly stationary solution* $\{(Y_t, \boldsymbol{X}_t)\}$. *Moreover,* $\mathsf{E}(Y_t) = \mathsf{E}(\mu_t) = \beta_0 + \sum_{i=1}^p \tau_i \sum_{j=0}^{\infty} \beta_{ij}$.

Proof. The proof is similar to the proof of Proposition 1. □

Clearly, the assumptions of Proposition 4 do not necessarily garantee the existence of higher order moments of linear GAM-PCA-VAR process. Indeed, the rth order moment $\mathsf{E}(Y_t^r)$ is finite if and only $\int_{-\infty}^{\infty} f^r(x)\varphi(x \,|\, \lambda_i)\mathsf{d}x < \infty$ for all i where $r \geq 1$.

4 Simulation study

In order to evaluate the effect on the parameter estimation of a GAM model in the presence of temporal correlation in the covariate $\{X_t\}$, a simulation study was conducted. The data were generated according to the model discussed in Example 1. Three estimation methods were considered: the standard GAM with only one covariate where the estimated parameters were β_0 and β_1 (M1); the standard GAM with two covariates, the original one and its 1-step lagged series, where the estimated parameters were $\beta_0, \beta_1, \beta_2$ and $\phi = -\beta_2/\beta_1$ (M2); the full GAM-PCA-VAR model by the procedure described in Section 2 where all parameters $\beta_0, \beta_1, \phi, \lambda$ were estimated (M3).

For the model discussed in Example 1 the data were generated under $\beta_0 = 0.2$, $\beta_1 = 1$, $\lambda = 2$ and three scenarios were considered as $\phi = -0.7, 0.3, 0.9$ to

model strong negative, small positive and strong positive correlations, respectively. In order to model the impact due to some unobservable variables, e.g., environmental ones in the context of the next section, independent $\mathcal{N}(0, 0.1)$ distributed random variables were added to the predictor of $\log \mu_t$ for all $t \in \mathbb{Z}$. The sample size $n = 1000$ and the number of Monte Carlo simulations was equal to 100. The empirical values of mean, bias and mean square error (MSE) are displayed in Table 1. All results were obtained by using R-code.

Table 1. Simulation results for model in Example 1

Estimation method	ϕ	Parameter	Mean	Bias	MSE
M1: GAM with X_t	-0.7	$\beta_0 = 0.2$	0.699	0.499	0.253
		$\beta_1 = 1$	0.507	-0.492	0.244
M2: GAM with X_t, X_{t-1}		$\beta_0 = 0.2$	0.204	0.004	0.001
		$\beta_1 = 1$	0.999	-0.001	0.0002
		$\phi = -0.7$	-0.7	0	0.0001
M3: GAM-PCA-VAR		$\beta_0 = 0.2$	0.205	0.005	0.001
		$\beta_1 = 1$	0.999	-0.001	0.0002
		$\phi = -0.7$	-0.695	0.004	0.0005
		$\lambda = 2$	2.003	0.003	0.008
M1: GAM with X_t	0.3	$\beta_0 = 0.2$	0.302	0.102	0.012
		$\beta_1 = 1$	0.905	-0.095	0.009
M2: GAM with X_t, X_{t-1}		$\beta_0 = 0.2$	0.209	0.009	0.001
		$\beta_1 = 1$	0.998	-0.002	0.0002
		$\phi = 0.3$	0.3	0	0.0002
M3: GAM-PCA-VAR		$\beta_0 = 0.2$	0.209	0.009	0.001
		$\beta_1 = 1$	0.999	-0.001	0.0002
		$\phi = 0.3$	0.306	0.006	0.0008
		$\lambda = 2$	1.995	-0.005	0.009
M1: GAM with X_t	0.9	$\beta_0 = 0.2$	1.002	0.802	0.651
		$\beta_1 = 1$	0.191	-0.809	0.655
M2: GAM with X_t, X_{t-1}		$\beta_0 = 0.2$	0.2	0	0.001
		$\beta_1 = 1$	1	0	0.0002
		$\phi = 0.9$	0.899	-0.001	0
M3: GAM-PCA-VAR		$\beta_0 = 0.2$	0.203	0.003	0.001
		$\beta_1 = 1$	1	0	0.0002
		$\phi = 0.9$	0.899	-0.001	0.0001
		$\lambda = 2$	2.007	0.007	0.0086

In the case of standard GAM estimation (M1) it can be seen that the estimate of β_1 is heavily affected by the autocorrelation structure present in the covariate, by presenting a negative bias which increases in absolute value as $|\varphi|$ increases. The estimated MSE also increases substantially with $|\varphi|$. On the other hand, it can also be seen that the fitted standard GAM model tends to severely overestimate β_0. Contrarily, the estimation methods M2 and M3 work equally well, the estimates of the parameters are very close to the true values with no-

ticeably small MSE. The undoubted advantage of method M3 against M2 is that an AR(1) model is also fitted for the covariate where the innovation variance λ is estimated and which can be applied later in the prediction. In this procedure firstly the covariate variable is predicted by equation (4) and then the response variable is predicted by the GAM using the link (5).

5 Application to air pollution data

In this study, the number of hospital admissions (Admissions) for respiratory diseases (RD) as response variable was obtained from the main childrens emergency department in the Vitória Metropolitan Area (called Hospital Infantil Nossa Senhora da Gloria), ES, Brazil. The following atmospheric pollutants as covariates were studied: particulate material (PM_{10}), sulphur dioxide (SO_2), nitrogen dioxide (NO_2), ozone (O_3) and carbon monoxide (CO). For details, e.g., descriptive statistics and basic time series plots, see [17]. The data analysed in this section can be obtained from

`http://wileyonlinelibrary.com/journal/rss-datasets`

The graphs of the sampling functions of the autocorrelations and partial autocorrelations in Figure 1 show that the series of the number of hospital admissions for RD possesses seasonal behaviour, which was to be expected for this phenomena. Another characteristic observed in the series was an apparently weak stationarity. Similar graphs for the pollutant series can be found in [17].

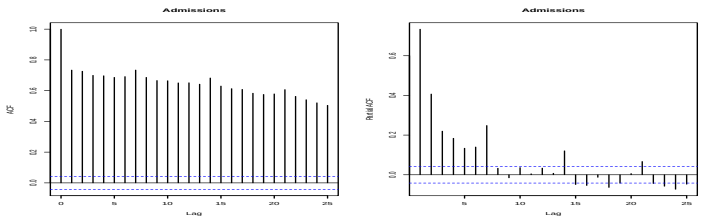


Fig. 1. Sample autocorrelation function (ACF) and partial autocorrelation function (PACF) of the response variable.

Figure 2 shows the sample cross-correlation functions (CCF) between the response and pollutant covariates. As we discussed in Remark 1 four CCF's among them present similar behaviour: the impact of pollutants CO and NO_2 is positive while the impact of SO_2 and O_3 are negative to the response variable at every lag. This observation is consistent with the PCA result presented in [17], see Table 5, where CO and NO_2 form a joint cluster for PC1. On the other hand, all CCF's possess seasonal behaviour as well.

Figure 3 shows the sample cross-correlation functions (CCF) between the response variable and the first three PCs derived from applying PCA for the

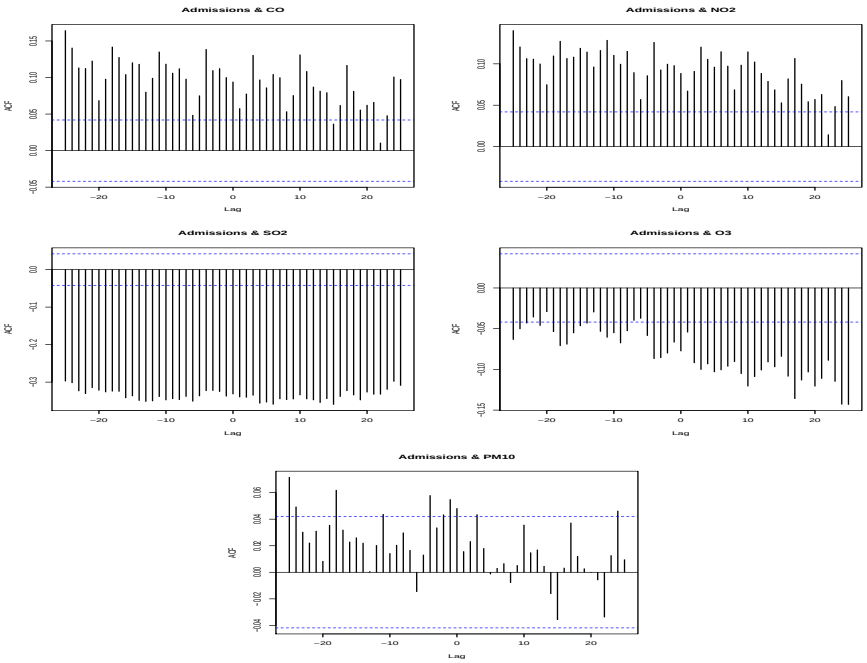


Fig. 2. Sample cross-correlation function (CCF) of the response and pollutant variables.

vector of pollutants. In Section 3.2 of [17], see Table 5 there, one can see that the first three components correspond to 83.2% of the total variability. The temporal behaviour of the PCs is also presented in the autocorrelation plots of [17, Figure 4]. The autocorrelations and the cross-correlations displayed here presented heavy seasonality as well. On the other hand, the shape of the CCFs for the response and PCs can also be classifed into similar groups to the CCFs in Figure 2. The CCF of PC1 is similar to the one of the PM_{10}. The CCF of PC2 displays only negative correlations similar to SO_2 and O_3, while the CCF of PC3 (Figure 3) displays only positive correlations, see CO and NO_2 in Figure 2.

In order to filter the vigorous seasonality both in the response and pollutant variables, seasonal ARMA filters with a 7-day period were applied. The pollutant vector time series and the one-dimensional response time series were filtered by $\text{SVAR}_7(1)$ and $\text{SARMA}_7(1,1)$ processes, respectively. The residuals obtained by these filters indicate remaining significant correlations, see the CCFs between these residuals in Figure 4. The significant cross-correlations and their respective lags are presented in Table 2. Clearly, the correlations which belong to the negative lags are spurious. However, the correlations which belong to the positive lags measure the true impact of a covariate. For example, there are sig-

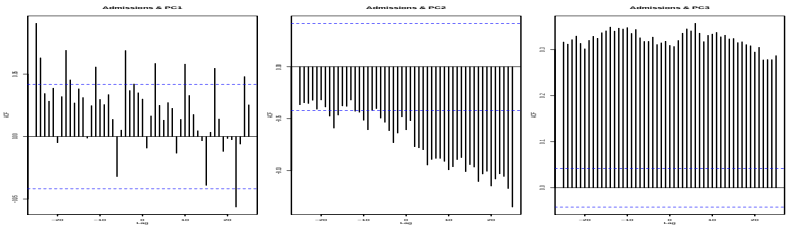


Fig. 3. Sample cross-correlation function (CCF) of the response and first three PCs.

nificant correlations at lag 2 for pollutants PM_{10}, NO_2 and CO equally which could mean that the influence of these pollutants to the response indicates 2 days delay. Contrarily, the influence of the pollutants SO_2 and O_3 presents far delays.

Table 2. Significant cross-correlations and their respective lags between the response and pollutants after the filtering

	RD×SO_2					RD×NO_2				
Lag	-19	-14	-6	12	23	-12	2	4	14	22
Value	-0.063	-0.062	-0.042	-0.047	-0.051	-0.044	-0.050	0.048	0.053	-0.044

	RD×PM10		RD×CO			RD×O_3	
Lag	2	23	-12	2	6	9	25
Value	-0.044	-0.043	-0.053	-0.048	0.045	0.054	-0.055

Figure 5 shows the sample CCF between the residuals of the response variable and the first three PCs after the filtering. The significant cross-correlations and its respective lags are presented in Table 3. It should be emphasized that there are strong coincidences in the lags between Table 2 and 3. For example, the lag 2 in PC1 corresponds to the pollutants PM_{10}, NO_2 and CO, the lag 6 in PC1 corresponds to the pollutant CO, while lag 25 in PC1 corresponds to the pollutant O_3. The lag 12 in PC2 corresponds to the pollutant SO_2. Finally, the lag 14 corresponds to the pollutant NO_2 and the lag 23 to the pollutants SO_2 and NO_2. These correspondences are compatible with the clustering derived in [17, Table 7]. The fitted GAM-PCA-VAR model with its goodness-of-fit measures are reported in [17] as well. We note that in this fitted model $f_{ij} = 0$ was chosen for all $j > 0$. In view of the above results the GAM-PCA-VAR model with link

$$\log \mu_t = \beta_0 + \sum_{i=1}^{p} \sum_{j \in \mathcal{I}_i} f_{ij}(Z_{i(t-j)})$$

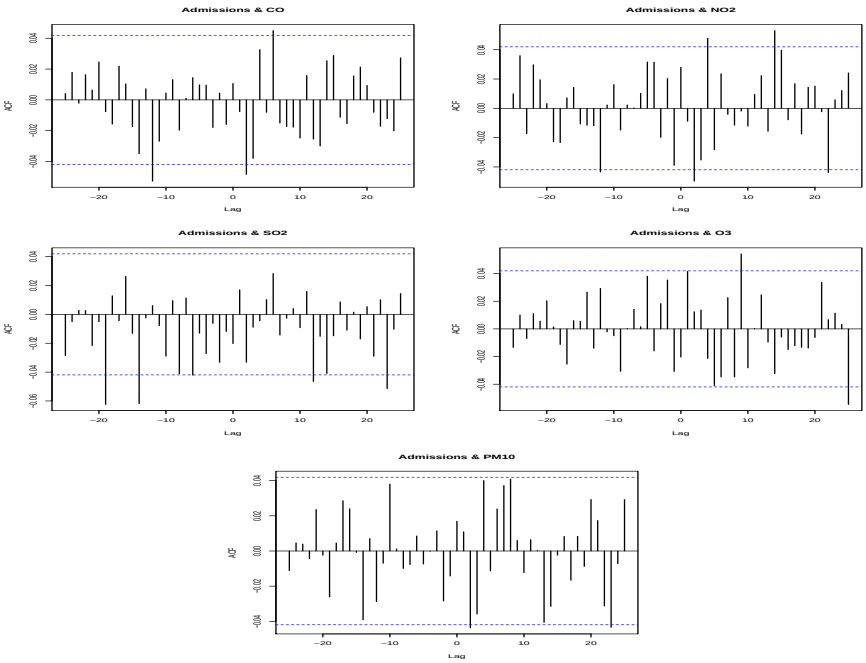


Fig. 4. Sample cross-correlation function (CCF) between the response and pollutant variables after the filtering.

can also be a possible candidate, where $\mathcal{I}_i$ denotes the set of lags which belong to the significant cross-correlation between the residuals of the response and the ith PC. This model can be fitted by using the procedure described in Section 2.

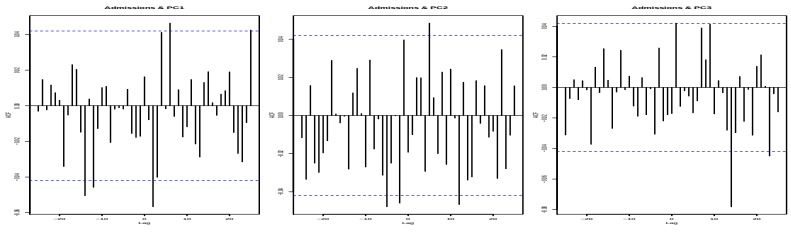


Fig. 5. Sample cross-correlation function (CCF) between the response and PCs after the filtering.

Table 3. Significant cross-correlations and their respective lags between the response variable RD and PCs after the filtering

	RD×PC1					RD×PC2				RD×PC3		
Lag	-14	-12	2	6	25	-5	-2	5	12	1	14	23
Value	-0.051	-0.046	-0.057	0.046	0.043	-0.048	-0.046	0.048	-0.047	0.042	-0.078	-0.045

6 Conclusions

A hybrid called GAM-PCA-VAR model composed by three statistical tools, the VAR model, PCA and the GAM, with Poisson marginal distribution, was developed in a more general framework than in [17]. A three-stage estimation method was proposed and studied by simulation for some examples. Some theoretical properties were also proved. The model was applied to describe the dependence between the number of hospital admissions for respiratory diseases and air pollutant covariates.

An extension of the proposed estimation method for the GAM-PCA-VAR model by a variable selection procedure which ensures that only the significant PCs with their respective lags are involved into the model will be pursed in future works.

Acknowledgments

The authors thank the following agencies for their support: the National Council for Scientific and Technological Development (Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq), the Brazilian Federal Agency for the Support and Evaluation of Graduate Education (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - CAPES), Espírito Santo State Research Foundation (Fundação de Amparo à Pesquisa do Espírito Santo - FAPES) and Minas Gerais State Research Foundation (Fundação de Amparo à Pesquisa do estado de Minas Gerais - FAPEMIG). Márton Ispány was supported by the EFOP-3.6.1-16-2016-00022 project.The project is co-financed by the European Union and the European Social Fund. Pascal Bondon thanks to the Institute for Control and Decision of the Université Paris-Saclay.

References

1. Al-Osh M. A., Alzaid A. A.: First-order integer valued autoregressive (INAR(1)) process. J. Time Ser. Anal. 8, 261–275 (1987)
2. Barczy M., Ispány M., Pap G., Scotto M. G., Silva M. E.: Additive outliers in INAR(1) models. Stat. Pap. 53, 935–949 (2012)
3. Benjamin, M. A., Rigby, R. A., Stasinopoulos, D. M.: Generalized autoregressive moving average models. J. Amer. Statist. Assoc. 98, 214–223 (2003)
4. Brockwell, P. J., Davis, R. A.: Time Series: Theory and Methods. Springer Series in Statistics. New York, Springer-Verlag (1991)

5. Chen, R. J., Chu C., Tan, J., Cao, J., Song, W., Xu, X., Jiang, C., Ma W., Yang, C., Chen, B., Gui, Y., Kan, H.: Ambient air pollution and hospital admission in Shanghai, China. J. Hazard. Mater. 181, 234–240 (2010)
6. Davis, R. A., Dunsmuir, W. T. M., Streett, S. B.: Observation-driven models for Poisson counts. Biometrika 90, 777–790 (2003)
7. Dionisio, K. L., Chang, H. H., Baxter, L. K.: A simulation study to quantify the impacts of exposure measurement error on air pollution health risk estimates in copollutant time-series models. Environ. Health 15:114 (2016)
8. Durbin, J., Koopman, S. J.: Time series analysis of non-Gaussian observations based on state space models from both classical and Bayesian perspectives. J. Roy. Stat. Soc. B 62, 3–56. (2000)
9. Ferland, R., Latour, A., Oraichi, D.: Integer-valued GARCH process. J. Time Ser. Anal. 27(6), 923–942 (2006)
10. Gamerman, D., Santos, T. R., Franco, G. C.: A non-Gaussian family of state-space models with exact marginal likelihood. J. Time Ser. Anal. 34, 625–645 (2013)
11. Hastie, T. J., Tibshirani, R. J.: Generalized Additive Models. London, Chapman and Hall (1990)
12. Jolliffe, I. T.: Principal Component Analysis. 2nd edn. New York, Springer (2002)
13. Nascimento, A. P., Santos, J. M., Mil, J. G., de Souza, J. B., Reis Júnior, N. C., Reisen, V. A.: Association between the concentration of fine particles in the atmosphere and acute respiratory diseases in children. Rev. Saude Publ. 51:3 (2017)
14. Ostro, B. D., Eskeland, G. S., Sánchez, J. M., Feyzioglu, T.: Air pollution and health effects: A study of medical visits among children in Santiago, Chile. Environ. Health Persp. 107, 69–73 (1999)
15. Roberts, S., Martin, M.: Using supervised principal components analysis to assess multiple pollutant effects. Environ. Health Persp. 114(12), 1877–1882 (2006)
16. Schwartz, J.: Harvesting and long term exposure effects in the relationship between air pollution and mortality. Am. J. Epidemiol. 151, 440–448 (2000)
17. de Souza, J. B., Reisen, V. A., Franco, G. C., Ispány, M., Bondon, P., Santos, J. M.: Generalized additive models with principal component analysis: an application to time series of respiratory disease and air pollution data. J. Roy. Stat. Soc. C-App., DOI: 10.1111/rssc.12239, (2017)
18. Souza, J. B., Reisen, V. A, Santos, J. M., Franco, G. C.: Principal components and generalized linear modeling in the correlation between hospital admissions and air pollution. Rev. Saude Publ. 48(3), 451–8 (2014)
19. Wood, S. N.: Generalized Additive Models: An Introduction with R. Second Edition, Chapman and Hall/CRC (2017)
20. Zamprogno, B.: PCA in time series with short and long-memory time series. PhD Thesis at the Programa de Pós-Graduação em Engenharia Ambiental do Centro Tecnológico, UFES, Vitória, Brazil. (2013)
21. Wang, Y., Pham, H.: Analyzing the effects of air pollution and mortality by generalized additive models with robust principal components. Int. J. Syst. Assur. Eng. Manag. 2, 253–259 (2011)
22. Zhao, J., Cao, J., Tian, S., Chen, Y., Zhang, Sh., Wang, Zh., Zhou, X.: A comparison between two GAM models in quantifying relationships of environmental variables with fish richness and diversity indices. Aquat. Ecol. 48, 297–312 (2014)

2 Management of air quality monitoring networks using robust principal component analysis

This article proposes a grouping methodology that applies robust PCA to identify air quality monitoring stations which present similar behavior for any pollutant or meteorological measure. To illustrate the usefulness of the proposed methodology, the robust PCA is applied to the management of the automatic air quality monitoring network of the Greater Vitória Region in Brazil consists of 8 stations. It was found that four components could explain 84% of the total variability and it is possible to create a group composed of at least two stations in each one of the components. Therefore, the redundant stations can be installed in a new site in order to expand the monitored area. This article proposed and applied a grouping methodology to identify monitoring stations which present similar behavior for a given pollutant. As a case of study, the AAQMN of GVR (Brazil) which monitors the PM_{10} pollutant was considered in order to enable better management of the local monitoring network. The methodology proposed consists of the application of robust principal component analysis and selecting the stations which presented higher contributions to the selected PCs. Then, a decision rule is to be applied to decide to keep the redundant station in the same place or to move it to a new area. In the case study, it was found the occurrence of possible outliers observations during the descriptive analysis of the the PM_{10} data which justified the comparison between the robust and standard PCA. It was found that Ibes, Enseada do Suá,and Vitória Centro presented similar behavior and thus can be grouped. Also, that Jardim Camburi and Enseada do Suá form another group. Therefore, two stations, Ibes and Enseada do Suá, are the candidates to be moved to a new site to enlarge the monitored area.

This paper was submitted to publication to the Environmental Monitoring and Assessment Journal.

Environmental Monitoring and Assessment manuscript No.
(will be inserted by the editor)

Management of air quality monitoring networks using robust principal component analysis

Higor Henrique Aranda Cotta · Valdério Anselmo Reisen · Pascal Bondon · Paulo Roberto Prezotti Filho

Received: date / Accepted: date

Abstract Air quality monitoring networks are essentials tools for monitoring air pollutants and, therefore, are important to protect the public health and the environment from the adverse effects of air pollution. It is possible that two or more stations monitor the same pollutant behavior. In this scenario, the equipment must be reallocated in order to provide a better use of public resources and to enlarge the monitored area. To identify which stations, the scientist may apply the principal component analysis (PCA) as a grouping technique. PCA is a tool comprehended by a set of linear combinations constructed to explain the variance-covariance structure of the original data. It is well known that outliers affect the

Higor Henrique Aranda Cotta · Valdério Anselmo Reisen · Paulo Roberto Prezotti Filho
Graduate program in Environmental Engineering, Departament of Statistics, Federal University of Espírito Santo, Brazil E-mail: higor.cotta@centralesupelec.fr

Higor Henrique Aranda Cotta · Valdério Anselmo Reisen · Pascal Bondon · Paulo Roberto Prezotti Filho
Laboratoire des signaux et systèmes, UMR8506, CNRS-Centrale Supelec-Univ. Paris Sud, France

covariance structure of the dataset. Since the components are computed by using the covariance or the correlation matrix, the outliers also affect the properties of the components. This article proposes a grouping methodology that applies robust PCA to identify air quality monitoring stations which present similar behavior for any pollutant or meteorological measure. To illustrate the usefulness of the proposed methodology, the robust PCA is applied to the management of the automatic air quality monitoring network of the Greater Vitória Region in Brazil consists of 8 stations. It was found that four components could explain 84% of the total variability and it is possible to create a group composed of at least two stations in each one of the components. Therefore, the redundant stations can be installed in a new site in order to expand the monitored area.

Keywords Air quality · Monitoring networks · Redundant stations · Robust principal component analysis · Outliers

1 Introduction

The concern about air pollution problems has increased considerably in the last 50 years. Especially in developing countries, the air quality has been degraded as a result of industrialization, population growth, high rates of urbanization, and inadequate or nonexistent policies to control air pollution. The problems caused by air pollution produce local, regional and global impacts. In this context, the particulate matter (PM), especially the PM_{10} which has an aerodynamic diameter less than 10 μm, is one of the most important pollutants with natural and anthropogenic sources. Its adverse impacts on humans health may lead to an increment of mortality rates, respiratory and cardiovascular problems for a shot and longterm

exposure at high concentrations (Beelen et al. 2014; Cesaroni et al. 2014; Hoek et al. 2013; Rückerl et al. 2011).

