Universidade Federal de Pernambuco
Centro de Ciências Exatas e da Natureza
Departamento de Estatística
Pós-Graduação em Estatística

Daniel Matos de Carvalho

Spatial Scan Statistics Based on Empirical Likelihood and Robust Fitting for Generalized Additive Models for Location, Scale and Shape

Recife
2021

Daniel Matos de Carvalho

Spatial Scan Statistics Based on Empirical Likelihood and Robust Fitting for Generalized Additive Models for Location, Scale and Shape

Tese apresentada ao Programa de Pós-Graduação em Estatística do Departamento de Estatística da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Doutor em Estatística.

Área de concetração:	Estatística Aplicada
Orientadora:	Profa. Dra. Fernanda De Bastiani
Co-orientador:	Prof. Dr. Getúlio José Amorim do Amaral

Recife
2021

Catalogação na fonte
Bibliotecária Monick Raquel Silvestre da S. Portes, CRB4-1217

C331s Carvalho, Daniel Matos de
Spatial scan statistics based on empirical likelihood and robust fitting for generalized additive models for location, scale and shape / Daniel Matos de Carvalho. – 2021.
114 f.: il., fig, tab.

Orientadora: Fernanda De Bastiani.
Tese (Doutorado) – Universidade Federal de Pernambuco. CCEN, Estatística, Recife, 2021.
Inclui referências.

1. Estatística aplicada. 2. Distribuição. I. De Bastiani, Fernanda (orientadora). II. Título.

310 CDD (23. ed.) UFPE- CCEN 2021 - 138

DANIEL MATOS DE CARVALHO

SPATIAL SCAN STATISTICS BASED ON EMPIRICAL LIKELIHOOD AND ROBUST FITTING FOR GENERALIZED ADDITIVE MODELS FOR LOCATION, SCALE AND SHAPE

Tese apresentada ao Programa de Pós-Graduação em Estatística da Universidade Federal de Pernambuco, como requisito parcial para obtenção do título de Doutor em Estatístitca.

Aprovada em: 04 de Agosto de 2021

BANCA EXAMINADORA

Profa. Dra. Fernanda De Bastiani
Universidade Federal de Pernambuco

Prof. Dr. Francisco Cribari Neto
Universidade Federal de Pernambuco

Prof. Dr. Alex Dias Ramos
Universidade Federal de Pernambuco

Prof. Dr. Miguel Angel Uribe Opazo
Universidade Estadual do Oeste do Paraná

Prof. Dr. Gilberto Alvarenga Paula
Universidade de São Paulo

I dedicate this work to my mother, Tereza, with affection and admiration; to my wife, Manu, with love and complicity; to my aunt Claúdia (in memoriam), with miss.

ACKNOWLEDGEMENTS

Completing a postgraduate course in Statistics at UFPE represents the realization of a dream and overcoming difficulties in the search for professional and personal growth. I would like to record my gratitude to the many people who hugged me during this walk, conveying serenity and support.

I would like to thank my mother for being my life example and guide, for all her love and affection and for all the sacrifice and effort to educate me. Thank you so much for your unremitting support and unconditional love.

I would like to thank my wife, Manu, for the affection, love, companionship and patience. Thank you very much for helping me and always being by my side in this stage of so many deprivations.

I would like to thank my daughters, Bia and Carol, my two loves who give meaning and happiness to my life.

I would like to thank my supervisor, professor Fernanda De Bastiani, an excellent researcher and person. Thanks for your availability, patience, constant and careful guidance always motivating me and transmit the necessary confidence for the development of this work.

I would like to thank my co-supervisor, professor Getúlio José Amorim do Amaral, for his dedication, example, for his own scientific collaborations. Thank you very much for the teachings and confidence.

I would like to thank to Robert Rigby and Mikis Stasinopoulos, for their help in the steps of this thesis.

I would like to thank my family, my grandmother Maria Cândida, my brothers Eli and Biel, my nephews Raissa and Matheus, thank you for everything you did for me and for always accompanying me at all times.

I would like to thank my wife's family, Dona Eliane, Mr. Félix, Edwin, Daniele, Thaisa, Renan, Dr. Danilo (in memoriam), Dona Penha, Mariana, Italo and all my nephews, for accompanying me over the years and being part of my life story.

I would like to thank the Brasilia family, mother Santana, Anayse, Dan, Julia and Guto who have always been with me in my heart.

I would like to thank my friends Cesar, Adenise, Any, Marcelo, Saul, for their welcome, friendship, support and willingness to help me and always accompany me.

I would like to thank my friends Thiago, Fábio, Max, João Paulo, Hemílio, Edwin, Renan, Italo, Paulinho and Rafael, I thank for the friendship and unforgettable moments of relaxation that we experienced.

I would like to thank the Federal Institute of Paraíba - IFPB for granting me a work permit to carry out this doctorate.

I would like to thank professor Hélder Alves de Oliveira, coordinator of the areas of Mathematics and Statistics at the Federal Institute of Paraíba - IFPB, for all the encouragement.

I would like to thank the IFPB colleagues, I express my gratitude for the words of encouragement during the walk.

I would like to thank the friends of the Weekly Sports Meeting Renan Azevedo (EDS), thank you for your friendship and for all the laughs.

I would like to thank the professors of the Department of Statistics at UFPE, for their teachings.

I would like to thank the participants of the examining board.

ABSTRACT

This thesis presents two independent themes with different background. The first theme presents a new method for detecting spatial clusters, that is, a method for detecting regions with a high concentration of spatial phenomena, compared to a expected number, given a random distribution of events. The main contribution is to present a nonparametric method based on empirical likelihood functions, as an alternative to traditional methods of using clusters (scan). In this way, no distribution family is required for the variable of interest. To evaluate the method, simulation studies were carried out considering the zero-inflated poisson model, comparing the results with the scan method proposed by Kuldorff. The results show that the new method reduces the error probabilities of the type I for zero inflated, with low power for cluster with less than 8 locations. A study was carried out for Measles data in São Paulo, Brazil, which present a excess of zeros. Only the Kulldorff scanning method identified the existence of a cluster, located and centered on the capital São Paulo. However, if a cluster is identified by the Kulldorff method in the presence of inflated and when not confirmed by the non-parametric approach, it is recommended that interpretations be performed with caution due to a high probability type error associated with Kulldorff method when model is not well specified. The second theme aims to present two new approaches to robust estimation for generalized additive models of location, scale and shape - GAMLSS, which focus on contamination situations in the tails of distributions. The main motivation is the scarcity of robust methods for GAMLSS models. The thesis were subdivided into two topics. The first topic presents a proposal that seeks transformations in order to limit the influence function associated with the probability distribution of interest, modifying the logarithm structure of the likelihood function, using concepts of censorship. It also features: the robust GAMLSS method proposed by Rigby et al. (2019), considering the gamma distribution, presenting the bias corrections for the estimators; a modification of the method proposed by Rigby et al. (2019), considering the weight of observations in the estimation; and, finally, a large simulation study to evaluate the proposals, using the gamma distribution and contamination in the right tail of the distribution. The second topic is based on a simple adaptive truncation, where observations identified as possible outliers are verified and, if necessary, removed by truncation of the response variable distribution. The simulation studies used the gamma and beta distributions, left and right tail contamination, and three distinct models: parametric models with and without covariates and non-parametric models. The results show that, compared to existing methods in the literature, the truncated adaptive method has a better performance with lower mean square error and lower variability in most simulated scenarios. The overall performances of the proposals are illustrated through three applications: brain image resonance data, using bivariate smoothing splines; extreme child poverty data; and data on acute viral infection of the respiratory system. excess of zeros. Only the Kulldorff scanning method identified the existence of a cluster, located and centered on the capital São Paulo. However, if a cluster is identified by the Kulldorff method in the presence of inflated and when not confirmed by the non-parametric approach, it is recommended that interpretations be performed with caution due to a high probability type error associated with Kulldorff method when model is not well specified.

Keywords: beta distribution; gamma distribution; robust GAMLSS; spatial cluster.

RESUMO

Esta tese apresenta contribuições para três tópicos distintos sobre dois temas independentes. O primeiro tema apresenta um novo método para detecção de clusters espaciais, ou seja, um método para detecção de regiões com alta concentração de fenômenos espaciais, comparado com um número esperado, dada uma distribuição aleatória de eventos. A principal contribuição é apresentar um método não paramétrico baseado nas funções de verossimilhança empírica, como alternativa para métodos tradicionais de varredura de clusters (scan). Desta forma, nenhuma família de distribuição é exigida para a variável de interesse. Os resultados mostram que o novo método reduz as probabilidades de erro do tipo I para observações inflacionadas de zero, com baixo poder para cluster com menos de 8 localizações. Foi realizado um estudo para dados de Sarampo em São Paulo, Brasil, que apresentam um excesso de zeros. Apenas o método scan de Kulldorff identificou a existência de um cluster, localizado e centrado na capital São Paulo. Entretanto, caso seja identificado um cluster pelo método Kulldorff na presença de observações inflacionadas e quando não confirmado pela abordagem não paramétrica, é recomendável que as interpretações sejam realizadas com cautela devido a alta probabilidade do erro do tipo I associado ao método Kulldorff quando o modelo não é bem especificado. O segundo tema tem como objetivo apresentar duas novas abordagens para estimação robusta para os modelos GAMLSS, que focam em situações de contaminação nas caudas das distribuições, devido a escassez de métodos. A tese apresenta diversas contribuições para este tema, que foram subdivididas em dois tópicos. O primeiro tópico apresenta uma proposta que busca transformações de modo a limitar a função de influência associada a distribuição de probabilidade de interesse, modificando a estrutura do logaritmo da função de verossimilhança utilizando conceitos de censura. Apresenta ainda: o método robusto GAMLSS proposto por Rigby et al. (2019), considerando a distribuição gama, apresentando as correções de viés para o estimadores; uma modificação do método proposto por Rigby et al. (2019), considerando o peso das observações na estimação; e, por fim, um amplo estudo de simulação para avaliação das propostas, utilizando a distribuição gama e contaminações na cauda direita da distribuição. O segundo tópico baseia-se em um truncamento adaptativo simples, onde observações identificadas como possíveis outliers são verificadas e, se necessário, removidas por truncamento da distribuição da variável de resposta. Apresenta também uma proposta adaptativa para definição da constante de sintonia, necessária para estimação do modelo. Além de propor uma nova abordagem para modelagem robusta, comparamos com métodos disponíveis na literatura. Os estudos de simulação utilizaram as distribuições gama e beta, contaminações na cauda esquerda e direita, e três modelos distintos: modelos paramétricos sem e com covariáveis e modelos não paramétricos. Os resultados mostram que o método adaptativo truncado apresenta melhor desempenho com menores valores no erro quadrático médio e menor variabilidade na maioria dos cenários simulados. O desempenho das propostas é ilustrado por meio de três aplicações: dados de ressonância de imagens cerebrais, usando splines de suavização bivariadas; dados de extrema pobreza infantil; e a dados de infecção viral aguda do sistema respiratório.

Palavras-chave: aglomerados espaciais; distribuição beta; distribuição gama; GAMLSS robusto.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

GAM Generalized Additive Models

AD Adapttive Truncation GAMLSS

AH Robust Fitting GAMLSS Propose by Aeberhardet al. (2021

CENS Robust Fitting GAMLSS based on Censoring)

CG Cole and Green Algorithm

EL Empirical likelihood

ELL Empirical Loglikelihood

ELLR Empirical Loglikelihood Ratio

ELS Empirical Likelihood Scan

G Nonrobust Fitting GAMLSS

GAM Generalized Additive Models

GAMLSS Generalized Additive Models for Location, Scale and Shape

GLM Generalized Linear Models

IF Influence Function

KS Kulldorff Scan

LM Linear Models

MLE Maximum Likelihood Estimation

MSD Mean Square Deviation

MSE Mean Square Error

NPV Negative Predictive Value

RG Robust Fitting GAMLSS proposed by Rigby et al. (2019)

RGW Robust Fitting GAMLSS proposed by Rigby et al. (2019) modified

RS Rigby and Stasinopoulos Algorithm

SS Sensitivity

ZIP Zero Inflated Poisson

CONTENTS

1 INTRODUCTION

1.1 PRELIMINARY

In this work was study two independents topics with different goals and theoretical approaches. Therefore, it is important to clarify the topics to be addressed, as well as to present the structure and content of the next chapters.

The first theme is a non parametric propose for detection of spacial areas that has a higher concentration of events compared to the expected number given a random distribution of events. The main contribution is to present a non parametric method based on empirical likelihood functions, as an alternative to traditional methods of using clusters (scan). The second theme is robust models for Generalized Additive Models for Location, Scale and Shape - GAMLSS, considering the presence of outliers.

The second theme is based on recent publication of the book by Rigby et al. (2019) where an alternative robust estimation method for GAMLSSs is presented. This alternative method seek transformations in order to limit the influence function associated with the probability distribution of interest. There are two main contributions in this theme, and they are presented in two chapters. In the first one, the robust GAMLSS method proposed by Rigby et al. (2019), considering the gamma distribution, presenting the bias corrections for the estimators; a modification in the method proposed by Rigby et al. (2019); and an unprecedented robust proposal using the idea of censored variables is presented. In addition, an extensive simulation study is carried out comparing the performance of the existing proposals. The other chapter aiming to develop a new approach to robustness that relies on a simple adaptive truncation approach where potential outlier contaminated observations are checked and if necessary removed by truncating the distribution of the response variable. The adaptive term is due to the fact that the choice of the tuning constant is performed automatically without the subjectivity of the user's choice. We show that this conceptual simple approach out performs the standard approaches.

The work is organized as follows. In this chapter will be introduced the motivation of the themes in the next section separately. Chapter 2 presents the non parametric Spatial Scan method and is organized as follows. Section 2 review the spatial scan statistic by Kulldorff and Section 2.1 presents the spatial scan statistics based on empirical likelihood. Numerical studies with simulated data are reported in Section 2.2. Section 2.3 shows the simulations results and Section 2.4 reports the application for measles data in São Paulo, Brazil. Finally, Section 2.5 provides the final remarks.

Chapter 3 and 4 present robust fitting for genearlized additive models for location, scale and shape, and they are organized as follows. Section 3.1 described briefly the methodology of classic regression. Section 3.2 defines the GAMLSS model. Section 3.3 present some concepts of robust estimation. Section 3.5 summarizes the robust GAMLSS model proposal by Aeberhard et al. (2021). Section 3.4 present the bases of the propose of robust fitting for a GAMLSS model, the robust fitting of a Gamma distribution and the the modification of Rigby et al. (2019) proposal. Section 3.6 introduces a new approach to robustness using censoring. Section 3.7 report simulation studies and Section 3.8 report an application a real data. Finally, Section 3.9 provides the final remarks. Chapter 4 introduce a new approach to robustness that relies on a truncation distribution. In Section 4.1 we present the motivation for the study of robust estimator of GAMLSS models. Section 4.2 introduces the GAMLSS model.

Section 4.3 discusses our robust proposal. Section 4.4 shown briefly reviews of outliers contamination and Section 4.5 reports the results of a simulation study. Two examples are discussed in Section 4.6 while conclusions are found in Section 4.7. Chapter 5 present the concluding remarks.

1.2 MOIVATION

1.2.1 Spatial Scan Statistics Based on Empirical Likelihood

Spatial cluster is a spatial analysis and mapping technique interested in the identification of clustering of spatial phenomena. A clustering can be defined as an area that has higher concentration of events compared to the expected number given a random distribution of events. Spatial cluster detection studies are important surveillance procedures in public health to prioritize and optimize resources to act against disease outbreaks. Others possibles applications are for instance: the spatial clustering of trees is studied in forestry; Risks of forest fire; and in astronomy to detect particular kind of star.

The use of spatial statistics for disease cluster has received considerable attention in the literature, and a large number of method have been proposed to test the presence of spatial cluster and identify their location. The first method proposed in the literature Besag and Newell (1991); Cuzick and Edwards (1990) can define a cluster by overlapping circles, however, most tests suffer from multiple testing problems due to one or two unknown parameters that must be set prior to their applications. The spatial scan statistics Kulldorff (1997); Kulldorff and Nagarwalla (1995), Tango's test Tango (1995, 2000) were the first methods able to control properly for the type I error.

The spatial scan statistics method identifies the most likely spatial cluster potentially violating the null hypothesis of no clustering. Power comparisons studies of disease clustering tests Kulldorff et al. (2003) show that the scan statistic has been the most powerful for detecting localized clusters.

Some review papers such as Fritz et al. (2013), Moore and Carpenter (1999), Chung et al. (2004), Elliott and Wartenberg (2004), Páez and Scott (2004), Kulldorff et al. (2003), Ozonoff et al. (2005), Aamodt et al. (2006), Duczmal et al. (2011) and Yao et al. (2011) report to innovations, method comparisons, practical issues and are beneficial for promoting knowledge transfer among users and developers in the use of spatial statistics for disease clusters. In addition, some modification of spatial scan statistics have been proposed: Jung (2009) proposed a multivariate adjustment, Loh and Zhu (2007) accommodated correlation, Zhang and Lin (2009) proposed a log-linear modeling, Zhang et al. (2012) presented a model to take into account overdispersion data and Cançado et al. (2011) accommodated zero inflation data. In all these cases, the tests are based on likelihood methods and a family of distribution has to be assumed. However, real data may present substantial departure from the underlying process assumed. One of the possible departures, for instance, is the zero inflated data that may produce biased inference (Gómez-Rubio and López-Quílez (2010) and Loh and Zhu (2007)) and the violation of the Poisson assumption, in the spatial scan statistic can cause excessive type I error probabilities.

