UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Exatas
Programa de Pós-Graduação em Estatística

Ana Júlia Alves Câmara

Counting Process and Derivations: An application for environmental and epidemiological data

Belo Horizonte
2023

Ana Júlia Alves Câmara

Counting Process and Derivations: An application for environmental and epidemiological data

Versão Final

Tese apresentada ao Programa de Pós-Graduação em Estatística da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Doutor em Estatística.

Orientador: Valdério Anselmo Reisen

Belo Horizonte
2023

Câmara, Ana Júlia Alves.

C173c Counting process and derivations [recurso eletrônico]: an application for environmental and epidemiological data / Ana Júlia Alves Câmara. – 2023.
1 recurso online (98 f. il, color.) : pdf.

Orientador: Valderio Anselmo Reisen.
Tese (doutorado) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística.
Referências: f. 65-74

1. Estatística – Teses. 2. Séries temporais de contagem – Teses. 3. Modelo GLARMA – Teses. 4. M-estimadores – Teses. 5. Bootstrap (Estatística) – Teses. 6. Doenças respiratórias – Epidemiologia – Teses. I. Reisen, Valderio Anselmo. II .Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística. III.Título.

CDU 519.2(043)

Ficha catalográfica elaborada pela bibliotecária Belkiz Inez Rezende Costa
CRB 6/1510 Universidade Federal de Minas Gerais - ICEx

UNIVERSIDADE FEDERAL DE MINAS GERAIS

PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

FOLHA DE APROVAÇÃO

"Counting process and derivations: An application for environmental and epidemiological data"

ANA JÚLIA ALVES CÂMARA

Tese submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em ESTATÍSTICA, como requisito para obtenção do grau de Doutor em ESTATÍSTICA, área de concentração ESTATÍSTICA E PROBABILIDADE.

Aprovada em 23 de fevereiro de 2023, pela banca constituída pelos membros:

Prof. Valderio Anselmo Reisen - Orientador
PPGEA-UFES

Prof. Pascal Thierry Bondon
- Laboratoire des Signaux et Systèmes - Université Paris-Saclay

Prof. Márton Ispány
University of Debrecen

Prof. Josu Arteche
(Universidad del País Vasco

Profa. Glaura da Conceicao Franco
DEST/UFMG

Belo Horizonte, 23 de fevereiro de 2023.

Esse trabalho é dedicado a todos que trabalham pela educação no Brasil.

Acknowledgments

First, I thank God for allowing me to walk this journey with health, courage, and determination.

To my mother and brother, Edméia and Igor, who have always been there, supporting me at all times and encouraging my academic life.

To my supervisor Valdério Reisen for the encouragement, friendship, advice and for always challenging me to be a better researcher.

To Professor Glaura Franco for her contributions and support.

To Professor Pascal Bondon for supervising and supporting me during my international mobility in France.

To my Ph.D. colleagues at UFMG and UFES for suggestions, encouragement, and for making this period lighter.

I also would like to thank the Brazilian agencies CAPES, CNPQ, FAPEMIG, and FAPES for providing financial support for my research. CentraleSupélec and Institut Dataia partially supported this work.

"*In God we trust, all others bring data.*"
(W. Edwards Deming)

Resumo

O modelo linear autorregressivo média móvel autorregressiva (GLARMA) tem sido utilizado em estudos epidemiológicos para avaliar o impacto de poluentes atmosféricos na saúde. Esse impacto é comumente quantificado por meio da medida de risco relativo (RR). Devido à natureza dos dados, é necessário atenção ao aplicar o modelo GLARMA em variáveis ambientais. Primeiramente, em geral, a inferência para o RR é baseada nas propriedades assintóticas do estimador de máxima verossimilhança, o que pode ser problemático para amostras pequenas. Em segundo lugar, os poluentes atmosféricos podem apresentar picos elevados, ou observações abruptas, que podem ser identificados como aditivos *outliers* e normalmente impactam as propriedades estatísticas das funções amostrais, como média, variância, autocorrelação e autocorrelação parcial. Além disso, a atmosfera é composta por uma mistura de gases, incluindo poluentes atmosféricos, que são séries temporais e apresentam propriedades complexas. Esses contaminantes também exibem a propriedade de multicolinearidade, que pode inflar a variância das estimativas e causar um viés significativo se ignorado. Esta tese propõe metodologias para o complexo sistema formado por variáveis ambientais utilizando o modelo GLARMA. Diferentes métodos bootstrap são estudados a fim de calcular intervalos de confiança para o RR sem qualquer suposição sobre a distribuição dos dados. Além disso, uma abordagem robusta para o modelo GLARMA é proposta para lidar com observações abruptas. Estudos numéricos são realizados para avaliar o desempenho das metodologias propostas considerando cenários distintos. Análises de dados reais são realizadas considerando variáveis atmosféricas e epidemiológicas nas cidades de Belo Horizonte, MG, e Vitória, ES, Brasil.

Palavras-chave: Séries temporais de contagem, GLARMA, Bootstrap, M-estimadores, Aditivo outliers, Doenças respiratórias, Poluição do ar.

Abstract

The generalized linear autoregressive moving average (GLARMA) model has been used in epidemiological studies to evaluate the impact of air pollutants in the atmosphere on human health. This impact is commonly quantified through the relative risk (RR) measure. Due to the nature of the data, care is required when the GLARMA model is applied to environmental variables. First, the inference for the RR is usually based on the asymptotic properties of the maximum likelihood estimator, which can be problematic for small sample sizes. Secondly, the air pollutants can present high peaks, or abrupt observations, that can be identified as additive outliers and typically cause consequences on the statistical properties of sample functions, such as mean, variance, auto-correlation, and partial autocorrelation. In addition, the atmosphere is composed of a mixture of gases, including air pollutants, which are time series presenting complex properties. These contaminants also display the multicollinearity property, which can inflate the variance of the estimates and can cause a significant bias if ignored. This thesis proposes methodologies for the complex system formed by environmental data using the GLARMA model. Different bootstrap methods are studied to calculate confidence intervals for the RR without any assumption about the data distribution. Besides, a robust approach for the GLARMA model is proposed to deal with outlying observations. Numerical studies are realized to evaluate the performance of the proposed methodologies considering distinct scenarios. Real data analyses are performed considering atmospheric and epidemiological variables in the cities of Belo Horizonte, MG, and Vitória, ES, Brazil.

Keywords: Count time series, GLARMA, Bootstrap, M-estimators, Additive outliers, Respiratory diseases, Air pollution

List of Figures

List of Tables

Contents

Chapter 1

Introduction

An increasing interest in count time series has been taking place recently in statistics. These non-Gaussian processes, composed of non-negative integers, can be found in many fields, such as economics, medicine, agriculture, social and physical sciences, sports, etc. Classic examples are the daily number of hospital admissions for a disease, the number of car accidents in a region, and the number of transactions of a given stock observed in one hour. The vast number of applications contributed to developing methodologies for analyzing count time series data. Statistical methods started to arise in the early 1970s with the introduction of generalized linear models (GLM) by Nelder and Wedderburn [1972a], an extension of the normal linear models for independent data and posteriorly developed by McCullagh and Nelder [1989]. The main idea was to expand the possibilities for the distribution of the response variable, which can assume distributions belonging to the exponential family, e.g., Normal, Poisson, Gamma, Negative Binomial, etc. In addition, the relation between the mean of the dependent variable (μ) and the linear predictor (η) can be more flexible, assuming any monotonous non-linear function. Nevertheless, the GLM can not capture the time dependency structure in the data. The earliest work considering correlated time series can be found in Cox [1981], where models are classified into two categories: observation and parameter driven. The main difference between them refers to how the dependence structure is added to the model.

Among the subsequent contributions stand out the following procedures. The integer-valued autoregressive (INAR), introduced by McKenzie [1985] and Al-Osh and Alzaid [1988]. Zeger and Qaqish [1988], proposed a quasi-likelihood approach to time series regression, posteriorly generalized by Benjamin et al. [2003], called generalized autoregressive moving average models (GARMA). Davis et al. [2003, 2005] proposed the generalized linear autoregressive moving average models (GLARMA). Heinen [2003] presented the ACP model class (Autoregressive conditional Poisson), and Ferland et al. [2006] proposed the Integer-valued GARCH process, both attractive for overdispersed counts. Fokianos and Tjosthein [2011] introduced the log-linear models for time series. Camara et al. [2021] recently proposed a model using the generalized additive model (GAM) with autoregressive moving average terms (GAM-ARMA) to model time structure and nonlinear associations between μ and η. In the Bayesian spectrum, Harvey and Fernandes [1989] applied state-space models with conjugate prior distributions, and Gamerman et al. [2013] proposed a family of non-Gaussian state-space models. Although many methods

have been developed for modeling correlated count series, all approaches have limitations, which contributed to the field did not develop a unified theory. However, Davis et al. [2021], in a recent review of these methodologies, addresses that despite the GLARMA model presenting some complexity in the estimation of general models, "this family is one of the most flexible and easily fit count methods that balance observation and parameter-driven models". In this procedure, an ARMA structure (see Box and Jenkins [1976]) is added to the GLM, allowing the modeling of correlated observations belonging to the exponential family. This method has been widely used in applications of distinct fields, see, e.g., Rydberg and Shephard [2003], in finances, Karami et al. [2017], in air pollution, Kim et al. [2018] in engineering, Ballesteros-Cánovas et al. [2018] and Peitzsch et al. [2021] in climate changes, among others.

Epidemiological data are frequently treated as time series of counts because they record the frequency of certain events in successive time intervals. In the last decades, many authors have been dedicated to studying the impact of air quality variables on human health. The primary sources of atmospheric pollutants in urban areas are industry and motor vehicles, which are clearly involved in the increase in hospitalizations and deaths by respiratory diseases. Pope et al. [1995], Dockery and Pope [1996], Villeneuve et al. [2003], and others indicated a positive association between mortality and particulate material (PM). Ostro et al. [1999], Schwartz [2000], and Chen et al. [2010] found a significant association between daily air pollutant concentration levels and hospital admission for respiratory and cardiovascular diseases. McGeehin and Mirabelli [2001], Ostro et al. [2009] and Hertel et al. [2009] analyzed temperature effects on mortality in USA and Germany. In 2022 the history of Ella Roberta Kissi-Debrah was told at the 27th United Nations Climate Conference (COP27). The British nine years old girl is the first person in history to have air pollution listed as a cause of death on a death certificate (Debrah [2022]). Those most susceptible to the effects of pollutants and climate variations are children, the elderly, and those with chronic diseases, especially cardiovascular and respiratory diseases.

Ozone (O_3), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), carbon monoxide (CO), and particulate matter (PM) are the main pollutants in the atmosphere, and even at concentrations within limits established by the World Health Organization (WHO) offer risk to human health (see Pope and Dockery [2018]and Lippmann [2014]). O_3 is a colorless and odorless gas. It can cause damage to the lung structure, reducing its capacity and decreasing resistance to infections. This air pollutant can also cause the aggravation of respiratory diseases, increasing the incidence of cough, asthma, and irritation in the upper respiratory tract and eyes (Baldacci et al. [2015]). NO_2 has a strong and irritating smell. It is very toxic and can cause inflammation, oxidative stress, and hyperreactivity in the airways (Guarnieri and Balmes [2014]). SO_2 is a colorless gas with a strong odor. Inhalation, even at very low concentrations, causes transient spasms of the smooth muscles of the pulmonary bronchioles. In progressively higher concentrations, they cause increased mucous secretion in the upper airways, severe mucosal inflammation, and reduced ciliary movement of the respiratory tract. It may also increase the incidence of rhinitis, pharyngitis, and bronchitis (Spann et al. [2016]). CO is a colorless, odorless, and tasteless

gas. It combines quickly with hemoglobin, taking the place of oxygen, and can lead to death by asphyxiation. Chronic exposure can cause damage to the central nervous, cardiovascular, pulmonary, and other systems. It can also affect fetuses causing reduced birth weight and delayed postnatal development (Roderique et al. [2015]). PM is a heterogeneous and complex mixture of particles that vary in size, weight, shape, chemical composition, solubility, and origin. The particles are classified according to their diameter as ultrafine (particles with a diameter smaller than 0.1μm - $PM_{0.1}$), fine (particles with a diameter between 0.1 and 2.5μm - $PM_{2.5}$) and coarse (particles with a diameter between 2.5 and 10μm – PM_{10}) (Anderson et al. [2012]). The $PM_{0.1}$ and $PM_{2.5}$ reach the lower respiratory system and can cause or aggravate respiratory diseases. The PM_{10} does not affect the lower respiratory tract. Still, it presents the most consistent association with premature mortality, increased hospital admissions for heart or lung causes, acute and chronic bronchitis, asthma attacks, and respiratory symptoms (Schwartz [2000]).

The statistical association between air quality variables and health effects must be computed with caution independently of the statistical regression and time series models used. This concern arises due to the following main factors: (1) In general, the response variable is a time series, and this should be taken into account; (2) The dynamic of the response variable and, therefore, the statistical functions that measure the impact of the pollutants on health, can not be fully explained by the response variable itself or by only one contaminant since the population under the study is exposed to a a complex mixture of pollutants and chemical compounds. Many authors have ignored this fact in the literature on epidemiological problems; (3) the covariates are, in general, time series and possess complex characteristics such as periodicity, missing values, and aberrant observations, among other phenomena. In addition, multicollinearity exists between these covariates, inflating the estimates' variance and can cause significant bias when ignored in the modeling strategy. Thus, ignoring these complex characteristics found in the response and covariates variables can lead to an erroneous model choice and a severe consequence on the analysis of the impact of the pollutants on the health of the population under study, such as, for example, a false-positive conclusion of the population health risk.

Recent works have paid attention to some of the issues mentioned above. For example, principal component analysis (PCA) has been proposed to mitigate multicollinearity in the covariate variables. Recently, Wang and Pham [2011] studied the combined effects of pollutants on daily mortality using a PCA and a robust method. The statistical estimates were more significant when the multivariate PCA technique was used. Nevertheless, the application of PCA generally requires the data to be obtained through independent replications. Souza et al. [2018] and Ispany et al. [2018] have considered the GAM model to estimate the RR to measure the impact of pollutants on health by combining multivariate time series and PCA technique. One of their main results showed that the estimation of the effects was more pronounced than previously suggested in the literature.

Focusing on problems in the environmental epidemiological area, in which the variables can have complex dynamics, this thesis aims to propose statistical count time series models and

tools, using GLARMA models, that are expected to be more accurate than the classical ones to quantify the association between respiratory diseases and air quality variables. For this, the following issues will be studied.

First, as relative risk (RR) is the standard statistical method to evaluate the correlation between air quality variables and health effects, researchers are frequently interested in interval estimation for this value. Confidence intervals (CIs) are usually computed based on asymptotic assumptions. However, in GLARMA models, theoretical properties were only established rigorously for a particular case: Poisson case, without covariate, and the time structure defined as a moving average process. Dunsmuir [2015] addresses that although asymptotic results have been developed for similar models in the last years, none of them applies to GLARMA given the model structure. Despite this, it is assumed that the central limit holds for inference purposes, which can be problematic for small sample sizes. Yet underexplored in studies involving air quality, the bootstrap CIs can be calculated without any assumption about the data distribution. The bootstrap procedure was proposed by Efron [1979] for independent observations but has been expanded to correlated data. Some important approaches considering dependent data can be found in Freedman [1984] and Efron and Tibishirani [1986], which adjusted a classic model-based method, where the model structure and independently and identically distributed (i.i.d.) sampling residues are used to generate bootstrap samples. Künsch [1989] proposed the blockwise bootstrap, a nonparametric procedure where blocks of successive observations are resampled, while Bühlmann [1997] proposed another nonparametric approach called sieve bootstrap, where an autoregressive process (AR) is adjusted to the data to approximate the real distribution. Härdle et al. [2004] extended the idea of model-based for semiparametric generalized models. Considering count correlated time series stand out the approaches for sieve bootstrap considering the integer-valued autoregressive (INAR) model, motivated by the fact that AR and INAR processes share the same autocorrelation function, see Cardinal et al. [1999] and Kim and Park [2006, 2008]. Jentsch and Weiss [2019] proposed a general INAR-type procedure where the INAR model is adjusted to the count time series, and a marginal distribution is assigned to the bootstrap innovations. Based on this, different bootstrap confidence intervals are studied in this thesis to calculate the interval estimation of the impact of atmospheric pollutants on human health for cases where the asymptotic properties are not easily established and/or not realistic in the context of the phenomena displayed in the data, that is, the data does not usually follow the required assumptions of the model.

The second main focus of this thesis is a phenomenon frequently ignored in the literature: abrupt observations, or high peaks, in the air pollutants. Extreme observations or outliers can strongly impact the parameter estimation, even in a small number. Consider a sample $\{y_t\}_{t=1}^{N}$, i.i.d., with N denoting the sample size, the mean is given by $\hat{\mu} = N^{-1}\sum_{t=1}^{T} y_t$, and assume that one replaces the observation y_N by the observation $y_N + \zeta$. The number of changes is small as only one out of the N original observations is replaced. However, the change is substantial, as ζ may diverge to infinity. So although the contamination fraction is small, the actual change in the

observations may be too significant.

According to Huber [1981], robustness indicates insensitivity to minor deviations from the assumptions. The foundations of this statistical approach can be found in Tukey [1960], Huber [1964], and Hampel [1968]. Robust models have the characteristic of fitting properly to most datasets. If the data has no abrupt observations, the robust method will behave approximately the same as the classic model. However, if the data is composed of a small percentage of outliers, the robust models will show results almost as good as the classic models applied to clean data. Usually, robust estimates depend on a dispersion function that varies more slowly in extreme values than the quadratic functions. For example, the ordinary least squares (OLS) method is not robust because its objective function, $\sum_{t=1}^{N} d_t^2$, can increase indefinitely in the presence of extreme values (d_t is the residual between the data point and the estimated fit).

Fox [1972] appears to be the first author to consider outliers within time series, proposing two types of outliers named Additive Outliers (AO), which affect only a single observation, and Innovation Outliers (IO) which affect succeeding observations. The presence of extreme values in correlated data can affect the autocorrelation structure of the series, affecting the statistical properties of the sample functions. Likewise, outliers can distort the estimated parameters of the model. Reisen et al. [2019] and Cotta et al. [2020] studied the influence of high peaks in air pollutant concentrations and robust estimation methods.

Due to the nonrobustness of the maximum likelihood estimator in generalized linear models (see Carroll and Welsh [1986], Künsch et al. [1989], Ruckstuhl and Welsh [1999], and others), Cantoni and Ronchetti [2001] is one of the most cited works related to robust methods for exponential family data. The authors proposed the Mallows' quasi-likelihood estimator (MQLE) based on the class of robust estimators introduced by Preisser and Qaqish [1999]. The MQLE is a natural generalization of quasi-likelihood functions and forms a class of M-estimators (Huber [1964]). Kitromilidou and Fokianos [2016] extended this method to count time series in the context of the log-linear Poisson model. They found that the MQLE behaved comparably to the classic log-linear model in the absence of perturbations, while in the presence of additive outliers, the MQLE provided more reliable results. Considering the previous discussion, the GLARMA model structure, based on GLM added by an ARMA process, and the nature of the data application, this thesis proposes a robust GLARMA Poisson model based on the MQLE estimator.

Numerical studies are carried out to evaluate the proposed methodologies, considering distinct scenarios and sample sizes once theoretical properties are unavailable for the general GLARMA model.

Real data analyses are performed considering environmental and epidemiological variables from two Brazilian regions, the metropolitan area of Belo Horizonte and the Great Vitória. Belo Horizonte is the capital of Minas Gerais state, with approximately 2,500,000 inhabitants, and the country's sixth most populated city. Studies concerning air pollution in this region are relatively rare, especially regarding air pollutant series with respiratory diseases. Therefore this

work also brings a practical contribution to the study of this critical problem that affects the health of the inhabitants of the region. The monthly numbers of Chronic obstructive pulmonary disease (COPD) cases in Belo Horizonte between 2007 and 2013 are studied. According to DATASUS, the department of information technology of the Brazilian public health system, each hour, three Brazilian citizens die due to this disease. Concentrations of the following air pollutants are examined: Particulate Material (PM_{10}), Nitrogen Monoxide (NO), Nitrogen Dioxide (NO_2), Carbon Monoxide (CO), and Ozone (O_3). The Great Vitória is a port and industrialized region, densely populated in the state of Espírito Santo, with approximately 1,900,000 inhabitants. Contrary to Belo Horizonte, studies concerning the effects of air pollution on health are pretty frequent, and the data related to pollutant concentrations are more structured. One of Brazil's most successful studies in this context, and the protagonist, is the AsmaVix, designed to quantify the influence of different air pollutants (gases and particulate matter) on asthma symptoms in children and adolescents. In this thesis, the impact of PM_{10} on the monthly deaths caused by respiratory diseases between 2011 to 2018 in Great Vitória is studied. This thesis reinforces the importance of evaluating these effects in the region and the limitation of the current classic methods. In both regions, epidemiological and environmental information was collected from DATASUS and IEMA, the State Environment and Water Resources Institute, respectively.

This thesis originated two papers, the first, denominated "*GLARMA model and Bootstrap approaches: An application to respiratory diseases and air pollutants*", submitted to *Applied Mathematical Modelling*, proposes to study bootstrap confidence intervals for the RR calculated from the GLARMA model. The second one, called "*Robust estimate for counting time series using GLARMA models*", still in compilation, proposes a robust approach for the GLARMA model.

The remainder of this manuscript is organized as follows. Chapter 2 presents the paper "*GLARMA model and Bootstrap approaches: An application to respiratory diseases and air pollutants*". Chapter 3 presents the second paper, "*Robust estimate for counting time series using GLARMA models*". Chapter 4 presents the final considerations and perspective of future works. This thesis is also composed of projects I worked on during my Ph.D. Due to this, Appendix A presents an application in the same field considering a longitudinal study conducted by the project AsmaVix, which evaluates air pollutants' impact on children's respiratory capacity in Great Vitória, ES, Brazil, between 2019 and 2020. Appendix B presents the paper called "*Generalized additive model for count time series: An application to quantify the impact of air pollutants on human health*", published in journal *Pesquisa Operacional*, which was a product of my Masters.

Chapter 2

GLARMA model and Bootstrap approaches: An application to respiratory diseases and air pollutants

Abstract The GLARMA model has been used in epidemiology to evaluate the impact of pollutants on health. These effects are quantified through the relative risk (RR) measure, which inference can be based on the asymptotic properties of the maximum likelihood estimator. However, for small series, this can be troublesome. This work studies different types of bootstrap confidence intervals (CI) for the RR. The simulation study revealed that the model parameter related to the data's autocorrelation could influence the intervals' coverage. Problems could arise when covariates present an autocorrelation structure. To solve this, using the VAR filter in the covariates is suggested.
Keywords: Time series of counts, INAR models, Integer-valued data, Respiratory diseases, Air pollution.

2.1 Introduction

Time series of counts are non-Gaussian processes composed of non-negative integers. These series can be found in different scientific areas, such as economics, medicine, agriculture, social and physical sciences, sports, among others. Some examples are the daily number of hospital admissions for a disease, the number of car accidents in a region, and the number of transactions of a given stock observed in one minute. In the last decades, various approaches emerged for modeling correlated count series. Recently, Davis et al. [2021] presented a review of these methodologies and discussed recent developments on this theme.

The impact of air pollution on population health has been the subject of many studies in the last decades, see Ostro et al. [1999], Schwartz [2000], Chen et al. [2010], Borhan et al. [2021] among others. The massive and continuous development of cities and communities leads

to urbanization and industrialization. Still, it can cause environmental and health problems as many activities generate residues that affect the inhabitants' quality of life (Barbera et al. [2010]). Epidemiological studies have consistently provided evidence of the association between daily pollutant concentration levels and hospital admissions, morbidity, and mortality, mainly caused by respiratory and cardiovascular diseases, see Souza et al. [2018] and Ispany et al. [2018] for references thereby.

Epidemiological data are frequently treated as time series of counts because they record the relative frequency of certain events in successive time intervals. In this context, many authors have been using the generalized additive model (GAM) [Hastie and Tibshirani, 1990] with Poisson marginal distribution to quantify the association between the effects of air pollution on health. Despite widespread use, care is required when the GAM is applied to time series. The GAM model assumes that errors are mutually independent and, therefore, can not capture the time dependency structure of the observations. One way to circumvent this is to use the generalized linear autoregressive moving average (GLARMA) model, proposed by Davis et al. [2003], which adds an ARMA structure to the generalized linear models (GLM) [Nelder and Wedderburn, 1972b]. The GLM is an extension of the Gaussian linear model where the distribution of the response variable belongs to the exponential family. Thus, the GLARMA model allows modelizing of correlated observations belonging to the exponential family. Although GLARMA presents some complexity in the estimation of general models, this methodology has been widely used in different applications see e.g., Rydberg and Shephard [2003], Jung et al. [2006], Jung and Tremayne [2011], Karami et al. [2017], Ballesteros-Cánovas et al. [2018], Peitzsch et al. [2021], among others. As addressed by Davis et al. [2021], the GLARMA family is one of the most flexible and easily fit count models that balance parameter and observation-driven models.

Besides the GLARMA model, different approaches to modeling count time series were also suggested in the literature. Zeger and Qaqish [1988] proposed a quasi-likelihood approach to time series regression, which was generalized by Benjamin et al. [2003] and called Generalized Autoregressive Moving Average models (GARMA). Heinen [2003] proposed the Autoregressive Conditional Poisson model (ACP) able to model overdispersion, and Fokianos and Tjosthein [2011] proposed log-linear models for time series. Recently, Camara et al. [2021] proposed a model using the GAM with autoregressive moving average terms (GAM-ARMA). This procedure can model the temporal correlation structure and estimate nonlinear associations between the covariates and the response variable. Following the Bayesian approach, Harvey and Fernandes [1989] used state-space models with conjugate prior distributions, where the counts are modeled as Poisson distribution, and Gamerman et al. [2013] proposed a family of non-Gaussian state-space models.

In epidemiology, one of the standard statistical measures used to quantify the relation between contaminant levels and adverse health effects is the relative risk (RR). Besides point estimation, epidemiologists are also interested in interval estimation for the RR. Although the asymptotic properties for the maximum likelihood estimators are not yet available for the general

case in GLARMA models, confidence intervals (CIs) for the relative risks are being computed based on the assumption of normality. It does not represent an issue for large sample sizes but can be troublesome for small samples. Still underexplored in studies related to air quality, bootstrap CIs can be calculated without any assumptions about data distribution. This procedure introduced by Efron [1979] initially for independent observations has been extended to more general situations. Freedman [1984], Efron and Tibishirani [1986], and Franke and Kreiss [1992] used a classic model-based approach, where bootstrap samples are generated using the estimated model structure and independently and identically distributed (i.i.d.) sampling residues. Härdle et al. [2004] used the classical idea of the residual-based bootstrap for semiparametric generalized models. A nonparametric bootstrap for the stationary process was proposed by Künsch [1989], the blockwise bootstrap, where blocks of consecutive observations are resampled. This method is robust against misspecification models, but the dependence between the blocks is neglected. Bühlmann [1997] proposed another nonparametric methodology, the sieve bootstrap, an approach to real-valued time series where an autoregressive process of order p, AR(p), is adjusted to the observations to approximate the real data distribution. In this procedure, the bootstrap samples are obtained by resampling from the centered residuals of the autoregressive process fitted.

Due to the popularity of the sieve bootstrap, and its simple implementation, in the last years, many authors proposed methodologies to adapt it to time series of counts. The most simple and probably obvious approximation is to centralize the counts and adjust the AR(p) process to the centralized observations. However, this approach leads to valid bootstrap approximation only for a limited number of cases (for details, see Jentsch and Weiss [2019]). In the last decades, the most prominent works that propose approaches for the sieve bootstrap considered the integer-valued autoregressive (INAR) time series, motivated by the fact that the AR and INAR process share the same autocorrelation function. In the INAR model, the dependent count observation, $\{Y_t\}_{t\in\mathbb{Z}}$, is expressed as a function of the p preceding values plus an innovation $\{\epsilon_t\}_{t\in\mathbb{Z}}$, with $\epsilon_t \sim G$, where G is a distribution with range $\mathbb{N}_0$. Cardinal et al. [1999] and Kim and Park [2006, 2008] proposed relevant alternatives to construct appropriate confidence intervals bootstrap for the INAR models. Recently, Jentsch and Weiss [2019] proposed an efficient method for bootstrapping INAR models. This paper addressed a critical literature review, highlighting the bootstrapping count time series challenge and the limitations of the techniques proposed so far considering INAR models. Jentsch and Weiss [2019] introduced a general INAR-type procedure, where the INAR(p) model is adjusted to the count time series, and a marginal distribution $\hat{G}$ is attributed to the bootstrap innovations. The authors proved the consistency of this procedure, assuming parametric and non-parametric distributions for the bootstrap innovations. In addition, they investigated the performance of this procedure by analyzing the coverage of 95% CIs for diverse statistics based on the observations. Their numerical simulation illustrates the superiority of the proposed method over the blockwise, sieve, and Markov bootstraps.