The main purpose of air quality management is to protect the public health and the environment from the adverse effects of air pollution. An adequate control of air quality involves a number of activities such as risk management, setting standards for emissions and air quality, implementation of control measures and risk communication (WHO 2005). The monitoring of air quality is essential for any air pollution control policy. The realization of efficient management of air quality is important for identifying and quantifying the pollutants found in a region and their sources. This is accomplished by using stations to monitor different pollutants according to the needs of the regions where the stations are installed.

In Brazil, although the limits for pollutants concentrations are clearly established by the federal legislation CONAMA 003/90 (Conselho Nacional do Meio Ambiente 1990), this decree does not contemplate guidelines on how to construct or how to manage monitoring networks and, thus, entrusting this task to each one of the 27 Federative Units. In this scenario, an actual overview of Brazil's air quality monitoring networks is given in a recent publication of the Brazilian Ministry of the Environment. The publication highlights that only 12 out of 27 unity members have an operational air quality monitoring network.

It is desirable that only one monitoring station operates in an area characterized by a specific pattern of air pollution. Pires et al. (2008a) indicate the number of stations that constitute a monitoring network must be optimized in order to reduce costs and expenses. If there are stations with similar patterns of pollution for a specific pollutant, the monitoring equipment could be properly relocated to another area of interest.

In this context, the principal component analysis has been successfully used in air pollution area for managing a network of monitoring stations in several studies, for instance, Zhao et al. (2015) applied PCA to verify reduntant air quality monitoring networks in Shanghai (China). Dominick et al. (2012) used PCA and Cluster Analysis (CA) to check the pattern of behavior of the pollutants carbon monoxide (CO), ozone (O_3), particulate matter of diameter $< 10\mu m$ (PM_{10}), sulfur dioxide (SO_2), nitric oxide (NO) and nitrogen dioxide (NO_2) in five different stations in Malaysia. Pires et al. (2009, 2008a,b) applied PCA to identify monitoring sites with similar concentrations of pollutants for PM_{10}, SO_2, CO, NO_2 and O_3 in the metropolitan area of Porto (Portugal). Lu et al. (2011) employed PCA to the network management of the air quality in Hong Kong for the pollutants of SO_2, NO_2 and Respirable Suspended Particulate (RSP). The authors found that the monitoring stations located in nearby areas are characterized by the same specific air pollution characteristics and suggested that redundant equipment should be transferred to other monitoring stations allowing for further enlargement of the monitored area. Other studies include Lau et al. (2009) and Gramsch et al. (2006).

The application of PCA is not exclusive to the management of air quality monitoring networks. Recently, Villas-Boas et al. (2017) used PCA and nonlinear PCA to assess redundancy of the parameters and monitoring locations of Piabanha water quality network in Brazil. Phung et al. (2015) applied PCA and other multivariate statistical tools to assess the river surface water quality and also redundant monitoring stations in Can Tho City (Vietnam).

At this point, PCA is one of the main multivariate statistical techniques. The goal of PCA is to explain the covariance structure of the data through auxiliary variables called components. These components are constructed from linear com-

binations of the original variables and are uncorrelated. Briefly, PCA calculates the eigenvalues and eigenvectors of the covariance or correlation matrix. The main application of PCA is to reduce the dimensionality of a correlated data matrix of n dimension to a m dimension, where $m < n$. The reduction is performed so that the new set of variables captures most of the variability contained in the original data. A review of the fundamentals of PCA using R (R Core Team 2018) can be found in Sergeant et al. (2016).

Besides the use for dimensionality reduction, the PCA technique can be used for clustering of the variables of a data matrix. Cadima and Jolliffe (1995) discuss the clustering of variables considering the eigenvectors of the PCA. The grouping of variables consists of choosing variables that have similar values for its eigenvectors in absolute value and are highly correlated to the principal component.

In the air pollution context, outliers may arise from different scenarios such as human-made disasters and natural catastrophes, measurement errors due to the failure of equipment or a sudden change in the atmosphere conditions, human errors, among others. Another important situation is when the observed pollutant is under control according to the legislation standards, but it may be considered as an atypical observation in the statistical analysis.

Furthermore, the PCA is sensitive to outliers since the estimation of the mean vector, the covariance matrix and the correlation matrix are directly influenced by outliers. As a consequence, the estimation of the eigenvalues and eigenvectors of the covariance or correlation matrix will be influenced by outliers present in the data, see, e.g., Filzmoser (1999). It is worthwhile to mention that even a single outlier may affect the classical statistics methods. Croux and Haesbroeck (2000)

indicate that conclusions obtained from principal component analysis calculated from a dataset with outliers may be misleading.

Under these circumstances, the common choice made by a wide range of scientists and practitioners to mitigate this problem is to delete the observations suspected to be outliers. As pointed out by Maronna et al. (2006, Chapter 1), the removal of an outlier observation may lead to many issues since the deletion is based on a subjective decision. A viable option to attenuate these problems is to use robust statistical methods since these methods still work well even when the presence of outliers is uncertain. Among the methods for robust estimation of the covariance or correlation matrix with time-independent datasets, there is the estimator proposed by Ma and Genton (2001). This estimator uses the so-called $Q_n(.)$ estimator proposed by Rousseeuw and Croux (1993), which is independent of the location measure of the dataset. In this paper, the central idea is to robustify the estimation of the covariance matrix before calculating its eigenvalues and eigenvectors in PCA.

As mentioned before, PCA based grouping technique is not a new tool for the air pollution area literature, however, other works have not considered the occurrence of outliers (a common issue in air pollution datasets) nor presented a clear methodology of how to perform the grouping through PCA and considered only a subjective approach of it. Therefore, to fulfill this gap, this article proposes a grouping methodology to identify air quality monitoring stations which present similar behavior for any pollutant or meteorological measure. The methodology proposed consists of the application of robust principal component analysis and selecting the stations which presented higher correlations to the selected PCs. Then, a decision rule is to be applied to decide to keep the redundant station

in the same place or to move it to a new area. This proposed methodology is also adequate when outliers are presented in the dataset. The PM_{10} data of the metropolitan area of the Greater Vitória Region (GVR), Brazil, is analyzed as an illustrative example.

The paper is structured as follows: Section 2 describes the data and the statistical model introducing the proposed estimation method and how to identify monitoring stations which present similar behavior; Section 3 presents the data analysis and its discussion comparing robust PCA to the standard one. Finally, Section 4 presents the closing remarks.

2 Data and methods

2.1 Sampling stations in the Greater Vitoria Region

The Greater Vitória Region is located on the southeast coast of Brazil (latitude 20^o 19S, longitude 40^o20W) with a population of approximately 1.900.000 inhabitants. The climate is tropical humid with average temperatures ranging from 24^oC to 30^oC. The region has many ports being an important cargo transport hub in Brazil. Also, there are many industries presented in the region, such as steel plants, iron ore pellet mill, stone quarrying, cement and food industry and asphalt plant.

The automatic air pollution monitoring network (AAQMN) of GVR is consisted by eight monitoring stations distributed in the cities of this region as follows: two stations in Serra (Laranjeiras and Carapina), three stations in Vitória (Jardim Camburi, Enseada do Suá and Vitória Centro), two stations in Vila Velha (Vila Velha Centro and Ibes) and one station in Cariacica (at the regional food distribution center, CEASA). The PM_{10}, in $\mu g/m^3$, is monitored in all stations. Figure 1

presents the geographical location of each station. The PM_{10} series corresponds to the daily average (over 24h period) observed at all stations from January 2005 to December 2009.

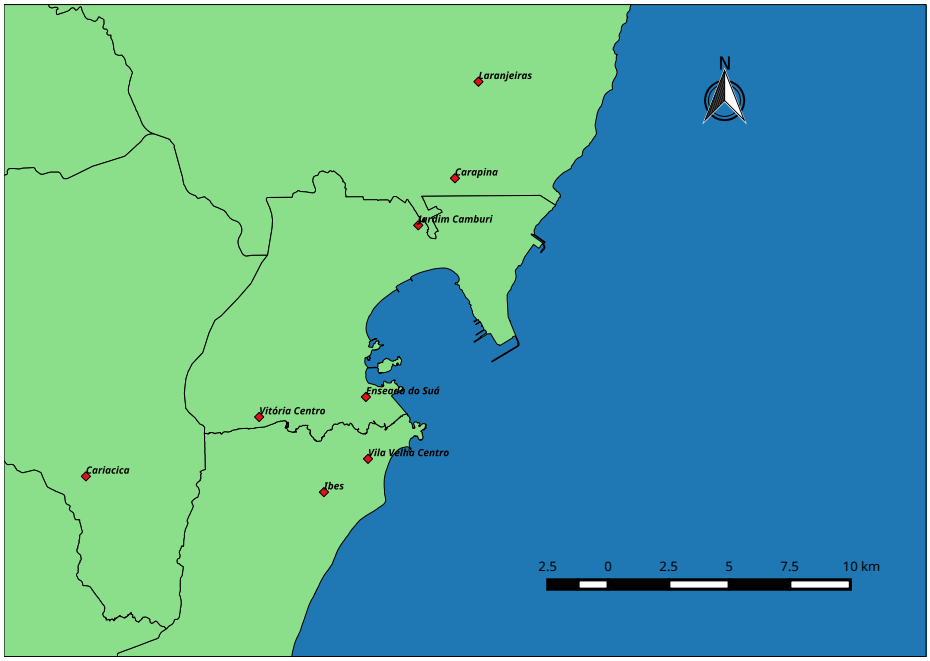


Fig. 1 Geographical location of the stations.

2.2 Principal component analysis

Most of the practitioners employ the standard PCA which is based on the sample covariance matrix and is summarized in the sequel. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be a sample of size n of an independent and identically distributed multivariate distribution with dimension p, mean vector $\boldsymbol{\mu}$, and covariance matrix $\boldsymbol{\Sigma}$. The method of moment estimator (MME) of $\boldsymbol{\Sigma}$ is

$$\hat{\boldsymbol{\Sigma}}_n = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}})(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}})', \tag{1}$$

where $\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i$. As stated by Jollife (2002), the big drawback of PCA tool based on covariance matrices is the sensitivity of the PCs to the units of measurement of the variables. Therefore, if large differences in the variances of variables are found, the variables with large variances will tend to dominate the first PCs. To avoid this problem, the use of PCA based on the correlation matrix is suggested. To this end, the sample correlation matrix $\hat{\boldsymbol{P}}$ can be obtained as $\hat{\boldsymbol{P}} = \hat{\boldsymbol{D}}\hat{\boldsymbol{\Sigma}}_n\hat{\boldsymbol{D}}$, where $\hat{\boldsymbol{D}} = \text{diag}(1/\sqrt{\hat{\boldsymbol{\sigma}}_{11}}, \ldots, 1/\sqrt{\hat{\boldsymbol{\sigma}}_{pp}})$, where $\hat{\sigma}_{ii}$, for $i = 1, \ldots, p$, is the sample covariance. It is straightforward to see that even one outlier will affect the sample mean, and, thus, the whole covariance (or correlation matrix).

Now, consider the random vector $\boldsymbol{X}' = [X_1, X_2, \ldots, X_p]$ with sample covariance matrix $\hat{\boldsymbol{\Sigma}}_n$ and its associated sample eigenvalues $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_p \geq 0$ with corresponding normed eigenvectors $\hat{\boldsymbol{a}}' = [\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_p]$. Let

$$\hat{Y}_i = \hat{\boldsymbol{a}}_i'\boldsymbol{X}. \tag{2}$$

Then, we have

$$\widehat{\text{Var}}(\hat{Y}_i) = \hat{a}_i'\hat{\boldsymbol{\Sigma}}_n\hat{a}_i = \hat{\lambda}_i, \quad i = 1, 2, \ldots, p, \tag{3}$$

$$\widehat{\text{Cov}}(\hat{Y}_i, \hat{Y}_k) = \hat{a}_i'\hat{\boldsymbol{\Sigma}}_n\hat{a}_k = 0, \quad i \neq k, i, k = 1, 2, \ldots, p, . \tag{4}$$

If some $\hat{\lambda}_i$ are equal, the choice of the corresponding eigenvectors $\hat{a}_i$ is not unique.

Associated with (2), it can be shown that

$$\sum_{i=1}^{p} \widehat{\text{Var}}(X_i) = \hat{\lambda}_1 + \hat{\lambda}_2 + \cdots + \hat{\lambda}_p = \sum_{i=1}^{p} \widehat{\text{Var}}(\hat{Y}_i). \tag{5}$$

Equation 5 states that the whole variability of $\boldsymbol{X}$ is retained by the principal components $\hat{\boldsymbol{Y}}$. Therefore, if the main goal of the use of PCA is to reduce the number of variables, the scientist may chose to retain only part of the total original variability.

2.2.1 Robust PCA

Outliers affect the estimation of the location (mean) and the scale (variance) of random variables. To address this problem, Rousseeuw and Croux (1993) proposed a robust estimator, Q_n, for the dispersion of a dataset. Let $X_1, \ldots, X_n$ be n i.i.d. copies of a random variable X, the estimator Q_n is the k-th order

$$Q_n(x) = d\ \{|X_i - X_j|; i < j\}_{\{k\}}, \tag{6}$$

where $i, j = 1, \ldots, n$, and d is a value for consistency of the estimator. The k-th order statistic is the integer value $k = \lfloor (\binom{n}{2} + 2)/4 \rfloor + 1$.

It is known that for any univariate second order random variables X and Y it is possible to compute the covariance between them as follows

$$\mathrm{Cov}(X, Y) = \frac{\alpha\beta}{4}\big(\mathrm{Var}(X/\alpha + Y/\beta) - \mathrm{Var}(X/\alpha - Y/\beta)\big), \tag{7}$$

for any $\alpha, \beta \in \mathbb{R}$, see, Huber (2004). In order to robustify (7), Ma and Genton (2001) proposed to use the estimator Q_n instead of the sample variance obtaining

$$\hat{\sigma}_{Q_n}(X, Y) = \frac{\alpha\beta}{4}\left[Q_n^2\left(\frac{X}{\alpha} + \frac{Y}{\beta}\right) - Q_n^2\left(\frac{X}{\alpha} - \frac{Y}{\beta}\right)\right], \tag{8}$$

where $\alpha = Q_n(X)$ and $\beta = Q_n(Y)$.

The correlation between the univariate second order random variables X and Y can be estimated by

$$\hat{\rho}_{Q_n}(X, Y) = \frac{Q_n^2\left(\frac{X}{\alpha} + \frac{Y}{\beta}\right) - Q_n^2\left(\frac{X}{\alpha} - \frac{Y}{\beta}\right)}{Q_n^2\left(\frac{X}{\alpha} + \frac{Y}{\beta}\right) + Q_n^2\left(\frac{X}{\alpha} - \frac{Y}{\beta}\right)}, \tag{9}$$

where X, Y, α and β are defined in (8).

Let $\boldsymbol{X}$ be a random vector of $p \geq 2$ variables. The robust sample covariance and correlation matrix of the random vector $\boldsymbol{X}$, namely, $\hat{\boldsymbol{\Sigma}}_{Q_n}$ and $\hat{\boldsymbol{P}}_{Q_n}$, respectively,

are obtained by estimating every covariance or correlation pairs between X_i and X_j, $i, j = 1, \ldots, p$. In this work, the robustified principal component analysis is achieved by replacing the standard covariance (or correlation matrix) with $\hat{\boldsymbol{\Sigma}}_{Q_n}$ and $\hat{\boldsymbol{P}}_{Q_n}$.

It is worthwhile to mention that the robust estimation procedure discussed above will provide similar results to the ones estimated using the standard sample estimator when there are no outliers presented in the dataset. Therefore, its usage is recommended.

2.2.2 PCA clustering and station selection

The PCA technique can also be used for clustering of the variables. A method for clustering variable using PCA is discussed Cadima and Jolliffe (1995). The grouping of variables consists of choosing variables that have similar values for its eigenvectors in module and are highly correlated to the principal component. The correlation between a retained PC group and the related full PC (containing all the variability of $\hat{Y}_i$) is given by

$$\hat{r}_k = \hat{\lambda}_j^{1/2} (\hat{\boldsymbol{a}}_j^{k'} \hat{\boldsymbol{\Sigma}}_k^{-1} \hat{\boldsymbol{a}}_j^k)^{1/2}, \tag{10}$$

where $\hat{\lambda}_j$ is eigenvalue of j-th component, $\hat{\boldsymbol{a}}_j^k$ is the clustered vector of $\hat{\boldsymbol{a}}_j$ containing k variables and $\hat{\boldsymbol{\Sigma}}_k^{-1}$ is a sub matrix of $\hat{\boldsymbol{\Sigma}}_n$, which involves lines and columns corresponding to the k grouped variables.

The main idea behind the method is to address the monitoring stations which present similar behaviors for the PM_{10} pollutant concentration (the method is easily expanded to any other pollutant or meteorological parameter). Thus, a

decision rule can be applied to decide to keep the redundant station in the same place or to move it to a new area.

As a possible decision rule, Pires et al. (2009) suggested three criteria: (i) sites should be representative monitoring the highest possible pollutant concentrations; (ii) the number of pollutants being monitored at each site should be maximized; and (iii) the distribution should maximize distances between sites.

In this context, the following methodology for addressing monitoring stations which present similar behavior for a given pollutant is proposed:

1. Perform a descriptive statistical analysis of the data to verify the occurrence of possible outliers and to check for different scale of the measured variables;
2. Compute the robust PCA using the covariance or the correlation matrix;
3. Select a desirable number of PCs to be retained, e.g., 80% or more of the total variability;
4. Arbitrarily choose a cutoff point for the absolute values of the eigenvectors;
5. Create a group of variables whose coefficient of eigenvectors are equal or greater than the cutoff point in the component;
6. Compute using (10) the correlation between the selected variables in the PC and the full component. If the chosen variables and the component are not correlated, e.g., greater than 65%, verify the cutoff point and redo the steps 4-6;
7. Apply the decision criteria of Pires et al. (2009) to decide to keep or to move to a new area the monitoring equipment of the pollutant considered in the study.

3 Data analysis and discussion

In this study, the robust PCA was applied as a classification tool to group monitoring sites with redundant measurements of PM_{10} concentrations from January 1st of 2005 to December 31 of 2009 ($n = 1826$). All the plots and analysis were performed using the computing environment R. $\hat{\boldsymbol{\Sigma}}_{Q_n}$ and $\hat{\boldsymbol{P}}_{Q_n}$ are available in the package *tsqn* (Cotta et al. 2017). The dataset and the R codes are available upon request.

Table 1 shows the descriptive statistics (i.e. the averages, standard deviations and quantile values, among others) of the variables considered. The concentrations of PM_{10} pollutants exceeded hourly and annually the guidelines suggested by the World Health Organization. It is observed a high range for all stations.

The boxplot of the data and the series of the PM_{10} are shown in Figures 2 and 3, respectively. From the boxplot and the plots of the series, one can observe high levels of PM_{10} pollutant. Although the high levels of PM_{10} are important information that should be considered in the context of the air pollution and its impact to human health, these observations can be identified, from a statistical point of view, as being outliers. Therefore, the high levels of PM_{10} presented in the series justify the use and comparison of the robust PCA.

Tables 2 and 3 show the correlations and the robust correlations (as in Section 2.2) between the monitoring stations in the study. From both tables we observe a strong correlations between the variables, e.g., 0.78 for Ibes and Enseada do Suá stations.

The grouping of stations with redundant measurements for the PM_{10} pollutant was carried out following the methodology proposed in Section 2.2.2. That is,

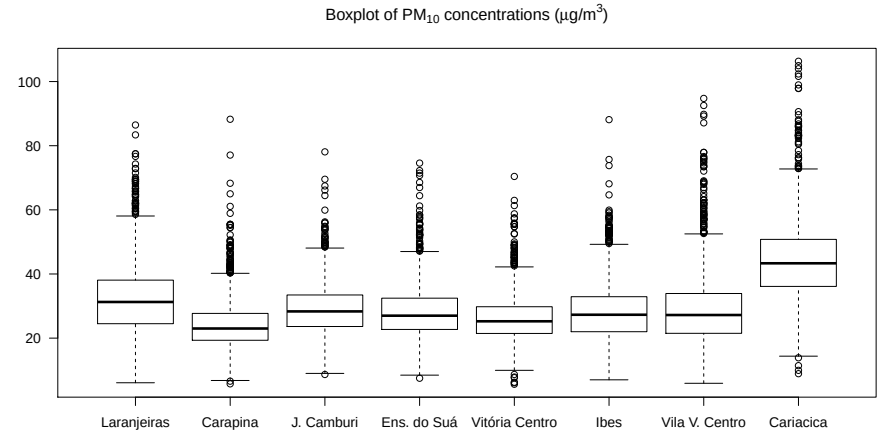


Fig. 2 Boxplot of PM_{10}'s concentrations of the AAQMN of the GVR.

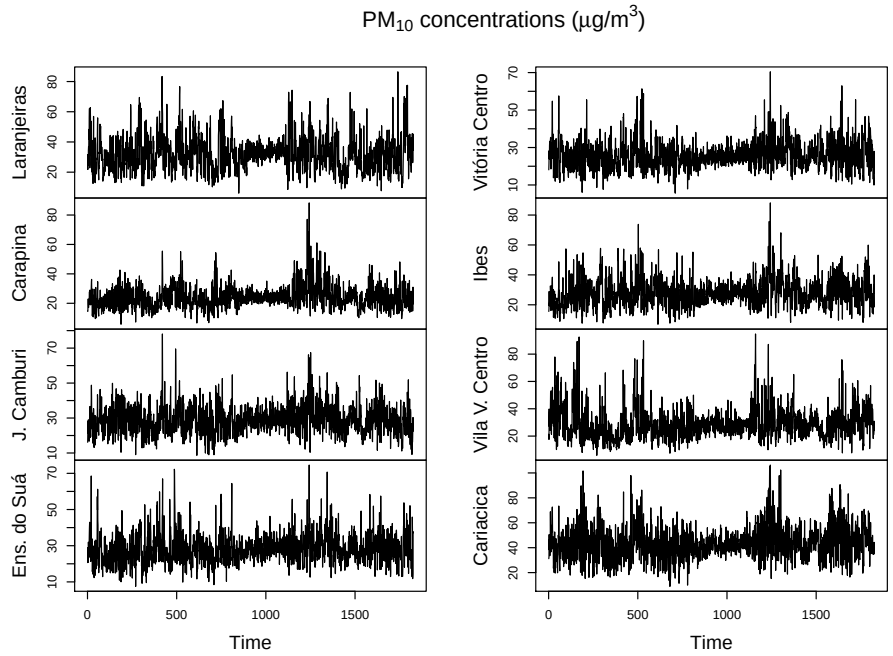


Fig. 3 PM_{10}'s concentrations of the AAQMN of the GVR.

Table 1 Descriptive statistics of PM_{10} data

	Laranjeiras	Carapina	Jardim Camburi	Enseada do Suá	Vitória Centro	Ibes	Vila Velha Centro	Cariacica
Mean	32.26	24.13	28.97	28.08	26.01	28.13	28.94	44.16
Std. Dev	11.29	7.67	8.01	8.29	7.37	9.20	11.33	13.12
Min.	6.08	5.75	8.67	7.50	5.62	7.00	5.92	8.92
25th perc.	24.50	19.33	23.64	22.71	21.46	22.01	21.51	36.14
50th perc.	31.27	23.00	28.33	27.00	25.25	27.29	27.21	43.33
75th perc.	38.07	27.71	33.46	32.46	29.78	32.91	33.92	50.79
Max.	86.46	88.25	78.08	74.58	70.42	88.12	94.75	106.30

Table 2 Correlation matrix ($\hat{P}$) between the stations

	Laranjeiras	Carapina	Jardim Camburi	Enseada do Suá	Vitória Centro	Ibes	Vila Velha Centro	Cariacica
Laranjeiras	1.00							
Carapina	0.35	1.00						
Jardim Camburi	0.52	0.55	1.00					
Enseada do Suá	0.53	0.54	0.53	1.00				
Vitória Centro	0.45	0.63	0.59	0.67	1.00			
Ibes	0.58	0.61	0.61	0.72	0.64	1.00		
Vila Velha Centro	0.38	0.49	0.44	0.46	0.61	0.46	1.00	
Cariacica	0.42	0.70	0.56	0.54	0.71	0.69	0.46	1.00

Table 3 Robust correlation matrix ($\hat{\boldsymbol{P}}_{Q_n}$) between the stations

	Laranjeiras	Carapina	Jardim Camburi	Enseada do Suá	Vitória Centro	Ibes	Vila Velha Centro	Cariacica
Laranjeiras	1.00							
Carapina	0.40	1.00						
Jardim Camburi	0.59	0.57	1.00					
Enseada do Suá	0.59	0.58	0.59	1.00				
Vitória Centro	0.45	0.65	0.61	0.71	1.00			
Ibes	0.66	0.61	0.62	0.78	0.66	1.00		
Vila Velha Centro	0.44	0.55	0.48	0.54	0.60	0.56	1.00	
Cariacica	0.46	0.70	0.55	0.60	0.73	0.69	0.51	1.00

stations having the same contribution in component will have similar values for their eigenvectors and they will also be correlated to the component.

In the PCA tool, the estimates of the eigenvalues and theirs corresponding eigenvectors using $\hat{\boldsymbol{P}}$ and $\hat{\boldsymbol{P}}_{Q_n}$ are given in Table 4. For both estimators, four components could explain approximately 85 % of the total variability of dataset leading to a dimension reduction of the data. It is observed that PCA computed by using $\hat{\boldsymbol{P}}_{Q_n}$ preserved a greater percentage of variability in the components.