This work proposes a non parametric scan method for cluster detection in any family distribution of data, based on empirical likelihood. The main contribution of this method applied in scan statistic is to be able to deal with the presence of overdispersion, zero inflated and other characteristics, that usually occur in real data. The results about this topic has been submitted and accepted for publication in Journal Communication in Statistics - Simulation and Computation. The paper can be accessed in `https://doi.org/10.1080/03610918.2021.1949470`.

1.2.2 A New Approach to Robust Regression Using Censoring

Regression analysis is one of the most popular and powerful statistical techniques that allows exploring and inferring the relationship betweens response variable with specific explanatory variables. The use of regression models is based on classic assumptions about normality, constant variance and

no correlation between the errors terms. The linearity of the relationship between response variable and the explanatory variables is unrealistic in many real situations. Generalized Linear Models - GLM and Generalized Additive models - GAM were introduced by Nelder and Wedderburn (1972a) and Hastie and Tibshirani (1990a), respectively, to solve some of the limitations of the standard linear model. Nevertheless, the GLM and GAM models assume that the response variable belongs to the exponential family. The increasing complexity, have demanded from researchers the development of even more sophisticated statistical methods capable of describing with greater degree of adequacy the interrelationships between variables. Therefore, Rigby and Stasinopoulos (2005) proposed a class of regression models called generalized additive models for location, scale and shape - GAMLSS. It is a univariate statistical modeling technique that allows to fit a wide family of continuous and discrete distributions for the response variable, using parametric and/or non parametric functions, of all parameters of the distribution of the response variable in relation to the explanatory variables. In the GAMLSS, the assumption that the distribution of the response variable belongs to the exponential family is not required, and different additive terms can be included in the predictor for each parameter distribution, which gives flexibility to the model. The generalized additive models for location, scale and shape is a more general class of regression models whose particular cases are the linear regression models, GLM and GAM. The GAMLSS is being widely applied in several fields: Smith et al. (2019)(Modeling spatio-temporal with GAMLSS), De Bastiani et al. (2018) (modelling and fitting of Gaussian Markov random field spatial components), Ramires et al. (2019) (semiparametric Weibull cure rate model), De Castro et al. (2010) (survival models for clinical studies), Glasbey and Khondoker (2009) (normalizing cDNA microarray), Rudge and Gilchrist (2005) (health impact of temperatures in dwellings) WHO (2006, 2007, 2009) (construction of the growth curves used by the World Health Organization) and Hossain et al. (2016) (uses a pulmonary index for the diagnosis of airway obstruction).

Deviations from the model can also occur for GAMLSS. The nature of possible deviations in the GAMLSS class of models are close to what one can see in the regression setting: outliers in the response (producing large residuals). To this end, robust regression is an alternative when data are contaminated with outliers or influential observations. In this sense, the evidence available in the literature, indicate that very little has been accomplished in terms of robust GAMLSS models. Aeberhard et al. (2021) propose a general approach to achieve robustness in fitting GAMLSS by limiting the contribution of observations with low log-likelihood values. This contribution is based on divergence measure approach of Eguchi and Kano (2001) that use a log logistic function that can be interpreted as a multiplicative robustness weight at the log likelihood level. In Aeberhard et al. (2021) was compared just the especial case of a GAM with the following methods: Alimadad and Salibian-Barrera (2011),Croux et al. (2012a) and Wong et al. (2014). The simulations in the special case of a GAM showed that Aeberhard et al. (2021) robust estimator report the best-performing. The recent publication of the book by Rigby et al. (2019) introduce an alternative robust estimation method for GAMLSSs. They robustly fit a GAMLSS model obtaining parameters estimators with bounded influence functions. The book shows a simulated a random sample of size 490 from a $BEo(5,5)$ distribution (beta original) and contaminated it with random samples of size 5 from each of two uniform distributions, $U(0,0.1)$ and $U(0.99,1)$. The simulations results showed that the contamination results in a distorted fit to the data and the method was able to accommodate the outliers. As mentioned by Rigby et al (2019), pag 259, the work is not complete and theoretical and computational aspects still need to be studied. Therefore, the main contributions in this study is purpose a modification in the method proposed by Rigby et al. (2019) and an unprecedented robust proposal using the idea of censored variables is presented. In addition, it also features the robust GAMLSS method proposed by Rigby et al. (2019) considering the gamma distribution, presenting the bias corrections for the estimators; a wide simulation study considering different types of models, sample sizes and contamination intensity were considered to evaluate the existing robust proposals.

1.2.3 A New Approach To Robust Regression Using Adaptive Truncation

In GAMLSS models, parametric terms (linear and non-linear) and additives are used to model p parameters, $\boldsymbol{\theta}^\top = (\theta_1, \ldots, \theta_p)$ of the density probability function $f(y|\boldsymbol{\theta})$, where $\boldsymbol{y}^\top = (y_1, \ldots, y_n)$ is the vector of the response variables. The maximum likelihood estimator of $\boldsymbol{\theta}$ is sensitive to the presence of extreme values (outliers), meaning that the estimated values can be distorted by the presence of outliers. To solve this issues in statistical modelling and data analysis, robust methods began to emerge in the 1960s [Heritier et al. (2009)], with the aim of minimize the impact of outliers and derive methods that produce reliable parameter estimates, with associated confidence intervals and tests.

In the literature, several proposals for robust methods can be found. However, only two works proposed a robust estimation based on GAMLSS model. The first, Aeberhard et al. (2021), introduces robustness by modifying the objective function following an idea introduced by Eguchi and Kano (2001). The second work is an alternative robust estimation method for GAMLSSs presented in the book of Rigby et al. (2019). This method achieves robustness by transformation of the observed response through (normalized) quantile residuals. The Rigby et al. (2019) method depends on a tuning constant that regulates the contributions of an observation to the objective function and the choice of tuning constant is subjective and made before fitting. The Aeberhard et al. (2021) method also depends on tuning constant, and propose a novel general criterion for the selection. It is simulation-based and relies on the heuristic idea of controlling how the robustness weights at the score level behave under data generated from the assumed model.

The main contribution of this study is to present a new approaches for the robust GAMLSS model based on truncation distribution, and propose a selection criterion for the tuning constant, which optimizes the quality of the fit of the data to the assumed model.

2 SPATIAL SCAN STATISTICS BASED ON EMPIRICAL LIKELIHOOD

Suppose that a region G has a partition $G_1,...,G_k$. Let n_i be the population and y_i be the disease count in area, $i = 1,\ldots,k$, hence, $N = \sum_i^k n_i$ and $C = \sum_i^k y_i$ are the population and the number of cases, respectively, in region G. Kulldorff (1997) proposed two possible stochastic models: the counts are independent binomial random variables with parameters n_i and p_i; the counts are independent Poisson random variables with expected value $n_i p_i$. Let the connected areas as a region of interest defined by subareas that share a geographic boundary so that, for a particular subarea i, there is at least one other subarea j that has a common boundary. Let Z be a candidate for a spatial cluster where Z is a subset of connected areas, $y_Z = \sum_{i\in Z} y_i$ be the number of disease cases in Z and $y_{\bar{Z}}$ the number of events outside Z. The risk population inside Z is given by $n_Z = \sum_{i\in Z} n_i$ and $n_{\bar{Z}}$ is the risk outside Z. The method evaluates if the probability of occurrence of an event inside and outside cluster Z is the same. Assume that $p_i = p$ when $i \in Z$ and $p_i = q$ when $i \notin Z$. Under the null hypothesis of no clustering, $p = q$, while under the alternative hypothesis of the presence of spatial cluster $p > q$. Assuming binomial counts in each area, the likelihood under the alternative hypothesis for a fixed cluster candidate is given by

$$L(Z,p,q)_{H_1} = p^{y_Z}(1-p)^{n_Z-y_z} q^{C-y_Z}(1-q)^{(N-n_Z)-(C-y_Z)}.$$

The maximum likelihood estimator (MLE) for p and q are $\hat{p} = \frac{c_Z}{n_Z}$ and $\hat{q} = \frac{C-c_Z}{N-n_Z}$. A likelihood ratio test statistic search for the most likely cluster, and the cluster Z with the highest likelihood ratio function is given by

$$K_Z = \frac{L(Z,\hat{p},\hat{q})_{H_1}}{L_{H_0}},$$

where $L_{H_0} = C^C(N-C)^{N-C}$. The maximum likelihood ratio test statistic for unspecified spatial cluster Z is given by

$$K = \max_{(Z\in\mathscr{Z})} K_Z,$$

where $\mathscr{Z}$ is the set of all possibles connected areas in G. The number of areas in $\mathscr{Z}$ is finite but is usually too big. This includes the individual regions as clusters and all the others possibilities. Although finding K requires the evaluation of K_Z only a finite number of times, this is unfeasible because the number of possible clusters is extremely large, except for very small number of areas. Kulldorff (1997) solved this problem defining a smaller class that contains a reasonable number cluster candidates. This class contains all circles centered on the centroid areas and with arbitrary radius r up to an upper limit defined by the operator. The centroid area is defined as the geometric center of subareas. The null distribution of K is analytically untractable (Kulldorff and Nagarwalla (1995)). To find the distribution of the test statistic under the null hypothesis Monte Carlo hypothesis testing is required Dwass (1957). The standard Spatial Scan Statistic algorithm is summarized as follow.

Step 1: Select an area and define a circle centered at the centroid with radius r with the nearest neighbor, i.e., the first cluster Z is defined;

Step 2: Compute the statistic K_Z for cluster Z;

Step 3: Increase the radius r of circle cluster including the next nearest neighbor up to an upper limit;

Step 4: Compute the statistic K_Z for each new Z cluster;

Step 5: Compute the maximum K for the centroid;

Step 6: Repeat steps 1-5 for each centroid;

Step 7: Compute the maximum likelihood ratio test statistic for all centroids, hence find the most likely cluster;

Step 8: Use the Monte Carlo method to estimate the null distribution of K.

The spatial scan statistic is one of the most important methods for detecting and monitoring spatial disease clusters. However, it has some limitations:

- the rigid geometry of the spatial cluster candidates. In practice, the cluster may have a elongated shape (factors as long river or main road). Other kind of geometry was proposed in the literature with different forms.
- It is assumed that disease cases follow a Poisson or Binomial spatial process but the cases count data sets frequently present an excess of zeros, overdispersion Zhang et al. (2012) and others process de Lima et al. (2015), resulting in violation of the assumptions, increasing type I error and leads to a incorrect inference for the model parameters.

The likelihood ratio test has good power properties and they are very efficient. The main disadvantage is that a family of distributions has to be assumed for the data. The next section presents a non parametric version of spatial scan statistics.

2.1 SPATIAL SCAN STATISTICS BASED ON EMPIRICAL LIKELIHOOD

2.1.1 Empirical likelihood for Two Samples

Empirical likelihood Owen (1988) is a type of nonparametric likelihood which can be used to obtain a nonparametric version of the theorem of Wilks (1938), where it does not assume a parametric family of distributions for the data. The method proposed in this thesis is centred around the concept of empirical likelihood to find the most likely cluster. A summary of the proposed method is presented as follows.

Let $X_1,\ldots,X_m$ and $Y_1,\ldots,Y_n$ be two independent and identically distributed samples from the random variables X and Y, respectively, with $E(X)=\mu_X$, $E(Y)=\mu_Y$, $Var(X)=\sigma_X{}^2$ and $Var(Y)=\sigma_Y{}^2$, where $\sigma_Y^2>0$, $\sigma_X^2>0$, $\mu_X\in\mathscr{R}$ and $\mu_Y\in\mathscr{R}$. Let $\theta=\mu_X-\mu_Y$ be the parameter of interest. Consider $p=(p_1,\ldots,p_m):p_i\in[0,1]$for $i=1,\ldots,m$ and $q=(q_1,\ldots,q_n):q_i\in[0,1]$ for $i=1,\ldots,n$ two sets of probability measure imposed over the two samples. The empirical loglikelihood (ELL) for difference of the means of two samples (θ) is

$$ELL(\theta)=\max_{(p,q)}\left(\sum_{i=1}^{m}\log(p_i)+\sum_{i=1}^{n}\log(q_i)\right), \tag{2.1}$$

subject to $p_i>0$, $q_i>0$, $\sum_i^m p_i=1$, $\sum_i^n q_i=1$ and $\sum_i^m p_iX_i-\sum_i^n q_iY_i=\theta$.

Suppose that it is necessary to test the hypotheses $H_0 : \theta = \theta_0$ and $H_1 : \theta \neq \theta_0$. The empirical loglikelihood ratio statistic is given by

$$ELLR_{\theta_0} = \frac{ELL_{H_0}}{ELL_{H_1}}, \tag{2.2}$$

where ELL_{H_0} and ELL_{H_1} is maximized using the Lagrange method under H_0 and H_1, respectively. Under H_1, the function to be maximized is given in Equation (2.1), subject to $\sum_i^m p_i = 1$ and $\sum_i^n q_i = 1$. The Lagrange multiplier method to maximize a function $f(w)$ subject to the constraint $j(w) = 0$ has the following steps.

Step 1 Calculate the value of $w = w_\lambda$, which solves $J(w) = \nabla f(w) - \lambda \nabla j(w) = 0$,where ∇ is the gradient operator.

Step 2 Obtain λ to solve $j(w_\lambda)$, where λ is the Lagrange multipliers.

Thus, the functions used in the Lagrange multiplier for the situation considered is

$$f(p,q) = \max_{(p,q)} \left(\sum_{i=1}^{m} \log(p_i) + \sum_{i=1}^{n} \log(q_i) \right), \quad j(p) = 1 - \sum_{i=1}^{n} p_i \text{ and } j(q) = 1 - \sum_{i=1}^{n} q_i.$$

The equation for J becomes

$$J(p_i, q_i) = \sum_{i=1}^{m} \log(p_i) + \sum_{i=1}^{n} \log(q_i) - \lambda_1 (1 - \sum_{i=1}^{m} p_i) - \lambda_2 (1 - \sum_{i=1}^{n} q_i).$$

The first step of the Lagrange multiplier method is to differentiate the J function and calculate the critical values of this function, where the critical values are the points where the derivative function is zero. The derivatives of J are

$$\frac{\partial J}{\partial p_i} = \frac{1}{p_i} - \lambda_1,$$

for $i = 1, \ldots, m$, and

$$\frac{\partial J}{\partial q_i} = \frac{1}{q_i} - \lambda_2,$$

for $i = 1, \ldots, n$. Solving these equations we obtain $p_i = \frac{1}{\lambda_1}$ and $q_i = \frac{1}{\lambda_2}$. Hence,

$$\sum_{i=1}^{m} p_i = \sum_{i=1}^{m} \frac{1}{\lambda_1} = 1 \rightarrow \lambda_1 = m,$$

and

$$\sum_{i=1}^{n} q_i = \sum_{i=1}^{n} \frac{1}{\lambda_2} = 1 \rightarrow \lambda_2 = n,$$

which yields

$$p_i = \frac{1}{m} \quad \text{and} \quad q_i = \frac{1}{n}.$$

The Lagrange multiplier method implies that $ELL_{H_1} = m^{-m} n^{-n}$. Hence, the likelihood ratio statistic on the parameter of interest, θ, is defined as

$$ELLR_\theta = \max_{(p,q)} \left(\sum_{i=1}^{m} \log(m \times p_i) + \sum_{i=1}^{n} \log(n \times q_i) \right),$$

subject to the constraints given in Equation (2.1). The J function used in Lagrange multiplier under H_0 is

$$\begin{aligned} J(p,q) = \; & \textstyle\sum_{i=1}^{m} \log(m \times p_i) + \sum_{i=1}^{n} \log(n \times q_i) - \lambda_1 \left[1 - \sum_{i=1}^{m} p_i\right] - \lambda_2 \left[1 - \sum_{i=1}^{n} q_i\right] \\ & -\lambda_3 \left[\sum_{i=1}^{m} p_i(X_i - \mu)\right] - \lambda_4 \left[\sum_{i=1}^{n} q_i(Y_i - \mu)\right]. \end{aligned}$$

The derivatives of the J function are

$$\frac{\partial J}{\partial p_i} = \frac{1}{p_i} - \lambda_1 - \lambda_3(X_i - \mu),$$

for $i = 1, \ldots, m$, and

$$\frac{\partial J}{\partial q_i} = \frac{1}{q_i} - \lambda_2 - \lambda_4(Y_i - \mu),$$

for $i = 1, \ldots, n$. Solving $\sum_{i=1}^{m} p_i \frac{\partial J}{\partial p_i} = 0$ and $\sum_{i=1}^{n} q_i \frac{\partial J}{\partial q_i} = 0$, we obtain, $\lambda_1 = m$ and $\lambda_2 = n$. Thus, setting $\frac{\partial J}{\partial p_i} = 0$ and $\frac{\partial J}{\partial q_i} = 0$,

$$p_i = \frac{1}{m + \lambda_3(x_i - \mu)},$$

where λ_3 solves

$$\sum_{i=1}^{m} \frac{(x_i - \mu)}{m + \lambda_3(x_i - \mu)} = 0,$$

and

$$q_i = \frac{1}{n + \lambda_4(y_i - \mu)},$$

where λ_4 solves

$$\sum_{i=1}^{n} \frac{(y_i - \mu)}{n + \lambda_4(y_i - \mu)} = 0.$$

The $ELLR_{\theta_0}$ can be obtained numerically. One of the important properties of $ELLR_{\theta_0}$ is that it can be used to obtain a nonparametric version of the Wilk's Theorem.