Based on the above discussion and the real data problem, this paper proposes to discuss

different approaches to compute confidence intervals for the relative risk using GLARMA models and evaluate their coverage rate. For this, three of those techniques cited previously were examined: the classic model-based approach, based on the work of Härdle et al. [2004], the well-known sieve bootstrap, and the INAR-type bootstrap. Confidence intervals assuming Gaussian distribution for the estimators were also calculated. Although Jentsch and Weiss [2019] have already verified the superiority of their methodology over sieve bootstrap for statistics based on the dependent count observations, here we aim to compare the performance of the investigated methods applied to the GLARMA model. An extensive simulation study was performed considering distinct sample sizes, types of covariates, and different autocorrelation structures, for a response variable autocorrelated and conditioned to the past, following a Poisson distribution. The objective was to verify if any change in the characteristics of the model's covariates or the complexity of the autocorrelation structure in the GLARMA model could impact the coverage or amplitude of the calculated intervals. A real time series was analyzed to illustrate the findings in the numerical simulations. We computed confidence intervals for the RR to quantify the impact of air pollutants on the number of chronic obstructive pulmonary disease cases in the metropolitan area of Belo Horizonte, Brazil. Apart from the motivation of the real problem, to the best of our knowledge, techniques using the INAR model in building bootstrap confidence intervals for measures based on parameters of the GLARMA model, such as the RR, are still not fulfilled in the literature. This paper aims to fill this gap.

This paper is organized as follows. Section 2 presents the GLARMA model and some essential properties regarding the model and the parameter estimation. Section 3 discusses the bootstrap for time series of counts, where three approaches are introduced. Simulation studies are performed in Section 4, considering different scenarios and sample sizes. A real data analysis is carried out in Section 5, and Section 6 presents the final considerations of the work.

2.2 The generalized linear autoregressive moving average model

The GLARMA models are a class of observation-driven state-space models. The state process consists of linear regression and observation-driven components comprising an autoregressive-moving average filter of past predictive residuals.

Let $\{Y_t\} := \{Y_t\}_{t\in\mathbb{Z}}$ be the observations and $\mathcal{F}_{t-1} = (Y^{(t-1)}, X^{(t)})$, where $Y^{(t-1)}$ is the past of the counting process and $X^{(t)}$ is the past and present of the regressor variables. Conditional on $\mathcal{F}_{t-1}$, the observations are independent and have a distribution in the exponential family with density

$$f(Y_t|W_t) = \exp\{Y_t W_t - a_t b(W_t) + c_t\}, \tag{2.1}$$

where $\{W_t\} := \{W_t\}_{t\in\mathbb{Z}}$ is the canonical parameter, called "state variable", which summarizes the information in $\mathcal{F}_{t-1}$, and a_t and c_t are sequences of constants. The conditional mean and variance of Y_t given $\mathcal{F}_{t-1}$ are respectively $\mu_t = E(Y_t|\mathcal{F}_{t-1}) = a_t\dot{b}(W_t)$ and $\sigma_t^2 = Var(Y_t|\mathcal{F}_{t-1}) = a_t\ddot{b}(W_t)$, where $\dot{b}$ and $\ddot{b}$ are the first and second derivatives with respect to its argument, see McCullagh and Nelder [1989].

Davis et al. [2003] considered that the specification of $\{W_t\}$ in (2.1) is given by

$$W_t = \mathbf{X}_t^T\beta + \sum_{i=1}^{\infty}\gamma_i e_{t-i}, \tag{2.2}$$

where $\mathbf{X}_t = (1, X_{1,t}, X_{2,t}, ..., X_{k,t})$ is the vector of covariates of dimension $k+1$, β is a $(k+1)$ x 1 vector of unknown coefficients to be estimated, and the standard residuals e_t are defined as

$$e_t = \frac{(Y_t - e^{W_t})}{e^{\lambda W_t}}, \tag{2.3}$$

for $\lambda \in (0, 1]$.

The infinite moving average weights γ_i in (2.2) can be specified in terms of an autoregressive-moving average (ARMA) filter:

$$\sum_{i=1}^{\infty}\gamma_i B^i = \frac{\theta(B)}{\phi(B)} - 1, \tag{2.4}$$

where the autoregressive and moving average components $\phi(B) = (1 - \phi_1 B - ... - \phi_p B^p)$ and $\theta(B) = (1+\theta_1 B+...+\theta_q B^q)$ are polynomials with roots outside the unit circle, γ is the parameter vector formed by ϕ's and θ's, and B is the backshift operator of the form $B^k(Z_t) = Z_{t-k}$. Defining $W_t = X_t^T\beta + Z_t$, where $Z_t = \sum_{i=1}^{\infty}\gamma_i e_{t-i}$, for $t \leq 0$ and $e_t = 0, Z_t = 0$, and for $t > 0$ the process Z_t is computed according to the following ARMA-like recursions

$$Z_t = \phi_1(Z_{t-1} + e_{t-1}) + ... + \phi_p(Z_{t-p} + e_{t-p}) + \theta_1 e_{t-1} + ... + \theta_q e_{t-q}. \tag{2.5}$$

Remark 1: Let $e_s = 0$ and $Y_s = 0$ for $s \leq 0$. Under these conditions, it is possible to show that e_t form a martingale difference sequence with zero mean, variance given by $E(\mu_t^{1-2\lambda}), t \geq 1$, and $\text{Cov}(e_t, e_s) = 0$ for $t \neq s$. In addition, for any $\lambda \in (0, 1]$, $E(W_t) = x_t^T\beta, Var(W_t) = \sum_{i=1}^{\infty}\gamma_i^2 E(\mu_{t-i}^{1-2\lambda})$, and, for $s = t+h, h > 0$, $\text{Cov}(W_t, W_{t+h}) = \sum_{i=1}^{\infty}\gamma_i\gamma_{i+h}E(\mu_{t-i}^{1-2\lambda})$. For more details see Section 2.2 in Davis et al. [2003].

Remark 2: For $\lambda = 1$, Davis et al. [2003] showed that considering the simplest model, where $Y_t|\mathcal{F}_{t-1} \sim \text{Poi}(\mu_t)$, and $W_t = \beta_0 + \gamma(Y_{t-1} - e^{W_{t-1}})e^{-\lambda W_{t-1}}$, assuming $p = 0$ and $q = 1$, the process $\{W_t\}$ has a unique stationary distribution and is uniformly ergodic. For $\frac{1}{2} \leq \lambda \leq 1$, $\{W_t\}$ is bounded in probability and therefore has a stationary distribution, yet the uniqueness of this distribution is currently unknown.

Define $\delta = (\beta_0, \beta_1, ..., \beta_k, \phi_1, ..., \phi_p, \theta_1, ..., \theta_q)^T$ as the parameter vector of the model (2). Let $y_1, y_2, ..., y_n$ be a sample from the process $\{Y_t\}$, the log-likelihood of $\{Y_t\}$ conditional to $\mathcal{F}_{t-1}$ is given by

$$L(\delta) = \sum_{t=1}^{n} \log f(y_t|W_t(\delta)). \tag{2.6}$$

For the exponential family, the log-likelihood becomes

$$L(\delta) = \textstyle\sum_{t=1}^{n} \{y_t W_t(\delta) - a_t b(W_t(\delta)) + c_t\}.$$

As a particular case, for $Y_t|\mathcal{F}_{t-1} \sim \text{Poi}(\mu_t)$, where $\mu_t = e^{W_t}$, the log-likelihood is given by

$$L(\delta) = \sum_{t=1}^{n} \left\{ y_t W_t(\delta) - e^{W_t(\delta)} - \log(y_t!) \right\}. \tag{2.7}$$

The log-likelihood function can be maximized using procedures such as Newton-Raphson iteration and Fisher scoring approximation. Defining the first and second derivatives of the log-likelihood by $d(\delta) = \partial L(\delta)/\partial\delta$ and $D_{NR}(\delta) = \partial^2 L(\delta)/\partial\delta\partial\delta^T$, we have

$$d(\delta) = \sum_{t=1}^{n} \left(y_t - a_t \dot{b}(W_t) \right) \frac{\partial W_t}{\partial \delta} \tag{2.8}$$

and

$$D_{NR}(\delta) = \sum_{t=1}^{n} \left(y_t - a_t \dot{b}(W_t) \right) \frac{\partial^2 W_t}{\partial\delta\partial\delta^T} - \sum_{t=1}^{n} a_t \ddot{b}(W_t) \frac{\partial W_t}{\partial \delta} \frac{\partial W_t}{\partial \delta^T}. \tag{2.9}$$

As at the true parameter of δ, $E(y_t - a_t\dot{b}(W_t))|\mathcal{F}_{t-1}) = 0$, the expected value of the first summation in (2.9) is zero, which motivates the Fisher Scoring approximation:

$$D_{FS}(\delta) = -\sum_{t=1}^{n} a_t \ddot{b}(W_t) \frac{\partial W_t}{\partial \delta} \frac{\partial W_t}{\partial \delta^T}. \tag{2.10}$$

Note that $E(D_{NR}(\delta)) = E(D_{FS}(\delta))$, however, these expectations can not be calculated in closed form. Thus, the Newton-Raphson (using D_{NR}) and the Fisher scoring (using D_{FS}) methods are used to maximize the log-likelihood function.

Davis et al. [2021] discussed that the consistency and asymptotic properties of the maximum likelihood $\hat{\delta}$ were proven only for the simplest model cited in Remark 2. For this special case, stationarity and ergodicity were established rigorously. In general, for inference purposes, it is assumed that the central limit theorem holds:

$$\hat{\delta} \approx N(\delta, \hat{\Omega}),$$

where the covariance matrix of the estimators $\hat{\Omega}$ is giving by $-[D_{NR}(\delta)]^{-1}$ using the Newton-Raphson iteration and $-[D_{FS}(\delta)]^{-1}$ using the Fisher scoring approximation.

In the epidemiology context, the impact of air pollutants on human health is evaluated by relative risk (RR). The RR of a variable $X_i = X_{i,t}$ is the change in the expected count of the

response variable per ζ-unit change in the X_i, keeping the other covariates fixed. Baxter et al. [1997] present its mathematical representation

$$\frac{E\left(Y|X_i=\zeta, X_j=x_j, i\neq j\right)}{E\left(Y|X_i=0, X_j=x_j, i\neq j\right)}.$$

For Poisson regression, the RR is given by

$$\mathrm{RR}_{X_i}(\zeta)=\exp(\beta_i\zeta) \tag{2.11}$$

and its approximate confidence interval (CI) at an α significance level in the GLARMA with Poisson marginal distribution is

$$\widehat{\mathrm{RR}}_{X_i}(\zeta)=\exp\left\{\zeta\left(\widehat{\beta}_i-z_{\alpha/2}\,\mathrm{se}(\widehat{\beta}_i);\widehat{\beta}_i+z_{\alpha/2}\,\mathrm{se}(\widehat{\beta}_i)\right)\right\}, \tag{2.12}$$

where $\widehat{\beta}_i$ is the conditional maximum likelihood estimator $\widehat{\beta}_{i,n}$ of β_i, $\mathrm{se}(\widehat{\beta}_i)$ is the estimated standard deviation of $\widehat{\beta}_i$, and $z_{\alpha/2}$ denotes the $(1-\alpha/2)$-quantile of the standard normal distribution.

2.3 Bootstrap for count time series

Initially proposed by Efron [1979] for independent variables, the bootstrap is a resampling method that attributes measures of accuracy to statistical estimates. It is a computer-based procedure that approximates the theoretical distribution by the empirical distribution of a finite sample of observations.

Let $\mathbf{X}=(X_1,X_2,...,X_n)$ and $\mathbf{x}=(x_1,x_2,...,x_n)$ denote the random sample and its observed realizations, respectively, from a distribution F. A bootstrap sample $\mathbf{x}^*=(x_1^*,x_2^*,...,x_n^*)$ is obtained by randomly sampling n times, with replacement, from the original data points $x_1,x_2,...,x_n$. That is, the bootstrap sample is drawn from the empirical distribution F^*. For any statistic u computed from the original sample data, it is possible to define a statistic u^* by the same formula but calculated using the resampled data.

Ignoring the temporal correlation of the time series can impact the estimations. In this context, many approaches have been proposed standing out the classic model-based (Freedman [1984], Efron and Tibishirani [1986], and Franke and Kreiss [1992]), the blockwise (Künsch [1989]) and the sieve (Bühlmann [1997]) bootstraps. In the residual model-based approach, the bootstrap replications are based on i.i.d. resampling of residuals. The blockwise and sieve bootstraps are nonparametric procedures, robust against misspecification models. In the first case, blocks of consecutive observations are resampled, and in the sieve bootstrap, the autoregressive process of order p, AR(p), is adjusted to the observations to approximate the actual data distribution.

However, the challenge arises when applying bootstrap to non-Gaussian responses. In the last decades, the bootstrap involving time series of counts has been studied by several authors. Härdle et al. [2004] proposed a methodology derived from the classic model-based for semiparametric generalized models. Cardinal et al. [1999], Kim and Park [2006, 2008], and Jentsch and Weiss [2019] proposed approaches to adapt the sieve bootstrap considering integer-valued autoregressive (INAR) time series.

In this section, three bootstrap procedures for count time series are discussed: The residual model-based approach of Härdle et al. [2004], an adaptation of sieve bootstrap for count observations, and the proposal of Jentsch and Weiss [2019] for INAR(p) process.

2.3.1 Classic model-based bootstrap

This procedure is used in regression problems in which it is assumed that the model is correctly specified and the error terms in the model are independent and identically distributed. The basic idea of the bootstrap with parametric assumptions is to obtain residuals using the estimated parametric model and then generate bootstrap samples using the estimated model structure and i.i.d. resampling of residuals. Assume that Y_t given the past history, $\mathcal{F}_{t-1}$, has a distribution in the exponential family with $\mu_t = g\left\{X_t^T\beta + Z_t\right\}, t \in \mathbb{Z}$, where g is a known link function and the component Z_t is defined in equation (2.5). The bootstrap procedure works as follows:

1. Compute the residuals $\hat{e}_t = (Y_t - \hat{\mu}_t)/\hat{\mu}_t^{\lambda}$;
2. Resample the estimated residuals with replacement generating the bootstrap residual samples $(e_1^*, ..., e_n^*)$.
3. Generate bootstrap observations $Y_1^*, ..., Y_n^*$ according to

$$Y_t^* = \hat{\mu}_t + e_t^*\hat{\mu}_t^{\lambda}.$$

This procedure is model-based, then misspecification problems can be observed, e.g., biased parameters and inconsistent standard errors. The sieve bootstrap, presented below, follows the same strategy of fitting a parametric model and resampling the residuals but approximating an infinite-dimensional non-parametric model by a sequence of finite-dimensional parametric models.

2.3.2 Sieve bootstrap

Let $\{Y_t\}_{t\in\mathbb{Z}}$ be a real-valued stationary process and denote $Y_1, Y_2, ..., Y_n$ a sample from this process. The basic idea of the sieve bootstrap (Bühlmann [1997]) is to adjust an autoregressive process of order p to the data:

$$Y_t = \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + ... + \alpha_p Y_{t-p} + \xi_t, \quad t \in \mathbb{Z}, \tag{2.13}$$

with increasing order p as the sample size n increases. The estimated coefficients $\hat{\alpha}_1, \hat{\alpha}_2, ..., \hat{\alpha}_n$ are then used to compute the residuals $\hat{\xi}_t$. These residuals are centered, $\tilde{\xi}_t = \hat{\xi}_t - (n-p)\sum_{t=p+1}^{n}\hat{\xi}_t$, and the bootstrap sample $Y_1^*, Y_2^*, ..., Y_n^*$ is constructed according to the recursion

$$Y_t^* = \hat{\alpha}_1 Y_{t-1}^* + \hat{\alpha}_2 Y_{t-2}^* + ... + \hat{\alpha}_p Y_{t-p}^* + \tilde{\xi}_t^*, \quad t \in \mathbb{Z}, \tag{2.14}$$

where $\tilde{\xi}_t^*$ is a random sample with replacement of the centered residuals.

The sieve bootstrap was initially proposed for real-valued data, but a simple approximation can be performed by ignoring the discrete nature of the process. Jentsch and Weiss [2019] cited this approach as follows

1. Compute the centered observations $X_t = Y_t - \bar{Y}$, where $\bar{Y} = 1/n\sum_{t=1}^{n} Y_t$.
2. Fit an AR(p) to the centered data $X_1, ..., X_n$

   $$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + ... + \alpha_p X_{t-p} + \xi_t,$$

   where the estimated AR coefficients $\hat{\alpha}_1, ..., \hat{\alpha}_p$ can be obtained, e.g., from Yule-Walker estimates.
3. Compute the estimated residuals $\hat{\xi}_t = X_t - \sum_{i=1}^{n}\hat{\alpha}_i X_{t-i}$, and center them obtaining $\tilde{\xi}_t$ as presented previously.
4. Generate bootstrap observations $X_1^*, ..., X_n^*$ according to

   $$X_t^* = \hat{\alpha}_1 X_{t-1}^* + \hat{\alpha}_2 X_{t-2}^* + ... + \hat{\alpha}_p X_{t-p}^* + \tilde{\xi}_t^*,$$

   where $\tilde{\xi}_t^*$ are randomly and uniformly resampling from the centered residuals.
5. The AR bootstrap sample of the original process $\{Y_t\}_{t\in\mathbb{Z}}$ can be calculated as $Y_t^* = X_t^* + \bar{Y}$.

Kreiss et al. [2011] showed that the sieve bootstrap captures the autocovariance structure of a process and will always be consistent for the sample mean and any statistic that depends exclusively on the autocovariance structure of the process under mild conditions.

2.3.3 INAR-type bootstrap

Jentsch and Weiss [2019] proposed the procedure based on the INAR model introduced by McKenzie [1985] and Al-Osh and Alzaid [1988], and extended by Alzaid and Al-Osh [1990] and Du and Li [1991].

The integer-valued autoregressive process of order p, denoted by INAR(p), express the value of the variable of interest at time t as a function of the p preceding values and an innovation in the following way

$$Y_t = \alpha_1 \circ Y_{t-1} + \alpha_2 \circ Y_{t-2} + ... + \alpha_p \circ Y_{t-p} + \epsilon_t, \quad t \in \mathbb{Z}, \tag{2.15}$$

where $\{Y_t\}_{t\in\mathbb{Z}}$, is a sequence of non-negative integer-valued variables, ϵ_t is an i.i.d non-negative integer-valued random variable with finite mean μ_ϵ and variance σ^2_ϵ, and $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_p)^T \in (0,1)^p$, such that $\sum_{i=1}^p \alpha_i < 1$. The operator "$\circ$" in (2.15) is called *binomial thinning operator* and $\alpha_i \circ Y_{t-i} \sim \text{Bin}(Y_{t-i}, \alpha_i)$, where $\text{Bin}(n, \pi)$ denotes the binomial distribution with parameters n and π. Due to the random thinning operator, the INAR models are non-linear, contrary to the sieve bootstrap, which belongs to the linear time series process class. It is important to observe that the INAR innovations ϵ_t and the sieve errors e_t were distinguished here. As these procedures share the same autocorrelation function, the INAR coefficients can be estimated by techniques from classical time series analysis such as Yule-Walker, Least-Squares, or Maximum-Likelihood estimators.

Du and Li [1991] proved that if $\sigma^2_\epsilon = \text{Var}(\epsilon_t) < \infty$ the autocorrelation of an INAR(p) process corresponds to that of an autoregressive process of order p, AR(p). However, differently from the sieve bootstrap, the procedure proposed by Jentsch and Weiss [2019] does not explicitly use the residuals from the fitted model. Let Y_t be a non-negative integer-valued time series. The general INAR bootstrap scheme is defined as follows:

1. Fit an INAR(p) process as in equation (2.15) to obtain the estimates of $\alpha_1, ..., \alpha_p$ for the INAR coefficients;

2. Specify the marginal distribution $\hat{G}$ for the innovations ϵ_t;

3. Generate the bootstrap samples according to

$$Y^*_t = \hat{\alpha}_1 \circ^* Y^*_{t-1} + ... + \hat{\alpha}_p \circ^* Y^*_{t-p} + \epsilon^*_t, \tag{2.16}$$

   where "$\circ^*$" denotes the mutually independent bootstrap binomial thinning operations and (ϵ^*_t) are i.i.d. random variables following the distribution $\hat{G}$.

The estimation of the parameters $\hat{\alpha}_1, ..., \hat{\alpha}_p$ and the choice of the distribution $\hat{G}$, in steps 1 and 2, respectively, can be calculated following parametric and semi-parametric approaches.

Here the parametric perspective was considered. In this case, under some assumptions related to the marginal distribution G of the innovation process (see Jentsch and Weiss [2019]), the bootstrap innovations $\epsilon_1^*, ..., \epsilon_n^*$ can be easily generated from $G_{\hat{\boldsymbol{\theta}}}$ following the steps above.

2.3.4 Bootstrap confidence intervals

Bootstrap confidence intervals can be computed using simple percentiles, bias-corrected percentile limits, bias-corrected and accelerated percentile (BCa), Student's t method, among other proposals. Here we will focus on the approach used by Jentsch and Weiss [2019] in their simulations. To compute the bootstrap confidence interval for a parameter θ, these authors first calculated a centering measure cent($\hat{\theta}^*$) and the centered bootstrap estimates $\hat{\theta}^*_{cent} := \hat{\theta}^* -$ cent($\hat{\theta}^*$). Then the bootstrap confidence interval was calculated using the $(1-\alpha/2)$- and $\alpha/2$-quantiles from $\hat{\theta}^*_{cent}$ as follows:

$$\left[\hat{\theta} - q_{1-\alpha/2}\left(\hat{\theta}^*_{cent}\right); \hat{\theta} - q_{\alpha/2}\left(\hat{\theta}^*_{cent}\right)\right], \tag{2.17}$$

where $\hat{\theta}$ is the estimate obtained from the sample.

As the main objective of this paper is to calculate bootstrap confidence intervals for the RR of the variable $X_i, i = 1, ..., k$, the CIs were constructed as follows:

$$\widehat{\text{RR}}_{X_i}(\zeta) = \exp\left\{\zeta\left(\hat{\beta}_i - q_{1-\alpha/2}\left(\hat{\beta}_{i\,cent}^{\,*}\right); \hat{\beta}_i - q_{\alpha/2}\left(\hat{\beta}_{i\,cent}^{\,*}\right)\right)\right\}, \tag{2.18}$$

where $\hat{\beta}_i$, is the i-th estimated coefficient, $\hat{\beta}_i^{\,*}$ is the bootstrap estimate, and $\hat{\beta}_{i\,cent}^{\,*} = \left(\hat{\beta}_i^{\,*} - \overline{\hat{\beta}_i^{\,*}}\right)$, with $\overline{\hat{\beta}_i^{\,*}}$ being the mean of $\hat{\beta}_i^{\,*}$. The interquartile variation of X_i is given by ζ, and α is the significance level.

2.4 Simulation study

A simulation study was conducted to evaluate and compare the performance of the bootstrap approaches presented in Section 2.3. As many studies use asymptotic confidence intervals for the RR based on the Gaussian distribution, this interval was also considered for comparison purposes. This analysis focused on the confidence intervals for the relative risk

calculated from the GLARMA(1,0) Poisson model, given by

$$Y_t|\mathcal{F}_{t-1} \sim \text{Poisson}(\mu_t)$$
$$\ln(\mu_t) = \beta_0 + \boldsymbol{\beta}\boldsymbol{X}_t + Z_t,$$

where Z_t is given by

$$Z_t = \phi[Z_{t-1} + (Y_{t-1} - \mu)\mu^{-\lambda}]. \tag{2.19}$$

2.4.1 Large samples

For this simulation study, three scenarios were considered:

- S1: $\boldsymbol{\beta} = (\beta_0, \beta_1)$ and $\boldsymbol{X}_t = (1, X_{1,t})^T$, where $X_{1,t} = t/n$.
- S2: $\boldsymbol{\beta} = (\beta_0, \beta_1)$ and $\boldsymbol{X}_t = (1, X_{1,t})^T$, where $X_{1,t}$ is an independent random vector in time.
- S3: $\boldsymbol{\beta} = (\beta_0, \beta_1)$ and $\boldsymbol{X}_t = (1, X_{1,t})^T$, where $X_{1,t} \sim$ AR(1), with the autoregressive parameter assuming values 0.2, 0.5 and 0.8.

The considered sample size n was equal to 1000. The values of ϕ in Equation (2.19) were fixed at 0.2, 0.4, and 0.6, and the nominal level for the confidence intervals was fixed at 95%. The parameter λ assumed the value 1.0. Although Davis et al. [2003] showed that, for $\lambda = 1$, only in the simplest model, the process $\{W_t\}$ has a stationary and ergodic distribution, the numerical simulations revealed that, even for complex models, this value of λ provides better estimates. For all scenarios, $\beta_0 = \beta_1 = 1.0$.

The classic model-based bootstrap considers the three steps presented in subsection 2.3.1 to construct the confidence intervals. For sieve and INAR(1) cases, the steps presented in subsections 2.3.2 and 2.3.3 were applied to the count time series Y_t, then the GLARMA Poisson model was fitted considering the bootstrap samples as the response variables. In addition, for the INAR(1) bootstrap, it was assumed an INAR process having Poisson-distributed innovations with $\epsilon_t \sim$ Poisson(λ), and $\hat{\lambda} = \bar{Y}(1 - \hat{\alpha})$, where $\hat{\alpha}$ is obtained by Yule-Walker estimation.

The Monte Carlo simulations were repeated 500 times with 500 bootstrap replications. The asymptotic confidence interval was estimated as in equation (2.12). All the codes were written in the R language and are available from the authors upon request.

- Scenario 1

Table 1a presents the mean and standard deviation of the 500 Monte Carlo estimates of parameters β_0, β_1 and ϕ in scenario 1 (S1). For β_0 and β_1 parameters, the mean of the estimates was close

to the real values, especially when $\phi = 0.2$ or 0.4. The standard deviation, on the other side, shows a consistent increase with the value of ϕ. Parameter ϕ is better estimated for small values. Table 1b presents the 95% confidence intervals for the RR of the covariate $X_{1,t}$. The values in square brackets are the average lower and upper limits of the calculated intervals. The classic and sieve bootstrap procedures presented lower coverage rates which decrease as the parameter ϕ value increases, although, for $\phi = 0.2$, the sieve bootstrap showed a coverage rate close to 0.95. The INAR(1) bootstrap and the asymptotic confidence intervals presented a coverage rate of approximately 95% for all values of ϕ.

Scenario 1 studied the impact of the same covariate considered in Davis et al. [2003], although the authors only evaluated cases where the time correlation is a moving average process of order 1. Real data sets commonly present an autoregressive autocorrelation structure. In this case, S1 showed that even when the time correlation structure is complex (e.g., $\phi = 0.6$), for deterministic covariates ($X_{1,t} = t/n$), the asymptotic theory and the INAR(1) bootstrap presented coverage rates close to the nominal level of the confidence intervals.

Table 1a: Parameter estimation (S1)

	ϕ=**0.2**		ϕ=**0.4**		ϕ=**0.6**	
	Mean	Sd	Mean	Sd	Mean	Sd
β_0	0.992	0.042	0.990	0.052	0.932	0.068
β_1	1.011	0.066	1.012	0.082	1.078	0.102
ϕ	0.200	0.030	0.392	0.026	0.553	0.102

Prepared by the author

Table 1b: Coverage and the average lower and upper limits of the 95% confidence intervals for the RR=exp($\zeta\beta_1$) (S1)

RR = 1.28	$\phi = 0.2$	$\phi = 0.4$	$\phi = 0.6$
Classic bootstrap	0.884 [1.218;1.370]	0.802 [1.226;1.385]	0.729 [1.222;1.401]
Sieve bootstrap	0.927 [1.217;1.368]	0.786 [1.227;1.378]	0.570 [1.265;1.419]
INAR(1) bootstrap	0.958 [1.196;1.386]	0.944 [1.193;1.427]	0.930 [1.161;1.435]
Asymptotic	0.962 [1.199;1.385]	0.948 [1.181;1.428]	0.934 [1.154;1.484]

Prepared by the author

- Scenario 2

In the parameter estimation presented in Table 2a, the mean of the estimates was close to the real values of all parameters, except when $\phi = 0.6$. It can also be seen that the standard deviations were much affected by the increase of ϕ. Table 2b presents the 95% confidence intervals for the RR in scenario 2 regarding the covariate $X_{1,t}$. Table 2b shows that for the classic bootstrap, the coverage rate was close to 1 for all values of ϕ, which means that almost 100% of the intervals contain the true relative risk value. Regarding the sieve bootstrap, for $\phi = 0.2$ and 0.4, the coverage rate was also close to 1. Meanwhile, for $\phi = 0.6$, the coverage rate decreased to 0.849. The INAR(1) bootstrap and the asymptotic intervals had similar performance, with coverages

close to 0.95 for $\phi = 0.2$ and 0.4 and considerably below the nominal level for 0.6. It should be pointed out that the INAR(1) bootstrap always presented coverages closer to 0.95 than the asymptotic interval, even for the $\phi = 0.6$ case.