For both PCAs, the cutoff point was selected to be 0.37 in absolute value which leaded to the highest correlation values. In the standard PCA, this cutoff leaded to a correlation between the selected PCs groups and the original PCs of 0.96, 0.88, 0.66 and 0.96, for the four PCs, respectively. In the case of robust PCA, correlations of 0.96, 0.89, 0.66 and 0.95 were found. The values are close in both standard and robust PCA.

Thus, for the method of moments estimator for the first component, it is possible to visualize the existence of a group of stations formed by Ibes, Vila Velha Centro , and Cariacica. In the second component, the group is formed by Laranjeiras and Carapina. For the third component Vila Velha Centro forms a group. Finally, the fourth component is the group formed by Jardim Camburi and Enseada do Suá.

For the grouping through robust PCA, in the first component, Ibes, Enseada do Suá, and Vitória Centro can be grouped. For the second component, Laranjeiras and Cariacica form a group. In the third component, Vila Velha Centro is the only station in the group. For the fourth component, the group is formed by Enseada do Suá and Jardim Camburi. Therefore, the proposed method allocated groups

Table 4 PCA results for PM_{10} of AAQMN of the GVR

Stations	PCA - $\hat{\boldsymbol{P}}$				PCA - $\hat{\boldsymbol{P}}_{Q_n}$			
	1	2	3	4	1	2	3	4
Laranjeiras	-0.3002	**0.7193**	-0.1756	0.1460	-0.3123	**0.6998**	0.0533	0.0683
Carapina	-0.3554	**-0.4004**	0.2628	0.1750	-0.3488	**-0.4144**	-0.1961	0.2701
Jardim Camburi	-0.3472	0.1700	0.0502	**0.7019**	-0.3446	0.2356	-0.2115	**0.7037**
Enseada do Suá	-0.3632	0.2163	0.0406	**-0.6118**	**-0.3722**	0.1519	-0.0045	**-0.5144**
Vitória Centro	**-0.3864**	-0.2265	-0.1026	-0.1629	**-0.3745**	-0.2867	-0.0211	-0.1276
Ibes	**-0.3869**	0.1787	0.2359	-0.2271	**-0.3863**	0.1902	-0.0881	-0.3395
Vila Velha Centro	-0.3055	-0.2942	**-0.8391**	0.0141	-0.3203	-0.1838	**0.8942**	0.1475
Cariacica	**-0.3721**	-0.2766	0.3542	0.0507	-0.3625	-0.3283	-0.3259	-0.0962
Eigenvalue	4.8971	0.7744	0.6282	0.4973	5.146	0.7568	0.5334	0.4612
Proportion	61.22	9.68	7.85	6.22	64.25	9.46	6.67	5.77
Cumulative	61.22	70.90	78.75	84.97	64.25	73.71	80.38	86.14

differently from $\hat{\boldsymbol{P}}$. However, base on boxplot (Figure 2) and descriptive statistics (Table 1), the grouping based on $\hat{\boldsymbol{P}}_{Q_n}$ is suggested here.

In order to visually confirm the grouping results for both estimators, the average daily profiles of PM_{10} daily averages for the groups are shown in Figure 4. It is seen that the grouping using $\hat{\boldsymbol{P}}_{Q_n}$ is superior since for the first principal the grouped stations have similar concentrations.

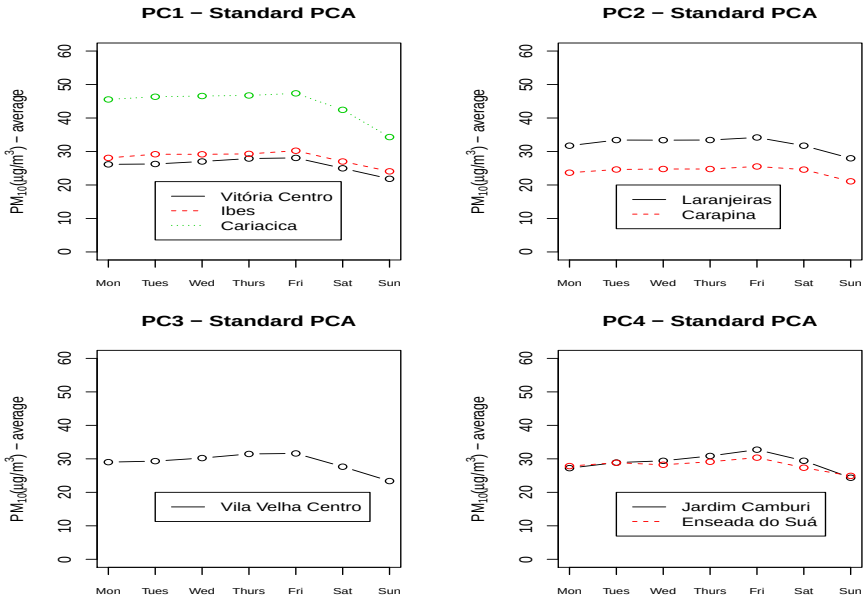
To end this analysis and continuing with the procedure of the methodology discussed in Section 2.2.2, the stations of Ibes and Enseada do Suá are selected to be moved to a new area to enlarge the total monitored area. It is highlighted that although Cariacica has no important contribution to the robust cluster, it is the only station located in Cariacica municipality and, therefore, must be kept.

4 Conclusions

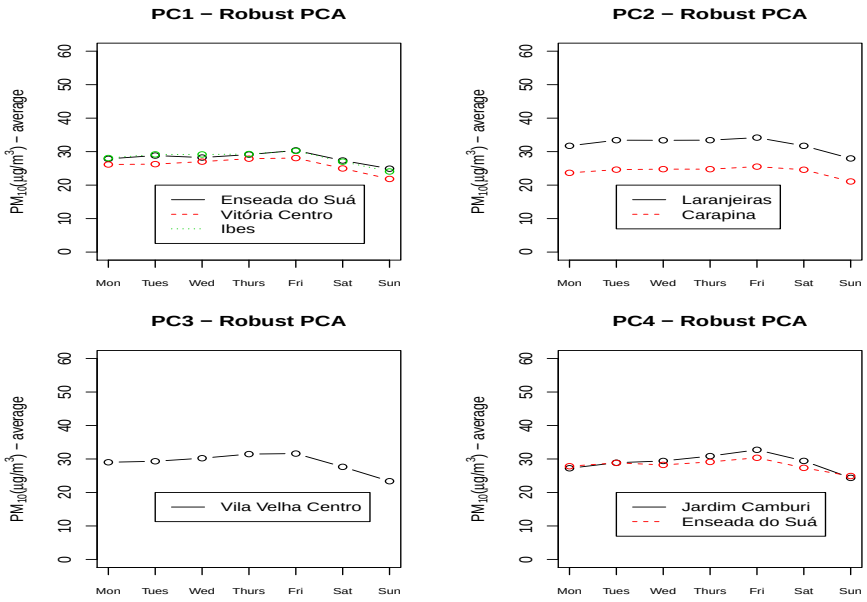
This article proposed and applied a grouping methodology to identify monitoring stations which present similar behavior for a given pollutant. As a case of study, the AAQMN of GVR (Brazil) which monitors the PM_{10} pollutant was considered in order to enable better management of the local monitoring network.

The methodology proposed consists by the application of robust principal component analysis and selecting the stations which presented higher contributions to the selected PCs. Then, a decision rule is to be applied to decide to keep the redundant station in the same place or to move it to a new area.

In the case study, it was found the occurrence of possible outliers observations during the descriptive analysis of the the PM_{10} data which justified the comparison between the robust and standard PCA. It was found that Ibes, Enseada do Suá,



(a) Standard PCA



(b) Robust PCA

Fig. 4 Average daily profile of the PM_{10} data.

and Vitória Centro presented similar behavior and thus can be grouped. Also, that Jardim Camburi and Enseada do Suá form another group. Therefore, two stations, Ibes and Enseada do Suá, are the candidates to be moved to a new site to enlarge the monitored area.

Acknowledgements The authors would like to thank the support from CNPq, ERASMUS, CAPES and FAPES.

References

Beelen R, Raaschou-Nielsen O, Stafoggia M, Andersen ZJ, Weinmayr G, Hoffmann B, Wolf K, Samoli E, Fischer P, Nieuwenhuijsen M, et al. (2014) Effects of long-term exposure to air pollution on natural-cause mortality: an analysis of 22 european cohorts within the multicentre escape project. The Lancet 383(9919):785–795

Cadima J, Jolliffe IT (1995) Loading and correlations in the interpretation of principle compenents. Journal of Applied Statistics 22(2):203–214

Cesaroni G, Forastiere F, Stafoggia M, Andersen ZJ, Badaloni C, Beelen R, Caracciolo B, de Faire U, Erbel R, Eriksen KT, et al. (2014) Long term exposure to ambient air pollution and incidence of acute coronary events: prospective cohort study and meta-analysis in 11 european cohorts from the escape project. BMJ 348:f7412

Conselho Nacional do Meio Ambiente (1990) Resolução CONAMA 003/90. Conama Brasília

Cotta HHA, Reisen VA, Bondon P, Lévy-Leduc C (2017) tsqn: Applications of the Qn Estimator to Time Series (Univariate and Multivariate). R Package version 1.0.0

Croux C, Haesbroeck G (2000) Principal component analysis based on robust estimators of the covariance or correlation matrix: Influence functons and efficiencies. Biometrika 87:603–618

Dominick D, Juahir H, Latif MT, Zain SM, Aris AZ (2012) Spatial assessment of air quality patterns in malaysia using multivariate analysis. Atmospheric Environment 60:172–181

Filzmoser P (1999) Robust principal component and factor analysis in the geostatistical treatment of environmental data. Environmetrics 10:363–375

Gramsch E, Cereceda-Balic F, Oyola P, Von Baer D (2006) Examination of pollution trends in santiago de chile with cluster analysis of PM_{10} and ozone data. Atmospheric environment 40(28):5464–5475

Hoek G, Krishnan RM, Beelen R, Peters A, Ostro B, Brunekreef B, Kaufman JD (2013) Long-term air pollution exposure and cardio-respiratory mortality: a review. Environmental Health 12(1):1

Huber P (2004) Robust Statistics. Wiley Series in Probability and Statistics–Applied Probability and Statistics Section Series, Wiley

Jollife IT (2002) Principal component analysis, 2nd edn. Prentice Hall

Lau J, Hung WT, Cheung CS (2009) Interpretation of air quality in relation to monitoring station's surroundings. Atmospheric Environment 43(4):769–777

Lu WZ, He HD, yun Dong L (2011) Performance assessment of air quality monitoring networks using principal component analysis and cluster analysis. Building and Environment 46(3):577–583

Ma Y, Genton MG (2001) Highly robust estimation of dispersion matrices. Journal of Multivariate Analysis 78:11–36

Maronna R, Martin D, Yohai V (2006) Robust statistics. John Wiley & Sons, Chichester

Phung D, Huang C, Rutherford S, Dwirahmadi F, Chu C, Wang X, Nguyen M, Nguyen NH, Do CM, Nguyen TH, et al. (2015) Temporal and spatial assessment of river surface water quality using multivariate statistical techniques: a study in can tho city, a mekong delta area, vietnam. Environmental monitoring and assessment 187(5):229

Pires JCM, Sousa SIV, Pereira MC, Alvim-Ferraz MCM, Martins FG (2008a) Management of air quality monitoring using principal component and cluster analysis-part i: SO_2 and PM_{10}. Atmospheric Environment 42(6):1249–1260

Pires JCM, Sousa SIV, Pereira MC, Alvim-Ferraz MCM, Martins FG (2008b) Management of air quality monitoring using principal component and cluster analysis-part ii: CO, NO_2 and O_3. Atmospheric Environment 42(6):1261–1274

Pires JCM, Pereira MC, Alvim-Ferraz MCM, Martins FG (2009) Identification of redundant air quality measurements through the use of principal component analysis. Atmospheric Environment 43(25):3837–3842

R Core Team (2018) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL https://www.R-project.org/

Rousseeuw PJ, Croux C (1993) Alternatives to the median absolute deviation. Journal of the American Statistical Association 88(424):1273–1283

Rückerl R, Schneider A, Breitner S, Cyrys J, Peters A (2011) Health effects of particulate air pollution: a review of epidemiological evidence. Inhalation toxicology 23(10):555–592

Sergeant CJ, Starkey EN, Bartz KK, Wilson MH, Mueter FJ (2016) A practitioner's guide for exploring water quality patterns using principal components analysis and procrustes. Environmental monitoring and assessment 188(4):249

Villas-Boas MD, Olivera F, de Azevedo JPS (2017) Assessment of the water quality monitoring network of the piabanha river experimental watersheds in rio de janeiro, brazil, using autoassociative neural networks. Environmental monitoring and assessment 189(9):439

World Health Organization (WHO) (2005) Air quality guidelines: global update 2005: particulate matter, ozone, nitrogen dioxide, and sulfur dioxide. World Health Organization

Zhao L, Xie Y, Wang J, Xu X (2015) A performance assessment and adjustment program for air quality monitoring networks in shanghai. Atmospheric Environment 122:382–392

3 Parameters influencing population annoyance due to air pollution

This paper investigates and quantifies annoyance caused by air pollution by means of surveys carried out during the winter and summer in Vitória, Espírito Santo, Brazil. Results show that 90% of the population reports nuisance by air pollution, from which 60% reported being “very” and “extremely annoyed”. Most respondents perceive air quality as ”important” or” very important” and they also feel “exposed” or “very exposed” to risk from air pollution. The form of air pollution that is mostly perceived by the respondents is dust, i.e., PM.

This paper will be submitted to Risk Analysis Journal.

Parameters influencing population annoyance due to air pollution

Milena Machado de Melo[a,b], Jane Meri Santos[b*], Antônio Fernando Pêgo e Silva [c], Neyval Costa Reis Junior[b], Pascal Bondon[e], Márton Ispány[f], Ilias Mavroidis[d], Paulo Roberto Prezotti Filho[a,c,e], Ana Teresa Lima[b], Valdério Anselmo Reisen[b,c,e]

[a] Instituto Federal de Ciência e Tecnologia do Espírito Santo, Guarapari, Brazil

[b] Department of Environmental Engineering, University Federal of Espirito Santo, Vitoria, Brazil

[c] Department of Statistics, University Federal of Espirito Santo, Vitoria, Brazil

[d] Hellenic Open University, School of Science and Technology, Greece

[e] Laboratoire des Signaux et Systmes (L2S), CNRS-Centrale Supélec-Université Paris-Sud, Gif sur Yvette, France

[f] University of Debrecen, Debrecen, Hungary

* Fax +55 (27) 4009-2648, e-mail: jane.santos@pq.cnpq.br

Abstract

Annoyance has been identified as a useful signal for potential health effects and loss quality of life of pollution in a community. This study deals with the determinants of perceived annoyance in the urban area of Vitoria (Brazil), whose inhabitants are particularly exposed to industrial air pollutants, especially particulate matter (PM). Questionnaire based surveys were conducted in winter and summer, totalling 2638 respondents. A logistic regression was used to identify the relationship between annoyance and air pollution expressed by PM concentrations, and between annoyance and qualitative questionnaire variables. Results show that 90% of the population reports nuisance by air pollution. About 80% of respondents frequently perceived air pollution by dust. People perceived particles (PM) more intensely during the summer, in sunny days and during the daytime. "Perceived importance of air quality", "perceived assessment of air quality", "perceived pollution by dust, the season", and "gender" are the variables most influencing the perceived annoyance. By exposure response relationship when the concentration level of PM increases, the probability of being annoyed also increases. These results can be useful for planning purposes, where policy makers usually do not have access to detailed information, especially for micro-management in a regional or city-planning level.

Keywords: *perceived annoyance, air pollution perception, survey, air quality, particulate matter, dust, exposure-response relationship*.

1. Introduction

Air pollutants consist of substances present in the atmosphere that in a given concentration promotes negative effects on human health, cause fauna, flora, and materials degradation, and also promotes nuisance in local population to the point that degrades public welfare [1]. Particulate matter (PM) originate from dust, soot, combustion of waste, vehicular and industrial emissions represents a major health issue in urban areas since inhalable particles can be a significant contributor to respiratory and cardiovascular diseases, mortality [2–4] and environmental nuisance [5–10]. Annoyance induced by PM is related to its perception, producing physiological and psychological negative effects [11], a state that has repercussions in terms of human welfare and well-being and goes against the World Health Organization (WHO) definition of health [12]. According to H. Orru et al. [13], annoyance has also been identified as a useful signal for potential health effects of pollution in a community. Annoyance caused by PM pollution qualifies as a public health problem, as it can be seen an ambient stressor causing stress and diseases and affect the quality of life [14].

Previous studies have shown that there is a complex relationship between the annoyance and perception of air pollution, health problems, concentration levels of air pollutants, individual and location within the urban sprawl [7,9,15–18]. Differences in people's opinions about levels of annoyance were observed even when exposed to the same levels of pollutants, showing that individual perceptions and attitudes towards the exposure are strongly influenced by factors such as gender, age, education level, health status and neighborhood characteristics. Although intricate, population surveys exploring such attitudes taking in consideration sociodemographic factors and location can help understanding and identifying respondents' reaction to different levels of annoyance when exposed to similar levels of environmental pollutants, e.g. [8,14–17,19]. According to Eek et.al. [20], individuals who assign nuisance to given environmental factors also report subjective health complaints, higher levels of stress, strain, more dissatisfaction with their work situation, and lower personal social support, compared with those who do not report annoyance. Therefore, it is important to understand how people perceive air pollution daily and its impact on the quality of life to reduce or minimize this inherent bias.

According to Holgate et al. [21], PM is usually classified as ultrafine particles ($PM_{0,1}$), fine particles or breathable particles ($PM_{2.5}$) that usually lodge in the terminal bronchiole, coarse particles (inhalable particles (PM_{10}) excluding breathable particles ($PM_{2.5}$)) that penetrate the respiratory system, but are retained in the upper respiratory system, and total suspended particles (TSP), usually larger particles that are in suspension in the atmosphere, with a wide particle size range (typically between 0.005μm and 100μm). All particles sediments and consequently cause annoyance, but particles greater than 10μm are subject to sedimentation in regions closer to the source, while smaller ones tend to sediment in longer time and farther away from the emitting sources [22].

The Metropolitan Region of Vitoria (MRV), state of Espirito Santo, Brazil is a densely inhabited and industrialized urban agglomeration. Pelletizing and steel industries together with the fleet of motor vehicles that has increased dramatically in recent years in MRV [23] have added the hospitalization and emergency care of local population connected to the effects of air pollution on health [24]. Recently, a survey found that about 90% of this population feels very to extremely annoyed by settled particles (dust) [14]. To deepen the previous results, here we use the MRV as proxy for identifying variables related to perceived annoyance caused by air pollution. The hypothesis is that, beyond the levels of pollutants, there are aspects of everyday life, which can explain annoyance due to air pollution reported by inhabitants. The current study aims to identify the parameters influencing annoyance and investigate the relationship between annoyance and concentration levels of TSP and PM_{10} measured in MRV from 2011 to 2014.

2. Materials and methods

2.1. The MRV monitoring network

The Metropolitan Region of Vitória is a highly industrialized urban area located in the Southeast coast of Brazil with 1,500,000 inhabitants. The main air pollutants sources in the region includes ports and steel, pellet making, quarry, cement, chemical, pharmaceutical, food and asphalt industries, among others [23].

Hourly averaged concentration of inhalable particles (PM_{10}) and total suspended particles (TSP) together with other air pollutants are automatically monitored by the

local environmental agency. The monitoring stations are positioned at strategic locations, taking in consideration areas of higher population density. These are Laranjeiras (M1), Carapina (M2), Jardim Camburi (M3), Enseada do Sua (M4), Vitoria Centre (M5), Vila Velha (M6), Ibes (M7) and Cariacica (M8) shown in Figure 1 which also presents the main highways (black lines) and industrial sources of particulate matter (red square). However, PM_{10} and TSP concentration data were not analysed at station M2 as its location (inside a cleared land surrounded by tall trees) interfere with the monitored data and therefore is not representative of the sub-region. Figure 2 shows the 24-hour averaged data of PM_{10} and TSP for the period of this study, from 2011 to 2014.

2.2. The surveys

The surveys consisted of 2638 in-person interviews from 2011 to 2014. The sample size was calculated by simple random sampling with proportional allocation around the coverage area of the eight air quality monitoring stations (called here as sub-regions) [25]. The selection of households was made randomly [26], but considering the spatial distribution of the sample in the sub-regions (in a radius of 1.5 km from each monitoring station). The structured questionnaire was originally developed and validate by [27] and consists of about 50 questions examining different variables (see supplemental material): annoyance (Questions A5, A6, A8 and A11), assessment and importance of air quality (Questions A1 to A4), air pollution perception due to dust (Question A9 to A12), consequences of air pollution (Questions B1 to B8), perceived emission sources (Questions C1 and C2), weather conditions (Questions D1 to D4), health effects (Questions B9 to B11), knowledge of the work of the environmental agency (Question E1), sociodemographic factors (gender, age, marital status, education level, children at home – Questions F3 to F8) and concentration levels of particulate matter (PM_{10} and TSP).

The questionnaire included different types of possibleanswers (see supplemental material). For instance, there was an option of binary response (yes and no) to answer the question “Do you feel annoyed by air pollution?”. And, also a Likert 5-options categorical scale to answer the question “How annoyed do you feel?”, namely: not annoyed, slightly annoyed, moderately annoyed, very annoyed and extremely annoyed [28]. The answer “do not know/do not answer” (NK/NA) was also provided/mentioned.

The purpose of the study was revealed after all questions were answered. Interviewers were trained on good survey practices. A pre-test pilot was carried out using 15 respondents in order to evaluate the questionnaire time (~25 minutes) and data analysis. Finally, a total of five environmental surveys were conducted: two during Winter (July 2011 and July 2013) and three during Summer (January 2012, January 2013 and November 2013).

2.3. Statistical analysis

A logistic regression [29] was chosen for statistical analysis, considering that the study comprised a large number of qualitative variables. This tool is suitable for the analysis of this type of data and also allows the estimation of the odds ratio (Equation 1) [30]. Here were selected as independent variables: sociodemographic factors, weather conditions, assessment and importance of air quality, air pollution perception due to dust, consequences of air pollution, health effects and knowledge of the work of the environmental agency. Annoyance is the dependent variable, with probability of success and failure. The odds ratio (OR) was used to determine whether a variable is determinant for annoyance and to compare the magnitude of influence of the variables of interest on the outcome variable (annoyance). For example, i) OR=1, indicates that the variable does not affect odds of annoyance; ii) OR>1 indicates that the variable is associated with higher odds of annoyance; iii) OR<1 indicates that the variable is associated with lower odds of annoyance [31]. The daily average relationships between PM_{10} and TSP concentrations and annoyance were calculated. “annoyance” was dichotomized as 0 (not annoyed) and 1 (at least slightly annoyed) and used in a logistic regression model. Equation 1 describes the estimated probability exposure–response relationships (already described in [14]). PM_{10} and TSP were chosen as air pollutants indicators as dust was identified in the survey as the main

To Equation 1

where $\hat{P}$ is the probability of perceived annoyance, $\widehat{\beta_0}$ and $\widehat{\beta_1}$ are the parameters estimated from the logistic regression adjustment using the PM_{10} or TSP concentration data and the levels attributed to perceived annoyance.

As shown in Machado et al. [14], the odds ratio can be written as

$\widehat{RR}(x\ ) \approx e^{x\beta}$ Equation 2

3. Results and discussion

3.1. Factors interfering with individual perceived annoyance

Knowledge of the work of the environmental agency
Health effects
Consequences of air pollution by dust
Air pollution perception due to dust
Assessment and importance of air quality
Perceived emission sources
Weather conditions
Sociodemographic factors
Perceived industrial risk verificar!!!! Na importance and assessment

Sociodemographic factors

The 2638 respondents consisted of 59% female and 41% male (Table 1). This represents well the region since the percentage of women is higher than that of men, according to 2010 census data [32]. Table 1 shows the profile of respondents for each sub-region. Regarding gender, there are about 62% of women in the sub-regions M5, M7 and M8 and on average 55% in the other areas. Regarding age, we can see that between 81% (M4) and 93% (M2) of respondents are between 16 and 54 years old. 69% (M3) and 91% (M2) of the respondents have a primary or secondary school diploma, and the percentage of respondents with incomplete primary level studies is about 20%. Most respondents are non-smokers, between 64% (M7) and 73% (M3). Regarding marital status, about 50% of the respondents in all sub-regions are married and most of the families have children living with them. The fact that there have been more female than male respondents may be also partly attributed to the regional socio-economic population profile. It is customary that women stay at home and look after the household, which characterises a typical patriarchal society [33,34].

Weather conditions

Previous studies carried out in the region have shown two distinct periods of predominant wind currents and precipitation levels, from September to March precipitation is high and the prevailing wind direction is North/Northeast (Summer) and from April to August precipitation levels are lower and the prevailing wind direction is southern (Winter) [23,35]. The selected seasons for carrying out the survey considered these weather conditions: Winter (July 2011 and July 2013) and three during Summer (January 2012, January 2013 and November 2013). To assess this seasonality about annoyance, the questionnaire listed the question "Today, how annoyed are you feeling about air pollution?". Figure 3 summarizes the overall results of annoyance according to season (summer and winter). The results reveal that respondents' opinion about perceived annoyance do not show significant differences between the surveys conducted in winter and summer, indicating that people are feeling annoyed by air pollution independent of the season. A similar study conducted by Jacquemin et al. [8] showed that seasonality was not highly correlated to perceived annoyance.