Non parametric version of the Wilk's Theorem (Wu and Yan (2012)): Suppose $\sigma_X^2 < \infty$, $\sigma_X^2 < \infty$ and $\frac{m}{o} \to \pi \in (0,1)$ as $o \to \infty$, where $o = m + n$. Then, $-2 \times ELLR_{\theta_0}$ converges in distribution to a χ_1^2 random variable with one degree of freedom, i.e.

$$-2 \times ELLR_{\theta_0} \xrightarrow{d} \chi_1^2.$$

The empirical likelihood theory described above is expressed under the alternative hypotheses - two side. Then, the Scan Statistic based on empirical likelihood has two new problems:

1. The Scan Statistic compares two samples and searches (Scan) for the most likely cluster based in fact that mean of cluster is largest of the other, i.e., the alternative hypotheses is an inequality $H_1 : \mu_X > \mu_Y$. Hence, how to maximize an empirical likelihood when subject to inequality hypotheses?

2. When the alterative hipothesis is $H_1 : \mu_X > \mu_Y$ the asymptotic distribution $-2 \times ELLR$ is not necessarily chi-squared (see Chapter 10,Owen (1988)). We cannot expect to find the distribution of the ratio in the closed analytical form. How to calibrate the likelihood ratios and make inference?

In the first problem our proposal is to use the augmented Lagrange multiplier method with a Sequential Quadratic Programming-SQP interior algorithm, proposed by Nocedal and Wright (2006), to maximize the empirical likelihood. In the second problem the propose is to use bootstrap method. Initiated by Efron in 1979, the basic bootstrap approach use re-sampling to generate an empirical estimate of the statistic sampling distribution. The EL method combined with Bootstrap method produces very accurate results (Owen (1988)).

2.1.2 Algorithm

Our proposal is based following the Kulldorff's approach and algorithm, distinguishing in the use of test statistics and obtaining the p-value. In the following, only the bootstrap algorithm for p-value is described because the algorithm is similar.

Suppose that a region G has a partition $G_1,\ldots,G_k$. For each G_i, let us to consider n_i be the population and y_i be the disease counts in area $i=1,\ldots,k$, hence, $N=\sum_i^k n_i$ and $C=\sum_i^k y_i$ are the population and the number of cases, respectively, in region G. Let Z be a candidate for a spatial cluster where z is a subset of connected areas. We assumed that $E(\frac{y_i}{n_i})=\mu_Z$ if $i\in Z$ and $E(\frac{y_i}{n_i})=\mu_{\bar{Z}}$ if $i\notin Z$. Let $\theta=\mu_Z-\mu_{\bar{Z}}$. Our interest is the hypothesis testing problem of the null hypothesis $H_0:\theta=0$ against the alternative hypothesis $H_1:\theta\geq 0$ using the statistic $ELLR$ and Bootstrap method.

The Bootstrap Test Algorithm

The basic concern when using bootstrap, is to formulate a re-sampling mechanism that leads to the distribution of the test statistic under the null hypothesis. Let $\mathbf{u}$ the union of the random samples of the cluster candidate $\mathbf{y_Z}=(y_1/n_1,\ldots,y_{k_1}/n_{k_1})$ and its complementary region $\mathbf{y_{\bar{Z}}}=(y_1/n_1,\ldots,y_{k_2}/n_{k_2})$, respectively, where $k_1+k_2=k$ and $k_1<k_2$. The empirical distribution function of $\mathbf{u}$ is defined as $\hat{J}_0$ and characterizes the probabilistic mechanism (Efron and Tibshirani (1994)). $\hat{J}_0$ is a non parametric estimate of the common distribution J_0 that would originate both $\mathbf{y_Z}$ and $\mathbf{y_{\bar{Z}}}$. Then, the natural estimation of probabilistic mechanism J under H_0 is obtained by bootstrap sampling $\mathbf{u}^*$ (for more details see cap. 16 of Efron and Tibshirani (1994)). In the following is described the algorithm of the computation of the bootstrap test statistic .

Step 1: On the basis of real observations $\mathbf{u}=(\mathbf{y_Z},\mathbf{y_{\bar{Z}}})=(u_1,\ldots,u_k)$, compute $ELLR_0$ (Equation (2.2)) and denote it by $ELLR_0^o$;

Step 2: Generate the bootstrap sample $\mathbf{u}^*=(u_1^*,\ldots,u_k^*)$;

Step 3: On the basis of bootstrap sample, perform Algorithm 1 to compute the simulated value $ELLR_0^*$;

Step 4: Repeat the step 2 and 3 B times;

Step 5: Denote $ELLR_{0,b}^*$ as the bth simulated value of $ELLR_0$ derived in step 2, 3 and 4. Then, the p-value of $ELLR_0$ equals $\#\{ELLR_0^o\geq ELLR_{0,b}^*:b=1,\ldots,B\}/B$ where $\#A$ represents the number of elements in set A.

2.2 SIMULATIONS STUDIES

In this Section the process of simulation of the data sets is presented, under the null hypothesis and the alternative hypothesis, used to estimate the distribution of statistic $-2\times ELLR_{\theta_0}$, to calculated the empirical power and size of the hypothesis test. The simulations of the data sets will be based

on the ZIP distribution, described above, and can be extended to any distribution family. The ZIP distribution is a discrete mixture with two components: zero with probability ϕ and a Poisson family distribution - $PO(\mu)$ - with probability $1-\phi$. The ZIP model is described as

$$Y_i = \begin{cases} 0 & \text{, with probability } \phi \\ \mathscr{P}(\mu_i) & \text{, with probability } 1-\phi \end{cases},$$

where $\mu > 0$ and $0 < \phi < 1$ is the zero inflated parameter, i. e., the probability of extra zeros. Hence, the probability function of ZIP is given by

$$P(Y = y|\mu,\phi) = \begin{cases} \phi + (1-\phi)\exp(-\mu) & y = 0 \\ (1-\phi)\frac{\mu^y \exp(-\mu)}{y!} & y = 1,2,3,\ldots \end{cases}$$

The mean and variance are given by Rigby et al. (2019) which are defined by

$$\begin{aligned} E(Y) &= (1-\phi)\mu \quad , \\ Var(y) &= (1-\phi)\mu + \phi(1-\phi)\mu^2 \quad . \end{aligned}$$

To describe this model in terms of null hypothesis, consider a region G partitioned into k disjoint areas. Let n_i be the population and y_i be the event counts in area $i = 1,\ldots,k$. Then, the hypothesis of clustering or no clustering under ZIP model can be express as

$H_0 : Y_i \sim ZIP(\lambda n_i, \phi), \forall i,$

$H_1 : Y_i \sim ZIP(\lambda_j n_i, \phi),\ j = 1,2$ where $\lambda_1 > \lambda_2$ for some set of i (connected areas),

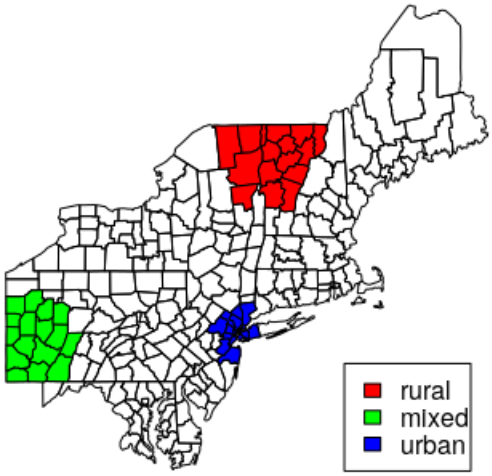
where λ represents the incidence rate of the event, which is unknown in general.

2.2.1 Cluster Models

The simulations were based on Kulldorff et al. (2003), which presented a collection of 1.220.000 simulated benchmark data sets generated under Poisson distribution, with 51 different cluster models and under the null hypothesis, to be used for power evaluations. The region considered is the 245 counties and county equivalents in the Northeastern United States, consisting of the states of Maine, New Hampshire, Vermont, Massachusetts, Rhode Island, Connecticut, New York, New Jersey, Pennsylvania, Delaware and Maryland, as well as the District of Columbia. The clusters were based on three different sets of local clusters: rural, urban and mixed area. Within each of these three sets, was constructed the different size clusters using 2, 4, 8 and 16 counties. The center of the rural cluster was Grand Isle County in northern Vermont, on the Canadian border. The center of the mixed cluster was Pittsburgh (Allegheny County) in western Pennsylvania. The center of the urban cluster was Manhattan (New York County) in New York City, closely surrounded by other very urban counties. Hence, there are 12 cluster types, where they differ in the number of cities in the cluster (2,4,8,16) and the cluster location (mixed, rural and urban). The clusters with 16 counties and the three location are shown in Figure 1. We considered these simulated data under the null hypothesis of the Poisson model to compare the effectiveness of the proposed Empirical Likelihood Scan method and the method proposed by Kulldorff and Nagarwalla (1995).

In the simulation of Zero Inflated Poisson distribution, as well as Kulldorff et al. (2003), we use the real female population in the 245 counties and county equivalents in the Northeastern United States to simulate a Zero Inflated Poisson model. As the population for each county we used the number of women living there according to the 1990 United States census. These data have been previously used to evaluate the existence of geographical clusters of breast cancer mortality in Kulldorff (1997)

Figure 1 – Map showing the urban cluster (blue) centered around Manhattan, New York, in the center, the mixed cluster (green) centered around Pittsburgh in the west and the rural cluster (red) centered around Grand Isle in the north.



Source: The author (2021)

2.3 SIMULATIONS RESULTS

Firstly, it was evaluated and compare the power functions between the ration test using empirical likelihood and the ration test using the likelihood based on Poisson model, without the scan process, that is, it was compared the hypothesis testing for two samples. In the simulations, it was compared the power functions of $ELLR$ and K generated under a Zero Inflated Poisson - ZIP with inflated proportion $\phi = 40\%$ and $\phi = 10\%$, based on samples with size $n_1 = 15$ and $n_2 = 5$, see Figure 2. We use 500 samples runs and 500 bootstrap to calculate the p-values for $\alpha = 5\%$. We used δ to measure the strength of the cluster from 1 to 20, with $\delta = 1$ indicating no spatial cluster effect and $\delta = 20$ indicating strong spatial cluster effect, that is, we assumed that $E[y_i] = 0.001 * \delta$, where 0.001 is the incidence rate.

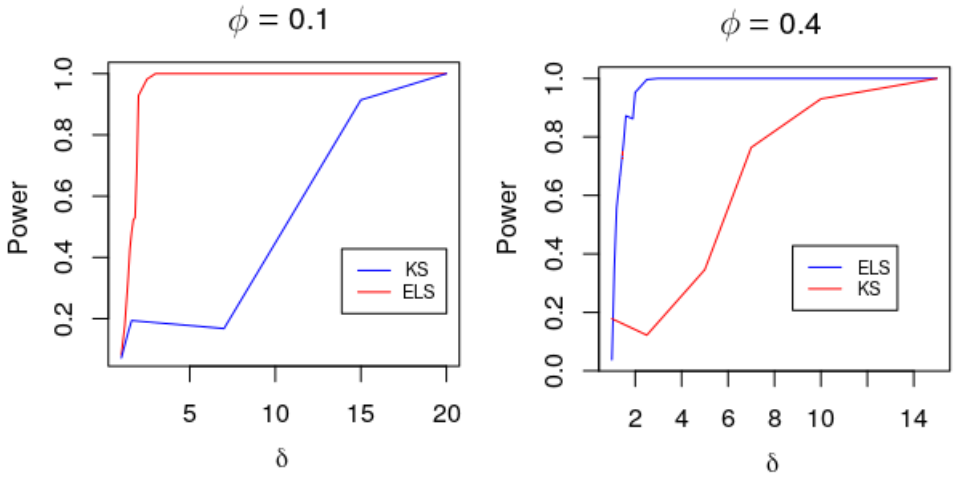
In the simulation, when $\phi = 0.1$, the type I error $(\delta = 1)$ is 0.082 and 0.072 and when $\phi = 0.4$ the type I error $(\delta = 1)$ is 0.04 and 0.18 for $ELLR$ and K, respectively. These results suggested that the type I error probabilities of $ELLR$ were not significantly affected by zero inflated effects. The comparison of the powers showed that the ration based on empirical likelihood is always higher than the Kulldorff's ratio in presence of zero inflated.

2.3.1 Power and Size Comparison

In this section, the results of the empirical power and the size of the tests based on the Poisson and Zero inflated Poisson models will be presented. Furthermore, two indicators were used to evaluate the accuracy of methods.

Costa and Assunção (2005), Tango and Takahashi (2005), Costa et al. (2012) evaluated the efficiency of the most likely cluster matched the real cluster, using sensitivity measures. Let $\hat{z}$ and z be the most likely cluster and the real cluster, respectively, with corresponding populations $n_{\hat{Z}}$ and n_Z. Sensitivity is the average proportion of the real cluster identified correctly. $SS = \frac{n_{\hat{Z} \cap Z}}{n_Z}$. The Negative Predictive Value - NPV is the average proportion of the detected cluster that is part of the complementary real cluster, i.e., the proportion of the detected cluster incorrectly identified, $NPV = \frac{n_{\hat{Z} \cap Z^c}}{n_{\hat{Z}}}$. A good method has large sensitivity and a low NPV. If the average proportion of SS and NPV is close to one and zero, respectively, the detected cluster is likely to cover a high proportion of the real clusters

Figure 2 – Simulations of power functions as functions of δ for selected ϕ, with samples size $n_1 = 15$, $n_2 = 5$ and $\alpha = 5\%$.



Source: The author (2021)

and with lower error.

2.3.1.1 Simulation Based on Poisson Model

Spatial Scan Statistic based on Empirical Likelihood were evaluated under the Poisson model. The type I error rates and the empirical power were compared with the Kulldorff's Scan. The results were based on 500 simulated data under the null and alternative hypothesis of each type of different cluster size (2,4,8 and, 16). To obtain the p-value for the Kulldorff Scan method each data were replicated 500 times using Monte Carlo method, and for the Empirical Likelihood Scan 500 bootstrap samples were generated for each data.

Table 1 presents the test size for Empirical Likelihood Scan - ELS and Kulldorff Spatial Scan - KS. The results are better for the ELS method, since the type I error probabilities is closer to the significance level for the ELS method than for the KS method, i.e. the ELS method can control better the false alarm rates (type I error probability). The results also show that the computation of the power of the test is more precise for the ELS method, when compared to the KS method.

Table 1 – Size of the Spatial Scan test considering 500 simulated data under the Poisson model.

ELS 5%	KS 5%	ELS 1%	KS 1%
3,4%	1,0%	0,4%	0,6%

KS - Kuldorff Statistic Scan ; ELS - Empirical Likelihood Scan; Source: The author (2021)

The results of empirical power and the indicators rates are reported in Table 2. The empirical power values are lower for the ELS method for all clusters than for the KS method. Moreover, this is expected by the fact that KS method is parametric and the ELS method is non parametric. Indeed, the power of empirical likelihood converges to the Kulldorf method when the size of the cluster increases. Note that the empirical power values are more precise for the ELS method, according to the results given in Table 1.

The results of *SS* and *NPV* indicators are presented in Table 3. It is shown that for the scenario considered, in average for all clusters, the Kulldorff method gives better results. The Empirical Likelihood Scan method identified correctly 62.9% of the cluster with 8 cities and the Kulldorff method has 82% of *SS*, while the *NPV* indicates that the proportion average of error is 36.2% for ELS method

Table 2 – Empirical Power values for ELS and KS method under Poisson model.

Cluster	$\alpha = 0,01$		$\alpha = 0,05$	
Size	ELS	KS	ELS	KS
2	0,182	0,872	0,380	0.904
4	0,338	0,822	0,648	0,892
8	0,646	0,860	0,826	0.920
16	0,754	0,874	0,897	0,918

KS - Kuldorff Statistic Scan ; ELS - Empirical Likelihood Scan; Source: The author (2021)

and 25.2% for KS method.

Table 3 – Sensitivity and Negative Predictive Value Indicators.

Cluster	SS		NPV	
Size	ELS	KS	ELS	KS
2	0,231	0,903	0,94	0,306
4	0,441	0,885	0,72	0,419
8	0,629	0,82	0,362	0,252
16	0,451	0,791	0,181	0,170

KS - Kuldorff Statistic Scan ; ELS - Empirical Likelihood Scan; Source: The author (2021).

2.3.1.2 Simulation Based on ZIP Model

The Spatial Scan Statistic based on Empirical Likelihood were evaluated from ZIP distribution, compared type I error rates and power between Kulldorff's Scan and the Empirical Likelihood Scan. Based on ZIP distribution with $\phi = 40\%$, we generated random ZIP counts (500 simulations runs), and calculate bootstrap p-values for both methods from 500 bootstrap samples. The regions were simulated using incidence rate of 100 cases per 100.000 persons and the clusters were simulated with incidence rate of 1.000 cases per 100.000 persons, according with the hypothesis described and the cluster of size 2, 4, 8, 16 as described in previous section.