Scenarios 1 and 2 showed that the coverage of INAR bootstrap and asymptotic approach were close to the nominal level for $\phi = 0.2$ and 0.4, cases where β_1 was appropriately estimated. In Tables 1a and 2a, the mean of this parameter was close to the true values, and although the standard deviation grew up in Table 2a, the coverage in scenario 2 was not impacted. However, for $\phi = 0.6$, the β_1 estimates were terrible, mainly in S2, and as the RR depends on β_1, the interval coverage was also impacted. Finally, it is essential to observe that the coverage intervals are unsuitable for classic and sieve bootstraps even when the β_1 estimates are reasonable.

Table 2a: Parameter estimation (S2)

	ϕ=**0.2**		ϕ=**0.4**		ϕ=**0.6**	
	Mean	Sd	Mean	Sd	Mean	Sd
β_0	0.995	0.026	1.010	0.285	1.201	1.046
β_1	1.002	0.015	0.992	0.139	0.951	0.299
ϕ	0.199	0.017	0.380	0.077	0.430	0.192

Prepared by the author

Table 2b: Coverage and the average lower and upper limits of the 95% confidence intervals for the RR=exp($\zeta\beta_1$) (S2)

RR = 1.94	$\phi = 0.2$	$\phi = 0.4$	$\phi = 0.6$
Classic bootstrap	0.990 [1.827;2.096]	1.000 [1.802;2.086]	0.992 [1.778;2.137]
Sieve bootstrap	0.986 [1.834;2.088]	0.989 [1.821;2.093]	0.849 [1.908;2.067]
INAR(1) bootstrap	0.953 [1.863;2.038]	0.948 [1.875;2.051]	0.732 [1.989;2.155]
Asymptotic	0.968 [1.860;2.041]	0.940 [1.887;2.049]	0.638 [2.005;2.143]

Prepared by the author

- Scenario 3

In epidemiology, it is common for air pollutants to present temporal correlation. To simulate this behavior, in scenario 3, the covariate $X_{1,t}$ followed an autoregressive process of order 1:

$$\ln(\mu_t) = \beta_0 + \beta_1 X_{1,t} + Z_t,$$

where $X_{1,t}$ is an AR(1) process with autoregressive parameter φ and Z_t is defined by equation (2.5). To evaluate the time structure's impact, the covariate's autoregressive parameter (φ) assumed values equal to $0.2, 0.5$ and 0.8. Table 3a shows the estimation of the parameters for $\varphi = 0.2$. Focusing on β_1, the mean of this parameter estimate is close to the real value for $\phi = 0.2$, while the estimation becomes worse for $\phi = 0.4$ and 0.6. In addition, it is possible to observe an increase in the standard deviation. Table 3b presents the 95% CI for the RR of the covariate $X_{1,t}$. The coverage rate for the classic model-based bootstrap was approximately 1 for all values of ϕ. The confidence intervals of the asymptotic approach, sieve, and INAR(1)

bootstraps presented a comparable performance, with the coverage decreasing for $\phi = 0.4$. For $\phi = 0.6$, all the methods presented a poor performance.

Table 3a: Parameter estimation (S3) $\varphi = 0.2$

	ϕ=**0.2**		ϕ=0.4		ϕ=0.6	
	Mean	Sd	Mean	Sd	Mean	Sd
β_0	0.978	0.038	0.811	0.275	1.246	0.795
β_1	1.020	0.031	1.134	0.147	0.907	0.300
ϕ	0.138	0.021	0.234	0.072	0.355	0.235

Prepared by the author

Table 3b: Coverage and the average lower and upper limits of the 95% confidence intervals for the RR=exp($\zeta\beta_1$) (S3) $\varphi = 0.2$

RR = 4.03	$\phi = 0.2$	$\phi = 0.4$	$\phi = 0.6$
Classic bootstrap	1.000 [3.428;4.312]	1.000 [3.419;4.528]	0.998 [3.386;4.440]
Sieve bootstrap	0.974 [3.960;5.152]	0.892 [3.815;5.052]	0.378 [3.529;4.389]
INAR(1) bootstrap	0.957 [3.850;4.171]	0.921 [3.877;4.167]	0.272 [3.439;3.671]
Asymptotic	0.956 [3.894;4.193]	0.908 [3.910;4.169]	0.182 [3.475;3.634]

Prepared by the author

Table 4a shows the estimation of the parameters for $\varphi = 0.5$. The mean of the β_1 estimate is still close to the true value for $\phi = 0.2$, but in comparison to Table 3a at the same value of ϕ, there is an increase in the standard deviation. For $\phi = 0.4$ and 0.6, the β_1 estimates worsened, and the standard deviation grew as the value of ϕ increased. Table 4b presents the coverage rate for the relative risk. A similar performance was observed in the classic and sieve bootstraps, with the coverage rate equal to 100% for $\phi = 0.2$, followed by a drop in coverage as ϕ grows. In addition, both procedures presented large confidence intervals. The confidence intervals obtained by INAR(1) and asymptotic approaches presented a coverage rate of approximately 95% for $\phi = 0.2$. For $\phi = 0.4$, these rates demonstrated decreases, while the INAR(1) bootstrap had the highest coverage. Once again, the ranges are very far from the nominal level when $\phi = 0.6$.

Table 4a: Parameter estimation (S3) $\varphi = 0.5$

	ϕ=**0.2**		ϕ=**0.4**		ϕ=**0.6**	
	Mean	Sd	Mean	Sd	Mean	Sd
β_0	1.028	0.178	1.101	0.326	1.970	2.195
β_1	0.974	0.159	0.936	0.216	0.657	0.696
ϕ	0.190	0.041	0.306	0.157	0.148	0.167

Prepared by the author

Table 5a shows the parameters' estimation when the covariate's autoregressive parameter (X_t) is $\varphi = 0.8$. Even for the smallest value of ϕ considered, the mean of β_1 estimate is not good, and the standard deviation is considerably large. For $\phi = 0.4$ and 0.6, the means of the β_1 parameters became terrible, and the standard deviation increased even more. Table 5b presents the coverage of the confidence intervals for the RR. For $\phi = 0.2$, the classic model-based and

Table 4b: Coverage and the average lower and upper limits of the 95% confidence intervals for the RR=exp($\zeta\beta_1$) (S3) $\varphi = 0.5$

RR = 4.7	$\phi = 0.2$	$\phi = 0.4$	$\phi = 0.6$
Classic bootstrap	1.000 [4.109;5.149]	0.977 [4.078;6.545]	0.802 [1.695;6.159]
Sieve bootstrap	1.000 [4.067;5.718]	0.941 [3.934;4.844]	0.382 [3.771;4.571]
INAR(1) bootstrap	0.921 [4.583;4.938]	0.767 [4.371;4.708]	0.145 [4.248;4.554]
Asymptotic	0.944 [4.574;4.947]	0.751 [4.369;4.711]	0.111 [4.316;4.550]

Prepared by the author

sieve bootstraps showed a range approximately equal to 1, while this rate was close to 0.93 for the INAR bootstrap and equal to 0.895 for the asymptotic approach. All methodologies in the study had a significant decline in the coverage rate for $\phi = 0.4$ and 0.6, being the CIs obtained by the asymptotic approach and the INAR(1) bootstrap being the most affected.

Table 5a: Parameter estimation (S3) $\varphi = 0.8$

	ϕ**=0.2**		ϕ**=0.4**		ϕ**=0.6**	
	Mean	Sd	Mean	Sd	Mean	Sd
β_0	1.127	1.038	2.048	1.509	6.593	3.359
β_1	0.962	0.257	0.448	0.734	-0.478	0.514
ϕ	0.175	0.067	0.052	0.098	0.001	0.008

Prepared by the author

Table 5b: Coverage and the average lower and upper limits of the 95% confidence intervals for the RR=exp($\zeta\beta_1$) (S3) $\varphi = 0.8$

RR = 8.3	$\phi = 0.2$	$\phi = 0.4$	$\phi = 0.6$
Classic bootstrap	1.000 [6.043;11.948]	0.791 [3.379;13.769]	0.227 [0.626; 5.914]
Sieve bootstrap	0.998 [6.899;9.078]	0.426 [4.657;5.989]	0.206 [2.730;3.145]
INAR(1) bootstrap	0.927 [8.132;8.721]	0.176 [5.444;5.773]	0.000 [1.170;1.212]
Asymptotic	0.895 [8.162;8.688]	0.154 [5.276;5.560]	0.000 [1.172;1.211]

Prepared by the author

The comparison between scenarios 2 and 3 indicates that time correlation in the covariates can impact the coverage rate of the confidence intervals; as the autoregressive structure becomes more complex, the interval coverage gets smaller. Tables 3a, 4a and 5a showed that the values of ϕ strongly impact the parameter estimation, and this effect gets worse as this autoregressive parameter increases in the direction of the nonstationarity region, either in the covariate as in the Z_t component. It is possible to verify that for any $\lambda \in (0, 1]$, $\text{Var}(W_t) = \sum_{i=1}^{\infty} \gamma^2 \text{E}(\mu_{t-i}^{1-2\lambda})$, where the covariate $X_{1,t}$ is an independent random vector in time. However, for $X_{1,t} \sim$ AR(1), the variability of the state process $\{W_t\}$ increases, which inflates the model estimates, directly impacting the coverage rates of the RR.

In addition to the empirical investigations discussed here, scenarios with more complex model structures, such as bivariate time series, were also considered. As expected, the coverage rate was not satisfactory. Thus, in practical situations where covariates are time series, the authors suggest the procedure proposed by Souza et al. [2018], which must be used before performing

the bootstrap approaches discussed here. This is addressed in Section 5, in the real data analysis, where the covariates follow a vector of time series data.

2.4.2 Small samples and ARMA covariate

This work also studied another scenario considering small samples for the GLARMA(1,0) Poisson model. In this case, the sample size was equal to 50, and the single covariate $X_{1,t}$ was an ARMA(p, q) process:

$$Y_t|\mathcal{F}_{t-1} \sim \text{Poisson}(\mu_t)$$
$$\ln(\mu_t) = \beta_0 + \beta_1 X_{1,t} + Z_t,$$

where $X_{1,t}$ is an ARMA process with autoregressive and moving average parameters φ and θ, and Z_t is defined by equation (2.19). Three different ARMA processes were considered. They were chosen due to their temporal structure, which is similar to some atmospheric pollutants in real data:

- (1) $X_{1,t} \sim$ ARMA$(1, 1)$, where $\varphi_1 = 0.8$ and $\theta_1 = 0.2$.
- (2) $X_{1,t} \sim$ ARMA$(1, 1)$, where $\varphi_1 = 0.8$ and $\theta_1 = 0.4$.
- (3) $X_{1,t} \sim$ ARMA$(2, 1)$, where $\varphi_1 = 0.5$, $\varphi_2 = 0.3$ and $\theta_1 = 0.4$.

Table 6 presents the 95% confidence intervals for the RR, where $X_{1,t} \sim$ ARMA(p, q), where $p = 1, 2$ and $q = 1$. Here, only the INAR(1) bootstrap and the asymptotic approach were compared, as these methodologies presented similar results in the previous simulation studies. The autoregressive parameter ϕ was fixed at 0.2 once the simulations presented the best adjustments at this value. In all cases, the INAR(1) bootstrap presented a coverage rate approximately equal to 0.95, while for the asymptotic approach, this rate was close to 0.90.

Table 6: Coverage rate of the 95% confidence intervals for the RR (n=50)

	ARMA(1,1) $\varphi_1 = 0.8$ and $\theta_1 = 0.2$	ARMA(1,1) $\varphi_1 = 0.8$ and $\theta_1 = 0.4$	ARMA(2,1) $\varphi_1 = 0.5$, $\varphi = 0.3$ and $\theta_1 = 0.4$
INAR(1) bootstrap	0.958	0.970	0.942
Asymptotic	0.896	0.906	0.910

Prepared by the author

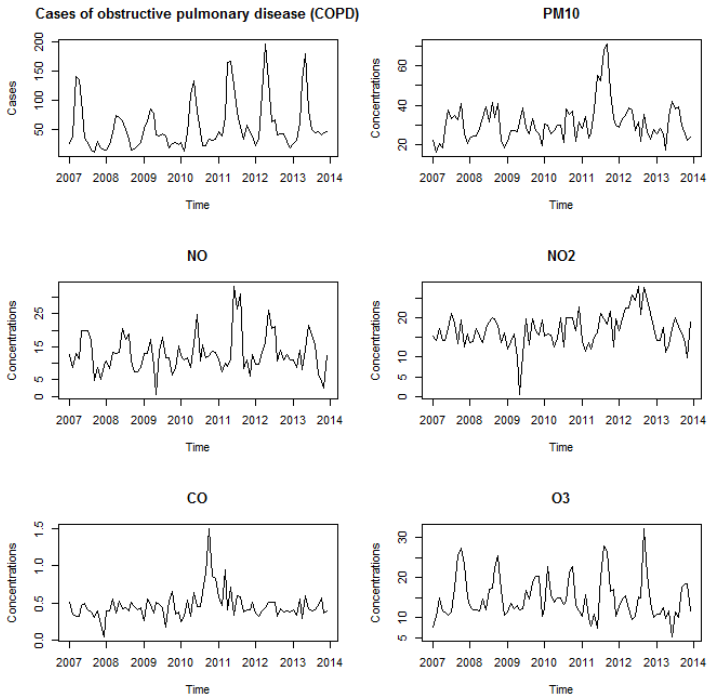
The INAR(1) bootstrap presented better results related to the coverage rate for the RR than the asymptotic approach considering the ARMA process as a covariate. This analysis is

relevant once in real data sets, in general, the sample size is not big, and the air pollutants present complex structures, e.g., time correlation, high volatility, and peaks. In addition, the coverage rates were close to 0.95 in the INAR(1) bootstrap even for high values of the autoregressive parameter ($\varphi = 0.8$), different from those observed in the simulations of scenario 3 (S3), in Section 4.1, where the covariate was an AR(1) process.

2.5 Real data analysis

A real data analysis was proposed to study the impact of air pollutants on the monthly number of Chronic Obstructive Pulmonary Disease (COPD) cases in Belo Horizonte, Brazil, between 2007 and 2013 ($n = 84$). Belo Horizonte is the capital of Minas Gerais state, with approximately 2,500,000 inhabitants, and the country's sixth most populous city. The following pollutants were considered: Particulate Matter (PM_{10}), Nitrogen Monoxide (NO), Nitrogen Dioxide (NO_2), Carbon Monoxide (CO) and Ozone (O_3). A significant correlation among some air pollutants was observed, while O_3 and NO presented the highest correlation with the response variable (COPD).

Figure 2.1: Time series of the number of COPD cases and concentrations of air pollutants in the metropolitan area of Belo Horizonte, Brazil.



Prepared by the author

A high number of covariates can lead to identification problems, and the correlation between them may imply multicollinearity. The Principal Component Analysis (PCA) is a possible solution to this problem. This methodology explains a random vector's variance and covariance structure through linear combinations of the original variables (Pearson [1901]). These combinations, called Principal Components (PC), are not correlated with each other. The PCA methodology requires independent observations; however, according to Zamprogno et al. [2020], if the covariates present time correlation, then the PCs are also autocorrelated. Based on this, the works of Souza et al. [2018] and Ispany et al. [2018] proposed a hybrid model called GAM-PCA-VAR, where the time dependency of data is removed through the VAR process. The PCs are derived from the residuals of VAR, and the GAM model is adjusted with PCs as explanatory variables.

This work estimated the impact of air pollutants on COPD occurrence according to the procedure cited above. The time correlation structure of the contaminants was removed by applying the VAR filter, and the PCs were derived from the residuals of VAR. The GLARMA model was fitted using the PCs as covariates. All the principal components were considered, although the first three correspond to almost 80% of the entire structure of variability. The GLARMA Poisson model adjusted was:

$$\begin{aligned}\ln(\mu_t) = \beta_1 * \text{PC1}_t + \beta_2 * \text{PC2}_t + \beta_3 * \text{PC3}_t + \beta_4 * \text{PC4}_t + \beta_5 * \text{PC5}_t \\ + \beta_6 * \text{sen6}_t + \beta_7 * \text{cos6}_t + \beta_8 * \text{sen12}_t + \beta_9 * \text{cos12}_t \\ + \beta_{10} * \text{trend} + Z_t,\end{aligned}$$

where

$$Z_t = \phi_2(Z_{t-2} + e_{t-2}).$$

The annual and semi-annual seasonality in the response variable were incorporated into the model with sine and cosine functions. The modeling also included the trend present in the data.

Table 7 presents the estimates of the adjusted model, with the corresponding standard errors. The best fit was obtained considering the autoregressive parameter of order 2. All coefficients were significant at the 5% level of significance.

Table 7: Parameter estimates of a GLARMA(2,0) model fitted to the COPD cases

Variable	Estimates	Standard Error	p-value
β_1	-0.0399	0.0105	0.0001
β_2	0.0347	0.0151	0.0221
β_3	-0.0560	0.0185	0.0025
β_4	0.0538	0.0173	0.0019
β_5	-0.0832	0.0206	0.0000
β_6	0.9277	0.0424	0.0000
β_7	-0.6710	0.0367	0.0000
β_8	-0.3772	0.0188	0.0000
β_9	-0.0695	0.0215	0.0012
β_{10}	0.0623	0.0005	0.0000
ϕ_2	0.1021	0.0017	0.0000

Prepared by the author

95% confidence intervals for the relative risk of air pollutants were calculated under the assumption of normality (asymptotic) and using the INAR(1) bootstrap. Table 8 shows that the CIs were similar, corresponding to the conclusions obtained in the simulation study. This real data analysis is equivalent to scenario 2, once the PCA covariates originated from the VAR process are not autocorrelated. For small values of the parameter ϕ, the numerical study in subsection 2.4.1 revealed that the confidence intervals provided by INAR bootstrap and the asymptotic approach are pretty close (see Table 2b).

Table 8: Comparison of the Relative Risk and 95% confidence intervals for an interquartile variation of the pollutant concentrations

	$\widehat{RR}$	Asymptotic CI	INAR(1) bootstrap CI
CO	0.9677	[0.9425 ; 0.9936]	[0.9372 ; 0.9933]
PM_{10}	1.0473	[1.0132 ; 1.0825]	[1.0174 ; 1.0783]
NO	0.9466	[0.9121 ; 0.9825]	[0.9172 ; 0.9772]
NO_2	1.1294	[1.0804 ; 1.1806]	[1.0721 ; 1.1876]
O_3	1.0442	[1.0078 ; 1.0819]	[1.0068 ; 1.0802]

Prepared by the author

2.6 Conclusions

This work proposed to study three bootstrap confidence interval approaches for the relative risk calculated from GLARMA models: the classic model-based, a procedure based on the specifications of the model, the sieve bootstrap well-known in the literature for real-valued process, and the recently proposed INAR(1) bootstrap based on the structure of the integer-valued autoregressive process.

An extensive numerical study was performed considering different scenarios and sample

sizes. For large samples, the analysis revealed that the model parameter (ϕ) could influence the estimates and, consequently, the coverage rate of the intervals. This impact is more significant as the complexity of the covariate's time structure increases. In general, the INAR(1) bootstrap and the asymptotic approach presented similar coverage rates, closest to 95%, than the other methodologies considered. The best performance of the INAR(1) bootstrap compared to the classic model-based and sieve bootstraps may be because this approach considers the data distribution. At the same time, the other procedures only consider the residuals in their implementation. Davis et al. [1999] discussed that typically in Poisson regression, using the residuals from the fit (based on observed counts minus the fitted values) would seriously underestimate the true serial dependence.

The observations from this study agree with the conclusions of Souza et al. [2018] and Ispany et al. [2018] once the best results were verified when X_1 does not present time correlation, which is equivalent to the variables filtered by the VAR process. This supports the importance of removing the temporal structure of the covariates before using any regression model.

Compared to the asymptotic confidence intervals, the INAR(1) bootstrap presented better coverage rates for small samples, even for more complex autocorrelation structures in the covariate.

In a real data set analysis, this work studied the impact of air pollutants on the monthly number of COPD cases in Belo Horizonte, Brazil. The best fit was obtained with the GLARMA(2,0) model. This work used the INAR(1) bootstrap and asymptotic approach to calculate the 95% confidence intervals for the relative risk of each contaminant. As observed in the empirical study, these intervals were equivalent.

Chapter 3

Robust estimate for counting time series using GLARMA models

Abstract The generalized linear autoregressive moving average (GLARMA) model has been used in epidemiological studies to evaluate the impact of air pollutants on human health, as frequently the response variable is a nonnegative integer-valued time series. The relative risk (RR) measure commonly quantifies these health effects. Due to the nature of the data, a robust approach for the GLARMA model is proposed based on the robustification of the quasi-likelihood function. In this method, outlying observations are bounded separately by weight functions on covariates and by the Huber loss function on the response variable. A numerical study was carried out to evaluate the performance of the proposed methodology for distinct sample sizes. In real data analysis, the impact of the particulate pollutant PM_{10} in the monthly number of deaths in Vitoria, Brazil, was investigated, showing that the parameter estimates involving the robust method are more reliable than the classic.
Keywords: Count time series, GLARMA model, M-estimators, Additive outliers, Respiratory diseases, Air pollution.

3.1 Introduction

The expansion of cities and communities in the last decades led to economic growth and urban development. However, it also originated environmental and health problems once many activities generate residues that affect the populations' quality of life. Ozone (O_3), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), carbon monoxide (CO), and particulate matter (PM) are the main pollutants in the atmosphere, and even at concentrations within limits established by the World Health Organization (WHO) offer risk to human health (Pope and Dockery [2018]and Lippmann [2014]). Epidemiological studies have shown evidence of an association between concentration levels of air pollutants and mortality, morbidity, and hospital admissions, mainly caused by respiratory and cardiovascular diseases (see Pope et al. [1995], Dockery and Pope

[1996], Ostro et al. [1999], Schwartz [2000], Ostro et al. [2009], Chen et al. [2010], Froes et al. [2016] among others).

Epidemiological data are frequently treated as counting time series as they record the frequency of events in successive time intervals. Count series are non-Gaussian processes formed by non-negative integers. They naturally arise in scientific areas such as the economy, medicine, agriculture, sports, among others. Examples are the monthly number of hospital admissions caused by a disease, the number of car accidents in a city, and the number of transactions of a given stock observed in one hour. Methodologies to deal with this kind of data started to emerge in the early 1970s. Initially, count time series were adjusted by generalized linear models (GLM), introduced by (Nelder and Wedderburn [1972a]), a procedure that expands the possibilities for the distribution of the response variable, which can assume distributions belonging to the exponential family, e.g., Normal, Poisson, Gamma, Negative Binomial, etc. In addition, the relation between the mean of the dependent variable (μ) and the linear predictor (η) can be more flexible, assuming any monotonous non-linear function. Nevertheless, the GLM can not capture the time dependency structure in the data. One of the earliest works considering correlated non-Gaussian time series can be found in Cox [1981], where models are classified into two categories: observation and parameter driven. The main difference between them is how the dependence structure is added to the model. Zeger and Qaqish [1988] proposed a quasi-likelihood approach to time series regression, generalized by Benjamin et al. [2003]. Davis et al. [1999] and Davis et al. [2003] introduced the generalized linear autoregressive moving average models (GLARMA). Fokianos and Tjosthein [2011] proposed log-linear models for time series. Other procedures can be found in Davis et al. [2021], which provided an overview of methodologies for counting time series. Although many methods have been developed in the field, they all present limitations that show the difficulty of attaining a unified theory. Despite this, Davis et al. [2021] reached the conclusion that the GLARMA family is "one of the most flexible and easily fit count models that balance parameter and observation-driven models". In this methodology, an ARMA structure (Box and Jenkins [1976]) is added to the GLM, allowing the modeling of correlated observations from the exponential family. Even though GLARMA presents some limitations regarding properties for general models, this method has been widely used in applications in distinct fields of knowledge; see e.g, Rydberg and Shephard [2003], in finances, Karami et al. [2017], in air pollution, Kim et al. [2018] in engineering, Ballesteros-Cánovas et al. [2018] and Peitzsch et al. [2021] in climate changes, among others.

Studying the statistical association between air pollutants and health effects is complex and must be analyzed cautiously regardless of the time series models used. In the epidemiological context, the response variable is usually time correlated, which should be considered. In addition, the dynamic of the response variable and, therefore, the statistical functions that measure the impact of the pollutants on health can not be fully explained by the response variable itself or by only one contaminant since the population under the study is exposed to a complex mixture of pollutants and chemical compounds. Many authors have been ignoring the fact that the

contaminants present multicollinearity. Souza et al. [2018] showed that if this characteristic is not treated properly, the association measures can be profoundly impacted, leading to false conclusions regarding the population's health risk in generalized additive models. Finally, covariates are time correlated and display complex behaviors such as periodicity, missing values, and extreme observations. High levels, or peaks, of pollutants are frequently observed in air quality variables and often ignored. However, they can affect the estimation of some characteristics of the data, like mean, variance, and correlation. In addition, many authors have been verifying that the presence of atypical observations (outliers) can seriously deteriorate the estimates of time series models (Reisen et al. [2017]).

Robustness indicates insensitivity to minor deviations from the assumptions (Huber [1981]). The foundations of this statistical approach can be found in Tukey [1960], Huber [1964], and Hampel [1968]. Robust models have the characteristic of fitting properly to most datasets. If the data has no abrupt observations, the robust method will behave approximately the same as the classic model. Nevertheless, if the data is composed of a small percentage of outliers, the robust models will show results almost as good as the classic models applied to clean data. Usually, robust estimates depend on a dispersion function that varies more slowly in extreme values than the quadratic functions. Outliers in time series can seriously affect the estimation and inference of parameters (Martin and Yohai [1985] and Bustos and Yohai [1986]). Fox [1972] appears to be the first author to consider outliers within time series, proposing two types of classes: the additive outliers, which affect only a single observation, and innovation outliers which affect succeeding observations. However, the additive outliers deserve special attention, as they usually cause more prejudice in practical problems. Ledolter [1989] showed that the ARMA models could be substantially affected by additive outliers. Chang et al. [1988] and Chen and Liu [1993] verified that the presence of additive outliers could bias the parameter estimates of the ARMA model. A similar conclusion was obtained by Reisen et al. [2017] and Sarnaglia et al. [2021] for fractionally integrated and periodic ARMA processes. Fokianos and Tjosthein [2011] verified that the maximum likelihood estimator in log-linear Poisson models is highly affected by additive outliers.

The nonrobustness of the maximum likelihood estimator in generalized linear models has been extensively studied in the literature (see Carroll and Welsh [1986], Künsch et al. [1989], Ruckstuhl and Welsh [1999], and others). Due to this, robust estimation procedures have been developed, e.g, Cantoni and Ronchetti [2001], Lo and Ronchetti [2009], and Valdora and Yohai [2014]. The work of Cantoni and Ronchetti [2001] is probably the most relevant which is based on the quasi-likelihood functions. The authors proposed the Mallows' quasi-likelihood estimator (MQLE) considering the class of M-estimators of Mallows' (Mallows [1975]). In this method, outlying observations are bounded separately by weight functions on covariates and by a loss function on the response variable. Although proposed for independent observations Kitromilidou and Fokianos [2016] extended this method to count time series in the context of the log-linear Poisson model. They found that the MQLE behaved comparably to the classic

log-linear model without perturbations. At the same time, in the presence of additive outliers, the MQLE provided more reliable results. Actually, procedures derived from M-estimators (Huber [1964]) are appropriate alternatives to modeling time series contaminated by outliers or generated by probability distribution with heavy tails (see Bai et al. [1992],Li [2008] and Wu [2007]). Thus, considering the previous discussion, the GLARMA model structure, and the nature of the data application, this paper proposes a robust alternative for the GLARMA Poisson model based on the MQLE estimator. To the best of our knowledge, robustified proposals for the GLARMA using M-estimators are still not explored in the literature. This paper aims to fill this gap. Due to the limitations regarding the asymptotic properties of the GLARMA model, we considered the development of asymptotic theory for the proposed robust approach beyond the scope of this work. In fact, Davis et al. [2021] claim that after all these years theoretical properties for the classic GLARMA model were only established for very restrictive special cases. However, although a general asymptotic theory has not yet been developed, the simulation study showed that asymptotic results corroborate that the estimators are consistent.