Meanwhile, there is a clear difference in particles deposition rate (dust) during these two seasons as shown by [14]. Thus, a question on whether the respondents perceive the changes on the amount of dust according to the time of the year showed that higher dust perception is detected during the summer (52.5%), followed by winter (36%) (Table S1). As mentioned earlier, summer is the season when the prevailing wind direction is Northeast [23], which transports particles from the industrial sources located North of MRV (Figure 1) directly into the well populated areas. The questionnaire results also suggest that there is higher dust perception in sunny days during daytime (Table S1), in fact suspended particulates in the atmosphere are more visible during sunny days and daytime [36,37].

Perceived emission sources

Annoyance was more related to source proximity than seasonality. Machado et al. [14] showed that more than 90% of respondents were at least slightly annoyed by dust, especially at M1, M3 and M4 where industrial source proximity can play a role on biasing residents' opinion. A similar result was found by Stenlund et al. [16], concluding that proximity to the industry affect air pollution perception.

Table S2 presents the main sources of particles reported by respondents, nonetheless, other sources are also responsible sources of TSP and PM_{10} in the MRV. Respondents identified sources of particles as industrial sources (39%), vehicular (37%) and construction work (11%). Construction work and vehicles have become important sources of particles due to population increasing and wealth improvement in the Country during the period of this research (2011-13). Respondents closer to M3 and M4, the regions closer to the main industrial sources, identified industry as the main pollutant source. At M5, M7, M6 and M8, the respondents reported vehicular sources as the main polluters (Table S2), in agreement with previous results on source apportionment reported by [23].

Furthering industrial weight on population nuisance, the respondents were asked to choose the main benefits and losses brought by local industries and 38% of the respondents selected “all options” of benefits versus 28% “all options” of losses. Despite being alert to air pollution and health deprivation issues, population still considers higher the benefits of having the industrial park close to the city centre.

Assessment and importance of air quality

Table S3 shows that about 90% of respondents consider air quality “very” (38%) and “extremely” (55%) important in all sub-regions of the MRV. More than 50% considered air quality as "extremely important" and about 40% as "very important" showing the awareness about air quality. Regarding the risk imposed by air pollution, about 80% reported feeling at least slightly exposed, especially at M4 where 26% of the respondents felt “extremely” exposed to risk. M4 is one of the prime areas of MRV, situated at the seashore, overlooking the main industrial polluting sources (Figure 1). Thus, plume visibility and air pollutant emissions are well present in residents’ moments of relaxation and fun, contributing to a higher perception of risk exposure. This is corroborated by [16] and [38], who found that degrees of perceived annoyance are positively associated with perceived air pollution and health risk.

Air pollution perception

Table S4 shows the different forms and frequency of perceived air pollution in different aspects: dust (77%), visibility (smog) (32%) and damage to vegetation (31%). Dust is in fact locally perceived by the population as an indicator of air quality. This is in

accordance with literature that shows that population nuisance is often associated with dust and vehicles exhaust odour [9,39].

Consequences of air pollution

Table S5 summarizes the daily actions taken by the population to minimize consequences of dust on their quality of life. Respondents considered "cleaning the house to remove dust" as the main consequence of air pollution, answering "always" (overall 88%) to the frequency that a given action had to be performed. Other consequences include "keep the window closed to prevent the entry of dust" (43%), "paint the house walls" (22%), "occurrence of health problems" (20%), and "avoid frequenting public places" (9%). Reactions are similar if considering each individual sub-region M1-8, however "cleaning" and "painting" the house are indicated mainly at M3 and M4 – the monitoring stations closer to the main industrial source – while "keep the windows closed" is a solution found mainly by M4, M7 and M8.

Health effects

"Health issues" are mainly pointed out at M3 (Table S5). These results confirm that the presence of dust interferes with peoples everyday routine and consequently causes nuisance and a deterioration of the quality of life, especially in close proximity to a source [16]. It should be noted that, although health-related symptoms vary depending on the type of pollutant, particulate material can cause respiratory diseases [40]. Eek [20] showed that people, who have reported environmental annoyance of some sort, also reported more health complaints and higher levels of stress.

For the respondents who reported having had at least rarely health problems caused by dust (Table S5), it was asked what type of disease / symptom. Figure 4 shows the percentage of main types of diseases/symptoms caused by air pollution (dust) reported by respondents. There is a clear tendency of respondents reporting symptoms/problems related to the respiratory system, as allergy, rhinitis, sinusitis, coughing, shortness of breath, etc. [38] studied associations between perceived pollution and health risk among communities in Kenya, and an important result found in their study was that cough/cold, difficulties in breathing, headache and eye problems were the most common health problems mentioned by respondents related to air pollution. This author found a similar

positive association among perceived levels of air pollution, annoyance caused by air pollution ad perceived health risks.

Knowledge of the work of the environmental agency

Finally, Table S6 shows the degree to which respondents are well-informed about the air quality monitoring data and the actions of the local environmental agency. It is interesting to note that more than 60% of respondents have never had information about air quality data and almost 80% does not know of any organization responsible for air quality in the MRV. This result is too important to the government to improve the ways to inform the community about air quality levels, about the limits stablished and about the goals to the local sources. Egondi et al. [38] showed that information about air pollution is an important tool to analyse people's opinions about air pollution perception.

3.2. Odds ratio of annoyance

To identify associations among "annoyance" and the variables selected in this work, a univariate and a multivariate logistic regression analysis was applied inspired in works developed by Oglesby et al. [17]. Rotko et al. [7] and recently [15,41] calculated odds ratio to analyse annoyance. The independent variables selected were: socio demographic factors (Questions F3 (gender), F6 (age), F1 (current occupation), F5 (children at home)), importance (Question A3) and assessment (Question A4) of air quality, perceived industrial risk (Question A7), consequences of air pollution by dust (Questions B1 to), weather conditions (Question D2), health effects (Question B10), perceived emission sources (Question XX), air pollution perception due to dust (Question A9). Each selected variable was quantified (as Likert scale), evaluated and added one by one while the dichotomized perceived annoyance was used as the outcome dependent variable of the univariate regression model. Also, a multivariate regression model (where the variables were added into the model at the same time) was then applied to analyse the combined effect of all selected variables (it is important to emphasize the variables are not correlated and there is no time correlation in those data).

In Table 2 is possible to compare the results of univariate and multivariate models. By the odds ratio value in univariate model all variables are significant, but when we add all variables together (multivariate model) the most significant variables were:

importance and assessment of air quality, air pollution perception due to dust and weather conditions. Thus, the multivariate model that considered the combined effect of all variables is possible to identify the variables that are not determinant to perceived annoyance.

Previous studies that used a similar technique found *perceived air pollution* [7] and *perceived dust* at work [17] as determinant variables of perceived annoyance. Regarding occurrence of health problems the odds of annoyance was not significant, it was not expected, for example, Orru et al [15] found that the odds of being annoyed significantly increased when a person had some respiratory problems (allergies, asthma, and cough). Regarding gender bias which was found significant in our study (Table 2 - univariate logistic regression), it is worth noting that we find studies that defend that females are more sensitive to air pollution than men [8,42] but also the contrary, where no unfairness is observed [7,17]. Gustafson [43] has also connected this difference between men and women due to their roles in society and the power relations that exist between them. This bias is also present in our study, maybe due to the local culture, where female social role is connected to raising children, keeping a clean and healthy household environment for their families.

3.5. Exposure-response relationship (annoyance versus PM_{10} and TSP concentrations)

To complete the possible determinants of perceived annoyance, we verified the relationship between perceived annoyance and exposure to actual PM_{10} and TSP air pollutants. For this, we limited evaluation questions to one: “How annoyed do you feel by air pollution?”. Possible answers varied between “not annoyed” to “extremely annoyed” (supplement questionnaire). Table 3 presents the descriptive statistical values of 24 hour averaged PM_{10} and TSP (maximum, minimum, mean, standard deviation, median and percentile 90. Mean TSP concentration levels are quite high when compared to PM_{10} concentration in all air quality monitoring stations indicating that there is a considerable amount of particles larger than 10 μm (Table 3). Yet, PM_{10} levels surpass those recommended by WHO in most stations (50 $\mu g/m^3$) [12]. The annoyance score was calculated according to Vallack & Shillito [22] and it shows a clear difference between sub-regions (Table 3).

Table 4 was then generated the parameters estimated by the logit regression model for PM10 and TSP separated by each sub-region (from data shown in Table 3). In all sub-regions, the odds ratio values (Exp($\bar{\beta}$)) were approximately equal to 1.1, which means that respondents who live in this regions have 1.1 times higher (10%) the odds of being "annoyed" by a 10 μg/m³ increase in the PM_{10} and TSP concentrations, thus there is no difference between both pollutants.

Table 5 summarizes parameters estimated for PM_{10} and TSP exposure-response logistic regressions for MRV (considering all sub-regions measurements in the same model). The population estimated probability of being annoyed by dust given an average residential TSP and PM_{10} exposure of 30μg/m³ was calculated according to Eq. (1). When the concentration level of TSP is equal to 30μg/m³ there is a probability of 57% of the respondents to be annoyed, while at the same 30μg/m³ of PM_{10} increases that annoyance to 70% (Table 5). This is reasonable since TSP encompasses all PM grain size, i.e. if we have a PM_{10} concentration of 30μg/m³, the corresponding TSP would be higher, justifying the larger levels of annoyance. Amundsen et al. (2008), in a similar study conducted in Norway, found that about 31% of respondents reported being annoyed when exposed to 30μg/m³ of PM_{10}. This result shows that, when compared to literature, Vitória's resident's sensitivity regarding PM perception is above average. This is possibly related to socioeconomic parameters and it reinforces the importance air pollutant emissions control in the region, considering that annoyance is a public health issue that affects the quality of life of citizens [12,14].

Exposure-response curves for TSP and PM_{10} are presented in Figures 5a and 5b respectively. The curve is cumulative and indicates the probability of annoyance caused by PM in the MRV. The bands in dotted lines indicate the 95% confidence intervals of the curves. These are the confidence intervals for the relationships, since at each exposure value there is quite large individual variation in the responses. Considering the WHO guideline for PM_{10} (equal to 50μg/m³ 24-hour mean) about 90% of the population living in the MRV perceive being "annoyed" by air pollution, and for the same concentration of TSP about 73% of the population perceive being "annoyed" by air pollution.

4. Conclusions

This paper investigates and quantifies annoyance caused by air pollution by means of surveys carried out during the winter and summer in Vitória, Espirito Santo, Brazil. Results show that 90% of the population reports nuisance by air pollution, from which 60% reported being “very” and “extremely annoyed”. Most respondents perceive air quality as "important" or" very important" and they also feel “exposed” or “very exposed” to risk from air pollution. The form of air pollution that is mostly perceived by the respondents is dust i.e. PM.

There is no significant difference between perceived levels of annoyance during the Winter and Summer, however people perceived particles (PM) more intensely during the summer, in sunny days and during the daytime. This is possibly due to the most visible presence of dust during the daylight and the fact that dust and particulate matter in general are affecting the quality of life of people more during the summertime, taking also into account that the main beach of Vitoria is located near the main industrial area. Further research could possibly examine perceived annoyance at different times of the day, during the weekend and through surveys performed during different seasons (and not only winter and summer).

Certain regions of the MRV reported higher levels of annoyance (M3 and M4), mostly due to source proximity. Curative air pollution measures need to be applied by citizens, such as “always” clean their houses and that there is a need to “sometimes” keep the windows closed. It should be noted that the main health problems reported by people are allergy, rhinitis and sinusitis, which are common symptoms of pollutants affecting the respiratory system, such as particulate matter (Figure 4). These respondents can also perceive higher levels (“very” and “extremely”) of annoyance from air pollution but, at the same time, reported not to be aware about air quality monitoring in the MRV or about institutions with environmental responsibility. This shows how important it is for the environmental agency not only to monitor air quality and provide policy-makers with the relevant information, but also to increase its visibility and the information provided to the public on air quality monitoring and on the management of atmospheric pollution.

Overall, univariate and multivariate regression models suggest that variables (common in both analysis) influencing perceived annoyance caused by air pollution in the MRV are the “perceived importance of air quality”, the “perceived assessment of air quality”, the “perceived pollution by dust and the season (summer)”, while the univariate analysis showed “gender (female)” was significant. Thus, it is possible to conclude that these main variables play an important role in predicting the perceived annoyance caused by air pollution and therefore in understanding the factors that affect perceived annoyance.

Finally, an exposure–response relationship between the levels of perceived annoyance by air pollution and air pollution concentrations (PM_{10} and TSP) was estimated. The exposure–response relationships for the MRV indicate that many people are annoyed at exposure levels that commonly occur in industrialized cities (daily mean 24h concentrations of PM_{10} and TSP for the last 30 days). This might be due to the industrial park being included in the main residential area of the MRV, adding to impact of other urban sources such as the port area and heavy traffic arteries. Therefore, these relationships are useful for planning purposes, where policy makers usually do not have access to detailed information on such determinant variables which are very useful especially for micro-management in a regional or city-planning level. Such an understanding can help policy makers to examine possible ways and measures to address such parameters to reduce annoyance form air pollution in a more targeted manner, in combination with the more general air pollution abatement measures.

Acknowledgements

The results presented here are part of the PhD thesis of the first author under supervision of J M Santos and V A Reisen in the department of Environmental Engineering at the Federal University of Espirito Santo in 2015. The authors would like to thank CNPq, CAPES, FAPES (Brazilian governmental agencies for technology development and scientific research) and to Centrale Supélec and Université Paris-Sud for its financial support.

References

1. Seinfeld, J. H.; Pandis, S. N. *Atmospheric chemistry and physics: from air pollution to climate change*; 3rd ed.; Wiley: New York, 2016; ISBN 978-1-118-94740-1.

2. Pitchika, A.; Hampel, R.; Wolf, K.; Kraus, U.; Cyrys, J.; Babisch, W.; Peters, A.; Schneider, A. Long-term associations of modeled and self-reported measures of exposure to air pollution and noise at residence on prevalent hypertension and blood pressure. *Sci. Total Environ.* **2017**, *593–594*, 337–346, doi:10.1016/J.SCITOTENV.2017.03.156.
3. Dockery, D. W.; Pope, C. A. Acute Respiratory Effects of Particulate Air Pollution. *Annu. Rev. Public Health* **1994**, *15*, 107–132, doi:10.1146/annurev.pu.15.050194.000543.
4. Schwartz, J. Air Pollution and Daily Mortality: A Review and Meta Analysis. *Environ. Res.* **1994**, *64*, 36–52, doi:10.1006/enrs.1994.1005.
5. Cantuaria, M. L.; Brandt, J.; Løfstrøm, P.; Blanes-Vidal, V. Public perception of rural environmental quality: Moving towards a multi-pollutant approach. *Atmos. Environ.* **2017**, *170*, 234–244, doi:10.1016/J.ATMOSENV.2017.09.051.
6. Klæboe, R.; Öhrström, E.; Turunen-Rise, I. H.; Bendtsen, H.; Nykänen, H. Vibration in dwellings from road and rail traffic — Part III: towards a common methodology for socio-vibrational surveys. *Appl. Acoust.* **2003**, *64*, 111–120, doi:10.1016/S0003-682X(02)00054-3.
7. Rotko, T.; Oglesby, L.; Künzli, N.; Carrer, P.; Nieuwenhuijsen, M. J.; Jantunen, M. Determinants of perceived air pollution annoyance and association between annoyance scores and air pollution (PM2.5, NO2) concentrations in the European EXPOLIS study. *Atmos. Environ.* **2002**, *36*, 4593–4602, doi:10.1016/S1352-2310(02)00465-X.
8. Jacquemin, B.; Sunyer, J.; Forsberg, B.; Gotschi, T.; Bayer-Oglesby, L.; Ackermann-Liebrich, U.; de Marco, R.; Heinrich, J.; Jarvis, D.; Toren, K.; Kunzli, N. Annoyance due to air pollution in Europe. *Int. J. Epidemiol.* **2007**, *36*, 809–820, doi:10.1093/ije/dym042.
9. Amundsen, A. H.; Klæboe, R.; Fyhri, A. Annoyance from vehicular air pollution: Exposure–response relationships for Norway. *Atmos. Environ.* **2008**, *42*, 7679–7688, doi:10.1016/j.atmosenv.2008.05.026.
10. Klæboe, R.; Kolbenstvedt, M.; Clench-Aas, J.; Bartonova, A. Oslo traffic study – part 1: an integrated approach to assess the combined effects of noise and air pollution on annoyance. *Atmos. Environ.* **2000**, *34*, 4727–4736, doi:10.1016/S1352-2310(00)00304-6.
11. Hyslop, N. P. Impaired visibility: the air pollution people see. *Atmos. Environ.*

2009, *43*, 182–195, doi:10.1016/j.atmosenv.2008.09.067.

12. World Health Organization *Air Quality Guidelines for Particulate Matter, Ozone, Nitrogen Dioxide and Sulphur Dioxide*; Geneva, 2006;
13. Orru, H.; Idavain, J.; Pindus, M.; Orru, K.; Kesanurm, K.; Lang, A.; Tomasova, J. Residents' Self-Reported Health Effects and Annoyance in Relation to Air Pollution Exposure in an Industrial Area in Eastern-Estonia. *Int. J. Environ. Res. Public Health* **2018**, *15*, 252, doi:10.3390/ijerph15020252.
14. Machado, M.; Santos, J. M.; Reisen, V. A.; Reis, N. C.; Mavroidis, I.; Lima, A. T. A new methodology to derive settleable particulate matter guidelines to assist policy-makers on reducing public nuisance. *Atmos. Environ.* **2018**, *182*, 242–251, doi:10.1016/j.atmosenv.2018.02.032.
15. Orru, K.; Nordin, S.; Harzia, H.; Orru, H. The role of perceived air pollution and health risk perception in health symptoms and disease: a population-based study combined with modelled levels of PM10. *Int. Arch. Occup. Environ. Health* **2018**, *91*, 581–589, doi:10.1007/s00420-018-1303-x.
16. Stenlund, T.; Lidén, E.; Andersson, K.; Garvill, J.; Nordin, S. Annoyance and health symptoms and their influencing factors: A population-based air pollution intervention study. *Public Health* **2009**, *123*, 339–345, doi:10.1016/j.puhe.2008.12.021.
17. Oglesby, L.; Künzli, N.; Monn, C.; Schindler, C.; Ackermann-Liebrich, U.; Leuenberger, P. Validity of annoyance scores for estimation of long term air pollution exposure in epidemiologic studies: the Swiss Study on Air Pollution and Lung Diseases in Adults (SAPALDIA). *Am. J. Epidemiol.* **2000**, *152*, 75–83, doi:10.1080/10473289.2000.10464156.
18. Atari, D. O.; Luginaah, I. N.; Fung, K. The Relationship between Odour Annoyance Scores and Modelled Ambient Air Pollution in Sarnia, "Chemical Valley", Ontario. *Int. J. Environ. Res. Public Health* **2009**, *6*, 2655–2675, doi:10.3390/ijerph6102655.
19. Blanes-Vidal, V.; Suh, H.; Nadimi, E. S.; Løfstrøm, P.; Ellermann, T.; Andersen, H. V.; Schwartz, J. Residential exposure to outdoor air pollution from livestock operations and perceived annoyance among citizens. *Environ. Int.* **2012**, *40*, 44–50.
20. Eek, F.; Karlson, B.; Österberg, K.; Östergren, P.-O. Factors associated with prospective development of environmental annoyance. *J. Psychosom. Res.* **2010**,

69, 9–15, doi:10.1016/j.jpsychores.2009.12.001.

21. Holgate, S. T.; Samet, J. M.; Koren, H. S.; Maynard, R. L. *Air pollution and health*; Academic Press, 1999; ISBN 9780123523358.
22. Vallack, H. .; Shillito, D. . Suggested guidelines for deposited ambient dust. *Atmos. Environ.* **1998**, *32*, 2737–2744, doi:10.1016/S1352-2310(98)00037-5.
23. Santos, J. M.; Reis Jr, N. C.; Galvão, E. S.; Silveira, A.; Goulart, E. V.; Lima, A. T.; Reis, N. C.; Galvão, E. S.; Silveira, A.; Goulart, E. V.; Lima, A. T. Source apportionment of settleable particles in a mining-impacted urban and industrialized region in Brazil. *Environ. Sci. Pollut. Res.* **2017**, doi:10.1007/s11356-017-9677-y.
24. de Souza, J. B.; Reisen, V. A.; Franco, G. C.; Ispány, M.; Bondon, P.; Santos, J. M. Generalized additive models with principal component analysis: an application to time series of respiratory disease and air pollution data. *J. R. Stat. Soc. Ser. C (Applied Stat.* **2018**, *67*, 453–480, doi:10.1111/rssc.12239.
25. Barnett, V. *Sample survey : principles and methods*; 3rd ed.; Wiley: Chichester, 2002; ISBN 9780470685907.
26. Cochran, W. G. *Sampling techniques - A Wiley publication in applied statistics*; 3rd ed.; Wiley: New York, 1977; ISBN 047116240X.
27. Calvo-Mendieta, I.; Flaquart, H.; Frere, S.; Gonthier, F.; Hellequin, A. P.; Blanc, A. Le *Perception du risque industriel par les populations du Dunkerquois, Rapport intermédiaire.*; Dunkerque, 2008;
28. Berglund, B.; Berglund, U.; Lindvall, T. Measurement and control of annoyance. In *Environmental annoyance: characterization, measurement and control*; Koelga, H. S., Ed.; Elsevier Sci B.V.: Amstredam, 1987; Vol. 15, pp. 29–43.
29. Abraham, B.; Ledolter, J. *Introduction to regression modeling*; Thomson Brooks/Cole, 2006; ISBN 0534420753.
30. Baxter, L. A.; Finch, S. J.; Lipfert, F. W.; Yu, Q. Comparing Estimates of the Effects of Air Pollution on Human Mortality Obtained Using Different Regression Methodologies. *Risk Anal.* **1997**, *17*, 273–278, doi:10.1111/j.1539-6924.1997.tb00865.x.
31. Domínguez-Almendros, S.; Benítez-Parejo, N.; Gonzalez-Ramirez, A. R. Logistic regression models. *Allergol. Immunopathol. (Madr).* **2011**, *39*, 295–305, doi:10.1016/j.aller.2011.05.002.
32. IBGE *Sensus 2010*; Instituto Brasileiro de Geografia e Estatística: Brasilia, 2010;

33. ONUBR População brasileira ainda é patriarcal, mostra pesquisa do IPEA apoiada pela ONU 2014.
34. Epstein, D. G. *Brasilia, Plan and reality: A Study of Planned and Spontaneous Urban Development*; University of California Press: Brasilia, 1973;
35. Salvador, N.; Reis, N. C.; Santos, J. M.; Albuquerque, T. T. A.; Loriato, A. G.; Delbarre, H.; Augustin, P.; Sokolov, A.; Moreira, D. M. Evaluation of weather research and forecasting model parameterizations under sea-breeze conditions in a North Sea coastal environment. *J. Meteorol. Res.* **2016**, *30*, 998–1018, doi:10.1007/s13351-016-6019-9.
36. Albuquerque, T. T. de A.; Andrade, M. de F.; Ynoue, R. Y. Characterization of atmospheric aerosols in the city of São Paulo, Brazil: comparisons between polluted and unpolluted periods. *Environ. Monit. Assess.* **2012**, *184*, 969–984, doi:10.1007/s10661-011-2013-y.
37. Castanho, A. D. A.; Artaxo, P. Wintertime and summertime São Paulo aerosol source apportionment study. *Atmos. Environ.* **2001**, *35*, 4889–4902, doi:10.1016/S1352-2310(01)00357-0.
38. Egondi, T.; Kyobutungi, C.; Ng, N.; Muindi, K.; Oti, S.; van de Vijver, S.; Ettarh, R.; Rocklöv, J. Community perceptions of air pollution and related health risks in Nairobi slums. *Int. J. Environ. Res. Public Health* **2013**, *10*, 4851–68, doi:10.3390/ijerph10104851.
39. Forsberg, B.; Stjernberg, N.; Wall, S. People can detect poor air quality well below guideline concentrations: a prevalence study of annoyance reactions and air pollution from traffic. *Occup. Environ. Med.* **1997**, *54*, 44–8, doi:10.1136/oem.54.1.44.
40. Kampa, M.; Castanas, E. Human health effects of air pollution. *Environ. Pollut.* **2008**, *151*, 362–367, doi:10.1016/j.envpol.2007.06.012.
41. Cantuaria, M. L.; Løfstrøm, P.; Blanes-Vidal, V. Comparative analysis of spatio-temporal exposure assessment methods for estimating odor-related responses in non-urban populations. *Sci. Total Environ.* **2017**, *605–606*, 702–712, doi:10.1016/j.scitotenv.2017.06.220.
42. Hellequin, A. P.; Zwarterook, I. Mauvais air : vivre au quotidien près d'usines polluantes et dangereuse. In *IIeme colloque international UMR 5600-ENTPE "Le risque industriel : une question de sciences humaines";* Lyon, 2010.
43. Gustafson, P. E. Gender Differences in Risk Perception: Theoretical and

Methodological Perspectives. *Risk Anal.* **1998**, *18*, 805–811, doi:10.1023/B:RIAN.0000005926.03250.c0.

4 Deconstruction of annoyance due to air pollution by multiple correspondence analyses

The aim of this study is to apply Multiple Correspondence Analysis (MCA) to investigate empirically the relationship among variables that can be considered determinants for the perception of annoyance caused by air pollution, and compare the results for two different samples, related to populations of Vitória, Brazil, and Dunkirk, France. The results show that inhabitant of Dunkirk perceived that the main sources of air pollution causing annoyance are related to industry, while in Vitoria the construction work and vehicle sources were indicated as the sources of air pollution, although both cities have similar industrial characteristics.For both cities,the MCA analysis showed a positive progressive correspondence between the levels of perceived annoyance and the variables' categories: importance of air quality, perceived exposure to industrial risk, assessment of air quality and perceived air pollution.