Table 4 shows the type I error probabilities for Empirical Likelihood Scan - ELS and Kulldorff Spatial Scan - KS. The Kulldorff Scan assumes that the number of disease cases in different locations have independent Poisson distributions leads to an increased rate of false positives with 100% and 100% for $\alpha = 5\%$ and $\alpha = 1\%$, respectively. Its expected that the type I error probabilities (false alarm rate) is the same as the significance level and this behavior is better in Empirical Likelihood method with 2% and 1% for $\alpha = 5\%$ and $\alpha = 1\%$, respectively. The Kulldorf's method is not able to control the zero inflated effects, with a false alarm rate of 100%.

Table 4 – Size of the Spatial Scan test by Bootstrap estimation (500 replicates). Using ZIP model ($\phi = 40\%$).

KS 5%	ELS 5%	KS 1%	ELS 1%
1%	2%	1%	1%

KS - Kuldorff Statistic Scan ; ELS - Empirical Likelihood Scan; Source: The author (2021)

The results of power are report in Table 5. The ELS method shows a low power for clusters of sizes 2, 4 and 8 for $\alpha = 0,01$ and clusters of sizes 2 and 4 for $\alpha = 0,05$. This can be explained by the fact that Kulldorf's Scan does not account zero inflated but Empirical Likelihood Scan does. Furthermore, the zero inflated effects is stronger in clusters of sizes 2 and 4.

Table 6 shows the *SS* and *NPV* results. On average, clusters with sizes 8 and 16 achieved the best *SS* indicating that the Empirical Likelihood Scan method identified correctly 100% and 90.3% of the

Table 5 – Empirical Power for $\delta = 10$ and $\phi = 0.4$

Cluster	$\alpha = 0,01$		$\alpha = 0,05$	
Size	ELS	KS	ELS	KS
2	0	1	0	1
4	0,066	1	0,162	1
8	0,052	1	0,656	1
16	1	1	1	1

KS - Kuldorff Statistic Scan ; ELS - Empirical Likelihood Scan; Source: The author (2021)

size of the real cluster, while the *NPV* indicate that the proportion average of error is 4% for both size cluster 8 and 16.

Table 6 – Sensitivity and Negative Predictive Value Indicators

Cluster	SS		NPV	
Size	ELS	KS	ELS	KS
2	0,253	0,5	0,90	0,5
4	0,782	0,75	0,42	0,0
8	1	0,875	0,04	0,0
16	0,905	0,898	0,046	0,085

KS - Kuldorff Statistic Scan ; ELS - Empirical Likelihood Scan; Source: The author (2021)

2.4 APPLICATIONS

Measles is an infectious disease, highly contagious and caused by viruses, can be contracted by people of any age, but has a higher incidence in children under one years old (Veronesi and Focaccia (2015)). Contamination is transmitted by air, caused by sneezing, coughing and other contact with contaminated secretions (Veronesi and Focaccia (2015)). It has a strong impact of socioeconomic aspects on the transmission and incidence of the disease.

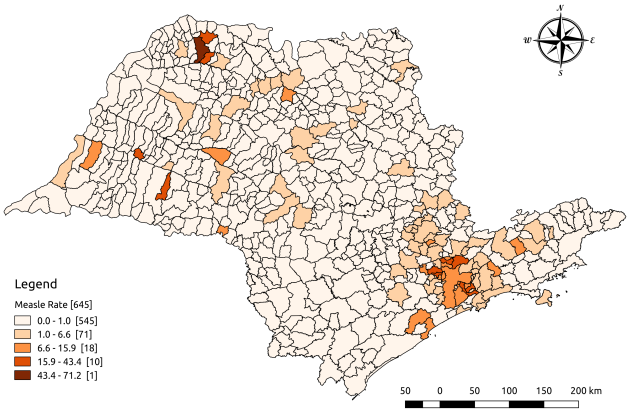
Brazil received from the Pan American Health Organization in 2016 the certificate of measles eradication. However, the outbreak in the north of the country in 2018 broke the cycle of absence of measles, evidencing the need for disease control, surveillance and eradication. In 2019, Brazil reported two outbreaks of the disease with 4.447 confirmed measles cases until September 18 (de Saúde do Estado de São Paulo (2019)). Until June 23, 2019, 3.906 cases of measles have been registered and 97.5% of these cases are concentrated in 153 counties of the state of São Paulo (de Saúde do Estado de São Paulo (2019)). With this scenario, an epidemiological investigation of measles outbreaks in the state of São Paulo is of fundamental importance for proper health surveillance.

São Paulo (SP) is a Brazilian state located in the Southeast region, bordering the state of Rio de Janeiro to the northeast; with Minas Gerais to the north; with Mato Grosso do Sul to the west; and with Paraná to the south. In its eastern and southeastern portion is bathed by the Atlantic Ocean. It has a territorial area of 248222 km^2, where approximately 46.670.000 people living (IBGE, 2019), totaling a demographic density of 166 inhabitants per square kilometer.

As mentioned earlier, São Paulo presented a large part of the occurrence of measles cases in Brazil in 2019. In this case study, the region of São Paulo was defined for the application of the proposed method and the data to be used are the 2982 confirmed measles cases from January 1st, 2019 to September 4th, 2019 provided by the São Paulo State Department of Health. Figure 3 presents the map of the administrative boundary of the 645 counties that make up the state of São Paulo and the spatial distribution of measles incidence in 2019 through September 4. Measles cases are concentrated in 111 counties of the state, with the largest occurring in its capital São Paulo, recording 67.15% of all cases.

The center of the high incidence of measles is concentrated in the metropolitan region of São Paulo, presenting incidence rates varying from 16 to 43 cases per 100 thousand inhabitants. Fernandópolis, in north of São Paulo is the county with the highest incident rate, 71 cases per inhabitants. The data presented a distribution with excess of zeros with 82.8% of the counties (534) having 0 (zero) occurrences of measles cases, featuring zero inflation.

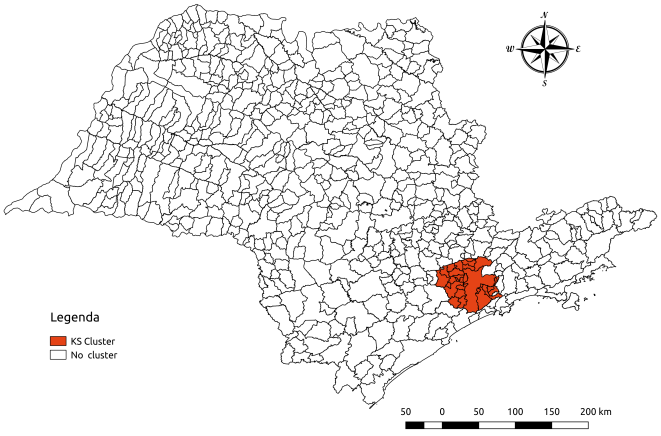
Figure 3 – Map showing the incidence rate of measles



Source: The author (2021)

The Kulldorff method identified a cluster with 28 regions (Figure 4) with an incidence rate of 14 per 100.000 inhabitants, while the region has a rate of 7 cases per 100.000 inhabitants. These results should be interpreted with caution due to the high occurrence of zeros in the whole region, which will not characterize the Poisson distribution, required in the application of the Kulldorff method. The ELS statistic was constructed using the measles occurrence rate as a variable of interest and with a maximum proportion of the population in the cluster of 50%. The statistical test did not indicate any significant cluster.

Figure 4 – Kuldorff Cluster of Measles cases in São Paulo in Brasil, 2019



Source: The author (2021)

2.5 CONCLUSIONS

In this chapter, a non-parametric cluster identification method based on Kuldorff's Scan Statistics and the Empirical Likelihood theory was proposed. The searching procedure between Kuldorff's method and the one based in empirical likelihood proposed in this thesis is similar, but the test statistic used in every scan is different.

In the context of scan statistics for simulated data and real data, the ELS method was efficient for clusters with larger number of regions, being able to reduce Type I error. False alarm rates should be taken into consideration when using scan methods. As shown in the analysis, the presence of zero inflation is associated with 100% type I error probability using the Kuldorff Scan method. In terms of Public Health management, this probability indicates that all cluster identify will be false alarms, while to the Empirical Likelihood the probability type I is controlled. Is expected that the type I error (false alarm rate) is the same as the significance level and this behavior is better in ELS method. This result indicates the presence of a cluster when in reality it does not exist. Hence, when the data are inflated with zeros the Spatial Scan by Kulldorff should be used with caution.

The Empirical Likelihood method is usually effective and powerful in dealing with populations with skewed distribution. The main contribution of the proposal is the possibility of application to any family of distributions and the ability to apply continuous variables duly associated with a population at risk.

The method proposed in that the empirical likelihood methods applied to spatial scan can be very promising and is a good candidate for a method for detecting clusters, but the mean is a non-robust statistic that is affected by the presence of outliers and this fact linked to the choice of the Poisson model inflated by zeros may have influenced the level of empirical power.

An improvement for future work could be to compare with other non parametric methods and to consider methods to deal with the non-circular form of the cluster and work with another parameter of interest.

3 A NEW APPROACH TO ROBUST REGRESSION USING CENSORING FOR GAMLSS

3.1 PRELIMINARY

This section briefly describes the methodology of classical regression and introduces the generalized additive models for location, scale and shape (GAMLSS).

The purpose of a regression is to establish a quantifiable dependency relationship between variables that can be expressed through a mathematical model, which has all its fixed components, or even by a statistical model, when was include at least one random component.

A specific class of a statistical model, called as multiple linear model regression - LM, is a statistical model that uses the relationship between two or more variables so that one of them can be described or its value estimated from the others, can be defined as:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_s x_{is} + \varepsilon_i,$$

where Y_i are random variables (response variables), for $i = 1, \ldots, n$, β_s are coefficients to be estimated, s is the number of explanatory variables, ε_i are the random errors independently distributed normal variables, with zero mean and constant variance σ^2, that is, $\varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$. In the following, we have another way to define the model: $Y_i \sim N(\mu_i, \sigma^2)$ and $\mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_s x_{is}$ for $i = 1, \ldots, n$. To avoid mathematical notation problems, we adopt the matrix form to define the models. The linear model in matrix form is defined as

$$\mathbf{Y} \overset{\text{i.i.d.}}{\sim} N(\boldsymbol{\mu}, \mathbf{I}\sigma^2)$$

$$\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta},$$

where $\boldsymbol{\beta}$ is the unknown vector of $s+1$ coefficients to be estimated and $\mathbf{X}$ is a design matrix $n \times (s+1)$ with s explanatory variables and a column representing the constant coefficient. The assumption of normality of random errors is an assumption which excludes a range of situations, such as, for example, binary response variables, proportion, counting or even categorical.

Nelder and Wedderburn (1972) proposed the class of Generalized Linear Models - GLM whose basic idea is to expand the possibility of other distributions for the response variable besides the distribution normal. In essence, an GLM is defined by a probability distribution that belongs to the exponential family for the response variable (random component), a set of independent variables describing the linear structure of the model (systematic component) and a monotonic link function $g(.)$ used in modelling the relationship between $\boldsymbol{\mu}$ and the explanatory variables. Data analysis through GLM is quite flexible, because for the same linear structure you can obtain several models depending on the random component and the chosen link function (Nelder and Wedderburn (1972a)).

Let $EF(\boldsymbol{\mu}, \boldsymbol{\Phi})$ the exponential family distribution, where μ and Φ are the vectors of the parameters of location and scale parameters. Distributions, such as normal, inverse normal, beta, Poisson, binomial and gamma are examples of distributions that belong exponential family and can be used to fit generalized linear models. Therefore, GLM model can be written as:

$$\mathbf{y} \overset{\text{i.i.d.}}{\sim} EF(\boldsymbol{\mu}, \boldsymbol{\Phi})$$

$$g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta},$$

where, $\mathbf{X}$ is design matrix and $\boldsymbol{\beta}$ is the unknown vector of coefficients. The $g(.)$ is the link function and must necessarily be twice differentiable and strictly monotonic. GLM are widely used, however, may also be unsuitable for some situations, such as example, when the relationship between the mean of the response variable and the explanatory variables is not linear.

In this context, the GAM, proposed by Hastie and Tibshirani (1990), add non-parametric smoothing functions to GLM, in order to leave their own data conduct their relationship with the predictor. The Generalized Additive Models can be definded as:

$$\mathbf{Y} \sim \text{EF}(\boldsymbol{\mu}, \boldsymbol{\Phi})$$

$$g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta} + \sum_{j=1}^{J} f_j(x_j)$$

where $\mathbf{X}$ is a matrix $n \times (s+1)$, f_j are non-parametric smoothing function applied to covariates x_j with $j = 1, \ldots, s$ and $\boldsymbol{\beta}$ is the unknown vector of $s+1$ coefficients. For more details on the LM, GLM and GAM mentioned above, see Nelder and Wedderburn (1972a) and Hastie and Tibshirani (1990a).

GLM and GAM are still limited, since only the location parameter (mean) of the distributions is modeled, and necessarily, the distribution needs to belong to the exponential family. There are situations where may require more flexibility to the distribution of the response variable, such as, it is desired that it has a large asymmetry and kurtosis.

To remedy the above restrictions, Rigby and Stasinopoulos (2005) proposed the generalized additives models for location, scale and shape - GAMLSS, a new class of models (semi) parametric regression models, which allow that all parameters of the response variable be modeled in a linear or non-linear function. In the next section, the GAMLSS model was presented as well as the process of estimation.

3.2 GENERALIZED ADDITIVE MODELS FOR LOCATION, SCALE AND SHAPE - GAMLSS

The GAMLSS regression model proposed by Rigby and Stasinopoulos (2005) allows the fitting of any distribution for the response variable, regardless of whether it belongs to a family of distributions. The GAMLSS model class also allows the systematic part of the model to be expanded, so that all parameters of location, scale and shape, of the chosen distribution, are fitted. That is, all parameters can be modeled according to the explanatory variables and, in addition, the predictors can also incorporate non-parametric smoothing functions, random effects, or other terms additions. Though presenting the LM, GLM and GAM models as particular cases, this model has the assumption that the observations of the Y response variable are independent. According to Stasinopoulos et al (2017), the GAMLSS can be defined as follows.

Let y_i, for $i = 1, \ldots, n$, be the response variable observations independent, with probability (density) function $f(y_i|\boldsymbol{\theta}^i)$, where ${\boldsymbol{\theta}^i}^\top = (\theta_1^i, \ldots, \theta_p^i)$ is a vector of p parameters that is related to the effects of explanatory variables and random effects through monotonic link function $g_k(\boldsymbol{\theta_k})$, for $k = 1, \ldots, p$. This function (g_k) is defined as additive model given by

$$g_k(\boldsymbol{\theta}_k) = \eta_k = \mathbf{X}_k\boldsymbol{\beta}_k + \sum_{j=1}^{J_k} \mathbf{Z}_{jk}\boldsymbol{\gamma}_{jk}, \tag{3.1}$$

where, $\boldsymbol{\theta}_k^\top = (\theta_k^1, \ldots, \theta_k^n)$, $\boldsymbol{\eta}_k^\top = (\eta_{1k}, \ldots, \eta_{nk})$ are vectors of length n, $\boldsymbol{\beta}_k^\top = (\beta_{1k}, \ldots, \beta_{b_k k})$ is a parameter vector of length b_k, X_k is a known design matrix of order $n \times b_k$, $\mathbf{Z}_{jk}$ is a fixed known

$n \times q_{jk}$ design matrix and $\boldsymbol{\gamma}_{jk}$ is a q_{jk}-dimensional random variable. The matrices $\boldsymbol{X}_k$ may or may not be equal, that is, it is the predictor of each parameter of the distribution can receive different explanatory variables (Rigby and Stasinopoulos (2005)).

Model (3.1) is called the GAMLLSS and is more general than the GLM or GAM. The distribution of the dependent variable is not limited to the exponential family and all parameters are modelled in terms of both fixed and random effects. See below some special cases of GAMLSS.

1. If $J_k = 0$ for $k = 1,2,\ldots,p$, there are no additive terms associated with the distribution parameters then model (3.1) reduces to a fully linear parametric GAMLSS model given by:

$$g_k(\boldsymbol{\theta}_k) = \eta_k = \mathbf{X}_k\boldsymbol{\beta}_k. \tag{3.2}$$

2. The fully linear parametric GAMLSS can be extended to allow the inclusion of terms nonlinear modeling of the k distribution parameters in the form:

$$g_k(\boldsymbol{\theta}_k) = \eta_k = h_k(\mathbf{x}_k, \boldsymbol{\beta}_k), \tag{3.3}$$

   where h_k for $k = 1,\ldots,p$ are non linear function and $\mathbf{x}_k$ is a explanatory vector assumed to be known.

3. If $\mathbf{Z}_{jk} = \mathbf{I}_n$, is an $n \times n$ identity matrix, and $\boldsymbol{\gamma}_{jk} = \mathbf{h}_{jk} = h_{jk}(\mathbf{x}_{jk})$ for all combination of j and k then Model (3.1) reduces to a linear semi parametric GAMLSS given by:

$$g_k(\boldsymbol{\theta}_k) = \eta_k = \mathbf{X}_k\boldsymbol{\beta}_k + \sum_{j=1}^{J_k} h_{jk}(\mathbf{x}_{jk}), \tag{3.4}$$

   where $\mathbf{x}_{jk}$, for $j = 1,2,\ldots,J_k$ and $k = 1,2,\ldots,p$, are vectors of length n, the function h_{jk} is an unknown function of the explanatory variable X_{jk}.