A Monte Carlo study was carried out to evaluate the impact of additive outliers in the response variable and covariates, considering the classic GLARMA (proposed by Davis et al. [2003]) and the robust proposal under distinct scenarios and sample sizes. Additionally, real data analysis was performed to study the effect of Particulate Material (PM_{10}) on the deaths caused by respiratory diseases in Vitoria, Brazil.

This work is organized as follows. Section 2 introduces the GLARMA model. Section 3 discusses robust estimation and proposes a robust approach for the GLARMA Poisson model. Section 4 presents a Monte Carlo empirical study to evaluate the performance of the proposed procedure. Section 5 presents a real data analysis, which is the primary motivation of this paper. Finally, Section 6 is composed of conclusions about the work.

3.2 The generalized linear autoregressive moving average model

The GLARMA models (Davis et al. [2003]) are a class of observation-driven non-Gaussian state space models in which the state process is linearly correlated to the explanatory variables and non-linearly to the past values of the observed process.

Let $\{Y_t\} := \{Y_t\}_{t\in\mathbb{Z}}$ be the observations on the response series, $\boldsymbol{X}_t = (X_{1,t}, X_{2,t}, ..., X_{k,t})^T$ the vector of k covariates observed for $t = 1, ..., n$, and $\mathcal{F}_{t-1} = \sigma\{Y_s, s \leq t-1\}$ the process history. The observation process Y_t conditioned on $\mathcal{F}_{t-1}$ is assumed exponentially distributed

with density

$$f(Y_t|W_t) = \exp\{Y_t W_t - a_t b(W_t) + c_t\}, \tag{3.1}$$

where $\{W_t\} := \{W_t\}_{t\in\mathbb{Z}}$ is the canonical parameter that summarizes the information in $\mathcal{F}_{t-1}$, and a_t and c_t are sequences of constants (for more, see Dunsmuir [2015] and Davis et al. [2021]). The conditional mean and variance of Y_t are $\mu_t = \mathrm{E}(Y_t|\mathcal{F}_{t-1})$ and $\sigma_t^2 = \mathrm{Var}(Y_t|\mathcal{F}_{t-1})$, respectively.

The specification of $W_t = g(\mu_t)$, where g is a link function, is given by

$$W_t = \boldsymbol{X}_t^T \boldsymbol{\beta} + Z_t, \tag{3.2}$$

where $\boldsymbol{\beta}$ is a $(k+1)$ x 1 vector of unknown coefficients, and the noise process $\{Z_t\}_{t\in\mathbb{Z}}$, which induces a serial dependence on the observation, is given by

$$Z_t = \sum_{i=1}^{\infty} \gamma_i e_{t-i}. \tag{3.3}$$

The parameters γ_i's are the coefficients in the power series expansion

$$\sum_{i=1}^{\infty} \gamma_i B^i = \frac{\theta(B)}{\phi(B)} - 1, \tag{3.4}$$

where the autoregressive and moving average components $\phi(B) = (1 - \phi_1 B - ... - \phi_p B^p)$ and $\theta(B) = (1 + \theta_1 B + ... + \theta_q B^q)$ are polynomials with no common zeroes and have all their zeros outside the unit circle. The parameter vector γ is formed by ϕ's and θ's, and B is the backshift operator of the form $B^k(Z_t) = Z_{t-k}$. From (3.3) and (3.4) Z_t can be calculated recursively with the difference equation

$$Z_t = \phi_1(Z_{t-1} + e_{t-1}) + \cdots + \phi_p(Z_{t-p} + e_{t-p}) + \theta_1 e_{t-1} + \cdots + \theta_q e_{t-q}. \tag{3.5}$$

The predictive residuals $\{e_t\}_{t\in\mathbb{Z}}$, in (3.3) are given by

$$e_t = \frac{Y_t - \mu_t}{\nu_t}, \tag{3.6}$$

where $\nu_t = \sigma_t$ for Pearson residuals. From (3.6), $\mathrm{E}(e_t|\mathcal{F}_{t-1}) = (\mathrm{E}(Y_t|\mathcal{F}_{t-1}) - \mu_t)/\nu_t = 0$. Under the initial conditions $e_s = 0$ and $Y_s = 0$, for $s \leq 0$, let $\mathcal{F}_{t-1}^e = \sigma(e_s, s \leq t-1)$. Equation (3.6) implies $\mathcal{F}_{t-1}^e \subset \mathcal{F}_{t-1}$, therefore

$$\mathrm{E}(e_t|\mathcal{F}_{t-1}^e) = \mathrm{E}[\mathrm{E}(e_t|\mathcal{F}_{t-1})|\mathcal{F}_{t-1}^e] = 0,$$

which means that $\{e_t\}$ are martingale differences, with $\mathrm{Cov}(e_t, e_s)$=0, for $t \neq s$. For Pearson residuals,

$$\mathrm{Var}(e_t) = \mathrm{E}(e_t^2) = \mathrm{E}[\mathrm{E}(e_t^2|\mathcal{F}_{t-1})] = \mathrm{E}\left[\mathrm{E}\left(\tfrac{Y_t-\mu_t}{\sigma}\right)^2 \,\middle|\, \mathcal{F}_{t-1}\right] = \mathrm{E}\left[\tfrac{\mathrm{E}(Y_t-\mu_t|\mathcal{F}_{t-1})^2}{\sigma^2}\right] = 1,$$

i.e. $\{e_t\}$ are weakly stationary white noise.

Considering n successive observations $y_1, y_2, ..., y_n$, the likelihood is constructed as the product of conditional densities of $\{Y_t\}$ given $\mathcal{F}_{t-1}$, corresponding to the following log-likelihood

$$L(\boldsymbol{\delta}) = \sum_{t=1}^{n} \left\{ Y_t W_t(\boldsymbol{\delta}) - a_t b(W_t(\boldsymbol{\delta})) + c_t \right\},$$

where $\boldsymbol{\delta} = (\boldsymbol{\beta}^T, \boldsymbol{\phi}^T, \boldsymbol{\theta}^T)^T$ is the parameter vector.

For the particular case of Poisson distribution, where $Y_t|\mathcal{F}_{t-1} \sim \text{Poisson}(\mu_t)$ with $\mu_t = e^{W_t}$, the log-likelihood is given by

$$L(\boldsymbol{\delta}) = \sum_{t=1}^{n} \left\{ y_t W_t(\boldsymbol{\delta}) - e^{W_t(\boldsymbol{\delta})} - \log(y_t!) \right\}. \tag{3.7}$$

The log-likelihood can be maximized using Newton-Raphson iterations or Fisher scoring procedure from suitable initial values by computing the first and second derivatives of the likelihood. According to Davis et al. [2005] the first derivative for (3.7) is given by

$$\frac{\partial L(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}} = \sum_{t=1}^{n} (y_t - \mu_t) \frac{\partial W_t(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}}, \tag{3.8}$$

and the second derivative is

$$\frac{\partial L^2(\boldsymbol{\delta})}{\partial \boldsymbol{\delta} \partial \boldsymbol{\delta}^T} = \sum_{t=1}^{n} (y_t - \mu_t) \frac{\partial^2 W_t(\boldsymbol{\delta})}{\partial \boldsymbol{\delta} \partial \boldsymbol{\delta}^T} - \sum_{t=1}^{n} \mu_t \frac{\partial W_t(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}} \frac{\partial W_t(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}^T}. \tag{3.9}$$

$E(y_t - \mu_t|\mathcal{F}_{t-1}) = 0$ at the true value of $\boldsymbol{\delta}$, which implies that the first summation in (3.9) is zero. This motivates the Fisher-scoring approximation based only on the first derivatives

$$\frac{\partial L^2(\boldsymbol{\delta})}{\partial \boldsymbol{\delta} \partial \boldsymbol{\delta}^T} = -\sum_{t=1}^{n} \mu_t \frac{\partial W_t(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}} \frac{\partial W_t(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}^T}. \tag{3.10}$$

Note that although the expectations of (3.9) and (3.10) are equal, they can not be calculated in closed form. Thus, the maximum likelihood estimator $\hat{\boldsymbol{\delta}}$ can be computed using the Newton-Raphson iterations (based on equations (3.8) and (3.9)) or the Fisher scoring approximations (equations (3.8) and (3.10)).

3.3 Robust estimation

Given a parametric model F_δ, a general M-estimate (Huber [1981]) of δ, say $\hat{\delta}$, is defined as a solution of

$$\arg\min_{\delta} \sum_{i=1}^{n} \rho(\xi_i, \delta), \tag{3.11}$$

where $\rho(\cdot)$ is the loss function. If $\rho(\cdot)$ is differentiable with respect to δ, where $\psi(\cdot) = \rho'(\cdot)$, then $\hat{\delta}$ is the solution of the estimating equations

$$\sum_{i=1}^{n} \psi(\xi_i, \delta) = 0. \tag{3.12}$$

There are several candidates for ρ. However, the Huber loss function, proposed by Huber [1964], and Tukey's biweight, given by (Beaton and Tukey [1974]) are the most used. Some other loss functions can be found in Hampel [1974], Andrews [1974], Dennis and Welsch [1978], and Maronna et al. [2006]. According to Huber [1981], $\rho(\cdot)$ must satisfy the following assumptions:
A1) $\rho(0) = 0$;
A2) $\rho(\xi) = \rho(-\xi) \forall \xi \in \mathbb{R}$, i.e. $\rho(\xi)$ is a symmetric function;
A3) $0 \leq \xi \leq \xi^* \Rightarrow \rho(\xi) \leq \rho(\xi^*), \forall (\xi, \xi^*) \in \mathbb{R}^2$;
A4) $\psi(\cdot)$ is bounded;
A5) $\rho(\cdot)$ has a second derivative almost everywhere.

Here we will focus on the Huber loss function

$$\rho_H(\xi) = \begin{cases} \frac{1}{2}\xi^2, & |\xi| \leq c \\ c|\xi| - \frac{1}{2}c^2, & |\xi| > c. \end{cases} \tag{3.13}$$

Its derivative, ψ-function, is given by

$$\psi_H(\xi) = \begin{cases} \xi, & |\xi| \leq c \\ c\text{sign}\,(\xi)\,, & |\xi| > c, \end{cases} \tag{3.14}$$

where the constant c must be prespecified and regulates the amount of robustness. This parameter regulates the trade-off between the efficiency and robustness of the estimators. Good choices for the constant value are in the range between 1 and 2. According to (Huber [1964]) $c = 1.345$ provides 90% efficiency when the data is normally distributed. Other specific values are also used in the literature, e.g., $c = 1.2$ (Cantoni and Ronchetti [2001]), $c = 1.25$ (Streett et al. [1988] and Chi [1994]). The choice of c should reflect the proportion of outliers in the data. Moreover, this value must be adjusted according to the data distribution.

3.3.1 Robust estimation for GLARMA models

To robustify the parameter estimation of the GLARMA model, we propose here an extension of the approach given by Cantoni and Ronchetti [2001] called the Mallows' Quasi-Likelihood Estimator (MQLE). Their approach is based on natural generalizations of quasi-likelihood functions, considering a general class of M-estimators of Mallows' type (Mallows [1975]), where the influence of deviations on response variable and covariates are bounded separately.

The MQLE for GLARMA family, denoted by $\hat{\delta}_{MQLE}$, is the solution of the estimating equation

$$S_n(\delta) = \sum_{t=1}^{n} \left[\nu(Y_t, \mu_t) w(\mathbf{X}_t) \mu_t' - \frac{1}{n} \sum_{t=1}^{n} \mathrm{E}\left(\nu(Y_t, \mu_t) | \mathcal{F}_{t-1}\right) w(\mathbf{X}_t) \mu_t' \right] = 0. \tag{3.15}$$

The $\hat{\delta}_{MQLE}$ is an M-estimator (Huber [1981]; Hampel et al. [1986]) characterized by the score function in (3.15). The term $a(\boldsymbol{\delta}) = \frac{1}{n}\sum_{t=1}^{n} \mathrm{E}(\nu(Y_t, \mu_t)w(\mathbf{X}_t)\mu_t'|\mathcal{F}_{t-1})$, in (3.15), is a bias correction used to ensure Fisher's consistency. Additionally, function $\nu(\cdot,\cdot)$ is chosen to control deviations on Y-space and leverage points on $\mathbf{X}$-space are down-weighted by $w(\cdot)$.

Künsch [1984] extended the definition of Influence Function (IF) of Hampel [1974] to time series for stationary process. Therefore, considering a cumulative distribution function F, the $\mathrm{IF}(Y_t; \boldsymbol{\psi}, F) = M(\boldsymbol{\psi}, F)^{-1}S_n(\delta)$, where $M(\boldsymbol{\psi}, F)^{-1} = -\mathrm{E}\left[\frac{\partial}{\partial \boldsymbol{\delta}}S_n(\delta)\right]$, for more, see Maronna et al. [2006]. Choosing a bounded function S_n leads to limits on the influence function, which ensures the robustness of the estimator. Thus, bounded functions $\nu(\cdot,\cdot)$ and $w(\cdot)$ must be chosen to restrict outlying values on the response variable and covariates, respectively.

Let $\nu(Y_t, \mu_t) = \psi_H(r_t)\frac{1}{Var(Y_t)^{1/2}}$, where, ψ_H is the Huber loss function defined in (3.14), $r_t = \frac{Y_t - \mu_t}{Var(Y_t)^{1/2}}$, are the Pearson residuals and $\mu_t' = \mu_t\frac{\partial W_t(\boldsymbol{\delta})}{\partial \delta}, t = 1, ..., n$. Replace it on (3.15), then $\hat{\delta}_{MQLE}$ of the GLARMA Poisson model is the solution of the following equation

$$S_n(\delta) = \sum_{t=1}^{n}\left[\frac{\psi_H(r_t)}{Var(Y_t)^{1/2}}w(\mathbf{X}_t)\mu_t\frac{\partial W_t(\boldsymbol{\delta})}{\partial \delta} - \frac{1}{n}\sum_{t=1}^{n}\mathrm{E}\left(\frac{\psi_H(r_t)}{Var(Y_t)^{1/2}}\,\middle|\,\mathcal{F}_{t-1}\right)w(\mathbf{X}_t)\mu_t\frac{\partial W_t(\boldsymbol{\delta})}{\partial \delta}\right] = 0, \tag{3.16}$$

where $Var(Y_t) = \mu_t$ and

$$\begin{aligned}\mathrm{E}\left(\psi_H(r_t)\,\middle|\,\mathcal{F}_{t-1}\right) = c\{&P(Y_t \geq j_2 + 1|\mathcal{F}_{t-1}) - P(Y_t \leq j_1|\mathcal{F}_{t-1})\} \\ &+\mu_t^{1/2}\{P(Y_t = j_1|\mathcal{F}_{t-1}) - P(Y_t = j_2|\mathcal{F}_{t-1})\},\end{aligned}$$

with $j_1 = \left\lfloor \mu_t - c\mu_t^{1/2}\right\rfloor$ and $j_2 = \left\lfloor \mu_t + c\mu_t^{1/2}\right\rfloor$, and c is the tunning constant.

A common choice for the sequence of weights $w(\mathbf{X}_t), t = 1, ..., n$, in (3.16) is $w(\mathbf{X}_t) = \sqrt{1 - h_{tt}}$, where h_{tt} is the tth diagonal element of the hat matrix $H = \mathbf{X}_t(\mathbf{X}_t^T\mathbf{X}_t)^{-1}\mathbf{X}_t^T$ (see Cantoni and Ronchetti [2001]). However, the hat matrix does not have breakdown points, i.e., the estimates are not reasonable if large atypical values contaminate the data. More sophisticated methods can be found in the literature based on the inverse of robust Mahalanobis distance.

Let μ and $\sum$ be the location parameter and the covariance matrix of $\mathbf{X}_t$, respectively. The squared Mahalanobis distance of each observation along a row in $\mathbf{X}_t$ from μ with respect to $\sum$ is

$$d_{\mu,\sum}(\mathbf{X}_t)^2 = (\mathbf{X}_t - \hat{\mu})^T\hat{\sum}^{-1}(\mathbf{X}_t - \hat{\mu}).$$

To robustify the Mahalanobis distance, the location parameters and the covariance matrix can be estimated using the minimum covariance determinant algorithm, the fast MCD (see Rousseeuw [1984], page 877 and Rousseeuw [1985] for more details). In this procedure, h observations (out n) are chosen whose classical covariance matrix presents the lowest determinant. Then the MCD estimate of location ($\hat{\mu}_{(\mathrm{MCD})}$) is the average of the h points, and their covariance matrix is the MCD estimate scatter ($\hat{\sum}_{(\mathrm{MCD})}$). In this paper, we use the weight function $w(\cdot)$

based on the MCD estimates. It is given by

$$w(\mathbf{X}_t) = \min\left[1, \left\{\frac{b}{(Y_t - \hat{\mu}_{(\text{MCD})})^T \hat{\sum}_{(\text{MCD})}^{-1}(Y_t - \hat{\mu}_{(\text{MCD})})}\right\}^{\alpha/2}\right], \tag{3.17}$$

where α and b are tuning constants. Simpson et al. [1992] evaluate some values for the constant α and claim that $\alpha = 1$ is usual for the class of generalized M-estimators. In addition, the authors set b equal to the $(1-\gamma)$-quantile of the chi-squared distribution with $k-1$ degrees of freedom, where k is the number of predictor covariates and $\gamma = 0.1$ and 0.05.

The quasi-likelihood function of $\{Y_t\}$ conditional to $\mathcal{F}_{t-1}$ is given by

$$Q(\boldsymbol{\delta}) = \sum_{t=1}^{n} Q_M(\delta), \tag{3.18}$$

where $Q_M(\delta)$ is given by

$$Q_M(\boldsymbol{\delta}) = \int_{\tilde{s}}^{\mu_t} \nu(Y_t, u)w(\mathbf{X}_t)du - \frac{1}{n}\sum_{j=1}^{n}\int_{\tilde{u}}^{\mu_j} \mathrm{E}[\nu(Y_j, u)w(\mathbf{X}_j)|\mathcal{F}_{t-1}]du,$$

with $\tilde{s}$ and $\tilde{u}$ defined such as $\nu(Y_t, \tilde{s}) = 0$ and $\mathrm{E}[\nu(Y_t, \tilde{u})] = 0$ (see Cantoni and Ronchetti [2001] for more details). The solution of equation (3.15) corresponds to minimize equation (3.18).

Parameter estimates can be obtained using Newton-Raphson or Fisher-scoring approximations. The first derivative of $Q(\boldsymbol{\delta})$ is

$$\frac{\partial Q(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}} = S_n(\boldsymbol{\delta}). \tag{3.19}$$

For $r_t \leq c$ the second derivative is

$$\frac{\partial^2 Q(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}\partial \boldsymbol{\delta}^T} = \sum_{t=1}^{n}\left[(Y_t - \mu_t)\frac{\partial^2 W_t(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}\partial \boldsymbol{\delta}^T}w(\mathbf{X}_t) - \mu_t \frac{\partial W_t(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}}\frac{\partial W_t(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}^T}w(\mathbf{X}_t) - a'(\boldsymbol{\delta})\right], \tag{3.20}$$

where

$$a(\boldsymbol{\delta}) = \frac{1}{n}\sum_{t=1}^{n} \mathrm{E}\left[\psi_H\left(\frac{Y_t - \mu_t}{\mu_t^{1/2}}\right) w(\mathbf{X}_t)\mu_t^{1/2}\frac{\partial W_t(\boldsymbol{\delta})}{\partial \delta}\,\middle|\, \mathcal{F}_{t-1}\right].$$

For $r_t > c$ the second derivative of $Q(\boldsymbol{\delta})$ with respect to $\boldsymbol{\delta}$ is

$$\frac{\partial^2 Q(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}\partial \boldsymbol{\delta}^T} = \sum_{t=1}^{n}\left\{c\,\text{sign}(r_t)w(\mathbf{X}_t)\left[\frac{1}{2}\mu_t^{1/2}\frac{\partial W_t(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}}\frac{\partial W_t(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}^T} + \mu_t^{1/2}\frac{\partial^2 W_t(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}\partial \boldsymbol{\delta}^T}\right] - a'(\boldsymbol{\delta})\right\}. \tag{3.21}$$

At the true parameter of $\boldsymbol{\delta}$, $E[(y_t - \mu_t)|\mathcal{F}_{t-1}] = 0$, the expected value of the first summation in (3.20) is zero, which leads to the Fisher Scoring approximation:

$$\frac{\partial^2 Q(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}\partial \boldsymbol{\delta}^T} = \sum_{t=1}^{n}\left[-\mu_t \frac{\partial W_t(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}}\frac{\partial W_t(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}^T}w(\mathbf{X}_t) - a'(\boldsymbol{\delta})\right]. \tag{3.22}$$

Details about the computation of the derivative $a'(\boldsymbol{\delta}) = \frac{\partial a(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}}$ and the second derivative of $Q(\boldsymbol{\delta})$ with respect to $\boldsymbol{\delta}$ can be found in Appendix 1.

3.4 Monte Carlo study

A simulation study was conducted to evaluate the performance of the robust estimation for the GLARMA Poisson model proposed in 3.3.1. The model is given by

$$Y_t|\mathcal{F}_{t-1} \sim \text{Poisson}(\mu_t) \tag{3.23}$$

$$\ln(\mu_t) = \beta_0 + \beta_1 X_{1,t} + Z_t. \tag{3.24}$$

Two scenarios were considered for the regressor variable; $(X_{1,t})$ is an independent $N(0,1)$ random variable, and $(X_{1,t})$ is an autoregressive process of order 1. We set $\beta_0 = 1$ and $\beta_1 = 0.5$. The Monte Carlo simulations were repeated 1000 times with sample sizes equal to $n = 100$ and $n = 1000$. The choice of the *tuning parameter* for the Huber function was $c = 1.345$. However, the cross-validation procedure applied to time series, using blocks, is also an option to choose the value of this constant (see Bergmeir and Benitez [2012] and Bergmeir et al. [2018]). Both options were considered in the numerical simulations and provided similar results.

3.4.1 Covariate contaminated by additive outliers

We added additive outliers to covariate $(X_{1,t})$ to disturb the linear predictor in equation (3.23). The contaminated version of $X_{1,t}$ is defined by $X^*_{1,t} = X_{1,t} + \omega\varphi_t$, where $\omega = 5$ is the magnitude of the outlier which impacts $X_{1,t}$ and φ_t indicates the presence or not of this outlier and its sign at time t, i.e., $\varphi_t = 0$ with probability $1 - \varphi$, $\varphi_t = 1$ with probability $\varphi/2$, and $\varphi_t = -1$ with probability $\varphi/2$, where $\varphi = 0.01$

Once Davis et al. [2003] only presented formal properties for the simplest case, where the time correlation structure is moving average, we will first show the scenario considering the GLARMA(0,1) model and then extend the simulations for the GLARMA(1,0) model.

3.4.1.1 Scenario 1: Moving average process - GLARMA(0,1)

The GLARMA(0,1) model is defined as equations (3.23) and (3.24), where $\{Z_t\}$ is a moving average process of order 1, defined as $Z_t = \theta(Y_{t-1} - e^{\eta_{t-1}})e^{-\lambda\eta_{t-1}}$ with $\theta = 0.2$ and $\lambda = 0.5$, which corresponds to Pearson residuals.

Table 1 presents the parameter estimation considering $X_{1,t}$ as an independent random vector in time, following a Normal(0,1) distribution. For $n = 100$, in the classic procedure, without outliers, the mean of $\hat{\beta}_0$ and $\hat{\beta}_1$ was close to the real values, while $\hat{\theta}$ was underestimated. The classic approach in the presence of additive outliers was impacted in the mean of $\hat{\beta}_1$ and $\hat{\theta}$, with both parameters being underestimated. The mean squared error (MSE) increased in the presence of outliers for all parameters in the study. The robust approach without any outlier presented parameter estimations similar to the classic in the same conditions. However, the mean of $\hat{\theta}$ was closer to the real value of θ, and consequently, the MSE was smaller than the error observed in the classic method. Finally, the proposed robust methodology was applied to contaminated data. The results showed that, differently from the classic GLARMA, the mean of the estimates was not affected, with values close to the real ones. The MSE was not impacted as well. Similar conclusions were observed for $n = 1000$, but the values of the MSE were smaller.

Table 1: Parameter estimation - $X_{1,t} \sim \text{N}(0, 1)$ - GLARMA(0,1)

		n=100				n=1000			
		no outlier		with outlier		no outlier		with outlier	
		Mean	MSE	Mean	MSE	Mean	MSE	Mean	MSE
Classic	$\hat{\beta}_0$	0.986	0.0057	1.055	0.0084	0.997	0.0006	1.029	0.0014
	$\hat{\beta}_1$	0.509	0.0036	0.403	0.0139	0.503	0.0003	0.403	0.0095
	$\hat{\theta}$	0.121	0.0096	0.079	0.0171	0.126	0.0057	0.094	0.0114
Robust	$\hat{\beta}_0$	0.968	0.0069	0.979	0.0066	0.984	0.0009	0.988	0.0008
	$\hat{\beta}_1$	0.539	0.0054	0.514	0.0049	0.533	0.0014	0.514	0.0005
	$\hat{\theta}$	0.168	0.0063	0.158	0.0069	0.175	0.0011	0.171	0.0013

Prepared by the author

In Table 2, $X_{1,t} \sim \text{AR}(1)$, with the autoregressive parameter assuming value 0.4. Similarly to Table 1, the classic GLARMA in the absence of contamination presented parameter estimates closer to the real values, except for $\hat{\theta}$, which was underestimated. In the presence of additive outliers, $\hat{\beta}_1$ and $\hat{\theta}$ were again affected. It is important to note that the impact in $\hat{\beta}_1$ was more prominent in this case, which suggests that for covariates with time correlation structure, the presence of perturbation in the data must be carefully treated. For the robust approach, the mean of the estimates in the study was close to the actual parameter values, with values of MSE slightly bigger than the classic method, except for $\hat{\theta}$. The robust GLARMA applied to the series contaminated by additive outliers provided parameter estimates close to the real ones and MSE values comparable to the results of the classic method in the absence of outliers. The same conclusions were observed for $n = 100$ and $n = 1000$.

Table 2: Parameter estimation - $X_{1,t} \sim \text{AR}(1)$ - GLARMA(0,1)

		n=100				n=1000			
		no outlier		with outlier		no outlier		with outlier	
		Mean	MSE	Mean	MSE	Mean	MSE	Mean	MSE
Classic	$\hat{\beta}_0$	0.980	0.0064	1.000	0.0055	0.997	0.0005	1.019	0.0009
	$\hat{\beta}_1$	0.493	0.0044	0.338	0.0281	0.501	0.0003	0.424	0.0060
	$\hat{\theta}$	0.111	0.0117	0.112	0.0124	0.120	0.0065	0.108	0.0087
Robust	$\hat{\beta}_0$	0.964	0.0078	0.977	0.0066	0.982	0.0009	0.987	0.0007
	$\hat{\beta}_1$	0.532	0.0063	0.479	0.0047	0.532	0.0014	0.517	0.0006
	$\hat{\theta}$	0.161	0.0072	0.160	0.0077	0.168	0.0015	0.166	0.0016

Prepared by the author

3.4.1.2 Scenario 2: Autoregressive process - GLARMA(1,0)

For the GLARMA(1,0) model, defined by equations (3.23) and (3.24), $\{Z_t\}$ is an autoregressive process of order 1, where $Z_t = \phi[Z_{t-1} + (Y_{t-1} - e^{\eta_{t-1}})e^{-\lambda\eta_{t-1}}]$, with $\phi = 0.2$, and $\lambda = 0.5$.