This paper was submitted to publication to Atmospheric Environment Journal.

Deconstruction of annoyance due to air pollution by multiple correspondence analyses

Milena Machado[a]*, Severine Frere[b], Phillipe Chagnon[b], Jane Meri Santos[c], Valdério Anselmo Reisen[d], Ilias Mavroidis[e], Neyval Costa Reis Junior[c], Pascal Bondon[f], Márton Ispány[g], Paulo Roberto Prezotti Filho[a,c,f]

[a] Instituto Federal de Ciência e Tecnologia do Espírito Santo, Guarapari -E.S.- Brazil

[b]Université du Littoral Côte d'Opale, Maison de la Recherche em Science de l´homme, Dunkerque, France

[c] Department of Environmental Engineering, Universidade Federal do Espírito Santo, Vitoria, Brazil

[d]Department of Statistics, Universidade Federal do Espírito Santo, Vitoria, Brazil

[e]Hellenic Open University, School of Science and Technology, Greece

[f]Laboratoire des Signaux et Systems (L2S), CNRS-Centrale Supélec-Université Paris-Sud, Gif sur Yvette, France

[g] University of Debrecen, Debrecen, Hungary

* Tel: +55 (27) *988527717*, e-mail: milenamm@ifes.edu.br

Abstract:
Annoyance caused by air pollution is an important public health issue since it can cause stress and ill-health, affects quality of life and increases mortality. In this context, the aim of this study is to apply Multiple Correspondence Analysis (MCA) to investigate empirically the relationship among variables that can be considered determinants for the perception of annoyance caused by air pollution. Face-to-face survey studies were conducted in two industrialized urban areas (Vitoria- Brazil and Dunkirk- France), because in these two regions the populations often report feeling annoyed by air pollution. The results show that habitant of Dunkirk perceived that the main sources of air pollution causing annoyance are related to industry, while in Vitoria the construction work and vehicle sources wereindicated as the sources of air pollution, although both cities have similar industrial characteristics.For both cities,the MCA analysis showed a positive progressive correspondence between the levels of perceived annoyance and the variables' categories: importance of air quality, perceived exposure to industrial risk, assessment of air quality and perceived air pollution. Finally, as an interesting conclusion,variables such as gender, age, occupation, local of residence, source of air

pollution, occurrence of heath symptoms and meteorological conditions can influence the elicited behaviour of individuals who have experienced annoyance due to air pollution exposure. Thus, it plays an important role in the annoyance perceived by the population, therefore affecting their quality of life.

Keywords:*air pollution, perceivedannoyance, behaviour, multiple correspondence analysis.*

1. Introduction

There is evidencethat air pollutioncan cause varioushealth impacts, such ashospital admissions, respiratory, cardiovascular, hypertension, cancer andmortality(WHO, 2005; Lercheretal., 1995;Llopetal.,2008; Reisen at al.,2018). Recent estimates from the World Health Organization (WHO) suggest that in 2012 approximately 7 million premature deaths were linked to air pollution (WHO, 2014). Apart from the direct health effects caused by exposure to pollutants, the perception of air pollution may cause annoyance (Orru et al., 2018b).

The concept of annoyance is complex and it is an extremely subjective variable that can be experienced as a perception, an emotion, an attitude or a mixture of these (Berglund, Berglund &Lindvall, 1987).Lindvall and Radford (1973) defined annoyance as "a feeling of displeasure associated with any agent or condition known or believed by individuals or groups to adversely affect them" and annoyance may be associated with other negative emotions (e.g., anger, disappointment, dissatisfaction, helplessness, anxiety, agitation) and behavioural/social changes (e.g., interference with intended activities) (Blanes-Vidal et al., 2012). Furthermore, WHO defines health as a "state of complete physical, mental and social well-being and not merely the absence of disease or infirmity". Therefore, annoyance caused by air pollution can be considered as a health problem and an ambient stressor that affects the quality of life.

Although there is alreadya significant number ofstudies linkingair pollution andhuman health risks (e.g. Oglesby, et al., 2000;Klæboe et al., 2000; Llop et al., 2008, Stenlund et al, 2009; Egondi et al., 2013), comparatively, there are few studies exploring perceived annoyance caused by air pollution as a health risk in metropolitan/ industrialized regions, specially, when it comes to different factors and variables affect the perception of annoyance. According to Jacquemin et al, (2007) there are factors or groups of qualitative variables, such as sociodemographic factors,health symptoms, location of residence, perception of dust levels, that can be determinants of the perceived annoyance.

One way tobetter understandthis relationshipamong qualitative variables isthrough the analysis ofcorrespondence. Multiplecorrespondenceanalysisis amultivariate analysis techniquefor categorical datathatallows to graphicallyassessthe differences, similarities and relationshipsbetween variablesandtheir response categories (Benzécri et al., 1973;Greenacre & Blasius, 2006).

The objective of the present work is to investigate the relationship among variables that can be considered determinants the perception of annoyance caused by air pollution, using a multivariate method, i.e. Multiple CorrespondenceAnalysis (MCA). Such a

technique considered more than two variables,has not been applied before in relation to air pollution annoyance, and therefore, the analysisofcorrespondencethrough specific conceptsand parameters presented here, can contributeto the disseminationof the technique instudiesof exploring qualitative variables related to air pollution annoyance and health risks.The MCAis developed and evaluated using datasetsfrom two surveysconductedin different urbanandindustrialized regions, namely Dunkirk (France) and Vitoria (Brazil) to allow comparison of the results and to enable further insight in the parameters affecting perceived annoyance from air pollution.

2. Materials and methods

2.1. Characteristicsof the regions

The study was conducted in twodistinct urban industrialized regions: Dunkirk(France) and Vitoria(Brazil), thus allowing comparison of the annoyance levels observed in two cities with similar characteristics and providing further insight on the relationship between annoyance and air pollution parameters. It should be noted that both cities are in coast, port andindustrialareaswithpotential sources ofair pollutants (see Figure 1 and Figure 2).

Despite thegeographic and socioeconomicdifferencesbetween theseregions, bothare exposed to air pollution and their populations often report to local authorities that they feel annoyed by air pollution. According to a report concerning the industrial risk perception in Dunkirk (Calvo-Mendieta et al., 2008), air pollution is cited as the firstenvironmental problemby its inhabitants, followed by water andsoil pollution. In Vitoria, according to Souza (2011), more than 24%of the complaints to the environmental agency refer to air pollution. Since 2009 the two cities signed an international cooperation to develop a variety of events, organizing projects in the realm of environment, culture, economy, port activities, urban development and universities (Les ateliers, 2010).

2.1.1. Dunkirk

The metropolitan region of Dunkirk (MRD) has about 210,000 inhabitants and is located on the northern coast of France in the region of Nord-Pas-de-Calais. The climate in MRDregionis oceanic. Theflatnessclearly explainsthehigh levelof precipitation, but with a distinctmaximumin the fall, typical of a coastal climate thatisstrongly influenced by thewind: summer breezesometimescontributes to increasedsunshine, but there are alsoepisodes of"squalls" accompanied bypenetratingwinterrains.Marinewinds from the north-east sector arequite common. These conditionsarefavourable,in general, for the quality of theairwhen the windsdispersepollutantstowards the sea.However,sea breezesandnortherly winds, which is fortunately a rarer wind direction sector,often result topollution episodes (PPA, 2002).

MRDis marked bythe presence of anindustrial portarea, which stretches for almost 20km,andincludesa high density ofindustrial facilities, most of which are largeemitters of airpollutants. In recent years Dunkirk's port expands (gaining market share with global traffic superior to port's neighbours) towards its hinterland with river and train connections(AGUR-DUNKERQUE, 2015).

2.1.2. Vitoria

The metropolitan region of Vitoria (MRV) has about 1,500,000 inhabitants (IBGE, 2010) and is located on the south-eastern coast of Brazil (Figure 2). The topography ischaracterizedbymountain rangesin thenorthand westernportions, plains andhighlandsin the northern partandlowlandsin the southern part. The land useis also variable, includinglarge areaswith vegetation coverand largepavedareasandsurroundingtowns. The proximity tothe ocean andthe topographyare factorsthat controlthe weather conditions, such as sea breezeand the formation ofrain. The climate in the MRV is classified astropical hot and humid.Thisclimate typeis characterized bylong summers(usuallyOctober to April) and hightemperatures, withmaximum temperatures occurringusuallyin December and January. Winter isweak,withaverage temperatureof the coldest monthabout 18°C, thecold sensationexistingoccasionally whenthere isoccurrence ofcold fronts. The prevailing wind direction is north-easterly (IEMA, 2011b, IEMA, 2013) and it contributes tothe dispersionof pollutants emittedfromindustrytowards the city.
MRV comprises the third largest port system in Latin America and has many industrial sites including a steel plant, an iron ore pellet mill, stone quarrying, cement, food, pharmaceutical and chemical industries, an asphalt plant, etc. In recent years, the MRV has experienced a process of economic growth and increased industrial production as well as urban development(IJSN,2015).

2.2.The surveys

In MRD, the survey was conducted in 2008 with a representative sample of 518 people (over 18 years old) interviewed using face-to-face questionnaires. For this survey, the urban community was grouped into 10 sub-regions (that were determined according to the density of the habitat but also of their more or less proximity to the industrial-port area): Bourbourg,BrayDunes/Leffrinckoucke, Téteghem, CoudekerqueBranche, Gravelines, StPolsurMer, GrandSynthe, Petite-Synthe, Dunkirk and Malo/Rosendael. The sample size was proportionally distributed according to: sex, geographic location (near or far from industries/pot area), and socio-professional category.

InMRV,the survey was conducted in 2011, the sample size was determined by using a simple random sampling with proportional allocation method (Cochran, 1977) totaling a representative sample of 515 individuals (over 16 years old), which were distributed proportionally in the sub-regions around the eightair quality monitoring station areas:Laranjeiras,Ibes, Jardim Camburi,Vitoria-Centre, Enseada do Sua, Cariacica andVilaVelha-Centre. The eight stations in Vitoria were determine also according to the population density and to the proximity to the main industrial sources, vehicular sources and port area.

The questionnaire developed, validatedand applied in MRD survey in 2008. In 2011 it wasadapted and appliedin the MRV. It was conducted a piloting questionnaire and a pre-test to ensuring stability over time and internal consistency of the questions.

The questionary contained questions concerning socioeconomicanddemographic factors suchas age, education level,occupation,daily habits, genderandlocation of residence. The questions on annoyance was based on the scientific literature in relation to categorical, qualitative and ordinal scales, such as in the studies by Passchier&Passchier (2000),Klaeboe et al. (2000), Llop et al. (2008) and Atari et al. (2009). Thus, here to

measure the perceived annoyance was applied the qualitative answers were then recorded in a categorical/ ordinal 4-point scale (1 for not annoyed; 2 for slightly annoyed; and 3 for very annoyed and 4 for extremely annoyed). Table 1 presents the questions from both surveys selected for this work, the variables from each question and the factor groups of variables. All respondents replied the survey, but foreachquestion was included the option of answers "not know" (NK), and in such cases, they are also presented, so it is important to analyse.

2.3.Measured air quality levels

Recent studies have shown that fine particles (≤2.5 μm, $PM_{2.5}$) and ultrafine particles (≤0.1 μm, PM0.1) are associated with diseases in the lower respiratory system (Farfel et al., 2005, Souza et al., 2017, Reisen et al, 2018). Regardless, inhalable particles (≤10 μm, PM_{10}) still pose severe public health concerns, such upper respiratory system disturbances and in some cases persistent nuisance (Vallack and Shillito, 1998). Thus, when considering the objective to evaluating perceived annoyance, the behaviour ofPM_{10} pollutant is presented in this study.

2.3.1. Dunkirk

The MRD comprises the third largest port in France and an industrialized city containing steel, food, pharmaceutical and chemical industries, an oil refinery andalso a nuclear power station for electricity production. These anthropogenic activities present potential sources of particulate matter, which is the main cause of complaints by the resident population in this region. Measurement of air pollutant concentrations have been carried out by air quality monitoring stations distributed according to the French national guidelines (ADEME, 2002), which implement European Union Directives 96/62/EC and 99/30/EC.

According to the protection planof the atmosphere (PPA, 2002) the inventory ofemissions for MRD clearly showsthat theindustrialsector is thelargest emitterof pollutants. For particulate matter (PM_{10}) the main sources in the region are industries,incineration plants, collective and individualheating, road transport.In PPA (2002) there are measures established for local industries,to prevent and reducethe dispersion ofparticles,such as spraying wateron stock piles, wateringpaths andstorage areas, changing the conditions ofdischarge.

Table 2 showsthe descriptive statistics of 24-hour-mean concentrations of PM_{10} measured in all the air quality stations during 2008 (Atmo Nord-Pas-de-Calais, 2009), i.e. the year that the survey took place. Concentrations measured differ greatly among the stations, for example at Fort Mardyck and St Pol Mer Nord presented the largest maximum values and the largest variability (as the standard deviation and the range values suggest) and in Malo-les-Bains and Petite-Synthe happen the opposite. In all locations, 24-hour-mean concentrations of PM_{10} have maximum values higher than the respective WHO annual air quality guideline for PM_{10} (WHO, 2005). Furthermore, although the mean values are less than 50 $\mu g/m^3$ in all stations, the other statistical parameters (standard deviation, 90% and 95% percentile values, maximum) show the occurrence of high concentration peaks during the period.

2.3.2. Vitoria

The MRV comprises a large port system, heavy vehicular traffic and an industrial park that includes steel production, pelletizing, quarry, cement and food industries, chemical industries and an asphalt plant (IEMA, 2011a) that are potential sources of particles.Tomonitor the air qualityin the MRV,eightair quality monitoringstations set in different locations are managedby the local environmentalagency(IEMA). According to the environmental agency (IEMA, 2011a), the majorcontributor sources of totalparticles arevehicular emissions(emission ofparticles fromheavytrafficarteries) followed by industrial emissions(mainly the pellets making and steelindustries).

Table 3 presentsthe descriptive statistics of 24-hour-mean concentrations of PM_{10} measuredduring 2011 at the eight air quality monitoring stations located in the Vitoria region (IEMA, 2011b), except to Vila Velha-centro station, which in 2011 did not register enough data (min70%) for analysis. The data show that the largest mean concentration value as well as the largest variability (as suggested by the standard deviation and the range) occurred at the Cariacica station. In all air quality stations, the maximum value is higher than the WHO annual air quality guideline forPM_{10} (WHO, 2005). As in the case of Dunkirk, although the mean values are less than 50 $\mu g/m^3$ in all stations, the other statistical parameters (standard deviation, 90% and 95% percentile values, maximum) show the occurrence of high concentration peaks during this year.

Details about quantitative analysis between perceived annoyance and particulate matter can be find in Machado (2018).

3. Multiple Correspondence Analysis (MCA)

Acoording to Le Roux & Rouanet (2010) the correspondence analysis was established in 1963 in France(see, e.g., Benzécri,1969) as a geometrical method,but only after 1980 was spread throughout the world by the books published in English, see, e.g.,Grenacre(1984), Benzécri, (1992), Le Roux &Rouanet, (2004). MCA can be viewed as an extension of simple correspondence analysis in that it is applicable to a large set of categorical variables (Greenacre, 2007). MCA, as the counterpart of PCA for categorical variables, became standard for the analysis of questionnaires (Le Roux &Rouanet, 2010).

MCA is amultivariate data analysis technique for categorical data, used to detect and represent data graphically (by the scatterplot) as a set of points with respect to two perpendicular coordinate axis: the horizontal axis often referred to as the x-axis and the vertical one as the y-axis. The objectiveof this techniqueis to analyse graphicallythe relationshipsamongvariables, response categoriesand objectsby reducing thedimensionalityof the data set (Crivisqui, 1995; Lebart et al., 1984).

To apply MCA, data areinitially representedbya table ofrespondentsversusquestions:the lines representthe respondentsparticipating in the surveyand the columns represent the questions from the questionnaire, so eachfilledcellis the response category (answer) chosen by each individualfor each question. The questions arecategorized variableswithfinite number ofresponse categories, for example, tothe question"Do you

feel annoyedbyair pollution?", the variable “annoyance” has the followingresponse categories(with theirencodings): not annoyed (ANNOY-1); slightlyannoyed (ANNOY-2); very annoyed (ANNOY-3); extremely annoyed (ANNOY-4) with also the possibility of “not known/response” (NK)” (ANNOY-9/99). Each respondentcan chooseone and only oneresponse categoryfor each variableorquestion. Thus, if the individual of line 1answered "very annoyed" to the above question thevariablecellwas filledwith the"ANNOY-4" category, and so on forallthe individualsfor each variable.

According to Le Roux &Rouanet (2010), interpretation of MCA outcome isbased on the observationof the cloud ofpoints, whichis defined asa finiteset of pointsin a geometricspace.The cloud of pointscanrepresent variables, response categoriesand individuals, so in this work the cloud of points represent the response categories.

The great advantageof MCAisthe possibility toreduce themulti-dimensional spacein an optimalsubspacethat allowsthe studythe scatterplot and the consequentanalysis and interpretation ofresults. The generatedgraphsallow tovisuallyassess whetherall variablesof interesthaveassociationsamong them and allow knowinghow to givethese associations.

The sizeof the scatterplot depends on the numberof information pieces in each rowor column,minus one.If the number ofcolumns isrelatedto the*K*categories ofresponses of the*Q* variables, the maximum dimensionalityof the scatterplot of categories isgiven byEq.01,

$$(K_1 - 1) + \cdots + (K_q - 1) + \cdots (K_Q - 1) = (K_1 + \cdots + K_q + \cdots + K_Q) + (-1)Q = K - Q \quad \text{Eq.01}$$

Thedimensionalityreductionis normallymade toR^2, to facilitate interpretationof the cloud ofpoints (scatterplot). Le RouxandRouanet(2010) define themiddle point of thecloud of point in the following way: Let *P* beany pointin space and$(M^k)_{(k=1,2,\ldots,K)}$the points of categoriesfor thescatterplot, so themidpointof the cloudpointisthe *G* point by the vector $\overrightarrow{PG}$ as Eq.02,

$$\overrightarrow{PG} = \frac{1}{n}\sum \overrightarrow{PM^k} \quad \text{Eq.02}$$

The point *G*does not dependon the choiceof thepoint*P*, i.e.,whatever thechosen point*P* is, point *G*is always the same. Thus,thepoint Gis definedas the averageof thecoordinates of all pointsgiven by Eq.03,

$$G = \frac{1}{n}\sum M^k \quad \text{Eq. 03}$$

The distancebetween pointsdepends on the differentchoicesofresponse categoriesfor each variable. Aslower the frequencies of the response categories are, as greaterthe distance betweenindividuals become. Let n'_{kk} be the number of subjectswho choseboth categories*k* and *k'*, then the square of thedistancebetweenM^k and $M^{K'}$ is given in Eq.04,

$$d^2\left(M^k M^{k'}\right) = \frac{n_k + n'_k - 2n'_{kk}}{n_k n'_{k/n}} \quad \text{Eq.04}$$

As more categories *k'* and *k* are chosenfor the sameindividuals, as shorterthe distancebetweenM^k and$M^{k'}$ is, and as closer two category points are, the stronger is the association between them. As lower the frequency for the category *k* is, the farther from the centre the point M^k. The lessfrequent the category ofresponse is, the more it contributes to the overallvarianceof the cloud of individual points. And the lessfrequentthe standardof responsesof an individual is,the more it contributes to the total variance(Le RouxandRouanet, 2010).

The first principalaxis of acloud of categoriescanbe defined as theline passing through themidpointof the cloud.The secondmain axisisperpendicular to the firstone and is also passing through the midpoint of thecloud of categories points. The same process is followed to define thethird axis, the fourth axis, andso on. There are no set rules forthe number ofaxis to beanalysed (Grenacre, 2006). In this study,it appears thatthe first twoaxisholdthe highest percentageof the total variabilityof the data. Therefore,scatter plotsof categories are formed from thefirst two axis.

Theresults of the MCAcanbe confusingdepending on thenumber ofvariables. Because ofthis, the valuesof the contributionsgenerated from the application ofMCA "collaborate"in the interpretationof the axis(LeRouxandRouanet, 2010). Therefore, in the present work the contribution of each category as well as their sum are analysed to identify the variable that most contributes to the interpretation of eachaxis.

To apply theMCA, itis possible to selectrows and columnsthat will generatethe active pointsand the illustrative or supplementary points. The active pointsare responsible for determiningthe orientationof the principal axis,providing the necessary information for theconstruction of theoptimalcloud of categories points. However, it ispossible to include moreinformation which isrepresented bysupplementary orillustrativepoints. The supplementarypoints maybe plottedon the mapalong with the active points, and they are useful in interpreting features discovered in the primary data, but do not contribute tothe construction of theaxis (Grenacre, 2007).Thus,supplementarypointsare used to representinformationabout the phenomenonunder studyandinvariableinformationover time, such as sex, race, or information forinfrequentcategories.

According to Le Roux &Rouanet, (2010) the contribution ofa category pointto constructanaxisdefinesthe importance of thispointfor this sameaxis.Through the coefficientsof this contribution, itispossible to identifywhich categories(orpoints) should be considered forthe interpretationof eachprincipal axis. Therelative contributionconstitutesthe axiscontributiontothe varianceof the individual point,so thatthe qualityof representation ofa pointcorrespondsto the sumof thesquaredcosines of axis1and axis2.Thetestvalues assist the interpretation, but they don't contribute to the totalvariance and are interpretedto diagnosehow wellrepresentedthe supplementarypoints(Greenacre, 2007).

4. RESULTS AND DISCUSSION

To generate each MCA stage,the active and supplementary variables weredefined, and the number of factors was setto compose the factorial plans. This decision was based on

the analysis of the composition of the populations study. There are no set rules defining how many factorial plans should be scanned in graphics (Le Roux &Rouanet, 2010). To facilitate interpretation, the first two factorial plans (axis 1 and 2) was selected to compose the correspondence graph (scatter plot). Table 1 presents the questions of interest, the selected variables to represent these questions and the factors represented as groups of variables/ questions. The sociodemographic andlocal variables are considered to compare differences and similarities between the respondents' opinions in the two study areas. It is important to note that in the first MCA (Figure 3) all active variables selected are nominal and ordinal while the supplementary variables (local) are nominal.

Table 4 presents the results obtained using the MCA from a matrixintersectionof1033individuals or respondents (rows) andtheir responsescategories to the five questions(columns) of the questionnaire (factor group named as "air pollution" in Table 1) from the surveys conducted in both study areas. Theresponse options (categories), encoding, frequency and percentages, coordinatesof the twoaxis (F1 and F2), contributionsin the constructionof the two axisand the squaredcosine valuesare also presented in Table 4 for each active variable. The coordinates for the axis F1 and F2 are the position of each category in the scatter plot (cloud of points). The proportion of the variance of the cloud due to the point is called the contribution of the point to the cloud. Thus, the sumof the category contributions for each variable in Table 4 shows that the activevariable"annoyance" contributes the most to thecloud and to eachaxis (annoyancecontributes27.9% to the axis F1and25.9%toF2).The representation'squality is express by the value of squared cosines (how high is the square cosine value higher is the quality of representation).

Figure 3 is the correspondence graph (scatterplot) with the coordinates of axis F1 and F2 generated for the active variables shown in Table 4 ("air pollution") and the supplementary variables shown in Table 5 ("local"). The axis F1 and F2 explain about 71% of the variability from the database, considering all active variablessimultaneously, which is considered as an excellent performance (Le Roux &Rouanet, 2010).

Analysing the direction from right toleft on the F1 axis there is aprogressive tendencyfor increased levels of annoyance as indicated by the respective categories (ANNOY1-not annoyed, ANNOY2- slightlyannoyed, ANNOY3-very annoyed and ANNOY 4-extremely annoyed). Thesame progressive tendencycan beobserved forthe variables: importance of air quality (IMP1-not important, IMP2- slightly important, IMP3-very important, IMP4-extremelyimportant); industrial risk perception (RISK1-not exposed, RISK2- slightly exposed, RISK3-very exposed, RISK4-extremelyexposed); assessment of air quality (AIRQ1-excellent AIRQ2-good, AIRQ3-bad, and AIRQ4-horrible); and air pollution perception (PPOL1-never, PPOL2-sometimes, PPOL3-often, PPOL4-always). Thus, the first axisF1 can be considered as defining (from the right to the left) a scale of "perceived annoyance", and the second axis F2 appears to oppose moderate response categories (lower side) to both extremely positive and extremely negative responses. The scatter plot can also be interpreted through the parabolic shape of the cloud of points of the chart from the bottom up to the centre setting categories for "slightly" and "very" levels, while the upper right corresponds to the "not" level and the upper left to the categories represented by the level "extremely". Such a pattern of response suggests what is known in the literature as the "Guttman effect" (Greenacre &

Blasius, 2006)or“horseshoeeffect” (Van Rijckevorsel, 1987 *in* Greenacre & Blasius, 2006) due to itsparabolicshape or arch. This is a structured form of the distribution of the categories of annoyance levels, which are arranged in a hierarchical way, from those who do not report nuisance (upper right), to those who express moderate annoyance (vertex of the parabolic) and arriving at extremely annoyed level (top left).