4. The linear semiparametric GAMLSS can be extended to allow the inclusion of terms nonlinear modeling of the k distribution parameters. If $\mathbf{Z}_{jk} = \mathbf{I}_n$, is the $n \times n$ identity matrix, and $\boldsymbol{\gamma}_{jk} = \mathbf{h}_{jk} = h_{jk}(\mathbf{x}_{jk})$ for all combination of j and k then model 3.1 reduces to a non linear semi parametric GAMLSS given by:

$$g_k(\boldsymbol{\theta}_k) = \eta_k = h_k(\mathbf{x}_k, \boldsymbol{\beta}_k) + \sum_{j=1}^{J_k} h_{jk}(\mathbf{x}_{jk}), \tag{3.5}$$

   where $\mathbf{x}_{jk}$ for $j = 1,2,\ldots,J_k$ and $k = 1,2,\ldots,p$, are vectors of length n, the function h_{jk} is an unknown function of the explanatory variable X_{jk}, $\mathbf{x}_{jk}$ for $j = 1,2,\ldots,J_k$ and $k = 1,2,\ldots,p$ are vectors of length n.

5. If $\mathbf{Z}_{jk} = \mathbf{I}_n$, is an $n \times n$ identity matrix, and $\boldsymbol{\gamma}_{jk} = \mathbf{h}_{jk} = h_{jk}(\mathbf{x}_{jk})$ for specific combinations of j and k in Model (3.1), then the resulting model contains parametric, nonparametric and random-effects terms.

3.2.1 Parameters Estimation

In this section, some techniques for estimating the parameters and hyperparameters of the GAMLSS models will be presented. The parameters estimation is based on two iterative estimation algorithms, the RS algorithm, proposed by Rigby and Stasinopoulos (2005), and CG, proposed by Cole and Green (1992).

GAMLSS models, previously defined in Equation (3.1), can be written as

$$\mathbf{Y} \sim \mathscr{D}(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\tau}, \boldsymbol{\nu})$$

$$\begin{aligned}
g(\boldsymbol{\mu}) &= \eta_1 = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{Z_{11}}\boldsymbol{\gamma}_{11} + \ldots + \mathbf{Z_{J_1 1}}\boldsymbol{\gamma}_{J_1 1}, \\
g(\boldsymbol{\sigma}) &= \eta_2 = \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{Z_{12}}\boldsymbol{\gamma}_{12} + \ldots + \mathbf{Z_{J_2 2}}\boldsymbol{\gamma}_{J_2 2}, \\
g(\boldsymbol{\tau}) &= \eta_3 = \mathbf{X}_3\boldsymbol{\beta}_3 + \mathbf{Z_{13}}\boldsymbol{\gamma}_{13} + \ldots + \mathbf{Z_{J_3 3}}\boldsymbol{\gamma}_{J_3 3}, \\
g(\boldsymbol{\nu}) &= \eta_4 = \mathbf{X}_4\boldsymbol{\beta}_4 + \mathbf{Z_{14}}\boldsymbol{\gamma}_{14} + \ldots + \mathbf{Z_{J_4 4}}\boldsymbol{\gamma}_{J_4 4},
\end{aligned}$$

where $\mathscr{D}(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\tau}, \boldsymbol{\nu})$ is a four parameter distribution, $\boldsymbol{\mu}$ is usually a location parameter, $\boldsymbol{\sigma}$ is often a scale parameter, $\boldsymbol{\nu}$ and $\boldsymbol{\tau}$ are the shape parameters of the distribution, generally associated with skewness and kurtosis, respectively. The fixed effect parameters are represented by $\boldsymbol{\beta} = (\boldsymbol{\beta_1}^\top, \boldsymbol{\beta_2}^\top, \boldsymbol{\beta_3}^\top, \boldsymbol{\beta_4}^\top)^\top$ are the fixed effects parameters, $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_{11}^\top \ldots, \boldsymbol{\gamma}_{J_1 1}^\top, \boldsymbol{\gamma}_{21}^\top \ldots, \boldsymbol{\gamma}_{J_4 4}^\top)^\top$ are the random effect parameters assuming that $\boldsymbol{\gamma}_{jk}$ have independent normal distributions with $\boldsymbol{\gamma}_{jk} \sim N_{q_{jk}}(\mathbf{0}, \mathbf{G}_{jk}^-)$, where $\mathbf{G}_{jk}^-$ is the generalized inverse of a q_{jk} x q_{jk} symmetric matrix $\mathrm{G}_{jk} = \mathrm{G}_{jk}(\boldsymbol{\lambda}_{jk})$, which depend on a vector of hyperparameters $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_{11}^\top \ldots, \boldsymbol{\lambda}_{J_1 1}^\top, \boldsymbol{\lambda}_{21}^\top \ldots, \boldsymbol{\lambda}_{J_4 4}^\top)^\top$ and regulates the degree of smoothing required in the fit.

When the model does not have random effects, that is, it does not have a smoothing function, then we have a parametric model that only requires the estimation of $\boldsymbol{\beta}$. In this case, the model parameters are estimated by maximum likelihood. As we assume a four-parameter model, the loglikelihood function is defined as

$$l = \sum_{i=1}^{n} \log\{f(y_i|\mu_i, \sigma_i, \tau_i, \nu_i)\}.$$

For a GAMLSS model with random effects, the penalized maximum likelihood estimation method is used, which refers to the estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ for constant $\boldsymbol{\lambda}$. Thus, the penalized likelihood function is given by

$$l_p = l - \frac{1}{2}\sum_{k=1}^{4}\sum_{j=1}^{J_k} \boldsymbol{\gamma}_{jk}^T G_{jk} \boldsymbol{\gamma}_{jk}. \tag{3.6}$$

Rigby and Stasinopoulos (2005) proposed two algorithms to maximize the likelihood function penalized (Equation (3.6)) and fit a GAMLSS for fixed values of hyperparameters, the CG and RS algorithms. The first, CG algorithm, is a generalization of the Cole and Green algorithm (1992), which uses the first derivatives and the exact values or approximate the second and cross derivatives. However, for many probabilities density functions the parameters have orthogonal information, that is, the values of the cross-derivatives of the likelihood function are equal to zero. In this case, it is used the RS algorithm which does not use the expected value of the cross derivatives. A third methodology, which mixes steps from both algorithms, starts the process by RS and finish with CG. The RS and CG algorithms lead to (penalised) maximum likelihood estimators for $\boldsymbol{\gamma}$ and esteem the parameters $\boldsymbol{\beta}$ for fixed hyperparameters $\boldsymbol{\lambda}$. The first proposals for hyperparameter estimation use the optimization method to minimize the Akaike criterion and thus select the best estimate for $\boldsymbol{\lambda}$. However, this process requires the adjustment of several models, resulting in a high computational cost, especially if the model has several smoothing functions (Rigby and Stasinopoulos (2013)). In this context, Rigby and Stasinopoulos (2014) proposed a way to automatically select the value of this parameter, which they call local estimation, because the method is applied in each step of the RS and CG algorithm, using the maximum likelihood theories. This alternative has the main advantage of being faster compared to the others. In addition, there are, at least, three different methodologies for estimating the smoothing hyperparameters: generalised cross validation, generalised akaike information criterion and maximum likelihood based methods (for more details on parameter estimation see Chapter 3 Stasinopoulos et al. (2017)).

3.3 ROBUST STATISTICS

Robust statistics aims at producing consistent and possibly efficient estimators, test statistics with stable level and power, when the model is slightly misspecified (Heritier et al. (2009)). Huber (1996) approaches the definition of robustness in three aspects: qualitative, quantitative and infinitesimal. The qualitative concept is based on the principle of fundamental continuity of robustness, which postulates that small disturbances in the underlying probability distribution should cause small changes in the performance of the statistical method used in the analysis. Let a cumulative probability distribution F_θ that captures the structural part as well as the random part of the model. A model is said to be misspecified if the data generating lies in a neighborhood of the true (postulated) model. This notion of a neighborhood, due originally to Huber (1965), is defined as:

$$F_\varepsilon = (1-\varepsilon)F_\theta + \varepsilon G, \tag{3.7}$$

where θ is a set of parameters of interest, G is an arbitrary distribution and $0 \leq \varepsilon \leq 1$. When $\varepsilon = 0$, then there is no model misspecification and the data-generating process is exactly the postulated model. This is the assumption in classical estimation based, for example, on the maximum likelihood estimator. A common approach of violating the probability distribution assumptions is the contamination of the sample with outliers.

The definition of quantitative robustness is based on the breakdown point concept, and measures the property of robustness of a statistic. The breaking point is defined as the amount maximum of bad specification of the probabilistic model that an estimator can resist before breakdown (Huber (1996)), that is, before produces inconsistent and inefficient estimators. The infinitesimal definition is based on the concept of the influence function that will be defined later. In this work, will be used the qualitative and infinitesimal aspects for formulate the new propose for robust fitting GAMLSS. More details can be found in Huber (1965, 2004), Heritier et al. (2009) andFarcomeni and Ventura (2012).

3.3.1 Influence Function

Influence function - IF is a useful concept for studying the robustness properties of an estimator. The influence function (IF) measures the effect on the estimator of an infinitesimal contamination on any observation in the sample and gives us the idea of how an estimator would look under point contamination. A particular case of the function F_ε, defined in (3.7), occurs when G is a distribution in which the value g occurs with probability 1. Thus, if X follows distribution G, then $P(X \leq x) = 0$ if $x < g$ and the mean of X is $E(X) = g$. The function in (3.7) is rewritten as $F_{\varepsilon,g} = (1-\varepsilon)F_\theta + \varepsilon G_g$. The purpose of defining function $F_{\varepsilon,g}$ is to observe how the value g affects the value of a function or estimator of the F_θ distribution, when g occurs with probability ε. It can be noted that when ε is small enough, the $F_{\varepsilon,g}$ and F_θ distributions are quite similar. The relative influence of the value g on an estimator $T()$, is given by

$$\frac{T(F_{\varepsilon,g}) - T(F_\theta)}{\varepsilon}$$

The influence function is the relative influence of g on an estimator F_θ, when the probability of contamination by g tends to zero

$$IF = \lim_{\varepsilon \to 0} \frac{T(F_\varepsilon) - T(F_\theta)}{\varepsilon}$$

It is an approximation to the behavior of $\hat{\theta}$ when the sample contains a small fraction ε of outliers. This influence function measures the impact of an infinitesimal contamination at y on the estimator. If an estimator is robust, IF should not be arbitrarily large for any value of y. In other words, IF should be bounded for all values of y if the estimator is robust. For a more discussion on influence function, see, for example, Hampel et al. (2011).

3.3.2 The Class of M-Estimators

Let $(y_1, y_2, \ldots, y_n)$ are i.i.d. observations generated from a distribution with cdf $F(y, \theta)$ and an unknown parameter θ. Huber (1996) proposed the class of M-estimators that naturally generalize the MLE. An M-estimator of θ is given by the solution $\hat{\theta}_{[M]}$ of the minimization problem

$$\min_{\theta} \sum_{i=1}^{n} \rho(y_i; \theta) \tag{3.8}$$

or, alternatively, by the solution for θ of

$$\sum_{i=1}^{n} \Upsilon(y_i; \theta) = 0 \tag{3.9}$$

for suitable ρ and Υ functions where $\Upsilon(y_i; \theta) = \frac{\partial \rho(y;\theta)}{\partial \theta}$. Note that, if $\rho = -\log(f(y|\theta))$ then M-estimator is the maximum likelihood estimator. In general, Υ needs not be the derivative of some ρ-function with respect to the parameter of interest, hence the Equation (3.9) is more general and is often referred as the proper definition of an M-estimator Heritier et al. (2009). The Huber function is one special case of M-estimator. It is defined by:

$$\rho_k(x) = \begin{cases} \frac{1}{2}x^2, & |x| \leq k. \\ k|x| - \frac{1}{2}k^2, & |x| > k. \end{cases} \tag{3.10}$$

Other classes of robust estimators are the R-estimators (R-estimators) and L-estimators (HUBER, 1996).

In statistical analysis, often outliers are observed and may have huge influences on the estimated model. In this way, robust statistics have the purpose of modeling the discrepancies, delimiting the influence of these outliers, to make them more stable, avoiding the model parameters from being under or overestimated.

3.4 ROBUST FITTING OF A GAMLSS MODEL

This work is centred around in the proposal presented in Rigby et al. (2019), that consists of transforming the IF of the maximum likelihood estimation. In general, maximum likelihood estimation leads to parameter estimators with unbounded influence function for some or all parameters. Hence, maximum likelihood estimation are generally vulnerable to the unbounded influence of even a single outlier y. For a robust estimator of parameter θ it is necessary that the influence function, for parameters $\boldsymbol{\theta_k}$ for $k = 1, \ldots, p$, be bounded as y moves into the left or right tail of the distribution of Y. The next lemma features the influence function of the MLE of $\boldsymbol{\theta}$, based on influence function of general M-estimates of Huber (2004), p45.

Lemma 1 - IF for the MLE (Rigby et al. (2019)) Let Y be a random variable with probability density function $f_Y(y; \boldsymbol{\theta})$ and cumulative distribution function $F_Y(y; \boldsymbol{\theta})$. Let $IF(y; \hat{\theta})$ for the maximum likelihood estimator $\hat{\theta}$ of parameter θ_k for $k = 1, \ldots, K$, assuming a random sample from $f_y(y; \theta)$. Let $IF(y; \hat{\boldsymbol{\theta}})$ be the corresponding vector of influence functions, i.e. $IF(y; \hat{\boldsymbol{\theta}}) = [IF(y; \hat{\theta}_1), \ldots,$
$IF(y; \hat{\theta}_k)]^\top$. Then,

$$IF(y, \hat{\boldsymbol{\theta}}) = \boldsymbol{A}^{-1} \frac{\partial l}{\partial \hat{\boldsymbol{\theta}}},$$

where $\boldsymbol{A} = -E\left\{\frac{\partial^2 l}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}\right\}$ is the expected information matrix and $\frac{\partial l}{\partial \boldsymbol{\theta}}$ is the vector of first derivatives of the log density function, i. e., $l = log f_Y(y; \boldsymbol{\theta})$. This lemma can be found in Heritier et al. (2009); Rigby et al. (2019).

Rigby et al. (2019) find the influence functions for the MLEs of all parameters of each of four example distributions for Y ($normal(NO)$, $beta(BE)$, $gamma(GA)$ and $tfamily(TF)$ distribution) and check whether the influence function bounded, for the each parameter, is bounded or unbounded as outlier value y moves towards one of the ends of the range of Y. Table 7 gives a summary of theses results considering the refer B as bounded an UB as unbounded.

Table 7 – Bounded - B or unbounded - UB influence functions for MLE of parameters of distributions.

	Left tail			Right tail		
	μ	σ	ν	μ	σ	ν
$NO(\mu,\sigma)$	UN	UN		UN	UN	
$BE(\mu,\sigma)$	UN	B		B	UN	
$GA(\mu,\sigma)$	B	UN		UN	UN	
$TF(\mu,\sigma,\nu)$	B	B	UN	B	B	UN

Source: Rigby et al pg. 251 (2019).

Consider a value y of a response variable Y, one outlier relative to the model is a value of y for which $F_Y(y;\theta)$ is very closed to 0 or 1. Some strategies can be used when dealing with outliers. The robust fitting of a GAMLSS model was initially proposed in Rigby et al. (2019) and exemplified only for the beta distribution and constant systematic component. In addiction, it was not done simulation studies or comparison with existing robust procedure. In this sense, we will carry out a simulation study for the beta distribution and extend the proposal to the gamma distribution

The proposal is based in bounded the influence function to obtain the parameter estimators. In the previous section was defined the general class of M-estimators, the proposal can be considered as an M-estimation Huber (2004) defined as

$$\rho(y,\beta,\gamma) = \psi(l(\beta,\gamma)) - b_\psi,$$

where $\psi(l(\beta,\gamma))$ is a bounded $l(\beta,\gamma)$ obtained by bounding y_i by setting

$$y_i^* = \begin{cases} y_i, & \Phi^{-1}(\alpha) \leq r_i \leq \Phi^{-1}(1-\alpha) \\ F_Y^{-1}(\alpha;\hat{\theta}_\mathbf{i}), & r_i < \Phi^{-1}(\alpha) \\ F_Y^{-1}(1-\alpha;\hat{\theta}_\mathbf{i}), & r_i > \Phi^{-1}(1-\alpha), \end{cases} \tag{3.11}$$

where α is a probability close to zero (e.g. $\alpha = 0.01$), $r_i = \Phi^{-1}[F_Y(y_i;\hat{\theta}_\mathbf{i}]$ is the normalized quantile residual Dunn and Smyth (1996a) and where, $\hat{\theta}_i = (\hat{\mu}_i, \hat{\sigma}_i, \hat{\nu}_i, \hat{\tau}_i)$. The value α can be seen as robustness tuning probabilities that regulates the contribution of observations based on the normalized quantile residual. The choice of α is made before fitting the model to data, however, there is no criterion for the choice of α. The term $b_\psi(\boldsymbol{\theta})$ guarantees that the estimator is Fisher consistent, that is, asymptotically unbiased under the postulated model (see, Heritege, 2009, section 2.3.2). This term can sometimes be difficult to compute and so is evaluated by numerical integration (Piessens et al. (2012)) . This term is defined as $E[\psi(l(\boldsymbol{\beta},\boldsymbol{\gamma})]$. Then in the GAMLSS fitting, for a given smoothing parameter $\boldsymbol{\lambda}$, the robust estimator for the $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ is defined by maximizing the rewriting penalized log-likelihood defined in (3.6) as:

$$l_{p_r}(\boldsymbol{\beta},\boldsymbol{\gamma}) = \psi(l(\boldsymbol{\beta},\boldsymbol{\gamma})) - b_\psi - \frac{1}{2}\sum_{k=1}^{4}\sum_{j=1}^{J_k} \boldsymbol{\gamma}_{jk}^T G_{jk} \boldsymbol{\gamma}_{jk}. \tag{3.12}$$

Rigby et al. (2019) suggest the following practical robustness procedures, iterate between (RE)fit the model robustly and identify and remove gross outliers. Then, procedure of robust fitting GAMLSS is as follow:

Step 1 Compute the estimation $\hat{\boldsymbol{\theta}}$ on the original data by maximizing (3.6) using a CG and RS algorithm and get global deviance and residual of the estimated model;

Step 2 For a given robustness tuning probabilities (α), with the estimation $\hat{\boldsymbol{\theta}}$, bounded the response variable y based on transformation defined in Equation (3.11), generating the new y values defined as y^*.