Table 3 presents the parameter estimation for $X_{1,t} \sim N(0,1)$. For both sample sizes, in the classic procedure, without perturbations, the mean of $\hat{\beta}_0$ and $\hat{\beta}_1$ was close to the real values, while $\hat{\phi}$ was underestimated. The parameter estimates of the classic approach in the presence of additive outliers were impacted in the mean and MSE of $\hat{\beta}_1$ and $\hat{\phi}$. Both parameters were underestimated. The robust proposal without contamination presented parameter estimations close to the classic in the same conditions, except for $\hat{\phi}$, which was closer to the actual value of ϕ, and displayed an MSE smaller than that observed in the classic method. When the proposed robust GLARMA model was applied to contaminated data, the mean of the estimates was not affected, with values close to the real ones. The MSE has practically not changed compared to the scenario without additive outliers.

Table 3: Parameter estimation - $X_{1,t} \sim \text{N}(1,0)$ - GLARMA(1,0)

		n=100				n=1000			
		no outlier		with outlier		no outlier		with outlier	
		Mean	MSE	Mean	MSE	Mean	MSE	Mean	MSE
Classic	$\hat{\beta}_0$	0.973	0.0071	1.029	0.0073	0.995	0.0006	1.020	0.0009
	$\hat{\beta}_1$	0.501	0.0029	0.404	0.0139	0.505	0.0003	0.421	0.0064
	$\hat{\phi}$	0.119	0.0100	0.082	0.0166	0.127	0.0057	0.115	0.0077
Robust	$\hat{\beta}_0$	0.965	0.0082	0.972	0.0078	0.983	0.0009	0.992	0.0007
	$\hat{\beta}_1$	0.536	0.0049	0.507	0.0041	0.530	0.0012	0.502	0.0003
	$\hat{\phi}$	0.169	0.0064	0.160	0.0069	0.178	0.0009	0.176	0.0011

Prepared by the author

Table 4 presents the parameter estimation for $X_{1,t} \sim \text{AR}(1)$. As observed in Table 3, the classical method in clear data presented parameter estimates closer to the real values, except for $\hat{\phi}$, which was underestimated. In the presence of additive outliers $\hat{\beta}_1$ and $\hat{\phi}$ were affected. Note that $\hat{\beta}_1 = 0.299$, while the real parameter value is 0.50, a significant reduction. For the robust approach, the mean of parameters in the study was close to the actual values, with values of MSE slightly bigger than the classic method, except for $\hat{\phi}$. The robust GLARMA applied to the series contaminated by additive outliers provided parameter estimates close to the real ones and MSE values similar to the classic method in the absence of outliers. Similar results were observed for $n = 100$ and $n = 1000$.

Table 4: Parameter estimation - $X_{1,t} \sim \text{AR}(1)$ - GLARMA(1,0)

		n=100				n=1000			
		no outlier		with outlier		no outlier		with outlier	
		Mean	MSE	Mean	MSE	Mean	MSE	Mean	MSE
Classic	$\hat{\beta}_0$	0.981	0.0069	1.005	0.0059	0.997	0.0006	1.041	0.0022
	$\hat{\beta}_1$	0.511	0.0042	0.299	0.0414	0.503	0.0003	0.376	0.0153
	$\hat{\phi}$	0.117	0.0099	0.135	0.0081	0.119	0.0068	0.108	0.0088
Robust	$\hat{\beta}_0$	0.967	0.0079	0.983	0.0068	0.982	0.0009	0.993	0.0006
	$\hat{\beta}_1$	0.549	0.0072	0.482	0.0038	0.528	0.0012	0.499	0.0003
	$\hat{\phi}$	0.169	0.0059	0.168	0.0066	0.167	0.0015	0.166	0.0016

Prepared by the author

3.4.2 Response variables Y_t contaminated by additive outliers

The effect of additive outliers in count time series was evaluated considering GLARMA(0,1) and GLARMA(1,0) models under scenarios where the covariate is an independent random variable or $X_{1,t} \sim AR(1)$. The contaminated version of Y_t is defined by $Y_t^* = Y_t + \omega\varphi_t$, where $\omega = 30$ is the magnitude of the outlier which impacts Y_t and φ_t indicates the presence or not of this outlier at time t, i.e., $\varphi_t = 1$ with probability φ, and $\varphi_t = 0$ with probability $1 - \varphi$, where $\varphi = 0.01$.

3.4.2.1 Scenario 3: Moving average process - GLARMA(0,1)

GLARMA(0,1) model is defined according equations (3.23) and (3.24), where process $\{Z_t\}$ is a moving average of order 1, given by $Z_t = \theta(Y_{t-1} - e^{\eta_{t-1}})e^{-\lambda\eta_{t-1}}$ with $\theta = 0.2$ and $\lambda = 0.5$.

Table 5 presents the parameter estimation for $X_{1,t} \sim N(0,1)$. Under the classical procedure, with clear data, the mean of $\hat{\beta}_0$ and $\hat{\beta}_1$ was close to the real values, and $\hat{\theta}$ was underestimated. However, all parameters were impacted by additive outliers on the count response variable. $\hat{\beta}_0$ was overestimated, while $\hat{\beta}_1$ and $\hat{\theta}$ were underestimated. The MSE increased in the presence of outliers for all parameters in the study. The robust approach without additive outliers on $\{Y_t\}$ presented parameter estimates similar to the classic in the same conditions. But, the mean of $\hat{\theta}$ was closer to the real value of θ, while its MSE was smaller than the in the classic method. The robust methodology applied to contaminated data showed that the mean of the estimates was not affected, with values close to the real ones. The MSE slightly decreased. Similar conclusions were observed for $n = 1000$.

Table 5: Parameter estimation - $X_{1,t} \sim \text{N}(0,1)$ - GLARMA(0,1)

		n=100				n=1000			
		no outlier		with outlier		no outlier		with outlier	
		Mean	MSE	Mean	MSE	Mean	MSE	Mean	MSE
Classic	$\hat{\beta}_0$	0.977	0.0071	1.121	0.0194	0.996	0.0006	1.137	0.0193
	$\hat{\beta}_1$	0.526	0.0047	0.388	0.0153	0.504	0.0003	0.468	0.0012
	$\hat{\theta}$	0.120	0.0099	0.009	0.0371	0.125	0.0058	0.031	0.028
Robust	$\hat{\beta}_0$	0.968	0.0077	0.977	0.0071	0.982	0.0009	0.997	0.0006
	$\hat{\beta}_1$	0.539	0.0053	0.526	0.0047	0.533	0.0014	0.531	0.0013
	$\hat{\theta}$	0.171	0.0073	0.168	0.0060	0.172	0.0012	0.165	0.0017

Prepared by the author

Table 6 presents the parameter estimates for $X_{1,t} \sim \text{AR}(1)$. The classic GLARMA(0,1), for $n = 100$, in the absence of contamination on $\{Y_t\}$ presented estimates closer to the real values, only for $\hat{\beta}_0$. $\hat{\beta}_1$ and $\hat{\theta}$ were underestimated. For $n = 1000$, only $\hat{\theta}$ was underestimated, while $\hat{\beta}_0$ and $\hat{\beta}_1$ were close to actual values. In the presence of outliers, all parameters were affected. Note that the impact for $n = 1000$ was more prominent in this case. For both sample sizes, in the robust approach, without additive outliers, the mean of parameters in the study was close to the real ones. In the presence of perturbations on the response variable, the robust approach provided parameter estimates close to the true values and MSE measures comparable to the results of the classic method in the absence of outliers.

Table 6: Parameter estimation - $X_{1,t} \sim \text{AR}(1)$ - GLARMA(0,1)

		n=100				n=1000			
		no outlier		with outlier		no outlier		with outlier	
		Mean	MSE	Mean	MSE	Mean	MSE	Mean	MSE
Classic	$\hat{\beta}_0$	0.993	0.0056	1.117	0.0184	0.998	0.0005	1.303	0.0921
	$\hat{\beta}_1$	0.447	0.0059	0.396	0.0134	0.502	0.0003	0.318	0.0331
	$\hat{\theta}$	0.116	0.0104	0.013	0.0351	0.125	0.0058	-0.005	0.0425
Robust	$\hat{\beta}_0$	0.982	0.0064	0.992	0.0066	0.983	0.0008	1.043	0.0024
	$\hat{\beta}_1$	0.505	0.0040	0.492	0.0037	0.536	0.0017	0.507	0.0004
	$\hat{\theta}$	0.167	0.0061	0.154	0.0067	0.172	0.0012	0.141	0.0039

Prepared by the author

3.4.2.2 Scenario 4: Autoregressive process - GLARMA(1,0)

The GLARMA(1,0) model is defined by equations (3.23) and (3.24), with the autoregressive process of order 1 $\{Z_t\}$ given by $Z_t = \phi[Z_{t-1} + (Y_{t-1} - e^{\eta_{t-1}})e^{-\lambda\eta_{t-1}}]$, with $\phi = 0.2$, and $\lambda = 0.5$.

Table 7 presents the parameter estimation for $X_{1,t} \sim N(0,1)$. In the classic procedure, without additive outliers, the mean of $\hat{\beta}_0$ and $\hat{\beta}_1$ was close to the real values, while $\hat{\phi}$ was underestimated. All the parameter estimates were impacted in the presence of additive outliers on $\{Y_t\}$. The MSE increased in the presence of outliers for parameters $\hat{\beta}_0$ and $\hat{\beta}_1$, for $n = 100$. For $n = 1000$, the MSE increased for all parameters. The robust approach without outlier presented parameter estimations close to the classic in the same conditions, except for $\hat{\phi}$, which was more relative to the true value of ϕ. As expected, for this parameter, the MSE was smaller than that observed in the classic method. Applying the robust procedure to the contaminated data, the mean of the estimates was not affected, with values close to the real ones.

Table 7: Parameter estimation - $X_{1,t} \sim \text{N}(0,1)$ - GLARMA(1,0)

		n=100				n=1000			
		no outlier		with outlier		no outlier		with outlier	
		Mean	MSE	Mean	MSE	Mean	MSE	Mean	MSE
Classic	$\hat{\beta}_0$	0.975	0.0068	1.229	0.0565	0.996	0.0006	1.183	0.0340
	$\hat{\beta}_1$	0.506	0.0024	0.423	0.0076	0.503	0.0002	0.431	0.0049
	$\hat{\phi}$	0.121	0.0093	0.111	0.0080	0.127	0.0056	0.010	0.0359
Robust	$\hat{\beta}_0$	0.963	0.0081	1.004	0.0065	0.985	0.0008	1.000	0.0005
	$\hat{\beta}_1$	0.527	0.0036	0.509	0.0028	0.528	0.0011	0.525	0.0009
	$\hat{\phi}$	0.167	0.0059	0.170	0.0050	0.176	0.0010	0.168	0.0014

Prepared by the author

Table 8 presents the parameter estimation for $X_{1,t} \sim \text{AR}(1)$. As observed in Table 7, the

classical method, without perturbation, presented parameter estimates closer to the real values, except for $\hat{\phi}$, which was underestimated. In the presence of additive outliers, all parameters were affected. For the robust approach, the mean of parameters in the study was close to the actual values, with values of MSE slightly bigger than the classic method, except for $\hat{\phi}$. The robust GLARMA applied to response variables perturbed by additive outliers provided parameter estimates close to the real ones and MSE values comparable to the classic method in the absence of outliers. Similar results were observed for $n = 100$ and $n = 1000$.

Table 8: Parameter estimation - $X_{1,t} \sim$ AR(1) - GLARMA(1,0)

		n=100				n=1000			
		no outlier		with outlier		no outlier		with outlier	
		Mean	MSE	Mean	MSE	Mean	MSE	Mean	MSE
Classic	$\hat{\beta}_0$	0.975	0.0067	1.28	0.0824	0.997	0.0006	1.118	0.0146
	$\hat{\beta}_1$	0.501	0.0032	0.413	0.0097	0.501	0.0003	0.448	0.0029
	$\hat{\phi}$	0.116	0.0105	-0.007	0.0439	0.117	0.0071	0.021	0.0319
Robust	$\hat{\beta}_0$	0.966	0.0077	0.988	0.0063	0.982	0.0010	0.993	0.0007
	$\hat{\beta}_1$	0.525	0.0043	0.530	0.0048	0.525	0.0009	0.523	0.0009
	$\hat{\phi}$	0.160	0.0069	0.166	0.0076	0.165	0.0016	0.162	0.0018

Prepared by the author

The empirical study showed that the classic GLARMA is impacted by additive outliers, independently of the perturbation affecting the covariate or the response variable. The proposed robust GLARMA approach provides similar results to the classic in the absence of additive outliers. In contaminated data, the robust approach was superior, providing parameter estimates closer to the true values with small MSE measures. In all scenarios analyzed, the MSEs for $n = 1000$ were smaller than $n = 100$, independently of the methodology applied.

3.5 Real data analysis

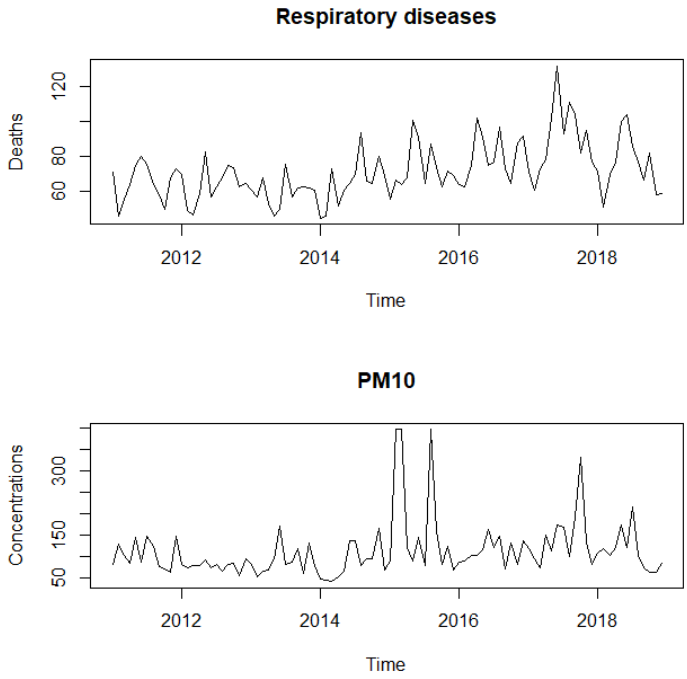
A real data analysis was carried out to evaluate the impact of the Particulate Material (PM_{10}) on the monthly number of deaths by respiratory diseases between 2011 to 2018 ($n = 96$) in the Great Vitoria region (GVR), Brazil, which is a port and industrialized region, densely populated in the state of Espírito Santo, with approximately 1,900,000 inhabitants. Although the atmosphere is composed of many gases and particulate matter, only PM_{10} was considered because the data quality of other contaminants during the period was too poor. The PM_{10} are microscopic solid particles and liquid droplets suspended in the air, with a diameter of 10 micrometers (μm) or less. This particle pollution mainly comes from motor vehicles, wood-burning heaters, and industry. It has been associated with premature mortality, increased hospital admissions for

heart or lung causes, acute and chronic bronchitis, asthma attacks, and respiratory symptoms (Schwartz [2000]).

A significant correlation was observed between the number of deaths and the maximum monthly concentrations of PM_{10} in the atmosphere ($\rho = 0.45$). Imputation data were performed before fitting the model to handle the missing observations presented in the PM_{10} series. We used the multivariate imputation by chained equation method (MICE), proposed by van Buuren and Oudshoorn [2000].

Figure 3.1 presents the series of deaths caused by respiratory diseases and concentrations of PM_{10}. The number of deaths shows a positive trend and seasonal behavior. Furthermore, the PM_{10} concentration also presents a positive trend and three peaks. These aberrant observations can be considered additive outliers.

Figure 3.1: Time series of the number of deaths by respiratory diseases and concentrations of PM_{10} in the metropolitan area of Vitoria, Brazil.



Prepared by the author

The modeling considered a slightly positive trend in the number of deaths, and to handle the annual seasonality, sine and cosine functions were incorporated. The model is written as

$$\eta_t = \beta_1 x_{t,1} + \beta_2 \text{trend} + \beta_3 \sin(2\pi t/12) + \beta_4 \cos(2\pi t/12) + Z_t, \tag{3.25}$$

where t is the month number, $x_{t,1}$ is the PM_{10} concentrations and Z_t is the autoregressive structure of the GLARMA model.

The classic GLARMA Poisson (Davis et al. [2003]) and the robust approach proposed in Section 3.1 were adjusted, and their parameter estimation was compared. Table 9 presents the

estimates $\hat{\beta}_i$'s of the parameters β_i's in the classic model. All the estimates were significant at the 5% level of significance, except β_1, the coefficient related to the PM_{10} levels in the atmosphere.

Table 9: Parameter estimates of the classic GLARMA model fitted to the number of deaths caused by respiratory diseases.

	Intercept	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
Estimate	4.0571	0.0003	0.0034	-0.0573	-0.0954
Standard error	0.0383	0.0002	0.0006	0.0228	0.0228
p-value	<2e-16	0.0706	<1.9e-06	0.0121	2e-05

Prepared by the author

Table 10 presents the estimates of the robust GLARMA model. All the estimates were significant at the 5% level of significance.

Table 10: Parameter estimates of the robust GLARMA model fitted to the number of deaths by respiratory diseases.

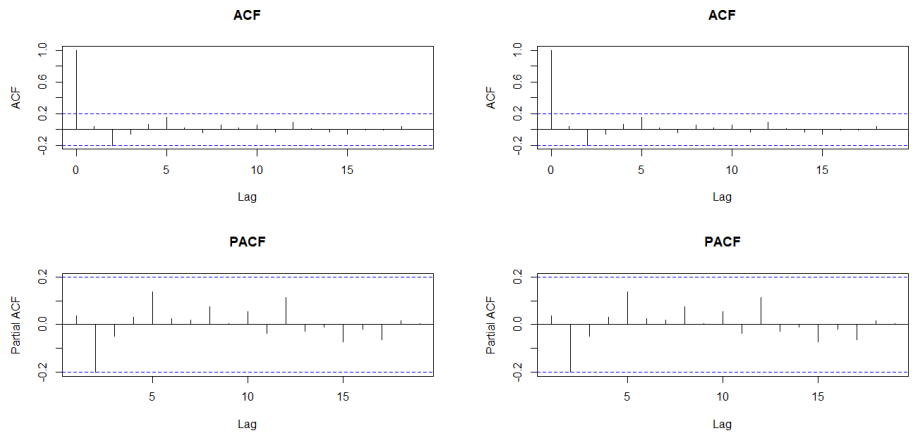
	Intercept	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
Estimate	4.0097	0.0009	0.0030	-0.0495	-0.0824
Standard error	0.0311	0.0001	0.0004	0.0177	0.0180
p-value	<2e-16	2e-07	1e-10	0.0053	4e-06

Prepared by the author

Figure 3.2 plots the sample autocorrelation function (ACF) and the sample partial autocorrelation function (PACF) of the residuals in the classic and robust GLARMA models. These plots show no difference with white noise, indicating a reasonable adjustment in both approaches.

Figure 3.2: Sample ACF and PACF of the residuals in the classic and robust GLARMA models.

(a) Classic GLARMA model. (b) Robust GLARMA model.



The parameter estimation in Tables 9 and 10 shows that although there is a significant correlation between PM_{10} concentrations and the monthly number of deaths in the period, in the

classic GLARMA model, the parameter related to the pollutant was not significant at 5% level of significance. However, in the robust approach, the impact of the PM_{10} was significant, which means that this pollutant contributes significantly to the increase in deaths caused by respiratory diseases. It is essential to observe that the value of parameter β_1 seems to be underestimated in the classic model ($\hat{\beta}_{1(classic)} = 0.0003$), once the robust estimate was three times this value ($\hat{\beta}_{1(robust)} = 0.0009$).

To evaluate the performance of the GLARMA robust proposal in this real data analysis, we removed the peaks in PM_{10} concentration measurements and replaced them with its mean value. Then the classic GLARMA model was adjusted to the series without the extreme values. The robust approach is expected to behave approximately the same as this model. The values of the parameter estimates in Table 11 are similar to those in Table 10, with emphasis on β_1, the main affected by the presence of outliers, confirming our expectation.

Table 11: Parameter estimates of the classic GLARMA model fitted to the number of deaths by respiratory diseases.

	Intercept	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
Estimate	4.0196	0.0008	0.0033	-0.0540	-0.0874
Standard error	0.0447	0.0003	0.0005	0.0220	0.0225
p-value	<2e-16	0.0274	2e-08	0.0143	0.0001

Prepared by the author

In the epidemiology context, the impact of air pollutants on human health is evaluated by relative risk (RR). This measure is an important information for the regulatory agencies to quantify the impact of these contaminants on the population's health. The RR of a variable $X_i = X_{i,t}$ is the change in the expected count of the response variable per ζ-unit change in the X_i, keeping the other covariates fixed.

For Poisson regression, the RR is given by

$$\hat{\text{RR}}_{X_i}(\zeta) = \exp(\hat{\beta}_i \zeta). \tag{3.26}$$

Table 12 presents the estimated RR and CIs for PM_{10}. The CIs were calculated with the bootstrap approach proposed by Camara et al. [2022] and the asymptotic approximation (equation (**??**)), with $\alpha = 5\%$. The $\hat{\text{RR}}$ was significant considering the classic and robust approaches (the value one does not belong to the intervals). In addition, the asymptotic and bootstrap CIs were equivalent, indicating that the classic model underestimated the relative risk. This result is in agreement with that observed in the simulation study.

Table 12: Estimated RR and 95% CI for PM_{10} in the classic and robust GLARMA models.

PM_{10}	Classic GLARMA	Robust GLARMA
$\widehat{RR}$	1.0187	1.0497
$\widehat{\text{CI}}$ asymptotic	[1.0001;1.0376]	[1.0305;1.0692]
$\widehat{\text{CI}}$ bootstrap	[1.0004;1.0379]	[1.0217;1.0794]

Prepared by the author

3.6 Conclusions

This work proposed a robust approach for the GLARMA model, introduced by Davis et al. [2003]. This methodology is based on the robustification of the quasi-likelihood function using M-estimator to control deviations on response variable and weight functions to limit leverage points on covariates.

The simulation study showed that additive outliers could widely affect the classic GLARMA. The robust proposal behaves approximately like the classical approach in the absence of outliers. At the same time, for contaminated data, the parameter estimation was almost as good as the classic method applied to clean observations.

The robust model was applied to the monthly number of deaths caused by respiratory diseases in Vitória, Brazil, to evaluate the impact of PM_{10} in the populations' health. This analysis showed that the RR is underestimated by the classic method, which means ignoring the impact of more than 100% of the exposure on the outcome. The numerical study agrees with this observation. The RR observed indicated that the PM_{10} contributed significantly to the increase of deaths by respiratory disease in the region.

Appendix 1

The derivative of constant $a(\boldsymbol{\delta})$ with respect to $\boldsymbol{\delta}$ is

$$a'(\boldsymbol{\delta}) = \frac{\partial a(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}} = \frac{1}{n}\sum_{t=1}^{n}\left[(\mathrm{A})(\mathrm{B}) + (\mathrm{C})(\mathrm{D})\right], \tag{3.27}$$

where

(A)

$$c\left\{\sum_{k=j_2+1}^{\infty}\frac{1}{k!}\left[e^{-e^{W_t}}(e^{W_t})^k\frac{\partial W_t(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}}(k-e^{W_t}))\right] - \sum_{m=0}^{j_1}\frac{1}{m!}\left[e^{-e^{W_t}}(e^{W_t})^m\frac{\partial W_t(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}}(m-e^{W_t}))\right]\right\}$$
$$+\left(\frac{1}{2}(e^{W_t})^{1/2}\frac{\partial W_t(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}}\right)\left[\left(\frac{e^{-\mu_t}\mu_t^{j_1}}{j_1!}\right) - \left(\frac{e^{-\mu_t}\mu_t^{j_2}}{j_2!}\right)\right]$$
$$+(e^{W_t})^{1/2}\left\{\frac{1}{j_1!}\left[e^{-e^{W_t}}(e^{W_t})^{j_1}\frac{\partial W_t(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}}(j_1-e^{W_t})\right] - \frac{1}{j_2!}\left[e^{-e^{W_t}}(e^{W_t})^{j_2}\frac{\partial W_t(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}}(j_2-e^{W_t})\right]\right\}$$

(B)

$$(e^{W_t})^{1/2}\frac{\partial W_t(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}}$$

(C)

$$c\left[\sum_{t=j_2+1}^{\infty}\frac{e^{-\mu_t}\mu_t^k}{k!} - \sum_{m=0}^{j_1}\frac{e^{-\mu_t}\mu_t^m}{m!}\right] + (e^{W_t})^{1/2}\left[\frac{e^{-\mu_t}\mu_t^{j_1}}{j_1!} - \frac{e^{-\mu_t}\mu_t^{j_2}}{j_2!}\right]$$

(D)

$$\frac{1}{2}(e^{W_t})^{1/2}\frac{\partial W_t(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}}\frac{\partial W_t(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}^T} + (e^{W_t})^{1/2}\frac{\partial^2 W_t(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}^T}$$

For $r_t \leq c$, the first derivative of $Q(\boldsymbol{\delta})$ with respect to $\boldsymbol{\delta}$ is given by

$$\begin{aligned}\frac{\partial Q(\delta)}{\partial \delta} &= \sum_{t=1}^{n}\left[\left(\frac{Y_t-\mu_t}{\mu_t^{1/2}}\right)\frac{1}{\mu_t^{1/2}}w(\mathbf{X}_t)\mu_t\frac{\partial W_t(\boldsymbol{\delta})}{\partial \delta} - a(\delta)\right]\\ &= \sum_{t=1}^{n}\left[(Y_t-\mu_t)\,w(\mathbf{X}_t)\frac{\partial W_t(\boldsymbol{\delta})}{\partial \delta} - a(\delta)\right].\end{aligned}$$

Thus, the second derivative of $Q(\boldsymbol{\delta})$ with respect to $\boldsymbol{\delta}$ is

$$\begin{aligned}\frac{\partial^2 Q(\delta)}{\partial \delta \partial \delta^T} &= \sum_{t=1}^{n}\left\{w(\mathbf{X}_t)\left[(Y_t-\mu_t)'\frac{\partial W_t(\boldsymbol{\delta})}{\partial \delta} + (Y_t-\mu_t)\frac{\partial^2 W_t(\boldsymbol{\delta})}{\partial \delta \delta^T}\right] - a'(\delta)\right\}\\ &= \sum_{t=1}^{n}\left\{w(\mathbf{X}_t)\left[-\mu_t\frac{\partial W_t(\boldsymbol{\delta})}{\partial \delta}\frac{\partial W_t(\boldsymbol{\delta})}{\partial \delta^T} + (Y_t-\mu_t)\frac{\partial^2 W_t(\boldsymbol{\delta})}{\partial \delta \delta^T}\right] - a'(\delta)\right\}\\ &= \sum_{t=1}^{n}\left\{(Y_t-\mu_t)\frac{\partial^2 W_t(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}\partial \boldsymbol{\delta}^T}w(\mathbf{X}_t) - \mu_t\frac{\partial W_t(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}}\frac{\partial W_t(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}^T}w(\mathbf{X}_t) - a'(\boldsymbol{\delta})\right\}.\end{aligned}$$

For $r_t > c$, the first derivative of $Q(\boldsymbol{\delta})$ with respect to $\boldsymbol{\delta}$ is given by

$$\begin{aligned}\frac{\partial Q(\delta)}{\partial \delta} &= \sum_{t=1}^{n}\left[c\text{sign}\left(\frac{Y_t-\mu_t}{\mu_t^{1/2}}\right)\frac{1}{\mu_t^{1/2}}w(\mathbf{X}_t)\mu_t\frac{\partial W_t(\boldsymbol{\delta})}{\partial \delta} - a(\delta)\right] \\ &= \sum_{t=1}^{n}\left[c\text{sign}\left(\frac{Y_t-\mu_t}{\mu_t^{1/2}}\right)\mu_t^{1/2}w(\mathbf{X}_t)\frac{\partial W_t(\boldsymbol{\delta})}{\partial \delta} - a(\delta)\right].\end{aligned}$$

Let $r_t = \frac{Y_t-\mu_t}{\mu_t^{1/2}}$, the $\frac{\partial^2 Q(\delta)}{\partial\delta\partial\delta^T}$ is given by

$$\sum_{t=1}^{n}\left(w(\mathbf{X}_t)\left\{\left[c\left(\frac{\partial}{\partial\delta}\text{sign}(r_t)\right)\left(\mu_t^{1/2}\frac{\partial W_t(\boldsymbol{\delta})}{\partial\delta}\right)\right] + c\text{sign}(r_t)\left[\frac{1}{2}\mu_t^{1/2}\frac{\partial W_t(\boldsymbol{\delta})}{\partial\delta}\frac{\partial W_t(\boldsymbol{\delta})}{\partial\delta^T} + \mu_t^{1/2}\frac{\partial^2 W_t(\boldsymbol{\delta})}{\partial\delta\partial\delta^T}\right]\right\}$$

where

$$\frac{\partial}{\partial\delta}\text{sign}(r_t) = 2\delta(r_t),$$

and $\delta(.)$ is the Dirac delta function. By definition, $\delta(r_t) = 0$ if $r_t \neq 0$. As $r_t > c, c > 0$, $\delta(r_t) = 0$. Then

$$\frac{\partial^2 Q(\delta)}{\partial\delta\partial\delta^T} = \sum_{t=1}^{n}\left\{w(\mathbf{X}_t)c\text{sign}(r_t)\left[\frac{1}{2}\mu_t^{1/2}\frac{\partial W_t(\boldsymbol{\delta})}{\partial\boldsymbol{\delta}}\frac{\partial W_t(\boldsymbol{\delta})}{\partial\boldsymbol{\delta}^T} + \mu_t^{1/2}\frac{\partial^2 W_t(\boldsymbol{\delta})}{\partial\boldsymbol{\delta}\partial\boldsymbol{\delta}^T}\right] - a'(\boldsymbol{\delta})\right\}.$$

Chapter 4

Conclusions and perspectives

This thesis focused on methodologies for counting time series using the GLARMA model. The main motivation for this work is the growing number of studies providing evidence of an association between concentration levels of air pollutants and mortality, morbidity, and hospital admissions mainly caused by respiratory and cardiovascular diseases. Due to the nature of the variables involved, analysis concerning this field must be realized with caution.