The joint progressionof annoyance levels and other active categories from right to left in axisF1 indicates that an individual who reported being extremely annoyed due to air pollution also thought that the air quality was extremely important, felt extremely exposed to industrial risks, assessed air quality as horrible and always perceived air pollution by dust/odour/air visibility. Thus, another possible pattern visible in Figure 3 is the "battery effect" it is often observed in survey analysis which the respondents choose similar answers without necessarily considering the content of the questions. However, we can’t affirm that it is a battery effect because these questions were not presented in the same order to the Dunkirk inhabitants and the Vitoria inhabitants as selected for this analysis. Furthermore, the response options presented for the question concerning the assessment of air quality were ordered in such a way that they express “opposed feelings” in relation to the response to other questions, causing therespondentto give due regardbefore answering.

Table 5 presents the categories for the sub-regions/areas where the respondents live, the code for each location, the frequencies and percentages of responses, the coordinates for each response category and the test values for the axisF1 and F2 related to the supplementary variables of the “Local” group. The test value is an indication of the significance of the obtained results (5% p-value or 1.96p-valuein absolute terms)(Crivisqui, 1995). The test value was calculated as the distancefrom each pointto the originof the axis F1 and F2 (in Figure 2), in numbersof standard deviations.

It is important to observe thenegative test values, since the negative values in axisF1 correspond to the local variables or the area wherepeople have reported to be very annoyed,whilepositive test values correspond tothe local wherepeople have reportedlittle or noannoyed by air pollution. According to the correspondence graph in Figure 2, the locations where respondents reported being very annoyed due to air pollution correspond to the localities that have negative test values (on the left part of the F1 axis): Grande Synthe (LOCAL-D7) because their location is at -2,699 standard deviations from the mean point (origin) on axis F1; Petite-Synthe (LOCAL-D8) test value = -3,11; Jardim Camburi (LOCAL-V4) test value = -4,138 and Enseada do Sua (LOCAL-V6) test value= -4,812.Thus, residents in this sub-regions/ area reported intense levels of annoyance and reported being “very exposed” to industrial risk, often assessed the air quality as “horrible” and perceived “high levels” of air pollution due to dust/odour/opacity in their neighbourhoods. It is very interesting to note that a comparison with the air quality results shown in Tables 2 and 3, suggests that, for both urban regions, the areas where the inhabitants reported higher levels of annoyance do not correspond with the areas where higher mean of particulate pollution were measured by the network of monitoring stations.Maybe there is a gap between the measurement of pollutants and the perception that does not correspond to the reality of measurements and that is important to investigate the other pollutants behaviour.

Figure 4 shows the correspondence graph between the active variables in the “air pollution” group and the supplementary variables in the “sociodemographic” group. In

this graph, the localvariableswere removedto facilitate visualizationof thecorrespondence withsociodemographicvariables: gender, age, occupation and level of education. Visually the most supplementary variables are close to theoriginof the graph.

Table 6 shows the results (frequency, coordinates and test value) for the correspondence graph in Figure 3 by each response category of sociodemographic variables. For MRV, it is possible to see that women reported being more annoyed than men, while in MRD this same association was not significantly. Consequently, in MRV women felt more exposed to industrial risk, assessed air quality as more important and perceived air pollution by dust/odour/opacity more than men. According to Fisher et al. (1991), these gender differences are noticeable especially in relation to environmental risks. Explanations are linked to the social roles of women in society, roles that are most often oriented towards health and children. Gustafson (1998) also discusses this difference between men and women in relation to their roles in society and the power relations that exist between them. For example, women are more sensitive to environmental risks because they take care of their homes and children and clean the house normally more frequently than men, especially in more conservative societies.

Regarding the correspondence graph (Figure 4), inMRDthere was no significant correspondence visible between the age categories and the annoyance categories. However, in MRV it is possible to visualise a progressive relation between age (AGE-V1, AGE-V2, AGE-V3, and AGE-V4) and levels of annoyance. As age increased, the levels of annoyance, importance of air quality, perceived exposure risk, assessment of air quality, and perceived air pollution also increased. This association can be confirmed considering the test values in Table 6. Respondents older than 34 years (AGE-V3, AGE-V4) are associated to being very orextremely annoyed more than young respondents (AGE-V1, AGE-V2). Normally, elderly people are more sensitive to health problems since they belong to the more sensitive population sub-groups (it should be noted that children do not participate in the survey and women responses were discussed above) and experience the effects of air pollution more often, such as when removing dust for house cleaning. Although this association is not so clear in MRD compared to MRV, it should be noted that the results of Lercher et al.(1995) and Klaeboe et al.(2000) suggest that older age is a determinant of perceived air pollution.

Regarding occupation, in MRDunemployed (OCCUP-D2) and student (OCCUP-D4) are on the right side of the F1 axis, thus, they are associated to being slightly or not annoyed by air pollution. While in MRVthe categories associated to being very and extremely annoyed by air pollution are the retired group (OCCUP-V3) and the unemployed (OCCUP-V2) on the left side of the F1 axis. This association is punctual for the retired group (test value = -3,44) in MRV, and can be justified because generally they are the group with older age that are also associate to being very and extremely annoyed by air pollution.

Considering the corresponding graphand the test values for level of education categories only in MRV, it is possible to see that the university group (EDUC-V4) are on the high levels of annoyance side of the axis F1. Although, Klaeboe et al.(2000), suggested that the education level was a determinant of perceived air pollution, in the present analysis there was no correspondence found with annoyance and levels of education for both surveys.

Figure 5 is the correspondence graph between “Air pollution” and “Health” groups. The supplementary variables selected were “health problems” and “health effects”. People who report “no” (HEFE-1) occurrence of health problems caused by air pollution are also the ones who are “less annoyed”. And people who answered “yes” (HEFE-2) to occurrence of health problems are associated to “being very annoyed”.

In Table 7 thetestvalue = -5,113for the variable health problems caused by air pollution, presented indicate the same association, people who answered “no”(HEFE-1) shown on the right part of the graph, tend not to feel annoyed by air pollution, while those who responded “yes”(HEFE-2) shown on the left sideof the graph, tend to report being extremely annoyed by air pollution. For this group of people that responded “yes”, the main problemsreported were eye irritation (HPROB-1)lung/respiratory (43%) but there is no evident association. The test value= -2,85 for health effects, indicate that people who reported being very/ extremelyannoyed by air pollution were associated to the ones who reported allergies(HPROB-3). Although, previous epidemiological studies have shown that certain levels of particulate matter concentrations and related pollutants can cause such health effects and increase the number of hospitalizations for respiratory problems (Pope III, 1991; Schwartz, 1991; Braga et al., 2001; Garçon et al., 2006; Llop et al., 2008), in this analysis was no correspondence found with annoyance and health effects.

To further explore these survey data, the MCA between the factors groups “Cause” and “Local” was carried out and the active variables“source”, “meteo”, “season” and “day/night” as well as the supplementary variables “Dunkirk” and “Vitoria” were selected for analysis. Table 8 presents the summation of the coordinates, contribution and squared cosine values for each response category. It can be observed thatthe categories grouped under the “METEO” and “SEASON” contributesignificantly toF1 axis. And the categories “SOURCE” contribute significantly to F2 axis. The summation of the contributions of the other response categories as well as the value of the squared cosines for the F1 and F2 axiscan confirm such affirmations.

Figure 6 presents the correspondence graph between the factor groups “Cause” and “Local”. Analysing the active variable “METEO”, the respondents that answered “no” to the question regarding the influence of meteorological conditions on the perception of air pollution (METEO-2) are located on the right part of the F1 axis, while the left part of the F1 axis indicates the respondents who answered “yes” (METEO-1). Analysing the active variable “SEASON,” the left part of the F2 axiscorresponds with the yes (METEO-1)categoryit is possible to see a progressive tendency from spring (SEASON-4), tosummer (SEASON-1), to autumn (SEASON-2), to winter (SEASON-3). That is, the respondents reported “yes” to the question about the influence of meteorological conditions on the perception of air pollution also perceived a progressive effect of seasonality from spring and summer to autumn and winter. As shown in others studies (e.g. Castanho&Artaxo, 2001; Albuquerque et. al., 2012) meteorological conditions have a major influence on the suspended particle concentrations, which can explain this association between seasons and perception of air pollution.

Regarding the active variable “SOURCE”, it is possible to observe the response category “source industry” (SOURCE-2) on the upper-right part of the graph and the categories “building” (SOURCE-4), “suspension of soil” (SOURCE-3), and “vehicle” (SOURCE-1) on the lower-middle of the graph with no association with the variable “METEO”. This clearly shows the important role of industry in relation to air pollution in the two examined industrial urban areas, especially as this is perceived by the population. The response categories “yes” (DN-2) and “no” (DN-1) for perceived air pollution changes between day and night are close to the origin and arenon-significant.

For the “LOCAL” supplementary variable, the lower part of the graph and on the middle and on left of the F1 axis corresponds to the areas in Vitoria showing that the people perceived an influence of weather changes on the quantity of particles or dust (corresponding to METEO-1). The upper part of the graph on the middle and right part of the F1 axisshows that in Dunkirk, there is no perceived association between the meteorological conditions and air pollution (corresponding to METEO-2). This result may be related to the fact that people who live close to industries are already accustomed to pollution and they do not feel the influence of weather changes on the perceived air pollution. Furthermore, as Figure 1 indicates, in Dunkirk industrial pollution sources are more interspersed within the urban area, while in Vitoria they are more at the boundaries of the city (especially northwest, but also partly southeast) and therefore the meteorological conditions – such as wind direction and wind speed – may influence on perceived air pollution.

Regarding the categories “Source” associated with “Local” the results in Figure 5 suggests that, for Vitoria, the locations far from the main industrial areas, like Cariacica (LOCAL-V7) and Vitoria-centro (LOCAL-V5), were found to be associated with the construction work/building source (SOURCE-4) and slightly with the vehicular source (SOURCE-1) of air pollution. It should be noted that in these areas, and especially Cariacica, high levels of particulate pollution were measured by the air quality monitoring stations (Table 3) and according to IEMA (2011) the main source of particulate matter in Cariacica is construction work. Currently, construction work in Vitoria is developing rapidly in terms of housing construction and paving streets and roads and the number of vehicles is also increasing, which can contribute to increasing air pollution and dust that cause annoyance, especially since these locations are further away from industrial sources. In Dunkirk, the locations close to the main industrial areas, St. Pol Sur Mer (LOCAL-D6), Grande Synthe (LOCAL-D7) and Petite-Synthe (LOCAL-D8) were found to be associated with the “industry” source (SOURCE-2), while the “construction work/building” and “vehicular” sources are not significant, as are the “industrial” sources which are located next to residential areas and visible by the main beaches. St. Pol Sur Mer is also the location in Dunkirk where the highest levels of particulate pollution were measured (Table 2) and very close to the industry sources.

The above results may also partly be explained due to the size of the population in Vitoria, which is larger than in Dunkirk,and the fact that public transport in Dunkirk is well developed; therefore, thenumber of vehicles circulatingin Dunkirk is not as high as in Vitoria. According to Rotkoet al.(2002) and Amundsen et al. (2008) heavy traffic is related to annoyance caused by air pollution, so this result can explain why people exposed to heavy traffic in Vitoria perceived vehicular sources more significant than Dunkirk, especially in areas which are not as influenced by industrial sources. Also, as noted above, currentlythe construction sectorin Vitoriaisdeveloping rapidlyboth interms

ofhousing construction and pavingstreets and roads, which can generate dust that may cause annoyance. As found by Nikolopoulou et al., (2011) the air quality is often considered to be poor at construction sites, which are burdened with higher PM concentrations.

Table 9 shows the coordinates and the test values for the "LOCAL" supplementary categories. Regarding the test values, there is a strong association (not contribution) with the variable that contributes to the F2 axis. The positive test values for the F2 axis are the Dunkirk locations in the upper quadrants of Figure 6, mean the residents which perceived urban air pollution from industries sources. The negative test values for the F2 axis are the Vitoria locations in the lower quadrants, which perceived urban air pollution from construction works and vehicular sources.

5. CONCLUSIONS

The purpose of this work was to explore relationship among variables that can influence behaviour of people about perceived annoyance caused by air pollution applying multiple correspondence analyses(MCA) technique. The data base is from a complex survey about air pollution, environmental issues and quality of life, developed in two metropolitan areas, Dunkirk (France) and Vitoria (Brazil), since people frequently report feeling annoyed by air pollution in both regions.

The results analysis showed a progressive relationship between levels of annoyance and the variables from "air pollution" factor group. Thus, as the levels of annoyance increased, the levels of the other qualitative variables (importance of air quality, perceived exposure to industrial risk, assessment of air quality, perceived air pollution) also increased. It is possible to conclude that people who reported feeling annoyed by air pollution also thought that air quality was very important, were very concerned about exposure to industrial risks, and assessed air quality often as horrible and frequently perceived air pollution by dust/odour/visibility. It is important to emphasize that this result cannot be considered "battery effect" since these questions were not applied in the sequence in which it was presents the results analysed.
In addition, the summary results of PM_{10} concentration measurements in the two regions showed values above the guidelines established by the World Health Organization, which can be an indicator of attention to the possibility of occurrence of health problems, quality of life effects and complaints about perceived annoyance. According to the correspondence graph, people who lives in areas close to industries, for example Petite-Synthe and Grande Synthe (in Dunkirk) and Enseada do Suá and Jardim Camburi(in Vitoria) have reported being very annoyed by air pollution (specially dust). Although PM_{10} measurements suggest that these are not the areas with the highest levels. Thus, location and the proximity to industrial sources of air pollution play an important role to explain people's behaviour.

This study has explored the association between respondents' perceptions and individual characteristics. With respect to gender and age were corresponded with perceived annoyance in Vitoria while in Dunkirk there was a much looser correspondence. In general, women were more annoyed than men, while people older than 55 years reported feeling more annoyed than people in other age ranges. Thus, we conclude that, for Vitoria, women are more annoyed than man and there is a progressive relationship between age ranges and the variables: level of annoyance, importance of air

quality, perceived exposure to industrial risk, assessment of air quality and perceived air pollution by dust/odour/visibility.

Socio-economic variables have been shown to be associated with the perception of local air quality, suggesting that these may be important factors in a study of perceived air quality (Kohlhuber et al., 2006). In our study, the relationship between different forms of occupation remains controversial. Was found that involvement of retired and unemployed was associated with high levels of annoyance in Vitoria but, in Dunkirk students and unemployed groups were associated with low or none level of annoyance. Previous studies have associated higher levels of education to be associated with higher annoyance level or poor air quality perceptions (Jacquemin et al., 2007; Kim et al, 2012). Though we did not find a significant association between levels of education and levels of annoyance was found that university people were more concerned about air pollution effects on quality of life than those with no or less than university level in Vitoria and the association was not significant in Dunkirk. The importance of socioeconomic factors in the context of air pollution research has been emphasized because they represent underlying aspects that affect susceptibility, exposure, or disease diagnosis and treatment (Bell et al. 2005). Therefore, there is need for careful choice and interpretation of socioeconomic factors depending on the location, and can be partly attributed to social/cultural difference and to the different weather conditions.

It was observed significant association between perceived health risks related to high level of annoyance caused air pollution, importance of air quality, perceived exposure to industrial risk, assessment of air quality and perceived air pollution by dust/odour/visibility. Previous studies reported a significant association between perceived air quality and self-reported health status (Kohlhuber et al., 2006; Llop et al, 2008). Though self-reported health and perceived health risk refers to different concepts. Perceived occurrence of health problems related to annoyance caused by air pollution was found to be associated with gender (feminine), age, level of education (university groups) and the type of occupation (retired groups) though it varied by study location. The difference between the two sites could be explained by the difference in the age, occupation and education levels of the residents. There were more people older, employed and with university level in Dunkirk as compared to Vitoria, while in Vitoria there are more young people, students and with primary level of education as compared to Dunkirk.

The main health effects reported are lung/respiratory, allergies and eye irritation, which are common symptoms for many urban air pollutants such as PM and NOx (WHO, 2005). But, was found association only with allergies and high levels of annoyance. It should be noted that the self-reported health status is often associated with perceived air pollution more than with measured air pollution.

The results of the MCA for the active variable “CAUSE” showed that people perceived that weather conditions and seasonal changes could affect perceived annoyance. This perception was more evident for Vitoria, where heavy industries are at the boundaries of the city and their effect is more influenced by the prevailing meteorological conditions, such as the wind speed and especially direction. In Dunkirk, people identified industrial sources as important cause of air pollution and did not perceive that air pollution annoyance changes with differences in weather. Furthermore, in Vitoria, the influence of building or construction and vehicular sources on the perception of air pollution was

evident. Thus, considering the geographic location of these two regions, the weather conditions could influence the perceived annoyance, and therefore the location can justify this differences.

Although the spatial patterns of the results from PM_{10} concentration measurements in the two cities do not coincide with the reported levels of perceived annoyance in each sub-region, they could provide a better insight to the behaviour of respondents. For example, there is an indication that sources related to construction works, for example, influence more the perception of annoyance from air pollution than the absolute levels of particulate matter measured in an area. Thus, respondents living in the areas with higher air pollution levels appear to distinguish a more significant influence of meteorology on air pollution, possibly since high levels of air pollution are associated to specific meteorological conditions that lead to the accumulation of pollutants from nearby sources in these areas. It is important to note that air pollution perceptions mark differences in the two study areas which indicates that perceptions in general may depend on an area's overall setting and availability of industries, other pollution sources or daily activities.

The results of this study have shown that Multiple Correspondence Analysis is a very useful tool in providing insight on environmental issues affecting the quality of life, such as the factors affecting the levels of air pollution annoyance of populations living in urban areas. Such tools can derive and synthesize important information from citizen surveys which can complement air quality measurements, to define the best mix of measures to address air quality issues. Such measures can include national or regional emissions reduction policies to meet the air quality objectives in background locations, as well as more site specific, short-term measures to address air pollution episodes in "hot spot" locations. Such combination of measures is often necessary for the protection of the public health and the improvement of the quality of life of citizens.

6. Acknowledgements

The authors would like to acknowledge the support of FAPES and CAPES (Brazilian governmental agencies for technology development and scientific research). They would also like to acknowledge COFECUB, IrénéeZwarterook and a research group studying industrial risk and the urban environment in the TVES laboratory, Université du Littoral Côte d'Opale - France.

7. References

Albuquerque T.T.A., Andrade M.F., Ynoue R.Y. (2012). Characterization of atmospheric aerosols in the city of São Paulo, Brazil: comparisons between polluted and unpolluted periods. Environmental Monitoring and Assessment, 84(2), 969-984.

ADEME (2002). Classification et critèresd'implantation des stations de surveillance de la qualité de l'air. Retrieved on 28 July 2013 from: http://www.oramip.org/pdf/ademe_typologies.pdf

Amundsen A.H., Klaeboe R.,Fyhri A. (2008). Annoyance from vehicular air pollution: Exposure–response relationships for Norway. Atmospheric Environment. 42, 679-688.

Atmo Nord-Pas-de-Calais (2009). Bilan 2008 des poussièressédimentables sur le Dunkerquois. Rapport d'études 01 – 2009- LC. Available at: http://www.atmo-npdc.fr

Bell, M.L., O'Neill, M.S., Cifuentes, L.A.; Braga, A.L.F., Green, C.,Nweke, A.,Rogat, J.,Sibold, K. Challenges and recommendations for the study of socioeconomic factors and air pollution health effects. *Environ. Sci. Policy* **2005**, *8*, 525–533

Benzécri,J.P.(1969) Statistical analysis as a tool to make patterns emerge from data. In S. Watanabe (ed.), Methodologies of patenrecognition.New York: Academic press.

Benzécri, J.P. (1973) L'Analyse des Données. Vol. 2: L'Analyse des Correspondances. Paris: Dunod.

Benzécri, J.P. (1992) Correspondence analysis handbook. New York: Dekker.

Berglund, B., Berglund, U., Lindvall, T. (1987). A study of response criteria in populations exposed to aircraft noise. Journal of Sound and Vibration, 41, 33–39.

Berglund, B., Berglund, U.,Lindvall, T. Measurement and control of annoyance. In Environmental Annoyance:Characterization, Measurement and Control; Koelga, H.S., Ed.; Elsevier: Amsterdam, The Netherlands, 1987;pp. 29–44.

Blanes-Vidal V., Suh H., Nadimi E. S., Løfstrøm P., Ellermann T., Andersen H. V., Schwartz J. (2012) Residential exposure to outdoor air pollution from livestock operations and perceived annoyance among citizens. Environment International, 40, 44-50.

Braga A.L.F.,Saldiva, P.H.N., Pereira, L.A.A., Menezes, J.J.C.,Conceição, G.M.S., Lin, C.L.,Zanobetti, A.,Shwartz, J., Dockery, D.W. (2001). Health effects or air pollution exposure on children and adolescents in São Paulo, Brasil. Pediatr. Pulmonol., 31, 106-113, 2001.

Castanho, A.D. A., Artaxo, P. (2001). Wintertime and summertime São Paulo aerosol source apportionment study. Atmospheric Environment, 35, 4889–4902.

Crivisqui, E. (1995). Apresentação da análise fatorial de correspondência simples e múltiplas. Programme de Recherche et D'Enseignement en Statistique Appliquée. PRESTA, Belgique: Université Libre de Bruxelles.

Egondi T., Kyobutungi C., Ng N., Muindi K., Oti S., Vijver S. V. De, Ettarh R., Rocklov J.(2013) Community perceptions of air pollution and related health risks in Nairobi Slums. Int. J. Environ. Res. Public Health, 10, 4851-4868.

Fischer G.W., Morgan M.G., Fischoff B., Nair I, and Lave L.B. (1991), "What risk are people concerned about?", Risk Analysis, 11, 303-314.

Garçon G., Dagher Z., Zerimech F., Ledoux F., Courcot D., Aboukais A., Puskaric E., Shirali P. (2006) Dunkerque city air pollution particulate matter-induced cytotoxicity,

oxidative stress and inflammation in human epithelial lung cells (L132) in culture. Toxicology in Vitro, 20, 519-528.

Grenacre, M. J. (1984) Theory and Applications of correspondence analysis. New York: Academic press.

Greenacre, M. & Blasius, J. (2006). Multiple correspondence analysis and related methods. Chapman & Hall/CRC Press, London.

Greenacre, M. (2007). Correspondence Analysis in Practice, Second Edition. London: Chapman & Hall/CRC.

Gustafson E. (1998). Gender differences in risk perception: theoretical and methodological perspectives, *Risk Analysis*, 18(6), 805-811.

IBGE (2010). Instituto Brasileiro de Geografia e Estatística – Senso 2010– Retrieved on 10 July 2011 from:www.ibge.gov.br

IEMA (2011a). Plano de Controle de Poluição Veicular do Estado do Espírito Santo.Instituto Estadual De Meio Ambiente, Vitoria, Espirito Santo, Brasil. Retrieved on 12 July 2013 from: www.iema.gov.br

IEMA (2011b). Inventário de Emissões Atmosféricas da Região da Grande Vitória. Acordo de Cooperação Técnica IEMA-ECOSOFT. Instituto Estadual De Meio Ambiente, Vitoria, Espirito Santo, Brasil. Retrieved on 11 July 2013 from: www.iema.gov.br

IEMA (2013). Relatório de Qualidade do Ar da Grande Vitoria 2013. Instituto Estadual De Meio Ambiente, Vitoria, Espirito Santo, Brasil. Retrieved on 11 July 2014 from:www.iema.gov.br.

Jacquemin B., Sunyer J., Forsberg B., Gotschi T., Oglesby L., Ackermann-Liebrich U., De Marco R., Heinrich J., Jarvis D., Toren K., Kunzli N. 2007. Annoyance due to air pollution in Europe. International Journal of Epidemiology, 36, 809–820.

Kohlhuber, M.,Mielck, A., Weiland, S.K., Bolte, G. (2006). Social inequality in perceived environmental exposures in relation to housing conditions in Germany. *Environ. Res.,101*, 246–255.

Kim M., Yi O., Kim H. (2012). The role differences in individual and community attributes in perceived air quality. Science of the Total Environment, 425, 20-26.

Klæboe, R., Kolbenstvedt, M., Clench-Aas, J., Bartonova, A. (2000). Oslo traffic study part 1: an integrated approach to assess the combined effects of noise and air pollution on annoyance. Atmospheric Environment, 34, 4727-4736.

Lebart L, Morineau A, Warwick K. (1984). Multivariate descriptive statistical analysis. Chichester, UK: Wiley.

Lercher, P., Schmitzberger, R., Kofler, W. (1995). Perceived traffic air pollution, associated behavior and health in an alpine area. Science of the Total Environment 169 (1–3), 71–74.

Le Roux, B., Rouanet, H. (2004). Geometric Data Analysis, From Correspondence Analysis to Structured Data Analysis. Dordrecht. Kluwer Academic Publishers.

Le Roux, B.,Rouanet, H. (2010). Multiple Correspondence Analysis, SAGE, Series: Quantitative Applications in the Social Sciences, CA:Thousand Oaks Paris.

Llop S., Ballester F., Estarlich M., Esplugues A., Fernández-Patier R., Ramón R., Marco A., Aguirre A., Sunyer J., Iñiguez C., on behalf of INMA-Valencia cohort. Ambient air pollution and annoyance responses from pregnant women. Atmospheric Environment, 42, 2982-2992, 2008.

Nikolopoulou M., Kleissl J., Linden P.F., Lykoudis S. (2011). Pedestrians' perception of environmental stimuli through field surveys: Focus on particulate pollution. Science of the Total Environment, 409(13), 2493-202.

Pope C.A. III, (1991). Respiratory hospital admissions associated with PM10 pollution in Utah, Salt Lake and Cache valleys. Arch. Environ. Health, 46, 90-97.