Step 3 Compute the estimation $\hat{\boldsymbol{\theta}}$ on the y^* data by maximizing (3.12).

Step 4 Stop if there is no change global deviance otherwise go back to step 2.

The first step of the procedure starts from using the residual of the model fit and the global deviance of the classical GAMLSS defined as $-2l(\hat{\boldsymbol{\theta}})$, where $l(\hat{\boldsymbol{\theta}})$ is the logarithm of the fitted likelihood function. The GAMLSS algorithm use the general idea of the local scoring algorithm that repeated weighted fits to a modified response variable using modified weights until convergence when the maximum is reached Stasinopoulos et al. (2017). Hence, following the suggest of the practical robustness procedures we defined a second proposal that weight observations that the residuals are too big, that is, set weights equal to zero when the normalized quantile residual exceed values of a second threshold. For a complete evaluation of the practical robustness procedure, these algorithms will be called robust fitting GAMLSS - RG and weighted robust fitting GAMLSS - RGW. Therefore, the first contributions of this work are

- ✓ Modify the structure of the robust process, set weights equal to zero using the normalized quantile residual, called as RGW;
- ✓ Present a broader simulation study for the beta distribution;
- ✓ Introduce the robust fitting GAMLSS model for the gamma distribution.
- ✓ Introduce a new approach to robustness using censoring for GAMLSS (will be presented in the Section 3.6)

3.4.1 A New Robust Fitting GAMLSS for Gamma Distribution

In this section will be presented the contribution of this thesis on GAMLSS robustness, described the robust fitting of a gamma distribution with parameters μ and σ and showing the bias correction for the parameters estimators of the gamma model. The functions will be available within one of the `gamlss` packages in R (R Core Team (2020)). Assume $Y \sim GA(\mu,\sigma)$ then:

$$f_Y(y;\sigma,\mu) = \frac{y^{\frac{1}{\sigma^2}-1}\exp\left[\frac{-y}{(\sigma^2\mu)}\right]}{(\sigma^2\mu)^{\frac{1}{\sigma^2}}\Gamma\left(\frac{1}{\sigma^2}\right)}, \tag{3.13}$$

where $y > 0$, $\mu > 0$ and $\sigma > 0$. The log-likelihood and first derivatives of the log density function, $l = \log(f_Y(y;\mu,\sigma))$, with respect to μ ans σ are given by:

$$l = \log(f_Y(y;\mu,\sigma)) = \left(\frac{1}{\sigma^2}-1\right)\log(y) - \frac{y}{\sigma^2\mu} - \frac{1}{\sigma^2}\log(\sigma^2) - \frac{1}{\sigma^2}\log(\mu) - \log\left(\Gamma\left(\frac{1}{\sigma^2}\right)\right).$$

$$\frac{\partial l}{\partial \mu} = \frac{1}{\sigma^2\mu^2}(y-\mu). \tag{3.14}$$

$$\frac{\partial l}{\partial \sigma} = \frac{2}{\sigma^3}\left[\frac{y}{\mu} + \Upsilon\left(\frac{1}{\sigma^2}\right) - \log\left(\frac{y}{\sigma^2\mu}\right) - 1\right], \tag{3.15}$$

where $\Upsilon(x) = \frac{\Gamma'(x)}{\Gamma(x)}$. The functions $\psi\left(\frac{\partial l}{\partial \mu}\right)$ and $\psi\left(\frac{\partial l}{\partial \sigma}\right)$ are bounded versions of $\frac{\partial l}{\partial \mu}$ and $\frac{\partial l}{\partial \sigma}$ obtained by setting

$$y_i^* = \begin{cases} y_i, & a_i \leq y_i \leq b_i \\ a_i, & y_i < a_i \\ b_i, & y_i > b_i \end{cases},$$

where $a_i = F_Y^{-1}(\alpha; \mu_i, \sigma_i)$, $b_i = F_Y^{-1}(1-\alpha; \mu_i, \sigma_i)$ and α is a small probability. Hence, within the GAMLSS algorithm, $\frac{\partial l_i}{\partial \mu_i}$ and $\frac{\partial l_i}{\partial \sigma_i}$ are replaced by

$$h_{1i} = \psi\left(\frac{\partial l_i}{\partial \mu_i}\right) - E\left[\psi\left(\frac{\partial l_i}{\partial \mu_i}\right)\right]$$

and

$$h_{2i} = \psi\left(\frac{\partial l_i}{\partial \sigma_i}\right) - E\left[\psi\left(\frac{\partial l_i}{\partial \sigma_i}\right)\right],$$

respectively, where the expected values provide the bias correction.

Bias Correction

Bias corrections for the estimators of μ and σ are presented below. The procedure for obtaining the bias was based on the methodology used in Rigby et al. (2019) but only developed for the beta distribution.

Bias correction for estimator of μ: Here we fit the gamma distribution ($GA(\mu, \sigma)$) robustly, showing the bias correction for $\hat{\mu}$. The first derivative of the log density function with respect to μ is given in Equation (3.14).

$$E\left[\psi\left(\frac{\partial l}{\partial \mu}\right)\right] = \int_{b_i}^{a_i} \frac{\partial l}{\partial \mu} f_Y(y_i; \mu_i, \sigma_i)\, dy + \alpha\left[\frac{\partial l}{\partial \mu_i}\bigg|_{a_i} + \frac{\partial l}{\partial \mu_i}\bigg|_{b_i}\right]$$

$$= \int_{b_i}^{a_i} \frac{1}{\sigma_i^2 \mu_i}(y_i - \mu) \frac{y_i^{\frac{1}{\sigma_i^2}-1} \exp\left[\frac{-y_i}{(\sigma_i^2 \mu_i)}\right]}{(\sigma_i^2 \mu_i)^{\frac{1}{\sigma_i^2}} \Gamma\left(\frac{1}{\sigma_i^2}\right)} dy + \alpha\left[\frac{1}{\sigma_i^2 \mu_i^2}(a_i - \mu_i) + \frac{1}{\sigma_i^2 \mu_i^2}(b_i - \mu_i)\right].$$

The integral in equation above is available as

$$= \frac{(\sigma_i \mu_i)^{-2}}{(\sigma_i^2 \mu_i)^{\sigma_i^{-2}} \Gamma\left(\frac{1}{\sigma_i^2}\right)} \int_{b_i}^{a_i} (y_i - \mu_i) y_i^{\sigma_i^{-2}-1} \exp\left(\frac{-y_i}{\sigma_i^2 \mu_i}\right) dy.$$

$$= \frac{(\sigma_i \mu_i)^{-2}}{(\sigma_i^2 \mu_i)^{\sigma_i^{-2}} \Gamma\left(\frac{1}{\sigma_i^2}\right)} \left[\int_{b_i}^{a_i} y_i^{\sigma_i^{-2}} \exp\left(\frac{-y_i}{\sigma_i^2 \mu_i}\right) dy - \int_{b_i}^{a_i} \mu_i y_i^{\sigma_i^{-2}-1} \exp\left(\frac{-y_i}{\sigma_i^2 \mu_i}\right) dy\right].$$

Using integration by parts,

$$\int_{b_i}^{a_i} y_i^{\sigma_i^{-2}} exp\left(\frac{-y_i}{\sigma_i^2 \mu_i}\right) dy = \frac{-y_i^{\sigma_i^{-2}} exp\left(\frac{-y_i}{\sigma_i^2 \mu_i}\right)}{\sigma_i^{-2} \mu_i^{-1}}\bigg|_{a_i}^{b_i} + \int_{b_i}^{a_i} \mu_i y_i^{\sigma_i^{-2}-1} exp\left(\frac{-y_i}{\sigma_i^2 \mu_i}\right) dy.$$

Hence,

$$E\left[\psi\left(\frac{\partial l}{\partial \mu}\right)\right] = \frac{1}{(\sigma_i^2 \mu_i)^{\sigma^{-2}} \Gamma(\frac{1}{\sigma^2}) \mu}\left[-b_i^{\sigma_i^{-2}} exp\left(\frac{-b_i}{\sigma_i^2 \mu_i}\right) + a_i^{\sigma_i^{-2}} exp\left(\frac{-a_i}{\sigma_i^2 \mu_i}\right)\right] + \alpha\left[\frac{(a_i - \mu_i)}{\sigma_i^2 \mu_i^2} + \frac{(b_i - \mu_i)}{\sigma_i^2 \mu_i^2}\right].$$

Bias correction for estimator of σ: Here we fit the gamma distribution ($GA(\mu,\sigma)$) robustly, showing the bias correction for $\hat{\sigma}$. The first derivative of the log density function with respect to σ is given by Equation (3.15).

$$E\left[\psi\left(\frac{\partial l_i}{\partial \sigma_i}\right)\right] = \int_{b_i}^{a_i} \frac{\partial l}{\partial \sigma_i} f_Y(y_i;\mu_i,\sigma_i)\,dy + \alpha\left[\frac{\partial l}{\partial \mu_i}\bigg|_{a_i} + \frac{\partial l}{\partial \mu_i}\bigg|_{b_i}\right].$$

The integral is not tractable, so it is evaluated by numerical integration. The numerical integration used here is based on series extrapolation methods. The central idea of these techniques was originally presented by Longman (1956), in which it is based on the slowly converging alternating series transformation of Euler. Many variations of extrapolation techniques were proposed after the approach of Longman. Among them, the algorithm (Wynn (1956)) stands out, which is an efficient recursive implementation of the Shanks Transform (Shanks (1955)) . For more details see Piessens et al. (2012).

3.5 ROBUST FITTING FOR GAMLSS B Robust Fitting for GAMLSS BY AEBEHARD ET. AL. (2021)

In order to compare and evaluate our proposal for a robust fitting GAMLSS model, it was considered the proposed by Aeberhard et al. (2021). To our knowledge, this is the only robust fitting GAMLSS model available in the literature besides Rigby et al (2019). The study is a general approach to achieve robustness in fitting GAMLSS by limiting the contribution of observations with low log-likelihood values, based on divergence measure approach of Eguchi and Kano (2001). Hence, the penalized log-likelihood function (3.6) is redefined as $l_{p_{ah}}$, is given by

$$l_{p_{ah}}(\boldsymbol{\beta},\boldsymbol{\gamma}) = \rho_c(l(\boldsymbol{\beta},\boldsymbol{\gamma})) - b_{\rho c} - \frac{1}{2}\sum_{k=1}^{p}\sum_{j=1}^{J_k} \boldsymbol{\gamma}_{jk}^T G_{jk}\boldsymbol{\gamma}_{jk}. \tag{3.16}$$

where,

$$\rho_c(z) = \log\frac{1+\exp(z+c)}{1+\exp(c)},\ c>0,$$

$$b_{\rho_c} = \sum_{i=1}^{n} b_{\rho_{ci}} = \sum_{i=1}^{n}\int \rho_c^*(\log f(y|\boldsymbol{\theta}^i))dy$$

is a correction term ensuring Fisher consistency and ρ_c^* is directly derived from the specified ρ_c through:

$$\rho_c^* = \int_{-\infty}^{z} exp(s)\rho_c^{'} ds.$$

In this case,

$$\rho_c^* = exp(z) - exp(c)log(1+exp(z+c)).$$

The robustness tuning constant c controls how early ρ_c starts to diminish the contribution of an observation to the objective function (3.16). To maximize Equation (3.16), they use the algorithm of Marra et al. (2017) to accommodate the robustified objective function and corresponding correction term b_{ρ_c}. Aeberhard et al. (2021) propose a novel general criterion for the selection of the tuning constant c based on median downweighting proportion (MDP) and studied in more details the robust fitting for a gamma distribution within GAMLSS framework.

3.6 A NEW APPROACH TO ROBUSTNESS USING CENSORING FOR GAMLSS

In this section it will be introduced the main contribution of this chapter, the idea for a new robust estimation method for the broad class of GAMLSS. This approach uses the same structure to bounded the influence function of the data, as set out in Section (3.4), however, the modeling is carried out using the concepts of censored data, which will be briefly presented in this chapter. Hence, will be unified the concepts of the transformation (3.4) and censored data to propose a new robust fitting GAMLSS method.

3.6.1 Censored Distributions

In several fields of science the interest is in situations where the random variable cannot be observed completely for all individuals in the experiment, but instead there is only one interval in which this variable is contained, characterizing what we call censored data (see Colosimo and Giolo (2006)).

There are three types of censoring, the most common being right censoring, which occurs when the interval observed is of the type $[a,\infty)$ for some finite constant known, that is, when we know that the true value of the variable of interest is greater than the observed value a. In several practical situations, censoring happens for reasons such as the limitations of the measuring equipment or the experimental design. For example, a digital scale that does not provide a reading above 200 kg, then it will shows 200 kg for all objects that weigh more than the limit.

The second type of censoring is on the left, when the true value of the interest is less than the observed value a. In this case, the observed range is of the type $(\infty,a]$, where a is a finite and known constant. In a school exam, the minimum percentage of correct answers for the approval is of 40% Breen (1996).

The last type of censoring is interval, which occurs when it is only possible to observe an interval finite type $[a,b]$ in which the true value of the variable is contained, with $|a| < \infty$, $|b| < \infty$ and $a < b$. Interval is a more general type of censoring that occurs, for example, in studies in which patients are followed up on periodic visits and it is known only that the event of interest occurred within a certain time interval.

The new robust fitting GAMLSS proposal is based on use of censored structure to generate a better representation of bounded the influence function to proposal a new robust estimation for GAMLSS.

3.6.2 Maximum Likelihood

Consider a random sample $t_1,\ldots,t_n$ from a T with probability distribution $f(t|\theta)$ and cumulative distribution function $F_\theta(t)$, where all observations are uncensored. The likelihood function for θ is given by

$$L(\theta) = \prod_{i=1}^{n} f(t_i,\theta)$$

Thus, L is a function of θ and finding the maximum likelihood estimator corresponds to finding the value $\hat{\theta}$ in which the function L reaches its maximum for the fixed sample. Note that the likelihood function structure does not allow for censored data. In this way, the likelihood function in the presence of censoring is defined in a way that each individual contributes to the likelihood function with specific information (Klein and Moeschberger (2006)).

An individual who, for example, has an exact failure time, contributes to the likelihood function with the probability of the event of interest occurring at this time. This contribution is given by the probability density function of T at this time. The contribution of each individual censored on the right is given by the T complementary cumulative distribution function, evaluated in the last visit time. Similarly, the contribution of a censored individual to left is given by the cumulative distribution

function of T assessed at the time of the first visit. Finally, the contribution of an individual who presents a time of failure in a certain interval is given by the probability that the time of occurrence of the event belongs to this interval. In summary, the general form of the likelihood function in the presence of censoring is

$$L(\theta) = \left[\prod_{i\in\mathscr{L}} F(t_i)\right] \times \left[\prod_{i\in\mathscr{R}} (1-F(t_i))\right] \times \left[\prod_{i\in\mathscr{I}} (F(t_{i2}) - F(t_{i1}))\right] \times \left[\prod_{i\in\mathscr{D}} f_T(t_i)\right]$$

where

- $\mathscr{L}$ - is the set of indices for which the $i-th$ observation was left censored;
- $\mathscr{R}$ - is the set of indices for which the $i-th$ observation was right censored;
- $\mathscr{I}$ - is the set of indices for which the $i-th$ observation was censored at an interval (t_{i1}, t_{i2});
- $\mathscr{D}$ - is the set of indices for which the $i-th$ observation failed.

In the next section, we will present our robust proposal based on the application of the concept of censoring in the normalized residual quantile.

3.6.3 Robust Modelling Using Censoring

Considering the normalized residual quantile of a classic GAMLSS model and considering that outliers may occur, we obtain interval censored observations in order to eliminate the impact of outliers in the estimation of the GAMLSS model.

Our interest is to consider the transformation defined in Equation (3.11) to generate censored observations with an interval criterion. Note that censored observations are those that the process identifies as a possible outlier. Thus, the contribution of each censored observation is given the likelihood function evaluated in the range defined by the tuning probabilities.