Chapter 2 investigates three bootstrap confidence interval approaches for relative risks calculated from the GLARMA model: the classic model-based, a procedure based on the model's specifications, the sieve bootstrap most used for real-valued process, and the INAR-type bootstrap based on the structure of the integer-valued autoregressive processes. A Monte Carlo study was performed under different scenarios and sample sizes. For large sample sizes, we verified that the model parameter related to the data's autocorrelation could influence the estimates and consequently affect the intervals' coverage. This effect is related to the complexity of the covariate, e.g., the impact is more significant as the complexity of the covariate increases. The INAR(1) bootstrap and the asymptotic approach generally presented better coverage rates than the other methods evaluated. The INAR(1) bootstrap for small sample sizes presented better coverages than the asymptotic approach, even considering covariates with complex autocorrelation structures. The preeminence of INAR(1) bootstrap compared to other procedures may be because this methodology considers the data distribution. In contrast, for the other proposals, only the residuals are considered. Davis et al. [1999] claim that using residuals from the adjustment in Poisson regression can underestimate the true serial dependence. The real data analysis evaluated the impact of air pollutants on the monthly number of chronic obstructive pulmonary disease (COPD) cases in Belo Horizonte, Brazil. The best fit was obtained with the GLARMA(2,0) model, and we computed the 95% confidence intervals using the asymptotic approach and the INAR(1) bootstrap. As observed in the empirical study, the intervals were pretty similar.

Chapter 3 proposes an approach to robustify the parameter estimation of the GLARMA model based on the Mallows' Quasi-likelihood Estimator (MQLE) given by Cantoni and Ronchetti [2001]. Based on the robustification of the quasi-likelihood function, this methodology controls deviations on the response variable using a M-estimator and leverages points on covariates using weight functions separately. A simulation study was performed to verify the

effect of additive outliers on the classic GLARMA model and the robust proposal. We found that the robust approach behaves similarly to the classical in the absence of perturbations. For contaminated data, the parameter estimation was almost as good as the classic method applied to clean observations. In all scenarios, the parameter estimation of the classic GLARMA was affected by the additive outliers, underestimating the parameter values. Due to the limitations regarding the asymptotic theory of the classic GLARMA model, asymptotic properties were not developed for the robust proposal in this paper. Nevertheless, the Monte Carlo study showed that asymptotic results corroborate that the estimators are consistent. A real data analysis was realized to evaluate the impact of particulate material (PM_{10}) on the monthly number of deaths caused by respiratory diseases in Vitoria, Brazil. The investigation revealed that the classic method underestimated the relative risk related to PM_{10} compared to the parameter values estimated by the robust proposal. This means ignoring the impact of more than 100% of the exposure on the outcome. The numerical study agrees with this observation. In addition, the analysis showed that the contaminant contributed significantly to the increase in deaths caused by respiratory diseases in the region.

Considering future research, some suggestions can be made. At first, we plan to construct the R code on the Mallows' Quasi-likelihood estimation for the GLARMA Poisson model (Appendix D) into an R package. In addition, a further problem would be to relax the Poisson assumption, expanding the model for other distributions of the exponential family, e.g., the Negative Binomial. Another interesting point regarding the robust proposal for the GLARMA model concerns the ψ function, used to limit outlying points on the response variable. Here, we used the Huber loss function. However, as addressed previously, some alternatives can also be used, e.g., Tukey's bisquare given by Beaton and Tukey [1974]. Regarding leverage points on the covariates, other weight functions can be considered. Such choices involve calculating the inverse of Mahalanobis distance using the minimum volume ellipsoid (MVE) introduced by Rousseeuw [1984] and the robust autocovariance estimator proposed by Ma and Genton [2000] in the context of time series. Considering the nature of the data, air pollutant concentrations are generally measured using some air quality stations. Due to this, an analysis considering spatial modeling could be very useful.

Lastly, regarding bootstrap confidence intervals, other procedures considering the GLARMA model structure can also be explored, e.g., conditional bootstrap resampling procedures (Figueiras et al. [2005]), which assume that the outcome in any observation is conditional to the covariate values.

Bibliography

M. Al-Osh and A. Alzaid. First-order integer-valued autoregressive (INAR(1)) processs. *J. Time Ser. Anal.*, 8(3):261–275, 1988.

A. Alzaid and M. Al-Osh. An integer-valued pth order autoregressive structure (INAR(p)) process. *Journal of Applied Probability*, 27(2):314–324, 1990.

J. Anderson, J. Thundiyil, and A. Stolbach. Clearing the air: a review of the effects of particulate matter air pollution on human health. *Journal of Medicine and Toxicology*, 8(2):166–175, 2012.

D. Andrews. A robust method for multiple linear regression. *Technometrics*, 16(4):523–531, 1974.

Z. Bai, C. Rao, and Y. Wu. M-estimation of multivariate linear 257 regression parameters under a convex discrepancy function. *Statistica Sinica*, 2(1), 1992.

S. Baldacci, S. Maio, S. Cerrai, G. Sarno, N. Baiz, M. Simoni, I. Annesi-Maesano, and G. Viegi. Allergy and asthma: Effects of the exposure to particulate matter and biological allergens. *Respir Med*, 109(9):1089–1104, 2015.

J. Ballesteros-Cánovas, D. Trappmann, J. Madrigal-González, N. Eckert, and M. Stoffel. Climate warming enhances snow avalanche risk in the western himalayas. *Proceedings of the National Academy of Sciences (PNAS)*, 115(13), 2018.

E. Barbera, C. Currò, and G. Valenti. A hyperbolic model for the effects of urbanization on air pollution. *Aplied Mathematical Modelling*, 34(8):2192–2202, 2010.

L. Baxter, S. Finch, F. Lipfert, and Q. Yu. Comparing estimates of the effects of air pollution on human mortality obtained using different regression methodologies. *Risk Anal.*, 17(3): 273–278, 1997.

A. Beaton and J. Tukey. The fitting of power series, meaning polynomials illustrated on band-spectroscopic data. *Technometrics*, 16(2):147–185, 1974.

M. Benjamin, R. Rigby, and D. Stasinopoulos. Generalized autoregressive moving average models. *J. Am. Stat. Assoc.*, 98(461):214–223, 2003.

C. Bergmeir and J. Benitez. On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191:192–213, 2012.

C. Bergmeir, R. Hyndman, and B. Koo. A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics and Data Analysis*, 120:70–83, 2018.

S. Borhan, P. Motevalian, J. Ultman, R. Bascom, and A. Borhan. A patient-specific model of reactive air pollutant uptake in proximal airways of the lung: Effect of tracheal deviation. *Applied Mathematical Modelling*, 91:52–73, 2021.

G. Box and G. Jenkins. *Time series analysis: forecasting and control*. Holden-Day, San Fracisco, 1976.

H. Brown and R. Prescott. *Applied Mixed Models in Medicine*. John Wiley Sons Ltd, England, 2006.

O. Bustos and V. Yohai. Robust estimates for arma models. *Journal of American Statistical Association*, 81:155–168, 1986.

P. Bühlmann. Sieve bootstrap for time series. *Bernoulli*, 3(2):123–148, 1997.

A. J. Camara, G. Franco, V. Reisen, and P. Bondon. Generalized additive model for count time series: An application to quantify the impact of air pollutants on human health. *Pesquisa Operacional*, 41:e241120, 2021.

A. J. Camara, V. Reisen, G. Franco, and P. Bondon. Glarma model and bootstrap approaches: An application to respiratory diseases and air pollutants. *In submission process*, 2022.

E. Cantoni and E. Ronchetti. Robust inference for generalized linear models. *J. Am. Stat. Assoc.*, 96(455):1022–1030, 2001.

M. Cardinal, R. Roy, and J. Lambert. On the application of integer-valued time series models for the analysis of disease incidence. *Statistic Medicine*, 18(15):2025–2039, 1999.

R. Carroll and A. Welsh. A note on asymmetry and robustness in linear regression. *American Statistician*, 42:285–287, 1986.

I. Chang, G. Tiao, and C. Chen. Estimation of time series parameters in the presence of outliers. *Technometrics*, 30(2):193–204, 1988.

C. Chen and L.-M. Liu. Joint estimation of model parameters and outlier effects in time series. *J. Am. Stat. Assoc.*, 88(421):284–297, 1993.

R. Chen, C. C., J. Tan, J. Cao, W. Song, X. Xu, C. Jiang, M. W., C. Yang, B. Chen, Y. Gui, and H. Kan. Ambient air pollution and hospital admission in shanghai, china. *J. Hazard. Mater.*, 181(1):234–240, 2010.

E. Chi. M-estimation in cross-over trials. *Biometrics*, 50(2):486–493, 1994.

R. Cleveland, W. Cleveland, J. McRae, and I. Terpenning. Stl: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics*, 6:3–73, 1990.

H. Cotta, V. Reisen, P. Bondon, and P. Prezotti. Identification of redundant air quality monitoring stations using robust principal component analysis. *Environmental Modelling & Assessment*, 25:521–530, 2020.

D. Cox. Statistical analysis of time series: Some recent developments. *Scandinavian Journal of Statistics*, 8:93–115, 1981.

J. D., W. Silny, A. Danczak-pazdrowska, A. Polanska, A. Osmola-mankowska, and K. Olek-hrab. Annals of agricultural and environmental medicine. 19:475–481, 2012.

I. Danilevicz, P. Bondon, V. Reisen, and F. Serpa. A longitudinal study of the influence of air pollutants on children. a robust multivariate approach. *Accept to publish: Journal of Applied Statistics*, 2023.

R. Davis, Y. Wang, and W. Dunsmuir. *Modelling Time Series of Count Data. In: S. Ghosh (ed) Asymptotics, Nonparametrics, and Time Series*. CRC Press, New York, 1999.

R. Davis, W. Dunsmuir, and S. Streett. Observation driven models for Poisson counts. *Biometrika*, 90(4):777–790, 2003.

R. Davis, W. Dunsmuir, and S. Streett. Maximum likelihood estimation for an observation driven model for Poisson counts. *Methodology and Computing in Applied Probability*, 7(2):149–159, 2005.

R. Davis, K. Fokianos, S. Holan, and H. Joe. Count time series: A methodological review. *Journal of the American Statistical Association*, 116(535):1533–1547, 2021.

R. A. K. Debrah. The ella roberta foundation. https://ellaroberta.org/, 2022. [Online; accessed 30-December-2022].

J. Dennis and R. Welsch. Techniques for nonlinear least squares and robust regression. *Communications in Statistics-simulation and Computation*, 7(4):345–459, 1978.

D. Dockery. *Outdoor Air Pollution*. Children's Environmental Health, Oxford, New York, 2014.

D. Dockery and C. Pope. *Epidemiology of Acute Health Effects: Summary of time series study. In Richard Wilson and John Spengler, eds., Particles in our air.* Harvard University Press, Cambridge, MA, 1996.

J.-G. Du and Y. Li. The integer-valued autoregressive (INAR(p)) model. *Journal of Time Series Analysis*, 12(2):129–142, 1991.

W. Dunsmuir. *from: Handbook of discrete-valued time series, Handbook of Modern Statistical Methods*. CRC Press, London, 2015.

B. Efron. Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

B. Efron and R. Tibishirani. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 1(1):54–75, 1986.

R. Ferland, A. Latour, and D. Oraichi. Integer-valued garch process. *Journal of Time Series Analysis*, 27(6):923–942, 2006.

A. Figueiras, J. Roca-Pardinas, and C. Cadarso-Suárez. A bootstrap method to avoid the effect of concurvity in generalised additive models in time series studies of air pollution. *Journal of Epidemiology Community Health*, 59:881–884, 2005.

K. Fokianos and J. Tjosthein. Log-linear poisson autoregression. *Journal of Multivariate Analysis*, 102(3):563–578, 2011.

A. Fox. Outliers in time series. *Journal of the Royal Statistical Society: Series B*, 34(3):350–363, 1972.

J. Franke and J. Kreiss. Bootstrapping stationary autoregressive moving average models. *Journal of Time Series Analysis*, 13(4):297–317, 1992.

D. Freedman. On bootstrapping two-stage least-squares estimates in stationary linear models. *Annals of Statistics*, 12(3):827–842, 1984.

C. Froes, V. Camara, P. Landrigan, and L. Claudio. Systematic review of children's environmental health in brazil. *Annals of Global Health*, 82(1):132–148, 2016.

D. Gamerman, T. Santos, and G. Franco. A non-gaussian family of state-space models with exact marginal likelihood. *Journal of Time Series Analysis*, 34(6):625–645, 2013.

A. Gowers, P. Cullinan, J. Ayres, H. Anderson, D. Strachan, S. Holgate, I. Mills, and R. Maynard. Does outdoor air pollution induce new cases of asthma? biological plausibility and evidence; a review. *Respirology*, 17(6):887–898, 2012.

R. Gruchalla, J. Pongracic, M. Plaut, R. Evans, C. Visness, and M. Walter. Inner city asthma study: relationships among sensitivity, allergen exposure, and asthma morbidity. *Journal of Allergy and Clinical Immunology*, (115):478, 2005.

M. Guarnieri and J. Balmes. Air pollution and asthma. *Lancet*, 3(383):1581–1591, 2014.

F. Hampel. Contributions to the theory of robust estimation. *Ph.D. thesis, University of California, Berkeley*, 1968.

F. Hampel. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393, 1974.

F. Hampel, E. Ronchetti, P. Rousseeuw, and W. Stahel. *Robust Statistics: The Approach Based on Influence Functions*. New York: John Wiley Sons, Inc, 1986.

A. Harvey and C. Fernandes. Time series models for count or qualitative observations. *Journal of Business and Economic Statistics*, 7(4):407–417, 1989.

T. Hastie and R. Tibshirani. *Generalized additive models*. Chapman & Hall, London, 1990.

A. Heinen. Modelling time series count data: An autoregressive conditional Poisson model. *Munich Personal RePEc Archive.*, 2003.

A. Hernández-Garcés, R. Cécé, A. Ferrer-Hernández, D. Bernard, U. Jáuregui-Haza, N. Zahibo, and J. González. Intercomparison of flexpart and calpuff dispersion models. an application over a small tropical island. 2020.

T. Hertel, H. Lee, S. Rose, and B. Sohngen. Modelling land use related greenhouse gas sources and sinks and their mitigation potential. *Economic Analysis of land use in global climate change policy*, 2009.

P. Huber. Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35(1): 73–101, 1964.

P. Huber. *Robust Statistics*. Wiley, New York, 1981.

W. Härdle, S. Huet, E. Mammen, and S. Sperlich. Bootstrap inference in semiparametric generalized additive models. *Econometric Theory*, 20(2):265–300, 2004.

M. Ispany, V. Reisen, G. Franco, P. Bondon, H. Cotta, P. Prezotti, and F. Serpa. *On Generalized Additive Models with Dependent Time Series Covariates. In: Rojas I., Pomares H., Valenzuela O. (eds) Time Series Analysis and Forecasting. ITISE 2017. Contributions to Statistics*. Springer, Cham, 2018.

C. Jentsch and C. Weiss. Bootstrapping INAR models. *Bernoulli*, 25(3):2359–2408, 2019.

R. C. Jung and A. R. Tremayne. Useful models for time series of counts or simply wrong ones? *Advances in Statistical Analysis*, 95:59–91, 2011.

R. C. Jung, M. Kukuk, and R. Liesenfeld. Time series of count data: Modelling, estimation and diagnostics. *Computational Statistics and Data Analysis*, 51:2350–2364, 2006.

S. Karami, M. Karami, G. Roshanaei, and H. Farsan. Association between increased air pollution and mortality from respiratory and cardiac diseases in tehran: Application of the glarma model. *Iranian Journal of Epidemiology*, 12(4):36–43, 2017.

H.-Y. Kim and Y. Park. Bootstrap confidence intervals for the INAR(p) process. *The Korean Communications in Statistics*, 13(2):343–358, 2006.

H.-Y. Kim and Y. Park. A non-stationary integer-valued autoregressive model. *Statistical Papers*, 49(485), 2008.

J.-Y. Kim, H.-Y. Kim, D. Park, and Y. Chung. Modelling of fault in rpm using the glarma and ingarch model. *Eletronics Letters*, 54(5):297–299, 2018.

S. Kitromilidou and K. Fokianos. Robust estimation methods for a class of log-linear count time series models. *Journal of Statistical Computation and Simulation*, 86(4):740–755, 2016.

J.-P. Kreiss, E. Paparoditis, and D. Politis. On the range of validity of the autoregressive sieve bootstrap. *Annals of Statistics*, 39(4):2103–2130, 2011.

H. Künsch. Infinitesimal robustness for autoregressive processes. *Annals of Statistics*, 12: 843–863, 1984.

H. Künsch. The jackknife and the bootstrap for general stationary observations. *Annals of Statistics*, 17(3):1217–1241, 1989.

H. Künsch, L. Stefanski, and R. Carroll. Conditionally unbiased bounded-influence estimation in general regression models with applications to generalized linear models. *Journal of the American Statistical Association*, 84:460–466, 1989.

N. Laird and J. Ware. Random-effects models for longitudinal data. *Biometrics*, 1982.

J. Ledolter. The effect of additive outliers on the forecast from arima models. *International Journal of Forecasting,*, 5:231–240, 1989.

T. Li. Laplace periodogram for time series analysis. *Journal of the American Statistical Association*, 103:757–768, 2008.

M. Lippmann. Toxicological and epidemiological studies of cardiovascular effects of ambient air fine particulate matter (pm2.5) and its chemical components: Coherence and public health implications. *Critical Reviews in Toxicology*, 44(4):299–347, 2014.

S. Lo and E. Ronchetti. Robust and accurate inference for generalized linear models. *Journal of Multivariate Analysis*, 100:2126–2136, 2009.

Y. Ma and M. Genton. Highly robust estimation of the autocovariance function. *Journal of time series analysis*, 2000.

C. Mallows. On some topics in robustness. *Technical Memorandum, Murray Hill, N.J., Bell Telephone Laboratories*, 1975.

R. Maronna, R. Martin, and V. Yohai. *Robust Statistics: Theory and Methods*. John Wiley & Sons, Chichester, 2006.

R. Martin and V. Yohai. *Robust identification of autoregressive moving average models. In Handbook of Statistics 5 (eds E. J. Hannan, P. R. Krishnaiah and M. M. Rao)*. Elsevier Science Publishers, Amsterdam, 1985.

P. McCullagh and J. Nelder. *Generalized Linear Models*. Chapman and Hall, London, 1989.

M. McGeehin and M. Mirabelli. The potential impacts of climate variability and change on temperature-related morbidity and mortality in the united states. *Environmental Health Perspect*, 109(2):185–189, 2001.

E. McKenzie. Some simple models for discrete variate time series. *Journal of the American Water Resources Association*, 21(4):645–650, 1985.

G. Molenberghs and G. Verbeke. A review on linear mixed models for longitudinal data, possibly subject to dropout. *Statistical Modelling*, 1:235–269, 2001.

A. Nascimento, J. Santos, J. Mill, J. Souza, N. Reis, and V. Reisen. Association between the concentration of fine particles in the atmosphere and acute respiratory diseases in children. *Rev Saude Publica*, (51):1–10, 2017.

A. Nascimento, J. Santos, J. Mill, and T. De. Association between the incidence of acute respiratory diseases in children and ambient concentrations of so2, pm10 and chemical elements in fine particles. *Environmental Research*, 2020.

J. Nelder and R. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society: Series A*, 135(3):370–384, 1972a.

J. Nelder and R. Wedderburn. Generalized linear model. *Journal of the Royal Statistical Society: Series A*, 135(3):370–384, 1972b.

B. Ostro, G. Eskeland, J. Sánchez, and T. Feyzioglu. Air pollution and health effects: A study of medical visits among children in santiago, chile. *Environmental Health Perspectives*, 107(1): 69–73, 1999.

B. Ostro, L. Roth, B. Malig, and M. Marty. he effects of fine particle components on respiratory hospital admissions in children. *Environmental Health Perspectives*, 117(3):475–480, 2009.

K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(11):559–572, 1901.

E. Peitzsch, G. Pederson, K. Birkeland, J. Hendrikx, and D. Fagre. Climate drivers of large magnitude snow avalanche years in the u.s. northern rocky mountains. *Scientific Reports*, 11 (10032), 2021.

C. Pope and D. Dockery. Health effects of fine particulate air pollution: lines that connect. *Journal of the Air Waste Management Association*, 23(2):307–318, 2018.

C. Pope, M. Thun, M. Namboodiri, D. Dockery, J. Evans, F. Speizer, and C. Heath. Particulate air pollution as a predictor of mortality in a prospective study of u.s. adults. *American Journal Respiratory Critical Care Medicine*, 151:669–674, 1995.

J. Preisser and B. Qaqish. Robust regression for clustered data with applications to binary regressio. *Biometrics*, 55:574–579, 1999.

V. Reisen, C. Levy-Leduc, and M. Taqqu. An m-estimator for the long-memory parameter. *Journal of Statistical Planning and Inference*, 187:44–55, 2017.

V. Reisen, A. Sgrâncio, C. Lévy-Leduc, P. Bondon, F. Ziegelmann, E. Z. Monte, and H. Cotta. Robust factor modeling for high-dimensional time series: an application to air pollution data. *Applied Mathematics and Computation*, 346:842–852, 2019.

J. Roderique, C. Josef, M. Feldman, and B. Spiess. A modern literature review of carbon monoxide poisoning theories, therapies, and potential targets for therapy advancement. *Toxicology*, 6 (334):45–58, 2015.

P. Rousseeuw. Least median of squares regression. *Journal of the American Statistical Association*, 79(388):871–880, 1984.

P. Rousseeuw. *Multivariate Estimation With High Breakdown Point. In: W. Grossmann, G. Pflug, I. Vincze and W. Wertz (eds) Mathematical Statistics and Applications.* Reidel Publishing Company, Dordrecht, 1985.

A. Ruckstuhl and A. Welsh. Robust fitting of the binomial model. *CMA, Australian National University, manuscript*, 1999.

T. Rydberg and N. Shephard. Dynamics of trade-by-trade price movements:decomposition and models. *Journal of Financial Econometrics*, 1(1):2–25, 2003.

A. Sarnaglia, V. Reisen, C. Levy-Leduc, and P. Bondon. M-regression spectral estimator for periodic arma models. an empirical investigation. *Stochastic Environmental Research and Risk Assessment*, 35:653–664, 2021.

J. Schwartz. Harvesting and long-term exposure effects in the relationship between air pollution and mortality. *Am. J. Epidemiol.*, 151(5):440–448, 2000.

J. Scire, D. Strimaitis, and R. Yamartino. *A user's guide for the CALPUFF dispersion model*. Earth Tech, Concord MA, USA, 2000.

F. Serpa, E. Zandonade, and J. Reis. Modelo linear misto com interações e componentes principais para avaliar o efeito múltiplo de poluentes e variáveis climáticas na saúde respiratória. *Ph.d., Universidade Federal do Espírito Santo*, 2019.

D. Simpson, D. Ruppert, and R. Carroll. On one-step gm estimates and stability of inferences in linear regression. *On One-Step GM Estimates and Stability of Inferences in Linear Regression*, 87(418):439–450, 1992.

J. Souza, V. Reisen, G. Franco, M. Ispány, P. Bondon, and J. Santos. 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*, 67(2):453–480, 2018.

K. Spann, N. Snape, E. Baturcam, and E. Fantino. The impact of early-life exposure to air-borne environmental insults on the function of the airway epithelium in asthma. *Annals of Global Health*, 82(1):28–40, 2016.

J. Streett, R. Carroll, and D. Ruppert. A note on computing robust regression estimates via iterative reweighted least squares. *Journal of the American Statistical Association*, 42(2): 152–154, 1988.

G. Trasande, L.and Thurston. Outliers, level shifts, and variance changes in time series. *Journal of Allergy and Clinic Immunology*, 115:689–699, 2005.

J. Tukey. *A survey of sampling from contaminated distributions. In Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling (I. Olkin et al., eds.)*. Stanford Univ. Press, 1960.

M. Valdora and V. Yohai. Robust estimators for generalized linear models. *Journal of Statistical Planning and Inference*, 146:31–48, 2014.

S. van Buuren and K. Oudshoorn. Multivariate imputation by chained equations:mice v1.0 user's manual. *TNO Prevention and Health*, 2000.

G. Verbeke and G. Molenberghs. *Linear Mixed Models for Longitudinal Data*. Springer-Verlag, New York, USA, 2000.

P. Villeneuve, G. Burnett, D. Aronov, S. Y., D. Krewski, M. Goldberg, C. Hertzman, and J. Brook. A time-series study of air pollution, socioeconomic status, and mortality in vancouver, canada. *Journal of exposure analysis and environmental epidemiology*, 13(6):427–435, 2003.

Y. Wang and H. Pham. Analyzing the effects of air pollution and mortality by generalized additive models with robust principal components. *International Journal of System Assurance Engineering and Management*, 2:253–259, 2011.

W. H. O. (WHO). *Air Quality Guidelines. Particulate Matter, Ozone, Nitrogen Dioxide and Sulphur Dioxide*. Global Update 2005. Summary of Risk Assessment. Geneva, Switzerland, 2006.

W. Wu. M-estimation of linear models with dependent errors. *The Annals of Statistics*, 35(2): 495–521, 2007.

B. Zamprogno, V. Reisen, P. Bondon, H. Cotta, and N. Reis Jr. Principal component analysis with autocorrelated data. *Journal of Statistical Computation and Simulation*, 90(12):2117–2135, 2020.

S. L. Zeger and B. Qaqish. Markov regression models for time series: A quasi-likelihood approach. *Biometrics*, 44(4):1019–1031, 1988.

Appendix A

AsmaVix Project: A longitudinal study

A.1 Introduction

Industrial development and economic growth in the last decades were responsible for a considerable increase in air pollutant emissions. The populational concentration in urban areas contributed to high human exposition to gases and particulate material, which led to a high prevalence of chronic respiratory diseases, e.g., asthma (Gruchalla et al. [2005],Trasande [2005], Gowers et al. [2012] and D. et al. [2012]). Air pollution causes harmful effects on human health even when pollutant levels are within the standards established by regulatory agencies and the World Health Organization (WHO), see Pope and Dockery [2018]and Lippmann [2014]. These effects range from physiological alterations to outcomes of more significant impact, such as death ([WHO]). Due to this, the air quality became a public health emergency.

Air pollution can affect human health from its birth until the elderly. In the last decades, many authors have been dedicated to studying the impact of air quality variables on health. The primary sources of atmospheric pollutants in urban areas are industry and motor vehicles, which are clearly involved in the increase in hospitalizations and deaths by respiratory diseases. Pope et al. [1995], Dockery and Pope [1996], Villeneuve et al. [2003], and others indicated a positive association between mortality and particulate material (PM). Ostro et al. [1999], Schwartz [2000], and Chen et al. [2010] found a significant association between daily air pollutant concentration levels and hospital admission for respiratory and cardiovascular diseases. McGeehin and Mirabelli [2001], Ostro et al. [2009] and Hertel et al. [2009] analyzed temperature effects on mortality in USA and Germany. Those most susceptible to the effects of pollutants and climate variations are children, the elderly, and those with chronic diseases, especially cardiovascular and respiratory diseases.