OglesbyL.,KunzliN.,Monn C., Schindler, C., Ackermann-Liebrich U.,LeuenbergerP. Validity of annoyance scores for estimation of long term air pollution exposure in epidemiologic studies: The swiss study on air pollution and lung diseases in adults (SAPALDIA). Am. J. Epidemiol. 2000, 152, 75–83.

PPA (2002). Plan de protection de l'atmosphère de l'aglomerationDunkerquoise. Retrieved on 20 April 2013 2013 from: http://www.nord-pas-de-calais.developpement-durable.gouv.fr/IMG/pdf/ppa-dunkerque.pdf

Rotko T., Oglesby L., Kunzli N., Carrer P., Nieuwenhuijsen M.J., Jantunen M. (2002). Determinants of perceived air pollution annoyance and association between annoyance scores and air pollution (PM2.5. NO2) concentrations in the European EXPOLIS study. Atmospheric Environment, 36, 4593–4602.

Souza, L. B (2011).Estudo de Correlação Entre a Percepção do Incômodo Causado Pelas Partículas Sedimentadas e seus Níveis de Concentração na Atmosfera. Dissertação de Mestrado - Universidade Federal do Espírito Santo, Vitória.

Stenlund T., Lidén E., Andersson K., Garvill J., Nordin S. Annoyance and health symptoms and their influencing factors: A population-based air pollution intervention study. Public Health, 123, 339-345, 2009.

Schwartz, J. (1991). Particulate air pollution and daily mortality in Detroit. Environmental Research, 56(2), 204-213.

Vallack H. W., Shillito D. E. (1998) Suggested guidelines for deposited ambient dust. Atmospheric Environment, 32(16), 2737-2744.

WHO (2005). WHO air quality guidelines global update 2005. Report on a WHO Working Group. Bonn, Germany, 18-20 October 2005. Retrieved on 28 July 2014 from: http://www.euro.who.int/Document/E87950.pdf.

WHO (2014). Burden of disease from the joint effects of Household and Ambient Air Pollution for 2012. WHO: Geneva. Retrieved on 28 July 2014 from: http://www.who.int/phe/health_topics/outdoorair/databases/FINAL_HAP_AAP_BoD_24March2014.pdf?ua=1

5 Spatial and temporal analysis of the effect of air pollution on children's health

The main objective of this study is to investigate the short-term association between air pollution and emergency care for respiratory diseases in children aged 0-6 years. The generalized additive model (MAG) of Poisson regression was used, the dependent variable was daily number of visits to the hospital emergency service of people with respiratory diseases.

The independent variables are the daily concentrations of the atmospheric pollutants (PM_{10}, SO_2, NO_2, O_3 and CO), temperature, humidity and precipitation. Table 1 shows the descriptive statistics of the variables used. Through the daily mean concentrations, estimates were made for all the RGV and "in-loco" analyses with the consideration of children resident around 2 km of the 8 air quality monitoring stations (RAMQAr).

As results we find that the increase of 10 micro grams per m^3 in the concentration levels of air pollutants increased the risk of emergency care due to respiratory disease. In the general region (Table 2), for PM_{10}, the increase was 2.43%, 2.73% and 3.29% in the accumulations of 5, 6 and 7 days, respectively. For SO_2, the increase was 4.47% on the day of exposure (lag 0), 5.26% two days after, 6.47%, 8.80%, 8.76% and 7.09% in the accumulated of 2, 3, 4 and 5 days, respectively. NO_2, CO and O_3, for the general region, did not cause statistical significant increases. CO showed a significant association for children living close to two monitoring stations and O_3 in only one (Table 3). We conclude that even within the limits established by the WHO, the pollutants PM_{10}, SO_2, NO_2 and O_3 are associated with a higher risk for respiratory diseases in children from 0 to 6 years and some effects were only identified in the disaggregated localities by region, that is , "In loco", which allows to capture greater variability of the data.

This paper was submitted to publication to the Journal Cadernos de Saúde Pública.

Análise espacial e temporal do efeito da poluição do ar na saúde de crianças

(Spatial and temporal analysis of the effect of air pollution on children health)

Emerson Pedreira Matos[1], Valdério Anselmo Reisen[1,2,3], Faradiba Sarquis Serpa[1,4], Paulo Roberto Prezotti Filho[1,3,5], Maria de Fátima Silva Leite[2]

E-mail: valderioanselmoreisen@gmail.com (Valdério A. Reisen)

1Programa de Pós-graduação em Engenharia Ambiental - PPGEA, Centro Tecnológico. Universidade Federal do Espírito Santo - UFES. Vitória, ES, Brasil
2 Departamento de Estatística. Universidade Federal do Espírito Santo - UFES. Vitória, ES, Brasil
3CentraleSupelec. Paris, France
4 Escola Superior de Ciências da Santa Casa de Misericórdia - EMESCAM. Vitória, ES, Brasil
5 Instituto Federal do Espírito Santo – IFES. Guarapari, ES, Brasil

RESUMO

Objetivo:investigar a associação de curto prazo entre a poluição do ar e atendimentos em emergência por doenças respiratórias, em crianças de 0 a 6 anos.

Métodos:estudo ecológico, espacial e temporal realizado na região da Grande Vitoria (RGV), ES, Brasil. Utilizou-se o modelo aditivo generalizado (MAG) de regressão de Poisson, com a variável dependente o número diário de atendimentos por doenças respiratórias e as variáveis independentes, concentrações diárias dos poluentes atmosféricos (MP_{10}, SO_2, NO_2, O_3 e CO), temperatura, umidade e precipitação pluviométrica. Por meio das médias diárias das concentrações, foram feitas estimativas para toda a RGV e análises “in loco” com a consideração de crianças residentes no entorno de 2 km das 8 estações de monitoramento da qualidade do ar (RAMQAr).

Resultados: o incremento de 10 μg/m^3 nos níveis de concentração de cada um dos poluentes atmosféricos aumentou o risco de atendimento em emergência por doença respiratória. Na região geral, para o MP_{10}, o aumento foi de 2,43%, 2,73% e 3,29% nos acumulados de 5, 6 e 7 dias, respectivamente. Para o SO_2, o acréscimo foi de 4,47% no dia da exposição (lag 0), 5,26% dois dias após, 6,47%, 8,80%, 8,76% e 7,09% nos acumulados de 2, 3, 4 e 5 dias, respectivamente. O NO_2, CO e o O_3, para a região geral, não causaram aumentos significativos. O CO apresentou associação

significativa para crianças residentes no entorno de duas estações e o O_3 somente em uma.

Conclusões: mesmo dentro dos limites estabelecidos pela OMS, os poluentes MP_{10}, SO_2, NO_2 e O_3 estão associados a maior risco para atendimento por doenças respiratórias em crianças de 0 a 6 anos e alguns efeitos só foram identificados nas localidades desagregadas por região, isto é, "in loco", que possibilita captar maior variabilidade dos dados.

DESCRITORES: Criança. Efeitos Adversos. Poluição do Ar. Doenças Respiratórias. Epidemiologia. Estudo ecológico.

INTRODUÇÃO

As doenças respiratórias representam a principal causa de morbimortalidade de crianças em todo o mundo[1]. Existem evidências de que os poluentes atmosféricos emitidos por indústrias e veículos automotores, mesmo em concentrações dentro dos limites estabelecidos pela Organização Mundial de Saúde (OMS), estão claramente envolvidos na gênese de sintomas respiratórios, maior número de hospitalizações e óbitos[1,2].

As emissões de material particulado (MP), óxidos de nitrogênio (NO_x), compostos orgânicos voláteis (VOCs), dióxido de enxofre (SO_2)e poluentes fotoquímicos como o ozônio (O_3), aumentaramnas últimas décadas devido ao crescimento da frota de veículos automotores e ao crescente processo de industrialização[2].Estudos realizados em grandes centros urbanos, em diversos países,mostraram associação significativaentre os níveis desses poluentes e o número de atendimentosem emergência ehospitalizações por causas respiratórias[3,4,5,6,7,8,9,10,11,12,13,14].

Para detectar essa associação é necessário empregar técnicas estatísticas que permitam isolar os efeitos da poluição do ar, uma vez que a ocorrência de doenças respiratórias está relacionada a diversos outros fatores como, temperatura e umidade. A dificuldade metodológica na análise de tais fenômenos consiste em detectar variações na ocorrência dos desfechos associados a eventos de poluição do ar de baixa intensidade. Os avanços das técnicas estatísticas têm possibilitado análises mais precisas desse tipo de associação. Particularmente, os modelos de séries temporais têm desempenhado um papel importante como ferramenta de análise nesses estudos com a consideração de áreas geográficas discriminadas, próximas de estações de monitoramento da qualidade do ar.

O modelo aditivo generalizado (MAG)[15] é amplamente utilizado como uma técnica flexível e efetiva para modelar por meio de regressão não-linear, dados coletados em função do tempo sobre efeitos dos poluentes na saúde. Esse modelo também é uma alternativa para o ajuste de relações não lineares não especificadas e mostra que essa classe de modelos constitui uma boa opção para

representar tanto a sazonalidade quanto a relação entre a variável resposta e os fatores de confusão[16].

Em geral, estudos epidemiológicos de séries temporais utilizam uma única estação de monitoramento fixa ou a média das estações para representar a exposição de toda a população aos poluentes. Essa abordagem não reflete a verdadeira exposição das pessoas[17]. Uma alternativa para uma exposição mais realista é considerar uma pequena área geográfica próxima da estação de monitoramento[18].

Poucos estudos epidemiológicos avaliaram a influência dos poluentes por área de abrangência das estações de monitoramento, embora existam algumas evidências de que a exposição de erro de classificação em análise de series temporais tende para um viés as estimativas para baixo, e nesse sentido, não limita a importância dos resultados para saúde pública[18].

Portanto, metodologias nas quais as áreas de abrangência sejam melhor discriminadas espacialmente tornam-se um fator importante para obter inferências mais precisas. Nesse contexto, o objetivo deste estudo foi avaliar, por meio do MAG, a relação de curto prazo entre o número de atendimentos de emergência por problemas respiratórios, em crianças menores de seis anos, com níveis de poluentes atmosféricos observados na Região da Grande Vitória (RGV), com a consideração de variáveis temporais medidas por localidade "in loco" (área em torno das redes de monitoramento) e por média entre as estações (média regional global).

MÉTODOS

Área de Estudo

Estudo ecológico, espacial e temporal, realizado na Região da Grande Vitória (RGV) no período de 01 de janeiro de 2005 a 31 de dezembro de 2010. A RGV écomposta por 7 municípios (Vitória, Vila Velha, Cariacica, Serra, Viana, Guarapari e Fundão), abrange uma área de 2.318.917 km^2, com uma de população de aproximadamente 1,7 milhões de habitantes, sendo um dos principais polos de desenvolvimento urbano e industrial do Espírito Santo[19]. As principais atividades poluidoras da região incluem as vias de tráfego, indústrias de diversos seguimentos (siderurgia, pelotização, mineração, cimenteiras), portos, aeroportos, emissões residenciais e comerciais[20].

Na RGV existem oito estações de monitoramento da qualidade do ar, que juntas compõem a Rede Automática de Monitoramento da Qualidade do Ar (RAMQAr), gerenciada pelo Instituto Estadual de Meio Ambiente e Recursos Hídricos (IEMA). Para a análise espacial, foram consideradas oito regiões no entorno das estações da RAMQAr, dentro de um círculo com raio de aproximadamente 2 km. (Figura 1)

Desfecho de Saúde

Foram levantados dados dos atendimentos diários por doenças respiratórias de crianças de 0 a 6 anos, atendidas nos serviços de emergência de dois hospitais da RGV, um da rede pública e outro da rede privada:Hospital Infantil Nossa Senhora da Glória e Centro Integrado de Atenção a Saúde da Unimed Vitória, respectivamente. As doençasrespiratórias foram codificadas de acordo com a 10ª revisão da Classificação Internacional de Doenças (CID-10: J00-J99). Foram levantados os dados dos atendimentos de crianças que residiam a uma distância de até 2 km das estações da RAMQAr. Os bairrosforam localizados no mapa de áreas de influência de cada uma das estações de monitoramento e à partir dessa informação, agrupou-se os atendimentos por área de influência.

Poluentes Ambientais e Variáveis meteorológicas

As concentrações diárias de MP_{10}, SO_2, NO_2, O_3 e CO e as variáveis meteorológicas temperatura (°C), umidade relativa do ar (%) e precipitação pluviométrica (mm) referentes ao período do estudo foram fornecidas pelo IEMA[20].

Dados Faltantes

As falhas no monitoramento dos poluentes na RAMQAr ocorridos durante o período estudado, tanto em dias isolados como em dias consecutivos, causaram lacunas nos registros das concentrações e foram corrigidas pelo método de imputação seguindo metodologia descrita por Braga[21]. Nesse método as estimativas obtidas são explicadas pela correlação espacial entre os diferentes níveis do mesmo poluente nos diferentes monitores e pela autocorrelação dos níveis do poluente no mesmo monitor, ao longo do tempo.

Análise estatistica

A estratégia de modelagem consistiu em definir um modelo central com todas asinformações conhecidas (tendência, sazonalidade, dias da semana, feriados e as condições meteorológicas), a fim de explicar a variabilidade do número de atendimentos por doenças respiratórias, exceto a concentração dos poluentes.

A escolha das variáveis e covariáveis para compor o modelo foram baseadas em testes e diagnósticos em cada etapa do processo de modelagem. Os diagnósticos foram baseados na análise residual e no critério Akaike (AIC)[15].

Modelo Aditivo Generalizado - MAG

O número diário de atendimentos médicos representa um processo de contagem e o modelo aditivo generalizado (MAG), com distribuição marginal de Poisson, foi a ferramenta estatística utilizada para estimar a forma da curva da relação entre desfecho de saúde e poluição do ar[5,13,14].

Seja $\{Y_t\} \equiv \{Y_t\}_{t\in\mathbb{Z}}$, uma série de contagem, ou seja, $y_t \in \{0, 1, \ldots\}$. A distribuição condicional de Y_t, dado o passado $\mathcal{F}_{t-1}$ que contém toda informação disponível até o momento *t*−1, é denotada por:

$$p(y_t;\mu_t|\mathcal{F}_{t-1}) = \frac{e^{-\mu_t}\mu_t^{y_t}}{y_t!}, \quad (1)$$

em que μ_t, representa o valor esperado (média) de Y_t. Assim, dada uma amostra $Y_1,\ldots,Y_n$, composta de "*n*" variáveis aleatórias mutuamente condicionalmente independentes, pertencentes à $\{Y_t\}$, a função de log-verossimilhança condicional é dada por:

$$l(\mu) = \sum_{t=1}^{n} \ln p(y_t;\mu_t|\mathcal{F}_{t-1}) \propto \sum_{t=1}^{n} (Y_t \ln\mu_t - \mu_t) \quad , \qquad (2)$$

onde o vetor $\mu = (\mu_1,\ldots,\mu_n)$ depende dos parâmetros e do processo $\{Y_t\}$. Seja $X_t = [X_{1t},\ldots,X_{pt}]^T$ o vetor de covariáveis de dimensão *p* no tempo *t*, onde T denota a transposta, que pode incluir valores passados de Y_t, e outras informações auxiliares, tais como poluentes e variáveis de confusão (tendência, sazonalidade e variáveis meteorológicas, entre outros). Neste estudo, a sequência $X_{1t},\ldots,X_{qt}$ denota as concentrações dos poluentes MP_{10}, SO_2, NO_2, O_3 e CO, portanto *q*=5, enquanto $X_{(q+1)t},\ldots,X_{pt}$ indica as variáveis de confusão no tempo *t*, (*p*>*q*).

A relação entre Y_t e o vetor $\mathbf{X}_t$ é dada por:

$$\ln(\mu_t) = \sum_{j=0}^{q} \beta_j X_{jt} + \sum_{j=q+1}^{p} f_j(X_{jt}), \quad \text{com } q \le p, \quad (3)$$

onde (β_0, β), com $\beta = (\beta_1,\ldots,\beta_q)^T$ é o vetor dos coeficientes a serem estimados (β_j é o coeficiente *j*-ésima covariável) e f_j é a função suavisadora para a *j*-ésima variável de confusão. Além disso, β_0 indica o intercepto da curva e está associado a X_{0t}= 1 para todo *t*. Todo o processo de modelagem foi realizado no software R (R Development Core Team, 2009) com o pacote ARES[22].

O risco relativo (RR) é frequentemente utilizado em estudos epidemiológicos para medir o impacto das concentrações de poluentes atmosféricos na saúde da população exposta e é definido como a razão entre as probabilidades de que um evento ocorra após certa exposição e a de o evento ocorrer sem ter havido a exposição ao fator de risco[23].

O RR de uma covariável poluente $X_j, j = 1,\ldots,q$, é dado como sendo a variação relativa na contagem esperada de eventos de doenças respiratórias pela variação ξ de unidade na covariável enquanto mantidas as outras covariáveis fixas. Mais precisamente, como definido em[24], fórmula (8), o RR é dado por:

$$RR_{X_j}(\xi) = \frac{E(Y|X_j = \xi, X_i, i \neq j)}{E(Y|X_j = 0, X_i, i \neq j)} \quad (4)$$

Para a regressão de Poisson o RR não depende dos valores x_i, $i \neq j$, das outras covariáveis e pode ser expresso como:

$$RR_{X_j}(\xi) = \exp(\beta_j \xi) \quad (5)$$

Para o modelo MAG com distribuição marginal de Poisson, o RR e o seu intervalo de confiança aproximado (IC), a um nível de significância α, de uma covariável X_j, j=1…q, é estimado da seguinte forma:

$$\hat{RR}_{X_j}(\xi)_{X_j} = \exp(\hat{\beta}_j \xi) \quad \text{e} \quad IC\left(RR_{X_j}(\xi)\right) = \exp(\hat{\beta}_j \xi \pm z_{\alpha/2} s(\hat{\beta}_j)\xi), \quad (7)$$

$\hat{\beta}_j$ é o coeficiente estimado associado ao poluente X_j em estudo com erro padrão $s(\hat{\beta}_j)$ e $z_{\alpha/2}$ é o quantil $\alpha/2$ da distribuição normal padrão. A um nível de significância α, a hipótese a ser testada é definida como $H_0: RR_{X_j} = 1$ contra $H_0: RR_{X_j} > 1$, onde $RR_{X_j} = RR_{X_j}(1)$ ou seja, RR da variação da unidade em X_j. A rejeição de H_0 implica estatisticamente que o respectivo poluente tem um efeito adverso significativo na saúde.

Neste estudo, os cálculos dos valores dos RR(x) correspondem ao aumento de k=1000 (μg/m^3) nos níveis de CO e de k=10 (μg/m^3) para os demais poluentes. Os resultados são apresentados em aumentos percentuais nos números de atendimentos médicos e são calculados através da expressão:

$$\%RR(x) = (RR(x) - 1) \times 100. \quad (8)$$

Defasagem (lag)

O estudo investigou os efeitos respiratórios associados aos níveis de poluição no dia do atendimento na emergência (lag0) e nos dias anteriores (lag1, lag2, lag3). O efeito acumulado foi avaliado com as médias móveis de dois a oito dias (MA01, MA02, MA03, MA04, MA05, MA06, MA07)[25].

O projeto foi aprovado no Comitê de Ética Profissional (CEP) do Centro de Ciência da Saúde da Universidade Federal do Espírito Santo, sob o número 04/11 em 14 de Maio de 2011.

RESULTADOS (700 palavras)

No período estudado foram registrados 46.421 atendimentos por doenças respiratórias de crianças de 0 a 6 anos, residentes nas áreas de abrangência das oito estações de monitoramento da RAMQAr. A média diária de atendimentos na RGVfoi de 21,19 (DP= 9,90) evariou de 1,72 a 4,84 atendimentos/dia nas diferentes regiões. O maior número de atendimentos foi de crianças residentes na região da Enseada do Suá.

A temperatura média no período variou de 20,85 a 29,36°C, a quantidade de chuva variou de 0 mm a 117,80 mm (média= 3,78 mm) e a umidade relativa do ar variou de 61,79% a 97,27% (média= 77,47%).

As concentrações dos poluentes não apresentaram comportamento uniforme entre as diferentes estações, o que pode ser justificado pelas atividades poluidoras específicas de cada região. As concentrações médias registradas no período estão destacadas naTabela 1.

As médias dos poluentes nas estações da RAMQAr são baseados em dados "in loco" e são, em geral, bem próximas e diferenciadas de forma significativa da média geral. Essa evidência empírica mostra que as médias locais $E(\mu_t^L)$, (*L*=1,…,8), são diferentes da média regional $E(\mu_t^R)$. Esse é um resultado esperado, pois o processo não é estacionário nos momentos, isto é, $E(\mu_t^L) \neq E(\mu_t^R)$, para *L*=1,…,8. Esse resultado justifica o estudo proposto no sentido de comparar o desfecho de saúde com relação aos poluentes nas áreas espacialmente discriminadas, ou seja, áreas em torno das RAMQAr com a região geral (RGV).

A série de contagens diárias de atendimentos, na região geral, foi suavizado por uma "spline" com 12 graus de liberdade, definida por meio do critério de modelagem AIC e da análise residual. O ajuste não paramétrico evidenciou sazonalidade e uma tendência decrescente ao longo do tempo, fatores de confusão que foramincluídos no processo de modelagem. O período de outono (março) para o inverno (junho) evidenciou uma sazonalidade com aumento do número de atendimentos por causas respiratórias.

A análise do diagnóstico do ajuste do modelo, descrito anteriormente, por meio dos resultados obtidos da Regressão de Poisson para estimação do efeito do MP_{10} da média móvel 7 dias para RGV está apresentada na Figura 2. Observou-se que não havia evidência empírica de mal ajuste do modelo, isto é, os resíduos não são correlacionados e são aproximadamente normais. Também, o periodograma comprova que os resíduos apresentaram características de um ruído branco, isto é, os valores do período estão distribuídos de forma uniforme em função da frequência. Essas análises gráficas residuais fornecem o suporte necessário para o bom ajuste do modelo e,

consequentemente, realizar inferências. Assim, a qualidade do modelo ajustados é garantida pelas propriedades empíricas mostradas pelos resíduos.

Os riscos relativos (RR) e intervalos de confiança (IC) para cada modelo ajustado "in loco" e na região geral foram calculados para cada poluente. A Tabela 2 apresenta os riscos relativos estimados para um aumento de 10 $\mu g/m^3$ nos níveis de concentração de cada um dos poluentes atmosféricos para a região geral. Ao analisar os padrões apresentados nessa tabela, os gradientes evidenciaram claramente os efeitos significativos de MP_{10} nos atendimentos, para defasagens cumulativas. Para o SO_2, na defasagem simples foram significativos os valores dos RR para o lag0 e o lag2. ONO_2, o CO e o O_3 não apresentaram significância estatística no RRpara a RGV.

Na Tabela3 é apresentado um resumo dos valores de RR dos efeitos de cada poluente cujos cálculos apresentaram significância estatísticae os resultados em negrito apresentam maiores magnitudes. É verificado que, para os poluentes MP_{10} e SO_2, os valores estimados de RR foram maiores quando calculados para cada estação separadamente do que os obtidos para toda a RGV, isto é, modelo ajustado onde as covariáveis correspondem às médias das concentrações. Como visto anteriormente, os efeitos dos poluentes NO_2, CO e O_3 para a região agregada, RGV, não apresentaram significância estatística no RR, o que pode ser considerado um resultado espúrio pois os efeitos desses poluentes, para as localidades desagregadas por região, levaram a resultados de RRs altamente significativos. Essa evidencia empírica também corrobora o estudo proposto neste artigo o qual tem como objetivo comparar ajustes de modelos de regressão, por meio do RR, para variáveis explicativas medidas "in loco" e com a utilização das médias dessas variáveis (estudo geral).

DISCUSSÃO

Foi observado neste estudo o efeito direto da poluição atmosférica na saúde de crianças da RGV, por meio da estimativa do Risco Relativo em um MAG. As concentração dos poluentes MP_{10}, SO_2, NO_2 e O_3, mesmo dentro dos padrões da legislação vigente e da OMS, mostraram associação significativa com o aumento do número de atendimentos em emergência por doenças respiratórias em crianças de 0 a 6 anos.

O número de atendimentos hospitalares foi maior entre os meses de março a junho, outono e início do inverno. Esse aumento esperado se deve a diferentes fatores, como as baixas temperaturas que predispõem o agravamento de doenças respiratórias pré-existentes, maior incidência de doenças

respiratórias virais e aumento na concentração dos poluentes primários determinada pela escassez de chuvas e ocorrência de inversão térmica.

Os poluentes que mostraram efeito mais consistente foram o MP_{10} e o SO_2.O MP_{10} apresentou um efeito bastante substancial sobre os atendimentos por doenças respiratórias em crianças em todas as defasagens e em quase todas as regiões analisadas, exceto na área de entorno da RAMQAr de Vila Velha Ibes. Nas demais áreas de entorno das estações de monitoramento e na estimativa para a RGV, observou-se um padrão de efeito mais acumulado variando entre 2,62% e 12,08% (Tabelas2 e 3). Esses efeitos mostraram magnitudes variadas sendo que nas áreas de abrangência de Jardim Camburi e Carapina os efeitos estimados foram de maior magnitude em relação a outras áreas analisadas. Esse padrão de efeito está de acordo com o observado em outros estudos com metodologia semelhante, como por exemplo em Itabira, Minas Gerais, onde os autores verificaram que o aumento de 10 $\mu g/m^3$ no nível de MP_{10} foi associadoa aumento nos atendimentos de pronto-socorro por doenças respiratórias de 4% no dia e no dia seguinte, para crianças menores de 13 anos, e de 12%, nos três dias subsequentes para os adolescentes entre 13 e 19 anos[8].