Consider the model defined in (3.1). For a given smoothing parameter $\boldsymbol{\lambda}$, the robust estimator for the $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ is defined by maximizing the rewriting penalized log-likelihood defined in (3.6) as

$$l_{p_{cens}}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = l_{cens}(\boldsymbol{\beta}, \boldsymbol{\gamma}) - \frac{1}{2}\sum_{k=1}^{4}\sum_{j=1}^{J_k} \boldsymbol{\gamma}_{jk}^T G_{jk} \boldsymbol{\gamma}_{jk}, \tag{3.17}$$

where

$$l_{cens} = \sum_{i=1}^{n} \log\left\{\left[f(y_i|\boldsymbol{\theta_i})\right]^{1-\delta_i-\xi_i}\left[\int_{-\infty}^{a_i} f(y_i|\boldsymbol{\theta_i})\right]^{1-\delta_i}\left[\int_{b_i}^{\infty} f(y_i|\boldsymbol{\theta_i})\right]^{1-\xi_i}\right\},$$

$$\delta_i = \begin{cases} 1, & r_i \geq \Phi^{-1}(\alpha) \\ ,0, & r_i < \Phi^{-1}(\alpha) \end{cases}, \tag{3.18}$$

$$\xi_i = \begin{cases} 1, & r_i \leq \Phi^{-1}(1-\alpha) \\ 0, & r_i > \Phi^{-1}(1-\alpha) \end{cases}, \tag{3.19}$$

$r_i = \Phi^{-1}[F_Y(y_i; \hat{\theta}_\mathbf{i}]$ is the normalized quantile residual (see Dunn and Smyth (1996a)) and $\hat{\theta}_i = (\hat{\mu}_i, \hat{\sigma}_i, \hat{\nu}_i, \hat{\tau}_i)$, $a_i = F_Y^{-1}(\alpha; \boldsymbol{\theta_i})$ and $b_i = F_Y^{-1}(1-\alpha; \boldsymbol{\theta_i})$ and α is a small probability and it is the same tuning probability defined in Section 3.4.

3.7 SIMULATIONS STUDIES

To investigate the finite sample properties of the proposed method and assess the robustness properties, we carry out three simulation studies. The first study simulates a parametric model without covariates in systematic components of μ and σ. The second study is fully linear parametric gamma model with one covariate, and the third study used a non parametric gamma model. In these three studies, was compared robust fitting GAMLSS proposed by Rigby et al. (2019), called RG; the modification of the Rigby method, called RGW; non robust fitting GAMLSS - G; robust fitting GAMLSS based on censoring, called CENS; and the robust fitting GAMLSS method propose by Aeberhard et al. (2021), called as AH method. The choice of the robustness tuning constant c of Aeberhard method was based on median downweight proportion using 0.95 efficiency propose by Aeberhard et al. (2021). The tuning probabilities α (defined in Section 3.4) regulate the contribution observation based on the normalized quantile residual. The tuning probabilites for *RG*, *CENS* and *RGW* was defined as $\alpha = 0.01$. For the *RG*, all values of normalized quantile residual that do not belong to the range defined in Equation (3.11) are transformed. For *RGW* method the values of observations set weights equal to zero when the values of normalized quantile residual do not belong to the range defined in Equation (3.11). The CENS method use $\alpha = 0.01$ in the functions (3.18) and (3.19) to accommodate possible outliers.

The evaluation of the performance of the estimates was carried out by investigating the Mean Square Error - $MSE(\hat{\theta},\theta) = \frac{1}{K}\sum_{i=1}^{K}(\hat{\theta}_i - \theta)^2$ where $\hat{\theta}_i$ is i-th Monte Carlo estimate K is he number of replications, mean square deviations for each sample defined as $MSD = \frac{1}{n}\sum_{j=1}^{n}(\hat{\theta}_j - \theta_j)^2$ where $j = 1,\ldots,n$ and the average bias across K replications defined as $\frac{1}{K}\sum_{j=1}^{K}(\hat{\theta}_{ij} - \theta_i)$ where $i = 1,\ldots,n$.

Our code was implemented in software R (R Core Team (2020)), based on the `gamlss` packages Stasinopoulos et al. (2017), and the computation of Aeberhard method was performed using the R packages `GJRM` Marra and Radice (2020). All computations are performed with 200 replicates for all scenarios. The simulated data are contaminated by choosing at random of the response and by adding a constant to their original values, following Aeberhard et al. (2021). The details of the contamination will be presented in the next sections.

3.7.1 Simulation under Parametric Gamma Model Without Covariates in Systematic Component

Let $\boldsymbol{y} = (y_1,\ldots,y_n)$ be a vector of variables from gamma distribution with parameters μ and σ. The probabilities density function, denoted by $GA(\mu,\sigma)$, is given in Equation (3.13). The parametric gamma GAMLSS model for μ and σ was simulated without covariates. Hence, the model can be defined with the following systematic components: $\eta_1 = \log(\mu) = \beta_{11}$ and $\eta_2 = \log(\sigma) = \beta_{12}$. The simulations study was based on 200 replicates of the distribution $GA(\mu = 2, \sigma = 0.5)$ with 3 different sizes ($n_1 = 100$, $n_2 = 200$ and $n_3 = 500$) and contaminated it with 4 levels: 0%, 2%, 5% and 10% of the sample. Each sample is contaminated by randomly selecting elements from the sample and adding a fixed value 15, the same methodology which is used by Aeberhard et al. (2021). This contamination process associates the outliers with the long tail of distribution and inflating estimates.

The investigation using the MSE shows that, without contamination (first result in Table 8), all methods have similar performance. When the data are contaminated, in all levels, the MSE of AH estimations are slight better for both parameters. Only the performance of RGW estimations of μ for sample size 200 and 500 and 10% of contamination are better than AH estimations. Note that, when the number of contaminated observations increases, the MSE of RG and Cens estimates also increase. Figures 5, 6 and 7 show the boxplots of estimates of μ and σ by sample size 100, 200 and 500, respectively. The red line indicate the real values of parameters. The results indicates the best performance for AH estimates. The simulations with samples of size 100 (Figure 5) indicate that the RGW estimations of μ has a good performance compared to the AH estimates that have

better performance, in all levels. However, for $\hat{\sigma}$ the results are better for AH estimates. Note that, the σ estimates for the RGW method are underestimated at all levels of contamination, while the σ estimates for the AH method are underestimated only at the 10% levels of contamination. When we consider samples of size 200 and 500 (Figures 6 and 7, respectively), the results has the same underestimation behavior.

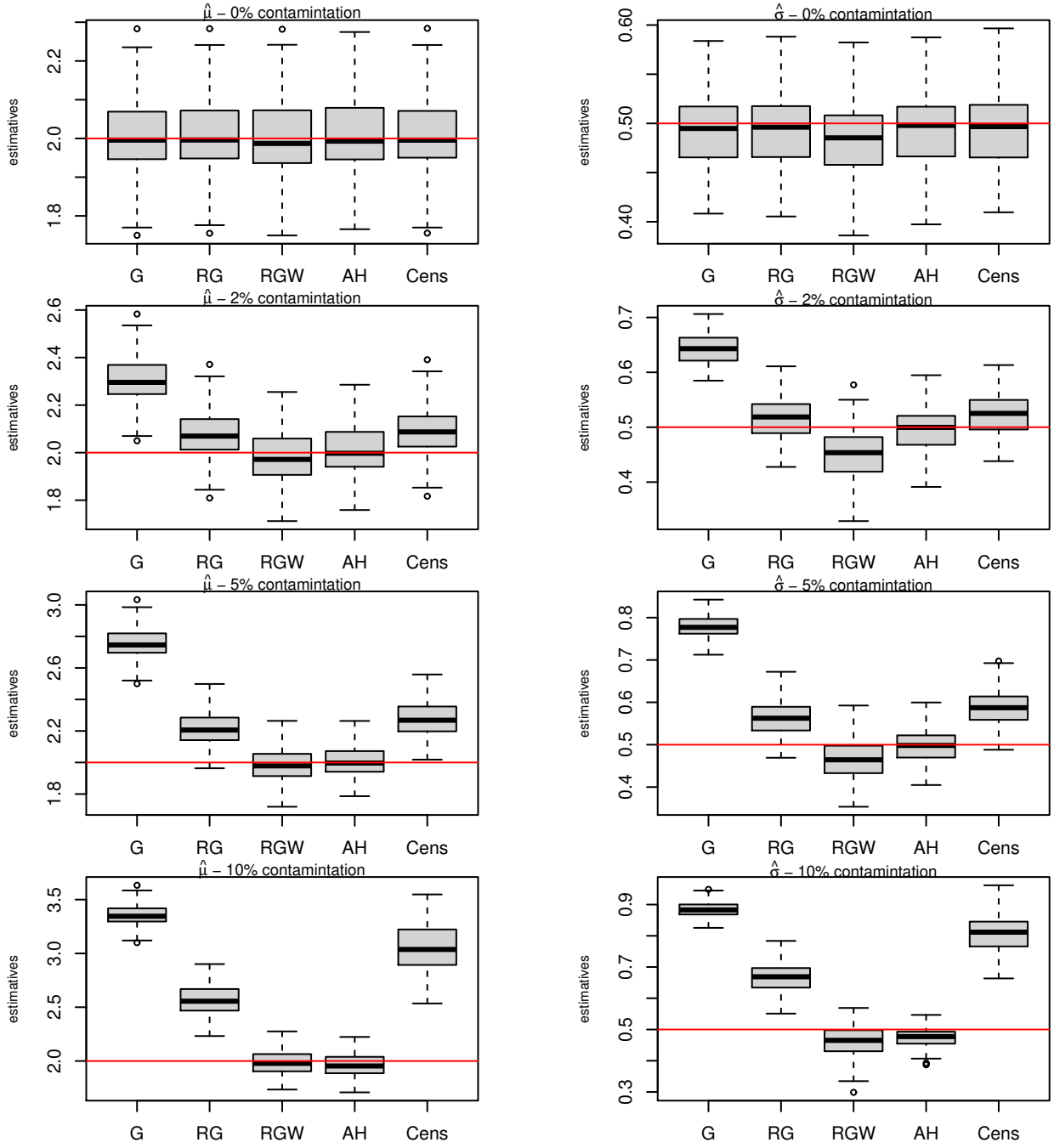
The results show a better performance for the AH method. However, the AH method presents a tuning constant selection criterion that can be considered an advantage compared to the RG, RGW and Cens methods in which the tuning constants were defined in a subjective way.

Table 8 – Mean squared error for $\hat{\mu}$ and $\hat{\sigma}$ with 0%, 2%, 5% and 10% contamination with constant systematic component based on the gamma distribution.

0% contamination													
$\hat{\mu}$	n	G	RG	RGW	AH	Cens	$\hat{\sigma}$	n	G	RG	RGW	AH	Cens
	100	0.001	0.001	0.01	0.01	0.01		100	0.001	0.001	0.002	0.001	0.001
	200	0.005	0.005	0.005	0.005	0.005		200	0.001	0.001	0.001	0.001	0.001
	500	0.002	0.002	0.002	0.002	0.002		500	0.000	0.000	0.000	0.000	0.000
2% contamination													
$\hat{\mu}$	n	G	RG	RGW	AH	Cens	$\hat{\sigma}$	n	G	RG	RGW	AH	Cens
	100	0.105	0.017	0.013	0.010	0.019		100	0.021	0.002	0.004	0.001	0.002
	200	0.096	0.010	0.007	0.005	0.015		200	0.023	0.001	0.004	0.001	0.002
	500	0.090	0.006	0.004	0.002	0.015		500	0.023	0.001	0.002	0.000	0.002
5% contamination													
$\hat{\mu}$	n	G	RG	RGW	AH	Cens	$\hat{\sigma}$	n	G	RG	RGW	AH	Cens
	100	0.585	0.060	0.012	0.010	0.091		100	0.078	0.005	0.003	0.001	0.009
	200	0.571	0.054	0.007	0.006	0.127		200	0.080	0.006	0.002	0.001	0.017
	500	0.561	0.047	0.003	0.002	0.213		500	0.080	0.005	0.001	0.000	0.031
10% contamination													
$\hat{\mu}$	n	G	RG	RGW	AH	Cens	$\hat{\sigma}$	n	G	RG	RGW	AH	Cens
	100	1.86	0.344	0.013	0.013	1.160		100	0.148	0.030	0.004	0.002	0.099
	200	2.043	0.446	0.008	0.009	2.600		200	0.157	0.040	0.002	0.001	0.188
	500	2.156	0.503	0.003	0.006	2.181		500	0.163	0.047	0.002	0.001	0.164

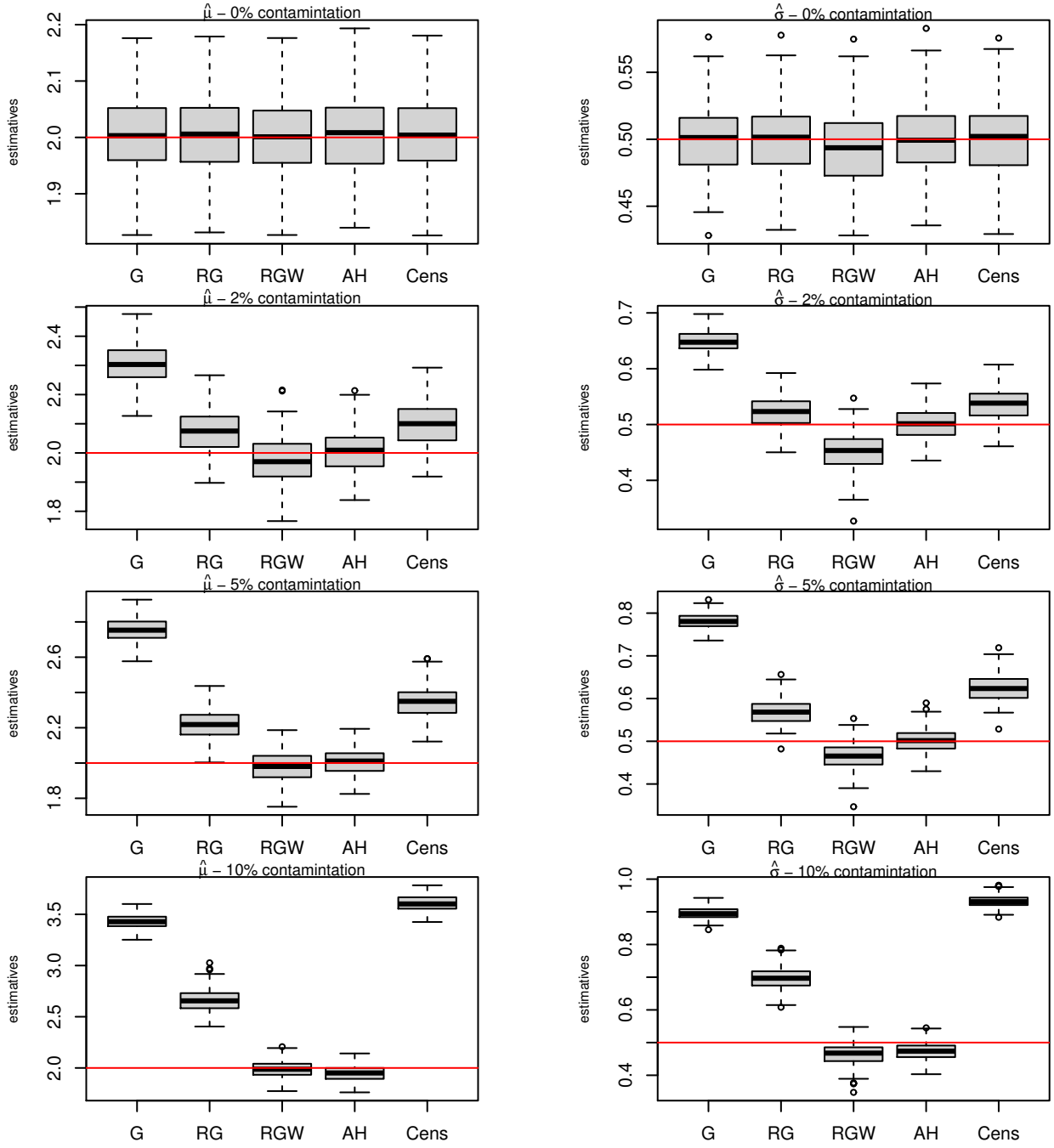
Source: The author (2021)

Figure 5 – Boxplots of μ and σ estimates, for the parametric model without covariates, sample size 100 and 2%, 5% and 10% levels contamination, based on the gamma distribution.