Ozone (O_3), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), carbon monoxide (CO), and particulate matter (PM) are the main pollutants in the atmosphere. The PM is a heterogeneous and complex mixture of particles that vary in size, weight, shape, chemical composition, solubility, and origin ([WHO]). The particles are classified according to their diameter as ultrafine (particles with a diameter smaller than 0.1μm - $PM_{0.1}$), fine (particles with a diameter between 0.1 and

2.5μm - $PM_{2.5}$) and coarse (particles with a diameter between 2.5 and 10μm – PM_{10}). The $PM_{0.1}$ and $PM_{2.5}$ reach the lower respiratory system and can cause or aggravate respiratory diseases. The PM_{10} does not affect the lower respiratory tract. Still, it presents the most consistent association with premature mortality, increased hospital admissions for heart or lung causes, acute and chronic bronchitis, asthma attacks, and respiratory symptoms (Schwartz [2000]). The other pollutants, O_3, NO_2, SO_2 and carbon monoxide CO, when inhaled, trigger a series of biological processes that compromise the perfect functioning of the respiratory system. Air pollution is a complex mixture of pollutants and other toxic and non-toxic chemical compounds. Consequently, the health effects derive from this mixture (Dockery [2014]) and can not be fully explained by only one contaminant. In this context, some authors have been studying and proposing multivariate approaches for the analysis (see, e.g., Souza et al. [2018] and Ispany et al. [2018]).

The linear regression model is the canonical procedure to evaluate the effect of multiple covariates simultaneously. This method is generally based on the assumption that the errors are independently and identically distributed (i.i.d) and the effects (covariates) are fixed, i.e., random variables with probability distribution are not considered. Therefore, the variance of the response variable in the regression model only depends on the variance of the errors. These assumptions make the classic regression models very limited in their application in different areas of knowledge, especially when the covariates have a temporal correlation and the measured observations of the response variable are serial. By accommodating fixed and random effects, mixed models (MM) are important approaches for the statistical modeling of correlated data in a linear and non-linear way, as they allow relaxing the hypothesis of independence of the variables involved and taking into account more flexibly, complicated data structures, see Laird and Ware [1982]. For example, repeated measurements were taken on a patient. In that case, i.e., longitudinal observations, MM allows specifying a standard for the correlation between these measurements, i.e., the patient is a random variable with a specified probability distribution. Additionally, the MM can model the covariance structure of the data, which means that the method can recognize the relationship between serial observations in the same unit.

The characteristics of the variables studied, that is, pollutant concentrations and lung function data, require a statistical model that allows the analysis of longitudinal data and that incorporates the dependence and correlation structure of errors like the LMM. The class of mixed models is very broad, an introduction to several aspects of MM applications in health problems can be found in Chapter 1 of the book Applied Mixed Models in Medicine (Brown and Prescott [2006]).

In Brazil, studies relating to the effects of air pollution on health are relatively deficient. In the same line of study, it is worth mentioning the works of Froes et al. [2016], Serpa et al. [2019], and Nascimento et al. [2017, 2020]. The AsmaVix is a multidisciplinary team of medical and air quality researchers focus on finding the correlation between air pollution exposure and asthma symptoms frequency and severity in children and adolescents in the Great Vitoria Region

(GVR), a heavily industrialized region with a high population density in the southeast of Brazil. The group composed of medical researchers is responsible for the health monitoring of the participants in the study to obtain longitudinal data necessary for the search for possible causal associations. Data collection is made directly in contact with individuals and their parents or guardians. The air quality group is responsible for monitoring and modeling of air pollutants of interest (PM_{10}, $PM_{2.5}$, CO, SO_2, O_3, and NO_X) to which the investigated population is exposed.

Different methods of exposure calculation are used in the project. Here we will focus on two of them: 1) the air quality automatic monitoring network (AQAMN) of GVR composed of eight monitoring stations, two in Serra (Laranjeiras and Carapina), three in Vitoria city (Jardim Camburi, Praia do Suá and Vix-Centro), two in Vila Velha (VV-Centro and Ibes), and one in Cariacica.. 2) An indirect method in which the concentration values are obtained using a dispersion model instead of measurements. The modeling was performed using the California Puff Model (CALPUFF), see Scire et al. [2000].

The medical parameters analyzed were the Forced Expiratory Volume in the first second (FEV1) and Peak Expiratory Flow (PEF). The FEV1 is the amount of air eliminated in the first second of forced expiration. It is the most clinically useful measure of lung function and allows for assessing the degree of airflow obstruction. The 80% of the predicted value is considered the lower limit of normal pulmonary function. Thus, the percentage value of FEV1 concerning the predicted value (FEV1$\%$) is the parameter used to assess the relationship between pulmonary function and concentrations of atmospheric pollutants. The PEF is the maximal rate a person can exhale during a short maximal expiratory effort after a full inspiration. In patients with asthma, the PEF percent predicted correlates reasonably well with the percent predicted value for the forced expiratory volume in one second (FEV1) and provides an objective measure of airflow limitation when spirometry is not available. The response variables FEV1 and PEF were measured daily from November 2019 to February 2020. Other noise covariates, such as age and gender, were also considered.

This study is organized as follows. Section A.2 presents the Linear Mixed Model, and real data analysis is performed in Section A.3, considering two distinct sources for the concentrations of air pollutants. Section A.3 presents the final considerations.

A.2 Linear Mixed Model

Mixed models are extensions of fixed effects models (classic regression models) once they include random effects, random coefficients, and/or covariates in the error variance matrix. For the linear mixed model (LMM), let $\boldsymbol{Y}_i = (Y_{i1}, ..., Y_{in_i})^T$ denotes the vector of repeated measures for subject i, where $i = 1, ..., m$. Molenberghs and Verbeke [2001] defines the general

LMM as

$$Y_i = X_i\boldsymbol{\beta} + Z_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i, \tag{A.1}$$

where $\boldsymbol{\beta} = (\beta_0, ..., \beta_{p-1})^T$ is a vector of p regression coefficients, called fixed effects, and $\boldsymbol{b}_i$ is the vector of q unobserved random effects. It is assumed that $\boldsymbol{b}_i \sim \mathrm{N}(\mathbf{0}, \boldsymbol{D})$, where $\boldsymbol{D}$ is the (q x q) covariance matrix. The $\boldsymbol{b}_i$ describes how the change of the ith subject deviates from the average change in the population. The design matrices $\boldsymbol{X}_i = [X_{i1}^T, ..., X_{in_i}^T]^T$ and $\boldsymbol{Z}_i = [Z_{i1}^T, ..., Z_{in_i}^T]^T$ are respectively the (n_i x p) and (n_i x q) matrices of subject i for fixed and random effects. The residual component $\boldsymbol{\varepsilon}_i$ are assumed to be independent $\mathrm{N}(\mathbf{0}, \sigma_e^2\boldsymbol{\Omega}_i)$. When $\boldsymbol{\Omega}_i$ is the identity matrix the components of $\boldsymbol{\varepsilon}_i$ are independent. However, it is also usually assumed that the residuals present time dependency, following an autoregressive process of order 1 (AR(1)). The vector $\boldsymbol{Y}_i$ is normally distributed with mean $\boldsymbol{X}_i\beta$ and covariance matrix $\sum_i = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^T + \sigma_e^2\Omega_i$.

In the context of this study, the vector $\boldsymbol{Y}_i$ is related to the measures of PEF and VEF1, and X_i is the matrix of covariates that represents the concentration of atmospheric pollutants. According to Verbeke and Molenberghs [2000] for each measurement Y_{ij} of $\boldsymbol{Y}_i$, where $j = 1, ..., n_i$, we can write

$$Y_{ij} = X_{ij}\boldsymbol{\beta} + Z_{ij}\boldsymbol{b}_i + \varepsilon_{ij}, \tag{A.2}$$

where X_{ij} and Z_{ij} are respectively (1 x p) and (1 x q) design vectors for the fixed and random effects.

A.3 Real data analysis

Real data analysis was realized to evaluate the impact of pollutants on the Peak Expiratory Flow (PEF) and Forced Expiratory Volume in the first second (FEV1) on children and adolescents in Vitoria, Brazil, between the days 11/02/2019 and 02/19/2020. Information related to age, gender, and pre-existing asthma were collected, besides days and shifts in which the measures of VEF1 and PEF were recorded. Daily concentrations of $PM_{2.5}$, PM_{10}, SO_2, CO, NO_2, NO_X were observed. Environmental variables were collected from the IEMA, the State Environment and Water Resources Institute, and the CALPUFF modeling system, a system for numerical simulation of air pollutants dispersion. Epidemiological variables were obtained by the AsmaVix project. Missing data were evaluated using multivariate imputation by chained equation method (MICE), proposed by van Buuren and Oudshoorn [2000] and Seasonally Decomposed Missing Value Imputation (Cleveland et al. [1990]). The VEF1 and PEF were used as response variables, while air quality variables were considered covariates in the Linear Mixed Model. The longitudi-

nal study involved 62 children and adolescents (27 girls and 35 boys) aged 8 to 14, inhabitants of Andorinhas, Maruípe, Enseada do Suá, and Maria Ortiz neighborhoods.

A.3.1 IEMA

For the IEMA data source were considered four air quality monitoring stations: Carapina, Jardim Cambruri, Enseada do Suá and Vitória Centro. Environmental information refers to the daily mean, minimum, and maximum two days before the spirometry test. The $PM_{2.5}$ concentrations were collected only from Enseada do Suá once the quality of measures in other stations was very poor. The Pearson correlation was calculated between each air pollutant and the response variables to verify which presented a significant association. Although many environmental covariates have been considered, only concentration levels of PM_{10} and SO_2, from Enseada do Suá station, presented a significative association with the PEF of the children in the study. No environmental variable of any station showed a significative association with the VEF1. Due to the nature of the variables, robust methods should be considered as addressed by Reisen et al. [2019], Cotta et al. [2020], and Danilevicz et al. [2023]. This last proposes a robust method for the LMM model. However, the use of robust techniques for this data did not provide more reliable results.

Table 1: Correlation among pollutants and VEF1

	VEF1	**PM10**	**SO2**
VEF1	1.000	-0.019	-0.015
PM10	-0.019	1.000	0.327
SO2	-0.015	0.327	1.000

Prepared by the author

Table 2: Correlation among pollutants and PEF

	PEF	**PM10**	**SO2**
PEF	1.000	-0.044	-0.075
PM10	-0.044	1.000	0.327
SO2	-0.075	0.327	1.000

Prepared by the author

Table 3: Descriptive statistics for the response variables in study

	Min.	1st qu.	Median	Mean	3rd qu.	Max.
VEF1	0.430	1.460	1.730	1.853	2.120	4.060
PEF	59.0	210.0	254.0	273.5	313.0	627.0

Prepared by the author

A.3.1.1 PM10

We verified an expressive effect of the maximum daily concentration of PM_{10} in Enseada do Suá station on the decrease of the peak expiratory flow of the patients in the study (Pearson correlation: $\rho = -0.070$, p-value: 0.0085).

Table 4 presents the point and interval estimates for the parameters of the LMM applied to gender and the maximum daily concentration of PM_{10} as covariates and PEF as the response variable. The random effect was the day, as the children and adolescents in the study were not evaluated on the same day. The AIC observed was 22212.84. The PEF of the patients decreased 0.112(verificar a medida) when the PM_{10} concentration increase $1\mu g/m^3$. Note that, although the concentration of PM_{10} was significant at the level of 10%, the confidence interval contains the value zero.

Table 4: Point and interval estimates for the parameters of the LMM - PM_{10}

Random: Day	Coefficient	Std. Error	p-value	Confidence Interval
Intercept	292.632	4.717	0.000	[283.379;301.884]
Gender (G)	-29.061	4.435	0.000	[-37.759;-20.363]
PM10	-0.112	0.065	0.085	[-0.240;0.015]

Prepared by the author

A.3.1.2 SO2

Table 5 presents the point and interval estimates for the parameters of the LMM applied to gender and the minimum daily concentration of SO_2 as covariates and PEF as the response variable. Again, the random effect was the day. We observed that the PEF decreased by 7.771 (ver a medida) as the concentration of SO_2 in the atmosphere increased $1\mu g/m^3$. Regarding gender, we note that the PEF decreased for the girls in the research, as boys are the reference category in the model. The value zero does not belong to any of the confidence intervals suggesting the significance of these variables. The AIC was 22199.44.

Table 5: Point and interval estimates for the parameters of the LMM - SO_2 - random effect: Day

Random: Day	Coefficient	Std. Error	p-value	Confidence Interval
Intercept	295.950	4.351	0.000	[287.416;304.484]
Gender (G)	-28.671	4.430	0.000	[-37.360;-19.981]
SO2	-7.771	2.588	0.002	[-12.849;-2.694]

Prepared by the author

This work also wanted to verify if there is any difference related to the effect of SO_2 in children with and without asthma. XX patients with asthma were observed, while YY did not present this health condition. Table 6 shows that the impact of the pollutant on PEF is also significant when the random effect of the LMM model is asthma. Due to this, we realized a second analysis considering subsets composed of children with and without the disease.

Table 6: Point and interval estimates for the parameters of the LMM - SO_2 - random effect: Asthma

Random: Ashtma	Coefficient	Std. Error	p-value	Confidence Interval
Intercept	307.594	23.272	0.000	[261.950;353.239]
Gender (G)	-37.587	4.513	0.000	[-46.439;-28.735]
SO2	-5.247	2.602	0.043	[-10.351;-0.142]

Prepared by the author

Table 7 presents the point and interval estimates for the parameters of the model, considering only patients with asthma. Again we observed that the covariates gender and the contaminant SO_2 were significant at 5% level, with the zero value out of all confidence intervals. Table 7 displays that the PEF decreased by 8.656 (ver a media) as the concentration of SO_2 in the atmosphere increased $1\mu g/m^3$. The AIC was 15903.53.

Table 7: Point and interval estimates for the parameters of the LMM - SO_2 - Children with Asthma

Random: Day	Coefficient	Std. Error	p-value	Confidence Interval
Intercept	294.997	4.650	0.000	[285.874;304.121]
Gender (G)	-52.725	4.940	0.000	[-62.417;-43.034]
SO2	-8.656	2.800	0.002	[-14.150;-3.163]

Prepared by the author

Table 8 presents the point and interval estimates for the parameters of the model, considering the patients without asthma. Contrarily observed previously, neither the gender nor the pollutant SO_2 were significant, indicating no impact of these variables on the PEF of this group of patients. The AIC was 5060.731.

The model considering the interaction between PM_{10} and SO_2 was also realized considering the hybrid model proposed by Souza et al. [2018] however, the contaminants did not display significance together.

Table 8: Point and interval estimates for the parameters of the LMM - SO_2 - Children without Asthma

Random: Day	**Coefficient**	**Std. Error**	**p-value**	**Confidence Interval**
Intercept	290.795	9.841	0.000	[271.448;310.142]
Gender	8.329	10.040	0.407	[-11.408;28.068]
SO2	6.852	6.126	0.264	[-5.191;18.895]

Prepared by the author

A.3.2 CALPUFF

Air quality modeling is an essential alternative for most air pollution studies. CALPUFF (Scire et al. [2000]) is an advanced system capable to simulate the effects of time- and space-varying meteorological conditions on pollutant transport, transformation, and removal. Many studies have been using this tool, and recently Hernández-Garcés et al. [2020] realized an overview of them. Contrarily IEMA data, the concentrations obtained by CALPUFF were calculated for each participant, considering their neighborhood conditions in the period of study. Nevertheless, environmental variables were only available for 21 children. These informations refer to the daily mean, minimum, and maximum collected on the same day the spirometry test was realized. Six air pollutants were analyzed but only concentration levels of $PM_{2.5}$ displayed a significative association with the PEF indicator. Again, no environmental variable showed a significative association with the VEF1. Due to the nature of the variables, robust analysis was also considered but did not provide distinct results.

A.3.2.1 $PM_{2.5}$

Table 9 presents the point and interval estimates for the parameters of the LMM applied to age and the minimum daily concentration of $PM_{2.5}$ as covariates and PEF as the response variable. The random effect was the day. The AIC observed was 7412.245. The PEF of the patients decreased by 1.023(verificar a medida) when the $PM_{2.5}$ concentration increase $1\mu g/m^3$. Note that, although the concentration of $PM_{2.5}$ was significant at the level of 10%, the bootstrap confidence interval contains the value zero.

Table 9: Point and interval estimates for the parameters of the LMM - $PM_{2.5}$ - CALPUFF

Random: Ashtma	**Coefficient**	**Std. Error**	**p-value**	**Confidence Interval**
Intercept	27.856	21.783	0.201	[-14.923;70.636]
Age	23.440	1.882	0.000	[19.744;27.136]
PM2.5	-1.023	0.520	0.049	[-2.045;-0.002]

Prepared by the author

A.3.2.2 PCA analysis

Considering the combination of contaminants the population is exposed we also realized a multivariate analysis considering all air pollutants as covariates and the PFE as the dependent variable. For this, a principal component analysis (PCA) was realized as the contaminants presented a significative correlation among them (Tables 10 and 11).

Table 10: Correlation among VEF1 and pollutants - CALLPUFF

	VEF1	**PM10**	**PM2.5**	**SO2**	**CO**	**NO2**	**NOX**
VEF1	1.000						
PM10	0.003	1.000					
PM2.5	0.008	0.164	1.000				
SO2	0.022	0.152	0.0009	1.000			
CO	0.033	0.490	0.095	0.792	1.000		
NO2	0.055	0.572	-0.020	0.280	0.433	1.000	
NOX	0.015	0.944	0.201	0.251	0.589	0.523	1.000

Prepared by the author

Tables 10 and 11 show that only the $PM_{2.5}$ presents a negative correlation with the PEF, which means that as the concentration of this pollutant increases the peak expiratory flows decreases.

Table 11: Correlation among PEF and pollutants - CALLPUFF

	PEF	**PM10**	**PM2.5**	**SO2**	**CO**	**NO2**	**NOX**
PEF	1.000						
PM10	0.050	1.000					
PM2.5	-0.066	0.164	1.000				
SO2	0.044	0.152	0.0009	1.000			
CO	0.061	0.490	0.095	0.792	1.000		
NO2	0.136	0.572	-0.020	0.280	0.433	1.000	
NOX	0.036	0.944	0.201	0.251	0.589	0.523	1.000

Prepared by the author

The first four principal components (PC) correspond to approximately 83.2% of the total

variability. The highest coefficients (in eigenvectors) of principal components 1, 2, 3 and 4 are those of the pollutants NOX, PM10, SO2, PM2.5 and NO2, respectively.

Table 12: Importance of components - CALPUFF

	PC1	PC2	PC3	PC4	PC5	PC6
PM2.5	-0344	0.379	**0.759**	0.388	0.089	-0.038
NO2	-0.406	0.068	-0.578	**0.697**	0.059	0.070
PM10	**-0.460**	0.271	-0.150	-0.319	-0.510	-0.573
SO2	-0.278	**-0.785**	0.248	0.149	-0.456	0.109
NOX	**-0.473**	0.194	-0.057	-0.416	-0.032	0.748
CO	-0.448	-0.350	0.002	-0.255	0.720	-0.303

Prepared by the author

Table 13 shows that PM10, SO2, NOX, and NO2 relate negatively with the response variable. Thus as the concentrations of these contaminates increase the PEF of the children in the study decreases. All confidence intervals contain the zero value. However, these intervals are too large. Figure A.3 shows that the variability of the residuals is too large which can impact the confidence intervals. Due to this, conclusions based on these confidence intervals can not be reasonable.

Table 13: Pouint and interval estimates for the parameters of the LMM - PCA - CALPUFF

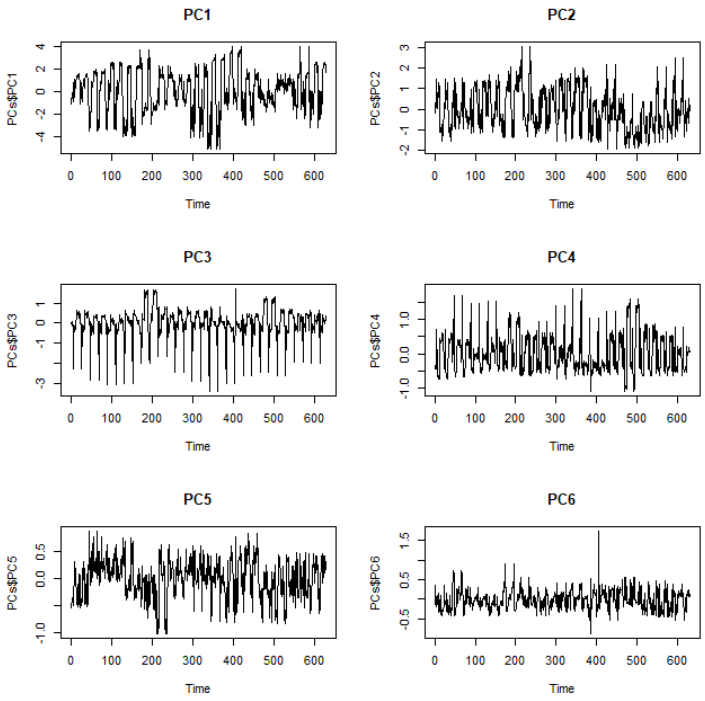
Random: Day	**Coefficient**	**CI**
Intercept	28.84	[-14.20;71.89]
Age	23.19	[19.46;26.91]
PM2.5	7.70	[-5.69 ;21.09]
NO2	-0.79	[-6.14 ;4.56]
PM10	-4.97	[-34.28;24.32]
SO2	-4.08	[-11.84; 3.66]
NOX	-9.88	[-22.99; 3.21]
CO	5.18	[3.39 ; 6.96]

Prepared by the author

Figure A.1 shows the time behavior of the principal components obtained from the pollutant concentration series, i.e., the original data.

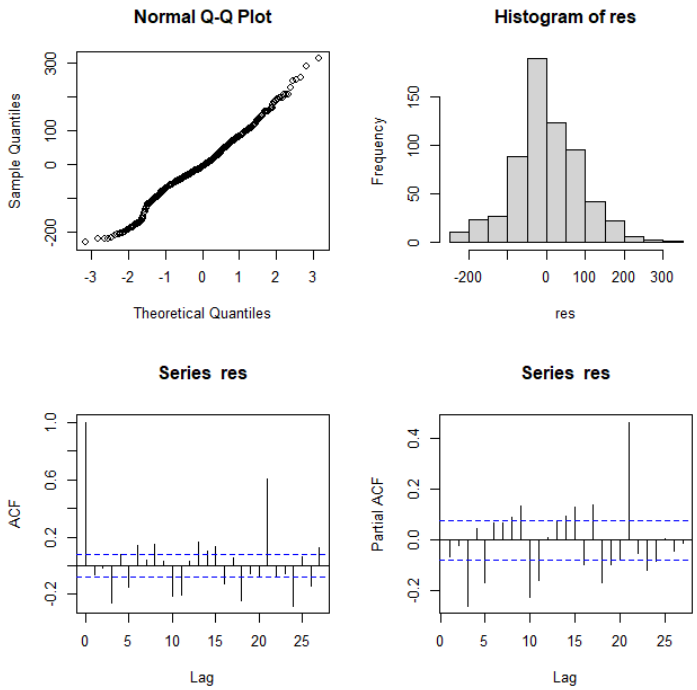
The residuals analysis shows that the residuals are approximately normally distributed. However, the residuals are not white noise as they present a correlated structure. The Figure of adjusted values versus residuals also does not show a random distribution of the residuals. In addition, the variability of the residuals seems to increase.

Figure A.1: PCs plots



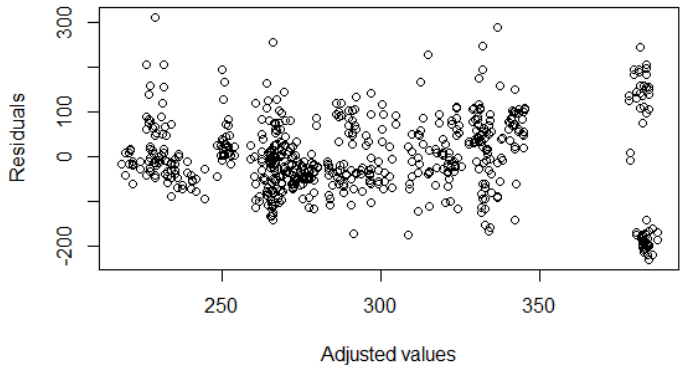
Prepared by the author

Figure A.2: Residuals analysis from the PCA model - CALLPUFF



Prepared by the author

Figure A.3: Residuals analysis - adjusted values versus residuals - CALLPUFF



Prepared by the author

A.4 Conclusions

This study analyzed the effects of air pollutants on the respiratory capacity of children and adolescents in Vitória, Brazil, participants of the study AsmaVix realized between 2019 and 2020. For this, the medical parameters VEF1 and PEF were collected daily for fifteen days for each patient two times a day. The contaminants, particulate material (PM_{10} and $PM_{2.5}$) and gases (CO, SO_2, O_3, NO_2 and NO_X) were collected from two sources, the IEMA and by the CALLPUFF method.

An analysis of real data involved the IEMA data source involved only the pollutants PM_{10} and SO_2, from the AQAMN Enseada do Suá. We observed that the contaminants were only correlated significantly with the PEF, a medical measure more sensitive to any alteration. In the univariate modeling, only the SO2 was significantly related to the PEF. However, in the residuals analysis, we observed that the variability of the residuals is too large, indicating that the confidence intervals should not be used to evaluate the significance of these relations. Both pollutants considered were negatively related to the response variable in the sense of an increase of $1\mu g/m^3$ of the concentration of PM_{10} decrease the PEF in $0.112l/s$, while the increase of $1\mu g/m^3$ in the concentration of SO_2 decrease the PEF in $7.771l/s$.

The analysis considering the CALLPUFF data source involved PM_{10} and $PM_{2.5}$,CO, SO_2, NO_2 and NO_X. In the univariate modeling, only $PM_{2.5}$ presented a negative and significant correlation with PEF, in the sense of an increase of $1\mu g/m^3$ of the concentration of $PM_{2.5}$ decrease the PEF in $1.023l/s$. Neither pollutant presented significative relation with the VEF medical measure. In the multivariate analysis, we realized a PCA analysis in which PM_{10}, SO_2, NO_2 and NO_X showed negative relation with the PEF. Again, the residuals displayed a large variability, and we did not use confidence intervals to evaluate the significance of the pollutants.

In general, the residual analysis showed that the model is not well adjusted. It is important to observe that only 21 participants were considered in this analysis. Thus, in a further study, we aim to have all data available for analysis.

Appendix B

Generalized additive model for count time series: An application to quantify the impact of air pollutants on human health

Abstract. The generalized additive model (GAM) has been used in many epidemiological studies where frequently the response variable is a nonnegative integer-valued time series. However, GAM assume that the observations are independent, which is generally not the case in time series. In this paper, an autoregressive moving average (ARMA) component is incorporated to the GAM. The resulting GAM-ARMA model is based on the generalized linear autoregressive moving average (GLARMA) model where some linear components are replaced by natural splines. Numerical simulations are presented and show that the ARMA component influences the estimation. In a real data analysis of the effects of air pollution on respiratory disease in the metropolitan area of Belo Horizonte, Brazil, it is shown that the proposed model presents a better fit when compared to the classical GAM approach, that does not take into account the autocorrelation of the data.

B.1 Introduction

Epidemiological data are frequently treated as time series of counts because they record the relative frequency of certain events that occur in successive time intervals and the observations are correlated.

Many epidemiological studies have been carried out to investigate the impact of ambient air pollution concentrations and meteorological conditions on human health. Kelsall *et al.* (1997), Ostro *et al.* (1999), Goldberg *et al.* (2003) and other authors found significant association between daily pollutant concentration levels and mortality. Alonso *et al.* (2010) studied the impact of atmosphere pressure, air humidity, and temperature on the number of hospitalizations. Besides that, Roberts (2004), Stafoggia *et al.* (2008) and other authors found the evidence of

interactive effects between temperature and air pollution (e.g., particulate matter and ozone) on mortality and adverse health outcomes. Such studies are an alert about the importance of controlling and reducing air pollutant emissions, and provide support for health departments in resource allocation.