As concentrações de SO_2, assim como observado para o MP_{10}, apresentaram associação com aumento no número de atendimentos por doenças respiratórias na população estudada. O RR na área de abrangência das estações de Jardim Camburi e Laranjeiras foram os mais elevados, o que pode ser explicado pela proximidade dessa região com o polo industrial siderúrgico existente dentro da malha urbana.

Os ajustes realizados no modelo, para cada uma das regiões no entorno das estações de monitoramento da RAMQAr, permitiu observar melhor o efeito de todos os poluentes sobre o número de atendimentos hospitalares por doenças respiratórias. Comparando as estimativas quando simulamos para toda a RGV com as análises no entorno das estações da RAMQAr, alguns efeitos só foram percebidos quando as análises foram feitas nas localidades desagregadas por região, sinalizando um efeito de maior magnitude em relação a estimativa para toda a região. A explicação para isso é que a análise pela média de todas as estações tende a suavizar os dados e assim diminuir a sua variabilidade, ocultando alguns efeitos. Assim, efeitos do NO_2, do CO e do O_3 sobre o número de atendimentos por doenças respiratórias foram identificados, especialmente na região de Jardim Camburi, Vila Velha e Laranjeiras.

Os efeitos observados para todos os poluentes na RGV estão de acordo com estudos prévios realizados na mesma região[13,14]. Nascimento et al.[14](2017) encontraram associação positiva entre a concentração de material particulado fino na atmosfera e o número de atendimentos hospitalares por doenças respiratórias agudas em crianças de até 12 anos no inverno e verão de 2013. Souza et al.[13](2018) desenvolveram uma modelagem híbrida com três ferramentas estatísticas, o modelo

Vetorial Autorregressivo (VAR), a Análise de Componentes Principais (ACP) e o Modelo Aditivo Generalizado (MAG), para relacionar os poluentes atmosféricos (MP_{10}, SO_2, NO_2, O_3 e CO) com o número de crianças com até 6 anos de idade atendidas em emergências da região devido a problemas respiratórios no período de 2005 a 2010 e observaram relação significativa entre os níveis de concentração dos poluentes e o número de atendimentos hospitalares. Em Palermo, Itália,foi observada associação entre número de atendimentos por doenças respiratórias em emergência e exposição aos poluentes MP_{10} (OR= 1,039, IC95% 1,020-1.059), SO_2 (OR= 1,068, IC95% 1,014-1,126), NO_2 (OR= 1,043, CI95% 1,021-1,065) e CO (OR= 1,128, CI95% 1,074-1,184)[26].Samoli et al.[10] na Grécia observaram que um incremento de 10 mg/m^{3}de MP_{10} e de SO_2 estava associado com um aumento de 2,54% e 5,98% no número de hospitalizações por doenças respiratórias, respectivamente.

Apesar da associação entre os diversos poluentes e o risco de atendimento em emergência por doença respiratória em crianças, observou-se uma tendência decrescente dos níveis de poluentes ao longo do tempo, o que pode ser explicado por maior controle local da poluição do ar nos últimos anos.

O maior número de crianças com doenças respiratórias residentes na região da Enseada do Suá pode indicar maior exposição das crianças nessa região. O poluente que mostrou-se com média mais alta foi o SO_2, entretanto, devido a falta de dados disponíveis, só foi calculado o RR para o MP_{10} nessa região, o que é uma falha deste estudo.

Para pesquisas futuras, outros grupos suscetíveis devem ser investigados na mesma região para que seja possível elaborar um quadro completo dos efeitos agudos da poluição atmosférica na saúde da população. Como metodologia alternativa, técnicas de bootstrap poderão ser utilizadas com o objetivo de obter intervalos de mesma precisão, mas com menor amplitude amostral. Outra metodologia a ser considerada é a estimação da variância por meio de modelos heterocedásticos. O modelos GLARMA[27] e PINAR[28], modelos com maior complexidade estrutural, são ferramentas estatísticas que podem ser abordadas no estudo proposto.

A consistência das associações e a magnitude dos efeitos observados nas regiões analisadas, mesmo em um ambiente com níveis de poluentes dentro dos padrões estabelecidos pelas agências regulatórias, se mostram extremamente relevantes em termos de saúde pública. Os resultados encontrados fornecem subsídios para a elaboração de medidas que visem a minimizar os riscos à saúde, contribuindo ainda com o planejamento de saúde ambiental e urbana no aperfeiçoamento de políticas públicas.

Os resultados apresentados são baseados na dissertação de mestrado do primeiro autor sob a orientação de Reisen VA, no PPGEA, UFES, defendida no ano de 2012.

Agradecimentos: os autores gradecem ao apoio dos órgãos de fomento à pesquisa (CNPq, CAPES, FAPES, FACITEC),Instituto Estadual de Meio Ambiente (IEMA), Hospital Infantil Nossa Senhora da Glória (HINSG) e Centro Integrado de Atenção à Saúde da Unimed Vitória (CIAS).

Conflito de Interesses: Os autores declaram não haver conflito de interesses.

REFERÊNCIAS

1. WHO. Inheriting a sustainable world? Atlas on children's health and the environment. Geneva: World Health Organization; 2017. Licence: CC BY-NC-SA 3.0 IGO.
2. Holgate S. Every breath we take: the lifelong impact of air pollution' - a call for action. *Clin Med*. 2017;17(1):8-12. DOI: 10.7861/clinmedicine.17-1-8.
3. Cifuentes LA, Vega J, Köpfer K, Lave LB. Effect of the fine fraction of particulate matter versus the coarse mass and other pollutants on daily mortality in Santiago, Chile. *J Air & Waste Manage Assoc*. 2000;50(8):1287–98.
4. Gouveia N, Mendonça GA, León AP, Correia JEM, Junger WL, Freitas CU, et al. Poluição do ar e efeitos na saúde nas populações de duas grandes metrópoles brasileiras. *Epidemiol Serv Saúde*. 2003;12(1):29–40.
5. Daumas RP, Mendonça GAS, León AP. Air pollution and mortality in the elderly in Rio de Janeiro: a time-series analysis. *Cad Saúde Pública*. 2004;20(1):311–9.
6. Bakonyi SMC, Danni-Oliveira IM, Martins LC, Braga ALF. Poluição atmosférica e doenças respiratórias em crianças na cidade de Curitiba, PR. *Rev Saude Publica*. 2004;38(5):695–700.
7. Nascimento LFC, Pereira LAA, Braga ALF, Módolo MCC, Carvalho Jr JA. Effects of air pollution on children's health in a city in southeastern Brazil. *Rev Saúde Pública*. 2006:40(1):77–82.
8. Braga ALF, Pereira LAA, Procópio M, André PA, Saldiva PHN. Associação entre poluição atmosférica e doenças respiratórias e cardiovasculares na cidade de Itabira, Minas Gerais, Brasil. *Cad Saude Publica*. 2007;23(4):570–8.
9. Chen R, Chu C, Tan J, Cao J, Song W, Xu X, et al. Ambient air pollution and hospital admission in Shanghai, China. *J Hazard Mat*. 2010;181(1-3):234–40.

10. Samoli E, Nastos PT, Paliatsos AG, Katsouyanni K, Priftis KN. Acute effects of air pollution on pediatric asthma exacerbation: Evidence of association and effect modification. *Environmental Research*.2011;111:418–24.
11. Almeida SP, Casimiro E, Calheiros J. Short-term association between exposure to ozone and mortality in oporto, portugal. *Environ Res*. 2011;111(3):406–10.
12. Souza JB, Reisen VA, Santos JM, Franco GC. Principal components and generalized linear modeling in the correlation between hospital admissions and air pollution. *Rev Saúde Pública*. 2014;48:451–8.
13. Souza JB, Reisen VA, Franco GC, Ispány M, Bondon P, Santos JM. Generalized additive models with principal component analysis: an application to time series of respiratory disease and air pollution data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*. 2018;67(2):453–80.
14. Nascimento AP, Santos JM, Mill JG, Souza JB, Reis Júnior NC, Reisen VA. Associação entre concentração de partículas finas na atmosfera e doenças respiratórias agudas em crianças. *Rev Saúde Pública*. 2017; 51:3. DOI:10.1590/S1518-8787.2017051006523
15. Hastie TJ, Tibshirani RJ. Generalized Additive Models. Monographs on Statistics and Applied Probability 43. Chapman & Hall: New York. 1990, 352p.
16. Dominici F, McDermott A, Zeger SL, Samet JM. On the use of generalized additive models in time-series studies of air pollution and health. *Am J Epidemiol*. 2002;156(3):193–203.
17. Namdeo A, Tiwary A, Farrow E. Estimation of age-related vulnerability to air pollution: Assessment of respiratory health at local scale. *Environ Int.* 2011 Jul;37(5):829-37. doi: 10.1016/j.envint.2011.02.002.
18. Zeger SL, Thomas D, Dominici F, Samet JM, Schwartz J, Dockery D, et al. Exposure measurement error in time-series studies of air pollution: concepts and consequences. *Environ Health Perspectives*. 2000;108(5):419-26.
19. Instituto Brasileiro de Geografia e Estatística (IBGE). Censo populacional 2010. Disponível em https://ww2.ibge.gov.br/home/estatistica/populacao/censo2010. Acessado em 19 de maio de 2018.
20. IEMA – Instituto Estadual de Meio Ambiente e Recursos Hídricos do Estado do Espírito Santo. Relatório da qualidade do ar na região da grande vitória 2013. Disponível em < http://www.meioambiente.es.gov.br/>. Acessado em 19 de maio de 2018.
21. Braga A, Zanobetti A, Schwartz J, The time course of weather-related deaths.*Epidemiology*.2001;12(6):662-7.

22. Junger W, Leon AP. *ares: Environment air pollution epidemiology: a library for time series analysis*, 2011. R package version 0.7.2.
23. Zou G. A modified poisson regression approach to prospective studies with binary data. *Am J Epidemiol*. 2004;159(7):702–6.
24. Baxter LA, Finch SJ, Lipfert FW, Yu Q. Comparing estimates of the effects of air pollution on human mortality obtained using different regression methodologies. *Risk analysis*. 1997;17(3):273–78.
25. Katsouyanni K, Schwartz J, Spix C, Touloumi G, Zmirou D, Zanobetti A, et al. Short term effects of air pollution on health: a european approach using epidemiologic time series data: the APHEA protocol. *J Epidemiol Community Health.* 1996;50(Suppl 1):S12–S18.
26. Tramuto F, Cusimano R, Cerame G, Vultaggio M, Calamusa G, Maida CM, et al. Urban air pollution and emergency room admissions for respiratory symptoms: a case-crossover study in Palermo, Italy. *Environ Health*. 2011;10(1):31. DOI: 10.1186/1476-069X-10-31.
27. Davis RA, Dunsmuir WTM, Streett SB. Maximum likelihood estimation for an observation driven model for poisson counts. *Methodology and Computing in Applied Probability*. 2005;7(2):149–59.
28. Monteiro M, Scotto MG, Pereira I. Integer-valued autoregressive processes with periodic structure. *J Statist Plan Infer. 2010;*140(6):1529–41.

Tabelas:

Tabela 1.Estatística descritiva dos atendimentos por doenças respiratórias, das médias diárias do PM_{10} ($\mu g/m^3$), SO_2 ($\mu g/m^3$) e NO_2 ($\mu g/m^3$), e das médias móveis de 8 horas do O_3 ($\mu g/m^3$) e CO ($\mu g/m^3$), em cada RAMQAr, 2005-2010.

RAMQAr	*Número de Atendimentos*		PM_{10}		SO_2		NO_2		O_3		CO	
	Média(total)	**Desvio(máx)**	**Média**	**Desvio**	**Média**	**Desvio**	**Média**	**Desvio**	**Média**	**Desvio**	**Média**	**Desvio**
Laranjeiras	2,05 (4482)	1,83 (12,00)	32,90	11,39	12,61	5,79	41,23	14,85	43,34	12,98	647,58	175,55
Carapina	2,31 (5072)	2,1 (17,00)	23,02	7,96	-	-	-	-	-	-	-	-
Jardim Camburi	2,24 (4899)	1,82 (12,00)	26,95	8,06	14,15	7,49	42,07	12,76	-	-	-	-
Enseada S.	4,84(10,6k)	2,86 (19,00)	29,39	9,18	**16,32**	7,91	44,10	14,29	38,66	10,99	783,30	276,71
V. Centro	3,5 (7658)	2,39 (15,00)	26,09	7,23	15,77	6,32	**55,78**	15,27	-	-	**1730,91**	715,34
Ibes	1,72 (3779)	1,48 (9,00)	29,24	9,67	10,73	6,20	38,39	11,15	**54,48**	16,38	657,24	267,07
VV Centro	2,25 (4924)	1,76 (12,00)	23,49	8,22	11,99	5,80	-	-	-	-	-	-
Cariacica	2,28 (4994)	1,91 (13,00)	**43,06**	15,94	5,50	2,62	50,65	19,35	37,79	12,59	609,79	266,86
RGV	21,19(46,4k)	9,9 (64,00)	29,27	9,73	12,44	3,11	45,37	11,05	43,57	10,83	885,76	231,24

Em negrito estão as concentrações máximas para cada poluente.

Tabela 2. Riscos Relativos para atendimentos por doenças respiratórias em crianças menores de 6 anos para um acréscimo de 10 μg/m^3 de MP_{10}, SO_2, NO_2, O_3 e CO na RGV, jan /2005 a dez/2010.

Exposição	PM_{10}			
	%RR	Inferior	Superior	p.valor
Dia corrente	0,99	-0,5	2,5	0,19
Defasagem de 1 dia	0,04	-1,35	1,46	0,95
Defasagem de 2 dias	0,81	-0,58	2,23	0,25
Defasagem de 3 dias	1,1	-0,29	2,51	0,12
Acumulado de 2 dias	0,69	-1,02	2,42	0,43
Acumulado de 3 dias	1,13	-0,82	3,12	0,26
Acumulado de 4 dias	1,77	-0,44	4,03	0,12
Acumulado de 5 dias	2,43	-0,05	4,97	0,05*
Acumulado de 6 dias	2,73	0	5,53	0,05*
Acumulado de 7 dias	3,29	0,31	6,36	0,03*
Acumulado de 8 dias	2,5	-0,67	5,77	0,12
Exposição	***S O2***			
	%RR	Inferior	Superior	p.valor
Dia corrente	4,47	-0,01	9,14	0,05*
Defasagem de 1 dia	4,2	-0,2	8,79	0,06
Defasagem de 2 dias	5,26	0,81	9,9	0,02*
Defasagem de 3 dias	1,43	-2,89	5,94	0,52
Acumulado de 2 dias	6,47	0,99	12,26	0,02*
Acumulado de 3 dias	8,8	2,55	15,44	0,01*
Acumulado de 4 dias	8,76	1,93	16,05	0,01*
Acumulado de 5 dias	7,09	-0,13	14,84	0,05*
Acumulado de 6 dias	3,7	-3,71	11,69	0,34
Acumulado de 7 dias	2,23	-5,45	10,53	0,58
Acumulado de 8 dias	0	-7,85	8,52	1
Exposição	***NO2***			
	%RR	Inferior	Superior	p.valor
Dia corrente	0,25	-0,89	1,4	0,67
Defasagem de 1 dia	-0,9	-2,04	0,25	0,12
Defasagem de 2 dias	-0,45	-1,56	0,67	0,43
Defasagem de 3 dias	-0,16	-1,25	0,94	0,78
Acumulado de 2 dias	-0,5	-1,9	0,93	0,49
Acumulado de 3 dias	-0,77	-2,39	0,88	0,36
Acumulado de 4 dias	-0,82	-2,63	1,01	0,38
Acumulado de 5 dias	-0,91	-2,86	1,09	0,37
Acumulado de 6 dias	-0,91	-3,02	1,25	0,41
Acumulado de 7 dias	-0,37	-2,63	1,95	0,75
Acumulado de 8 dias	-0,31	-2,72	2,16	0,81
Exposição	***O3***			
	%RR	Inferior	Superior	p.valor
Dia corrente	0,54	-0,54	1,62	0,33
Defasagem de 1 dia	0,46	-0,59	1,52	0,39
Defasagem de 2 dias	0,07	-0,97	1,13	0,89
Defasagem de 3 dias	-0,32	-1,36	0,73	0,55
Acumulado de 2 dias	0,74	-0,55	2,05	0,26
Acumulado de 3 dias	0,69	-0,78	2,19	0,36
Acumulado de 4 dias	0,45	-1,2	2,12	0,6
Acumulado de 5 dias	0,34	-1,46	2,17	0,71
Acumulado de 6 dias	0,92	-1,04	2,91	0,36
Acumulado de 7 dias	0,66	-1,42	2,78	0,54
Acumulado de 8 dias	0,6	-1,6	2,86	0,6
Exposição	***CO***			
	%RR	Inferior	Superior	p.valor
Dia corrente	1,83	-3,25	7,17	0,49
Defasagem de 1 dia	-3,61	-8,21	1,23	0,14
Defasagem de 2 dias	0,39	-4,37	5,39	0,87
Defasagem de 3 dias	4,6	-0,26	9,7	0,06
Acumulado de 2 dias	-1,69	-7,76	4,76	0,6
Acumulado de 3 dias	-1,24	-8,3	6,37	0,74
Acumulado de 4 dias	2,15	-5,97	10,98	0,61
Acumulado de 5 dias	3,23	-5,74	13,05	0,49
Acumulado de 6 dias	2,53	-7,11	13,17	0,62
Acumulado de 7 dias	4,89	-5,72	16,7	0,38

Acumulado de 8 dias	5,83	-5,6	18,65	0,33

Tabela 3.Aumento percentual e intervalo de confiança de 95% dos atendimentos pediátricos de emergência por sintomas respiratórios. RGV, 2005-2010.

Exposição	MP_{10}			
	RR	**(IC 95%)**	**p-valor**	**RAMQAr**
Acumulado de 5 dias	2,43	(-0,05; 4,97)	0,05	RGV
Acumulado de 6 dias	2,73	(-0,00; 5,53)	0,05	RGV
Acumulado de 7 dias	3,29	(0,31; 6,36)	0,03	RGV
Defasagem de 1 dia	4,49	(1,45; 7,62)	0	Laranjeiras
Acumulado de 2 dias	4,5	(1,04; 8,08)	0,01	Laranjeiras
Acumulado de 3 dias	5,17	(1,35; 9,13)	0,01	Laranjeiras
Acumulado de 4 dias	4,66	(0,60; 8,89)	0,02	Laranjeiras
Defasagem de 1 dia	4,66	(0,48; 9,02)	0,03	Carapina
Acumulado de 6 dias	8,36	(0,16; 17,23)	0,05	Carapina
Acumulado de 7 dias	**12,27**	(3,15; 22,19)	0,01	Carapina
Acumulado de 8 dias	**11,5**	(1,85; 22,06)	0,02	Carapina
Acumulado de 6 dias	**10,59**	(1,79; 20,16)	0,02	Jardim Camburi
Acumulado de 7 dias	**12,08**	(2,63; 22,40)	0,01	Jardim Camburi
Acumulado de 8 dias	**11,82**	(1,87; 22,73)	0,02	Jardim Camburi
Defasagem de 3 dias	2,41	(0,17; 4,71)	0,03	Enseada do Suá
Dia corrente	4,08	(0,54; 7,76)	0,02	Vitoria Centro
Acumulado de 6 dias	6,58	(-0,03; 13,64)	0,05	Vitoria Centro
Acumulado de 7 dias	7,13	(0,01; 14,75)	0,05	Vitoria Centro
Acumulado de 7 dias	9,05	(1,54; 17,12)	0,02	VV Centro
Acumulado de 8 dias	9,01	(1,19; 17,42)	0,02	VV Centro
Defasagem de 3 dias	2,62	(0,58; 4,71)	0,01	Cariacica
Acumulado de 7 dias	4,53	(0,36; 8,88)	0,03	Cariacica
Acumulado de 8 dias	4,77	(0,40; 9,34)	0,03	Cariacica
Exposição	SO_2			
	RR	**(IC 95%)**	**p-valor**	**RAMQAr**
Dia corrente	4,47	(-0,01; 9,14)	0,05	RGV
Defasagem de 2 dias	5,26	(0,81; 9,90)	0,02	RGV
Acumulado de 2 dias	6,47	(0,99; 12,26)	0,02	RGV
Acumulado de 3 dias	8,8	(2,55; 15,44)	0,01	RGV
Acumulado de 4 dias	8,76	(1,93; 16,05)	0,01	RGV
Acumulado de 5 dias	7,09	(0,13; 14,84)	0,05	RGV
Dia corrente	**10,68**	(1,09; 21,17)	0,03	Laranjeiras
Defasagem de 2 dias	9,71	(1,68; 18,37)	0,02	Jardim Camburi
Acumulado de 7 dias	**13,52**	(-0,24; 29,18)	0,05	Jardim Camburi
Dia corrente	8,7	(1,12; 16,85)	0,02	VV Centro
Acumulado de 2 dias	**11,31**	(1,23; 22,40)	0,03	VV Centro
Defasagem de 3 dias	**11,9**	(0,04; 25,17)	0,05	VV Centro
Defasagem de 2 dias	7,57	(1,17; 14,38)	0,02	Cariacica
Exposição	NO_2			
	RR	**(IC 95%)**	**p-valor**	**RAMQAr**
Defasagem de 2 dias	3,85	(0,99; 6,80)	0,01	Jardim Camburi
Acumulado de 6 dias	3,56	(0,35; 6,88)	0,03	Vitoria Centro
Acumulado de 7 dias	3,88	(0,46; 7,42)	0,03	Vitoria Centro
Acumulado de 8 dias	4,32	(0,70; 8,06)	0,02	Vitoria Centro
Exposição	**CO**			
	RR	**(IC 95%)**	**p-valor**	**RAMQAr**
Defasagem de 3 dias	1,78	(-0,02; 3,62)	0,05	Laranjeiras
Defasagem de 2 dias	2,24	(0,72; 3,79)	0	Ibes
Defasagem de 3 dias	1,97	(0,46; 3,51)	0,01	Ibes
Acumulado de 4 dias	3,01	(0,60; 5,48)	0,01	Ibes
Acumulado de 5 dias	2,96	(0,34; 5,65)	0,03	Ibes
Acumulado de 6 dias	2,99	(0,20; 5,86)	0,04	Ibes
Exposição	**O3**			
	RR	**(IC 95%)**	**p-valor**	**RAMQAr**
Dia corrente	3,23	(0,17; 6,37)	0,04	Laranjeiras
Acumulado de 4 dias	4,13	(-0,08; 8,51)	0,05	Laranjeiras

Figura 01. Área de abrangência do estudo.

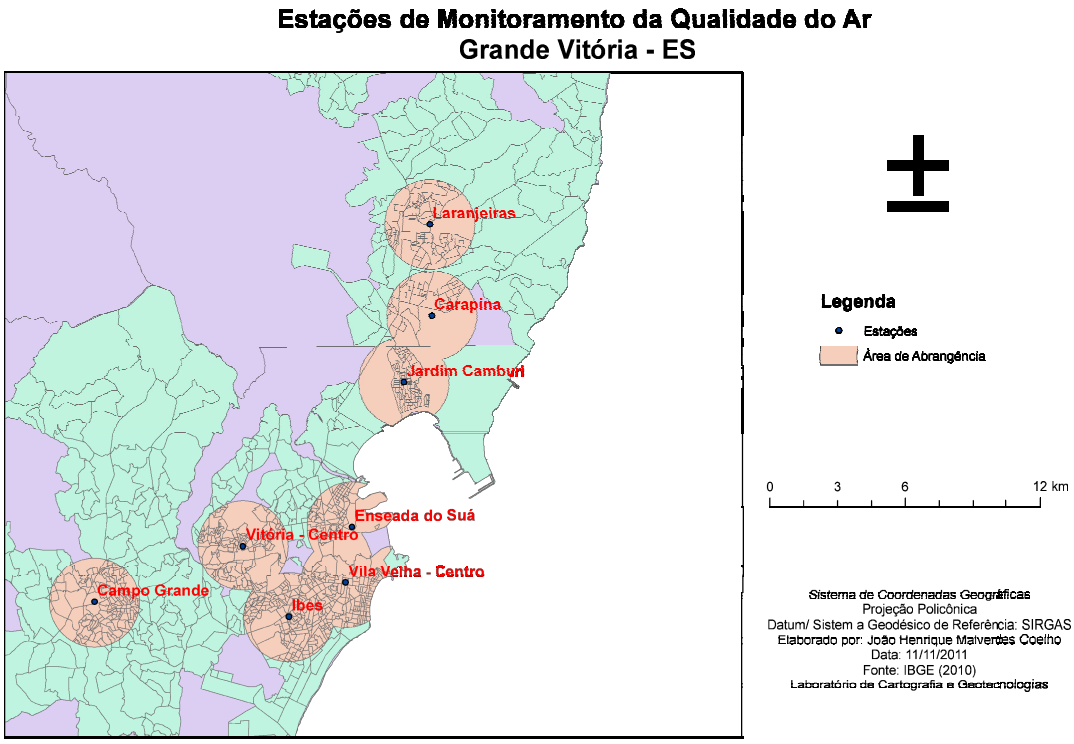


Figure 02. Diagnóstico do Modelo Central para RGV, da esquerda para direita e de cima para baixo, seguem nessa ordem: o gráfico dos valores previstos, os resíduos contra o tempo, a distância de Cook, a função de correlação parcial, o periodograma dos resíduos e os quantis dos resíduos contra quantis da distribuição normal.