Nevertheless, most of these studies try to model the relation between the occurrence of a disease and the air pollutants using procedures that are not able to capture the dependence inherent to the observations, such as the generalized linear model (GLM) (Nelder and Wederburn, 1972) and GAM (Hastie and Tibishirani, 1990). New methodologies were then proposed to model time series of counts. Shephard (1995) introduced the GLARMA model, then generalized by Davis $et\ al.$ (2003). This methodology adds an ARMA structure to the GLM and is able to model time series belonging to the exponential family. In the same vain, Benjamin $et\ al.$ (2003) proposed the generalized ARMA model. Mckenzie (1985) and Al-Osh and Alzaid (1987) introduced the integer-valued autoregressive model. Heinen (2003) proposed the autoregressive conditional Poisson model for counting data with time dependency and over-dispersion. Gamerman *et al.* (2013) proposed a family of non-gaussian state space models that allows the marginal likelihood to be calculated in an exact way.

The above models assume that the relation between the response variable and the covariates is linear. The GAM offers more flexibility and has been used by many authors to solve real problems in the environmental context, see e.g. Schwartz (2000), Aldrin and Haff (2005), and Belusic *et al.* (2015). Despite its widespread use, care is required when GAM is used in time series due to the serial correlation present in the data. Very few works are concerned with this issue, in particular Yang $et\ al.$ (2012) who proposed GAM with autoregressive terms. Souza $et\ al.$ (2018) have also proposed a hybrid model, including GAM, principal component analysis, and vector autoregression to address the multicolinearity problems that can occur when including several air pollutants in the analysis.

In this work a more general model for count data is proposed, which is able to handle both the autocorrelation structure of the time series and the nonlinearity existing in the covariates. This model is composed of a GAM with an ARMA component and is called a GAM-ARMA model. The non-parametric components are estimated through some smoothed functions, such as splines. Numerical simulations are performed to access the accuracy of parameter estimation in small sample size series following a Poisson distribution. Finally, a real-time series is analyzed without taking and taking into account the autocorrelation of the data. The example includes the fit of a GAM-ARMA model to evaluate the impact of air pollutants and meteorological variables on the number of chronic obstructive pulmonary disease cases in the metropolitan area of Belo Horizonte, Brazil.

The paper is organized as follows. Section B.2 presents the GAM-ARMA model, detailing some properties and the inference procedure. Section B.3 shows the simulation study. Section B.4 presents the analysis of a real series of pulmonary disease counts. Section B.5 concludes the work.

B.2 The GAM-ARMA model

B.2.1 Presentation of the model

We combine the GAM with the ARMA model proposed by Box and Jenkins (1976) to model linear and nonlinear relations between the response variable and the covariates, and the time correlation of the response. The advantage of this methodology is the possibility to adjust semiparametric and non-parametric models to the data, capturing either linear and non-linear relationships, and thus obtaining better estimates.

As in the GLARMA model, the conditional distribution of the observation y_t given the past information $\mathcal{F}_{t-1}^y = \sigma\{y_s, s \leq t-1\}$ follows a Poisson distribution, i.e.,

$$y_t \mid \mathcal{F}_{t-1}^y \sim (\mu_t), \tag{B.1}$$

where $\mu_t = (y_t \mid \mathcal{F}_{t-1}^y)$. Here, the predictor $\eta_t = \ln(\mu_t)$ follows the model

$$\eta_t = \beta_0 + \sum_{j=1}^{k} \beta_j x_{t,j} + \sum_{j=1}^{l} s_j(w_{t,j}) + Z_t, \tag{B.2}$$

where $(x_{t,1}, \ldots, x_{t,k})$ denotes the covariates related linearly to η_t, $(w_{t,1}, \ldots, w_{t,l})$ denotes the covariates related to η_t via smooth functions $s_1, \ldots, s_l$, and Z_t modelises the time correlation. Following Davis *et al.* (2003),

$$Z_t = \sum_{i=1}^{\infty} \tau_i \varepsilon_{t-i}, \tag{B.3}$$

where, for some $\lambda \in (0, 1]$,

$$\varepsilon_t = (y_t - \mu_t)\mu_t^{-\lambda} = (y_t - e^{\eta_t})e^{-\lambda\eta_t}, \tag{B.4}$$

and the parameters τ_i's are the coefficients in the power series expansion

$$\sum_{i=1}^{\infty} \tau_i z^i = \left(1 - \sum_{i=1}^{p} \phi_i z^i\right)^{-1} \left(1 + \sum_{i=1}^{q} \theta_i z^i\right) - 1, \quad |z| \leq 1, \tag{B.5}$$

where the polynomials $\phi(z) = 1 - \phi_1 z - \cdots \phi_p z^p$ and $\theta(z) = 1 + \theta_1 z + \cdots \theta_q z^q$ have no common zeroes and have all their zeros outside the unit circle. It follows from (B.3) and (B.5) that Z_t can be calculated recursively with the difference equation

$$Z_t = \phi_1(Z_{t-1} + \varepsilon_{t-1}) + \cdots + \phi_p(Z_{t-p} + \varepsilon_{t-p}) + \theta_1\varepsilon_{t-1} + \cdots + \theta_q\varepsilon_{t-q}. \tag{B.6}$$

According to (B.4), $(\varepsilon_t \mid \mathcal{F}_{t-1}^y) = \mu_t^{-\lambda}((y_t \mid \mathcal{F}_{t-1}^y) - \mu_t) = 0$. Now, let $\mathcal{F}_{t-1}^{\varepsilon} = \sigma\{\varepsilon_s, s \leq t-1\}$, (B.4) implies that $\mathcal{F}_{t-1}^{\varepsilon} \subset \mathcal{F}_{t-1}^y$. Therefore,

$$(\varepsilon_t \mid \mathcal{F}_{t-1}^{\varepsilon}) = [(\varepsilon_t \mid \mathcal{F}_{t-1}^y) \mid \mathcal{F}_{t-1}^{\varepsilon}] = 0,$$

which shows that (ε_t) is a martingale difference sequence. Hence, $(\varepsilon_s, \varepsilon_t) = 0$ for $s \neq t$, and the variance of ε_t is

$$(\varepsilon_t) = (\varepsilon_t^2) = [(\varepsilon_t^2 \mid \mathcal{F}_{t-1}^y)] = (\mu_t^{-2\lambda}[(y_t - \mu_t)^2 \mid \mathcal{F}_{t-1}^y]) = (\mu_t^{1-2\lambda}). \tag{B.7}$$

Now, (B.2), (B.6) and (B.7) imply that

$$(\eta_t) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{t,j} + \sum_{j=1}^{l} s_j(w_{t,j}),$$

$$(\eta_t) = \sum_{i=1}^{\infty} \tau_i^2 (\mu_{t-i}^{1-2\lambda}),$$

and

$$(\eta_t, \eta_{t+h}) = \begin{cases} \sum_{i=1}^{\infty} \tau_i \tau_{i+h} (\mu_{t-i}^{1-2\lambda}), & \text{if } h \geq 0, \\ \sum_{i=1}^{\infty} \tau_i \tau_{i-h} (\mu_{t+h-i}^{1-2\lambda}), & \text{if } h < 0, \end{cases}$$

When $\lambda = 0.5$, (ε_t) are the Pearson residuals and the covariances of (η_t) do not depend on t, even if (η_t) is not strictly stationary.

B.2.2 Parameter estimation

There are several approaches in the literature to estimate functions s_j's. Recent studies have used reduced rank approaches due to the low computational cost and facilities to obtain good estimators of the s_j's. Wood (2006) presents a review of methods for choosing the s_j's using the GAM methodology and some approaches as thin plate regression splines (Wood, 2003), B-splines and basis splines (De Boor, 1978; Dierckx, 1993), among others.

In this work, the B-spline curves were used given their simplicity to obtain flexible smoothing. B-splines are constructed from polynomial pieces, joined at control points called knots. By definition, the B-spline $B_{i,d}$ depends on the knots $t_i \leq \cdots \leq t_{i+d+1}$, where d is the order of the polynomial. If the knot vector is $(t_1, t_2, \ldots, t_{m+d+1})$ for some positive integer number m, it is possible to form m B-splines $B_{1,d}, \ldots, B_{m,d}$ of degree d associated with this knot vector. A spline function s_j is a linear combination of B-splines, i.e.,

$$s_j = \sum_{i=1}^{m} \alpha_{i,j} B_{i,d}, \tag{B.8}$$

where the reals $\alpha_{1,j}, \ldots, \alpha_{m,j}$ are called the B-spline coefficients of s_j. For more properties, see De Boor (1978). Here, we take $d = 3$ and we use natural cubic splines. In this case, the polynomials before the first knot and after the last knot are modeled through linear functions,

which means that the second derivative at the two end points are zero. General accounts about splines can be found in the books by Hastie *et al.* (2008), and Ahlberg *et al.* (1967). The choice of the optimal number of knots is based on the work of Harrell (2004) and depends on the sample size n. Typically, when $n \leq 100$, three or four knots usually generate good fitting and a balanced model in relation to flexibility and loss of accuracy. For large n, five knots is a good starting point. The Akaike's information criterion (AIC) can be used to choose the number of knots, see Akaike (1973).

Combining (B.2) and (B.8), and dropping $d = 3$ in the notation, the model of the predictor can be written as

$$\eta_t = \beta_0 + \sum_{j=1}^{k} \beta_j x_{t,j} + \sum_{j=1}^{l} \sum_{i=1}^{m} \alpha_{i,j} B_i(w_{t,j}) + Z_t, \tag{B.9}$$

where Z_t is given by (B.6). Thus, for a fixed integer m and fixed knots $(t_1, t_2, \ldots, t_{m+4})$, the parameter vector of the GAM-ARMA model is defined by

$$\delta = (\beta_0, \ldots, \beta_k, \alpha_{1,1}, \ldots, \alpha_{m,l}, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q).$$

According to (B.1), the conditional log-likelihood function is

$$L_n(\delta) = \sum_{t=1}^{n} (y_t \eta_t(\delta) - e^{\eta_t(\delta)}),$$

where $\eta_t(\delta)$ is given by (B.9) and $Z_t(\delta)$ is obtained by (B.6). The maximization of $L_n(\delta)$ can be performed by Newton's method initialized with zero values for all parameters. In practice, the convergence occurs approximately within 10 iterations.

Goodness-of-fit measures for the proposed methodology can be calculated with the AIC and the bayesian information criterion (BIC) defined by

$$= -2 \ln(L_n(\widehat{\delta}_n)) + r \ln(n),$$

where $\widehat{\delta}_n$ are the parameter values that maximize $L_n(\delta)$ and r is the number of parameters estimated by the model.

The relative risk (RR) is widely used to measure the impact of air pollution on human health, see Baxter *et al.* (1997). RR for the pollutant covariate $x_j = (x_{t,j})$ in (B.9) is the relative change in the expected count of respiratory disease event per ξ-unit change in x_j while keeping the other covariates fixed, and is given by

$$\widehat{}_{x_j}(\xi) = \exp(\widehat{\beta}_j \xi).$$

RR and its confidence interval (CI) of level $1 - \alpha$ are estimated as follows,

$$\widehat{}_{x_j}(\xi) = \exp(\widehat{\beta}_j \xi), \tag{B.10}$$

$$\widehat{\mathrm{CI}}\{_{x_j}(\xi)\} = \exp(\widehat{\beta}_j \xi \pm z_{\alpha/2} \, \mathrm{se}(\widehat{\beta}_i)\xi), \tag{B.11}$$

where $\widehat{\beta}_j$ is the conditional maximum likelihood estimator $\widehat{\beta}_{j,n}$ of β_j, $\mathrm{se}(\widehat{\beta}_j)$ is the estimated standard deviation (s.d.) of $\widehat{\beta}_j$, and $z_{\alpha/2}$ denotes the $(1 - \alpha/2)$-quantile of the standard normal distribution.

B.3 Simulation study

In our numerical experiment, the sample size is $n = 100$, the number of replications is $N = 1000$, $\lambda = 0.5$ in (B.4), $(p, q) = (1, 0)$ in (B.6) and $(k, l, m) = (2, 1, 3)$ in (B.9). The predictor model is given by

$$\eta_t = \beta_0 + \beta_1 x_{t,1} + \beta_2 x_{t,2} + \alpha_1 B_1(w_t) + \alpha_2 B_2(w_t) + \alpha_3 B_3(w_t) + Z_t, \tag{B.12}$$

where the B_i's compose the B-spline basis for natural cubic splines and

$$Z_t = \phi[Z_{t-1} + (y_{t-1} - e^{\eta_{t-1}})e^{-\eta_{t-1}/2}]. \tag{B.13}$$

The covariates $(x_{t,1}, x_{t,2})$ are simulated (one time) with the ARMA models, $x_{t,1} = 0.42x_{t-1,1} + u_t + 0.13u_{t-1}$ and $x_{t,2} = 0.30x_{t-1,2} + v_t - 0.76v_{t-1} - 0.17v_{t-2}$ where (u_t, v_t) is a sequence of independent Gaussian random variables with zero-mean and unit variance. The covariate (w_t) is the real time series of daily minimum temperature in Vitória, Brazil, between April 10, 2005 and July 19, 2005. The parameter values are

$$\beta_0 = 0.8, \quad \beta_1 = 0.1, \quad \beta_2 = -0.2, \quad \alpha_1 = 0.5, \quad \alpha_2 = -1.0, \quad \alpha_3 = 0.8,$$

and three different values of ϕ are considered, $\phi = 0.1, 0.4, 0.6$ corresponding respectively to increasing values of the autocorrelation in the response variable.

In Table 1, $\widehat{\mu}_{\widehat{\delta}_j}$ represents the average of the N estimates of the parameter δ_j and the corresponding mean squared errors (MSE) in parenthesis for $\phi = 0.1, 0.4, 0.6$. We see that the estimates are close to the true values of the parameters. In general, the values of MSE are small, but increase as ϕ increases.

Figure B.1 presents the histograms of the N estimates of ϕ and the β_j's for $\phi = 0.1, 0.4, 0.6$. While the empirical distribution of the estimates of ϕ is approximately symmetric about the true value when $\phi = 0.1, 0.4$, this distribution is asymmetric when $\phi = 0.6$. The empirical distribution of the estimates of β_0 is asymmetric about the true value for all values of ϕ. Concerning β_1 and β_2, the distributions are approximately symmetric about their true values, even when $\phi = 0.6$.

B.4 Results

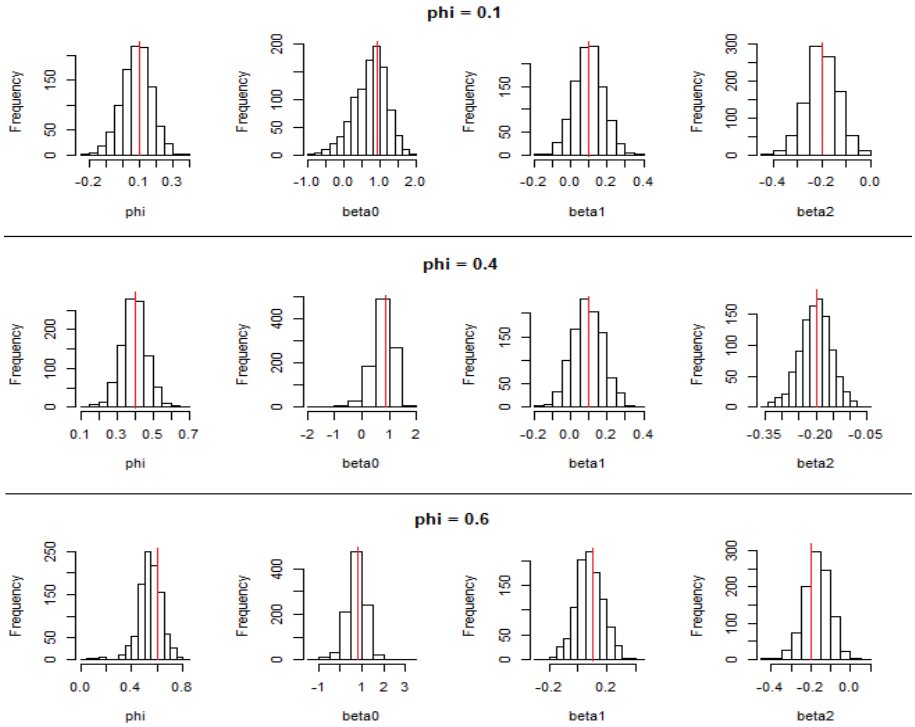
Here, we fit a GAM-ARMA model to the monthly number of chronic obstructive pulmonary disease (COPD) cases, popularly known as acute bronchitis, in the metropolitan area

Table 1: Parameter estimates in Model (B.12)–(B.13) with MSE in parenthesis.

	$\widehat{\mu}_{\widehat{\phi}}$	$\widehat{\mu}_{\widehat{\beta}_0}$	$\widehat{\mu}_{\widehat{\beta}_1}$	$\widehat{\mu}_{\widehat{\beta}_2}$	$\widehat{\mu}_{\widehat{\alpha}_1}$	$\widehat{\mu}_{\widehat{\alpha}_2}$	$\widehat{\mu}_{\widehat{\alpha}_3}$
$\phi = 0.1$	0.0845 (0.0074)	0.7637 (0.1795)	0.0986 (0.0078)	-0.1953 (0.0036)	0.5455 (0.1264)	-0.9812 (0.9071)	0.8134 (0.0917)
$\phi = 0.4$	0.3927 (0.0055)	0.6907 (0.1852)	0.0945 (0.0067)	-0.1956 (0.0028)	0.5332 (0.1236)	-0.8401 (0.7623)	0.9035 (0.0985)
$\phi = 0.6$	0.5311 (0.0128)	0.7078 (0.2084)	0.0362 (0.0145)	-0.2443 (0.0049)	0.2491 (0.2183)	-0.6168 (0.9690)	0.9289 (0.1587)

Prepared by the author

Figure B.1: Histograms of parameter estimates of ϕ and the β_j's in Model (B.12)–(B.13).



Prepared by the author

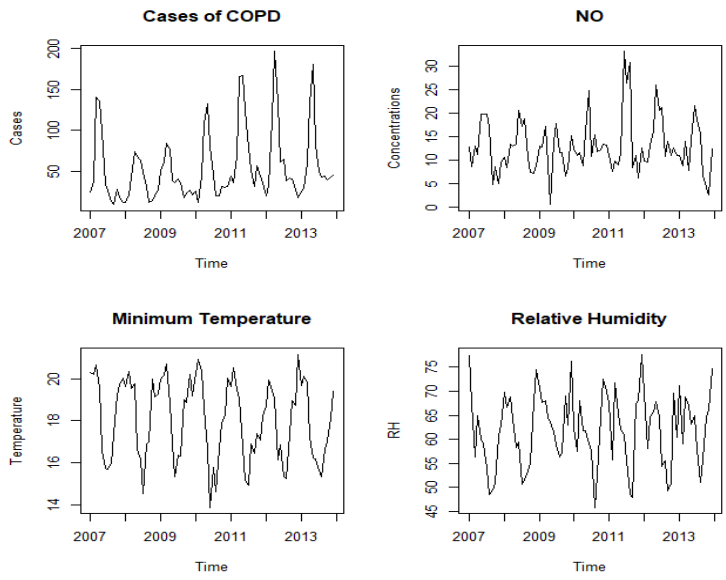
of Belo Horizonte, Brazil, between January 2007 and December 2013 ($n = 84$). According to the department of information technology of the Brazilian public health system, each hour three Brazilian citizens die as a result of this disease. The objective of this analysis is to evaluate the association among the concentration of atmospheric pollutants and meteorological conditions with the occurrence of COPD in Belo Horizonte.

Studies concerning air pollution in Belo Horizonte are relatively rare, even rarer regarding the relation between pollutant series and respiratory diseases. Information about the concentration of pollutants in this region is very limited, with all the series presenting missing observations. Among the pollutants measured at the state environment and water resources institute, we

select the nitrogen monoxide (NO) as the explanatory variable in this study since it presents the largest significative correlation coefficient $\rho = 0.3$ related to COPD. Some data imputations are performed before fitting the model, in order to handle the missing observations. We use a robust procedure for imputation in time series using Kalman smoothing and state space model (Harvey, 1989) and the package "imputeTS" from software R (Moritz, S., Package "imputeTS" - Time series missing value imputation).

Figure B.2 presents the time series of COPD cases, NO concentration, minimum temperature (T_{min}) and relative humidity (RH) of the air. A positive trend can be detected in the number of COPD cases and NO concentration. Furthermore, all time series present a seasonal behaviour. Table 2 contains some descriptive statistics of the data, where Q1 and Q3 denote the first and third quartile, respectively.

Figure B.2: Number of COPD cases, concentration of NO, minimum temperature and relative humidity of the air in the metropolitan area of Belo Horizonte, Brazil, between January 2007 and December 2013.



Prepared by the author

Table 2: Descriptive statistics of the data.

	Min	Max	Q1	Q3	Mean	Median	s.d.
Cases	10	196	27	66	54.93	41	42.3
NO ($\mu g/m^3$)	0.57	33.11	9.05	15.52	12.97	12.01	5.80
T_{min}(°C)	13.87	21.15	16.43	19.62	17.89	18.27	1.90
RH (%)	45.83	77.63	56.17	67.55	61.60	61.60	7.48

Prepared by the author

In our model, NO concentration is related linearly to η_t, while T_{min} and RH have a non-linear relation with η_t. Besides these explanatory variables, a trend component and sine and

cosine functions are also incorporated in the model. The trend is included to modelise the slight positive trend in the cases of COPD. The sine and cosine functions are necessary to handle the annual and semi-annual seasonality in the response variable. Therefore, the model writes

$$\begin{aligned}\eta_t = \beta_1 x_{t,1} + \beta_2 \sin(2\pi t/12) + \beta_3 \cos(2\pi t/12) + \beta_4 \sin(2\pi t/6) + \beta_5 \cos(2\pi t/6) + \beta_6 t + \\ + \alpha_{1,1} B_1(w_{t,1}) + \alpha_{2,1} B_2(w_{t,1}) + \alpha_{3,1} B_3(w_{t,1}) + \\ + \alpha_{1,2} B_1(w_{t,2}) + \alpha_{2,2} B_2(w_{t,2}) + \alpha_{3,2} B_3(w_{t,2}) + Z_t,\end{aligned} \tag{B.14}$$

where t is the month number, $x_{t,1}$ is the NO concentration, $(w_{t,1})$ is $T_{\min}$ and $(w_{t,2})$ is RH. A simple GAM model where Z_t is removed in (B.14) is also adjusted, to show the benefict of modeling the data autocorrelation through Z_t in the GAM-ARMA model. The choice of the optimal number of knots is based on the sample size. Thus, as recommended in Section B.2, three and four knots are tested, and comparing the AIC, the best model is obtained with three knots.

Table 3 presents the estimates $\widehat{\beta}_i$'s of the parameters β_i's in the fitted GAM model with the corresponding standard errors given by the software R. All estimates are significant at 5% level of significiance. On the other hand, the value of BIC is 1297.514 and the in-sample MSE between the fitted values and the observed values of COPD cases (see figure B.4) is 531.642.

Table 3: Parameter estimates of a GAM model (B.14) ($Z_t = 0$) fitted to the COPD cases.

Parameter	β_1	β_2	β_3	β_4	β_5	β_6
Estimate	0.0545	0.2470	-0.5562	-0.3137	-0.2522	0.0096
Standard error	0.0032	0.0415	0.0611	0.0306	0.0326	0.0007

Prepared by the author

Figure B.3(a) plots the sample autocorrelation function (ACF) and sample partial autocorrelation function (PACF) of the residuals in the GAM model. Some correlation is still present in these residuals, indicating the need for a more elaborated model.

Applying the GAM-ARMA methodology, the best fit is obtained with a GAM-AR(1) model. Table 4 shows the estimates $\widehat{\beta}_i$'s and $\widehat{\phi}$ of the parameters β_i's and ϕ in the fitted GAM-AR(1) model with the corresponding standard errors given by the software R. Again, all estimates are significant at 5% level of significiance. The value of BIC is 1155.059 and the in-sample MSE between the fitted values and the observed values of COPD cases (see figure B.4) is 356.169. Both values are smaller than the corresponding values obtained with the GAM model. Furthermore, the sample ACF and PACF plots in figure B.3(b) show no difference with a white noise which reveals a good adjustment of the GAM-AR(1) model.

Figure B.4 shows that the GAM-AR(1) model fits better the observed number of COPD cases than the GAM model.

Figure B.3: Sample ACF and PACF of the residuals in the GAM and GAM-AR(1) models.

(a) GAM model. (b) GAM-AR(1) model.

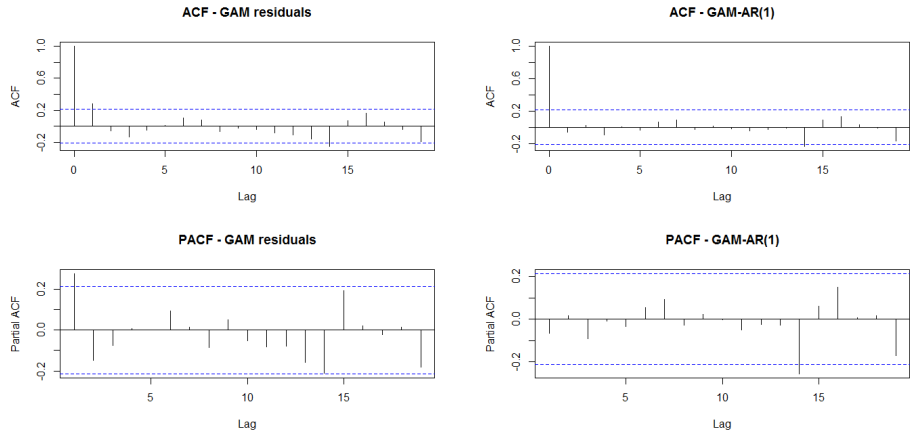
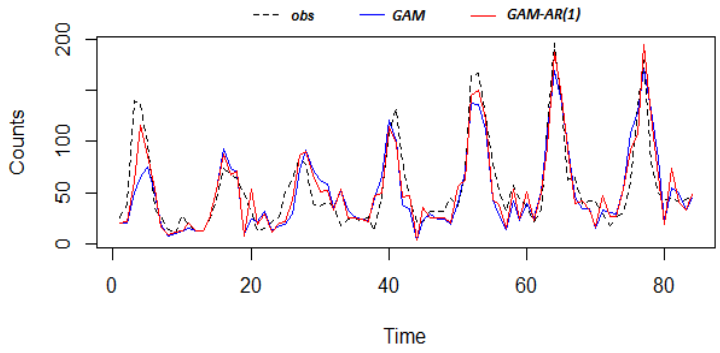


Table 4: Parameter estimates of a GAM-AR(1) model (B.14) fitted to the COPD cases.

Parameter	β_1	β_2	β_3	β_4	β_5	β_6	ϕ
Estimate	0.0515	0.3271	-0.5221	-0.3068	-0.2324	0.0127	0.0700
Standard error	0.0029	0.0499	0.0615	0.0349	0.0362	0.0010	0.0053

Prepared by the author

Figure B.4: Fits of GAM and GAM-AR(1) models to the number of COPD cases.



Prepared by the author

The RR for the NO is an important information for the regulatory agencies to quantify the impact of this pollutant on the population health. Table 5 presents the estimated RR and CI for the NO,ˆand $\widehat{\text{CI}}$ given by (B.10) and (B.11) where $\alpha = 5\%$, respectively, obtained with the GAM and GAM-AR(1) models. In both cases,ˆis significant which means that NO contributes significantly to the increase in the number of COPD cases;ˆis slightly smaller for the GAM-AR(1) model. Although ˆare comparable in the two models, the adjustment with the GAM-AR(1) model is the best in view of the measures of BIC and MSE, and the correlation of the residuals.

Table 5: Estimated RR and 95% CI for the NO in the GAM and GAM-AR(1) models.

NO	GAM	GAM-AR(1)
$\widehat{RR}$	1.0627	1.0591
$\widehat{\text{CI}}$	[1.0553;1.0702]	[1.0524;1.0658]

Prepared by the author

B.5 Conclusions

In this work, a new methodology called GAM-ARMA was proposed, based on the GLARMA model introduced by Davis *et al.* (2003). The GAM-ARMA model allows the fitting of semiparametric models, accommodating covariates with linear and non-linear relation with the response variable in count data with time correlation.

A numerical simulation study showed that the estimates of the parameters are close to the true values for a moderate sample size of $n = 100$, and that the preciseness of the estimation degrades as the correlation in the data increases.

The model was applied to the monthly number of COPD cases in Belo Horizonte, Brazil, to quantify the impact of NO concentrations and meteorological variables on the occurrence of this disease. The best fit was obtained with a GAM-AR(1) model. This model presented white noise residuals and smaller measures of BIC and MSE compared to the GAM. The RR analysis revealed that NO contributed significantly to the increase of COPD cases.