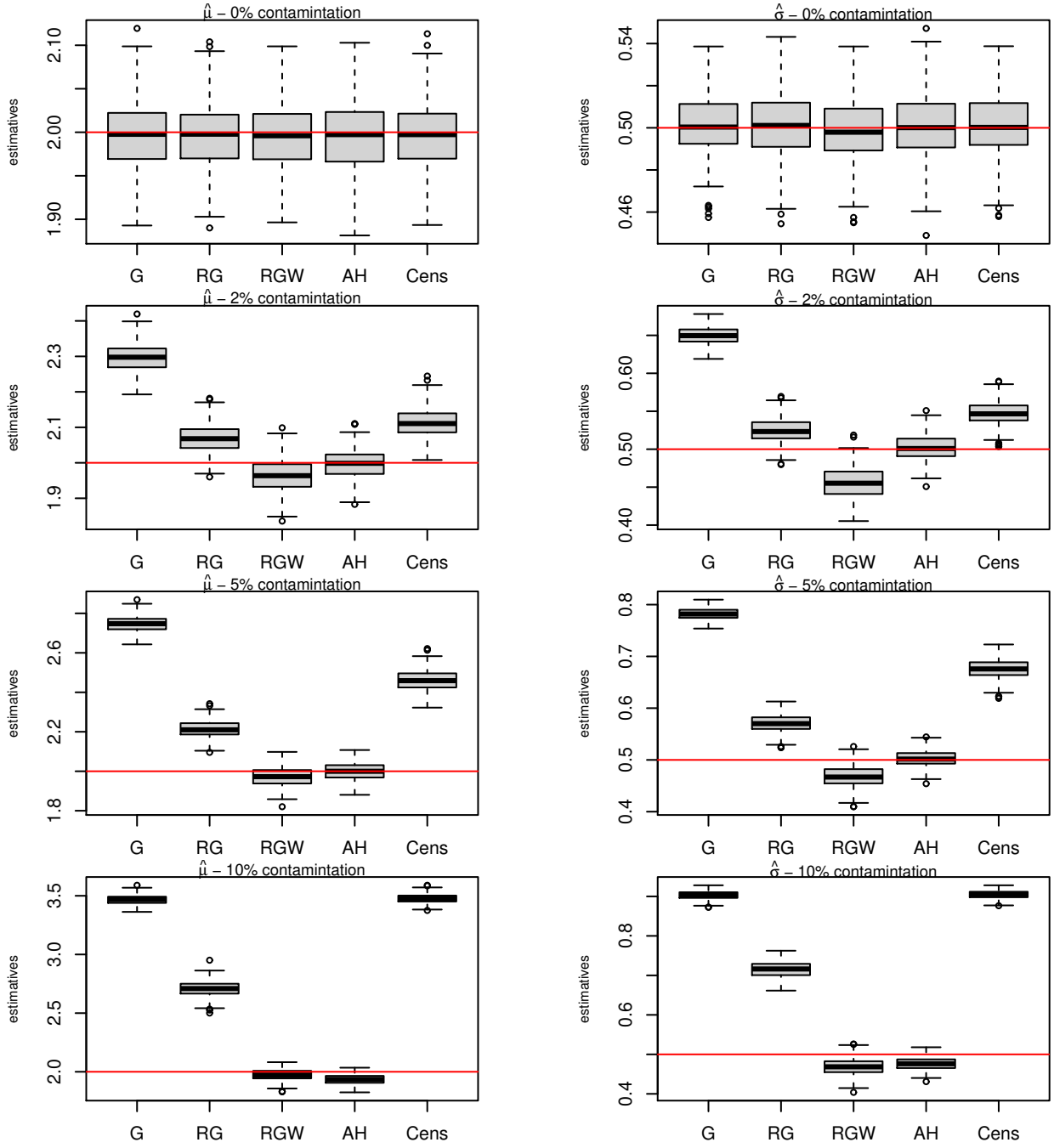
Source: The author (2021)

Figure 6 – Boxplots of μ and σ estimates, for the parametric model without covariates, sample size 200 and 2%, 5% and 10% levels contamination, based on the gamma distribution.



Source: The author (2021)

Figure 7 – Boxplots of μ and σ estimates, for the parametric model without covariates, sample size 500, 2%, 5% and 10% levels contamination, based on the gamma distribution.

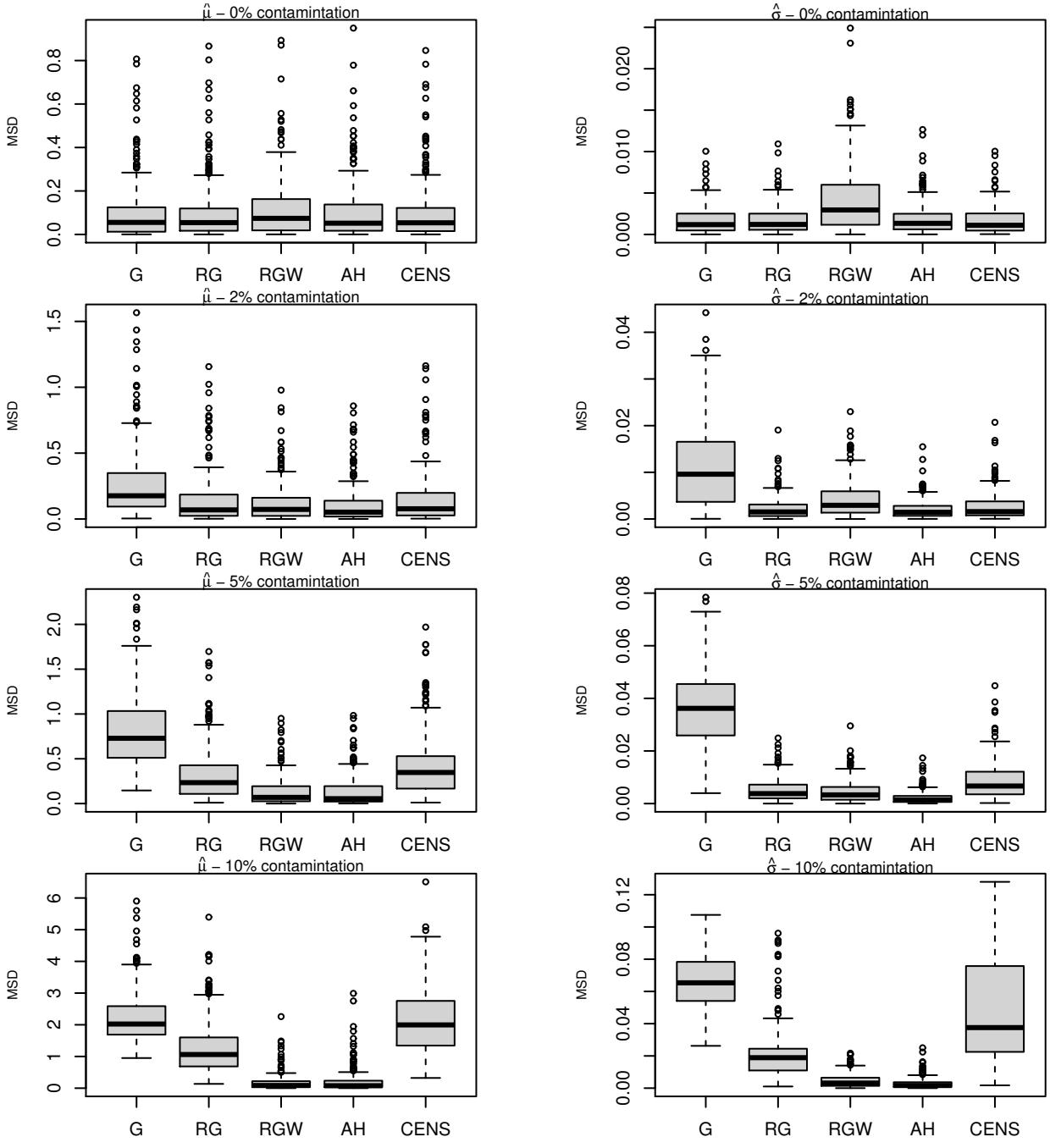


Source: The author (2021)

3.7.2 Simulation Under Parametric Gamma Model with Covariates in Systematic Component

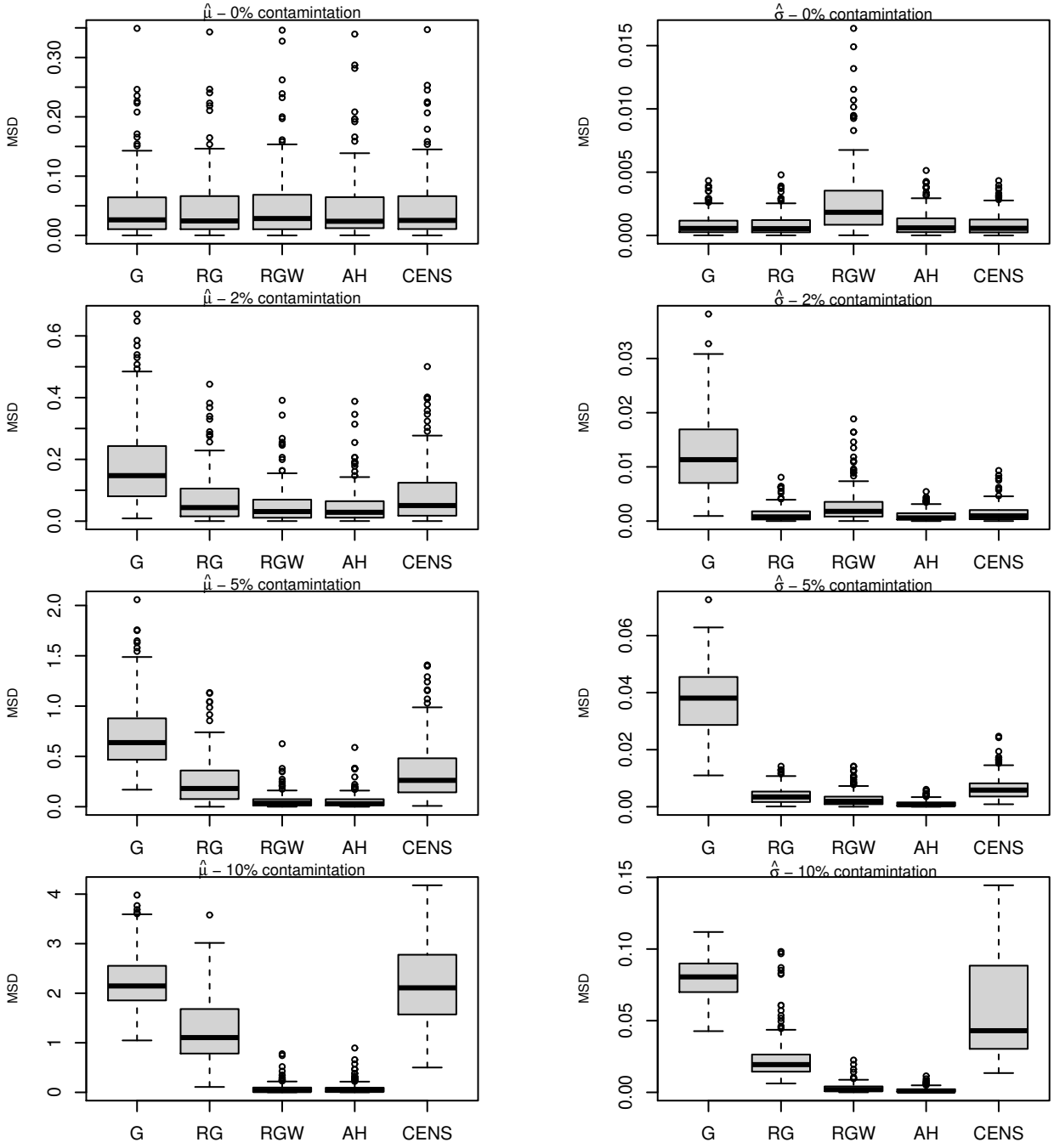
The parametric gamma GAMLSS model for μ and σ was simulated with one covariates. Hence, the model can be defined with the following systematic components: $\eta_{i1} = \log(\mu_i) = \beta_{11} + X_i\beta_{21}$ and $\eta_{i2} = \log(\sigma_i) = \beta_{12} + X_i\beta_{22}$, where $\beta_{11} = \beta_{21} = 1$, $\beta_{12} = -1.5$, $\beta_{22} = 1$ and the covariate X_i were fixed in all replicates and defined as a uniform distribution $Uniform(0,1)$. Here, was simulated 200 replicates of a random samples with 3 different samples sizes ($n_1 = 100$, $n_2 = 200$ and $n_3 = 500$) and contaminate it with 4 levels: 0%, 2%, 5% and 10% of the sample. Each sample is contaminated by randomly selecting elements from the sample and adding a fixed value 15, the same methodology used in Aeberhard et al. (2021).

Figure 8 – Boxplots of the Mean Squared Deviation of $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ estimates, for the parametric model with covariates, simple size 100, all levels contamination and based on the gamma distribution.



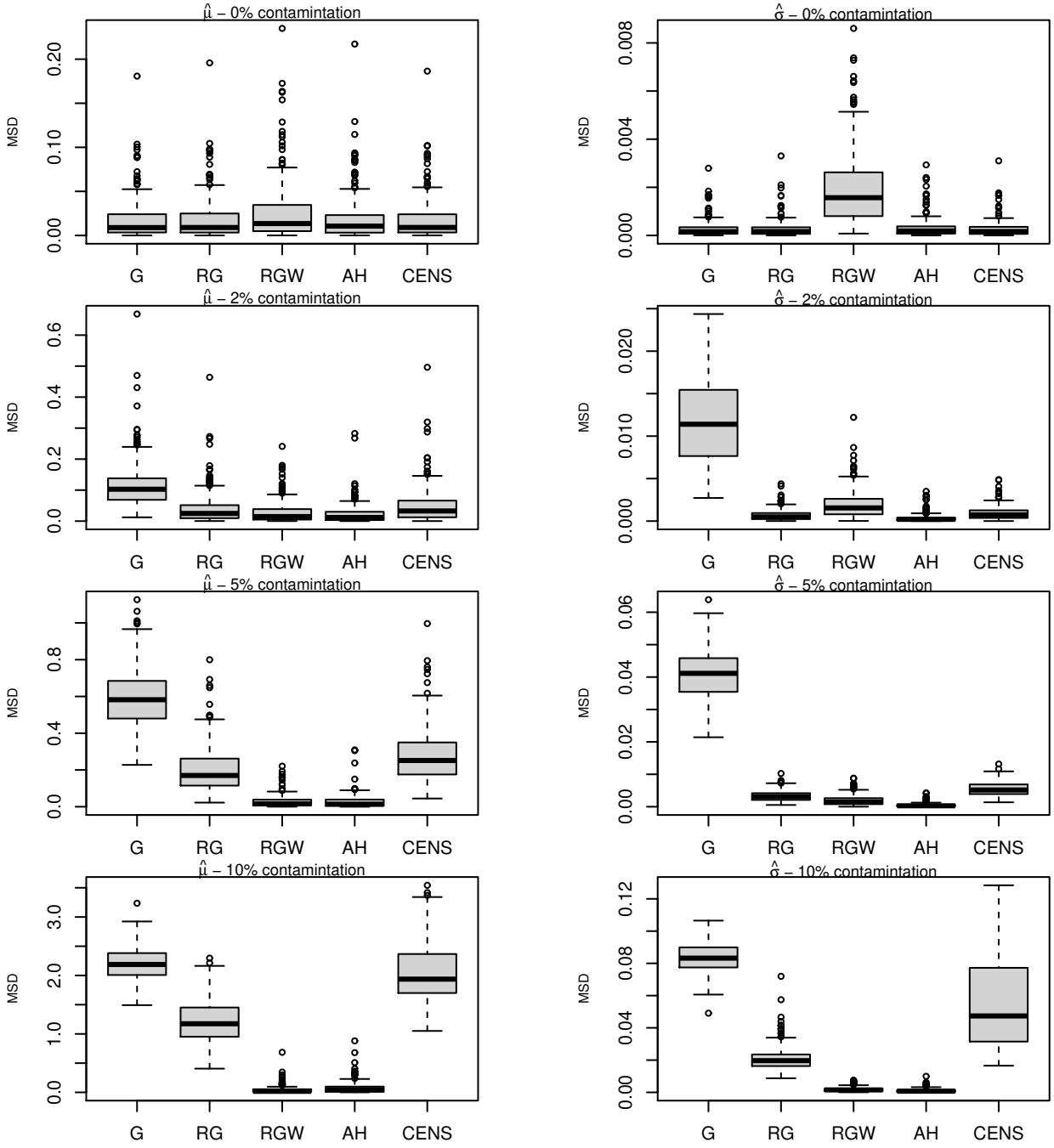
Source: The author (2021)

Figure 9 – Boxplots of the Mean Squared Deviation of $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ estimates, for the parametric model with covariates, simple size 200, all levels contamination and based on the gamma distribution.



Source: The author (2021)

Figure 10 – Boxplots of the Mean Squared Deviation of $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ estimates, for the parametric model with covariates, simple size 500 and all levels contamination based on the gamma distribution.



Source: The author (2021)

Figures 8, 9 and 10 below presents Boxplots of the Mean Squared Deviation - MSD of $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ estimates of all combination scenarios, for samples sizes 100, 200 and 500, based on 200 replicates. In each figure, the lines represents the contamination levels and the columns the parameters of models (first column $\boldsymbol{\mu}$ and second column $\boldsymbol{\sigma}$). For sample sizes 100 (Figure 8), without contamination (first line) the methods show similar results for both parameters, with values a little higher for the RGW $\boldsymbol{\sigma}$ estimates. At 2% contamination the robust methods are competitive with a better performance level for the AH estimates. At 5% and 10% contamination the results indicate a better performance for the $\boldsymbol{\mu}$ of the AH and RGW estimates, while for the estimates of $\boldsymbol{\sigma}$ the results of the AH estimates are better. Note that, when the percentage of contamination increases, the variability of the mean squared error of the Cens estimates also increases. The results for samples of sizes 200 and 500 (Figures 9 and 10, respectively) indicate the same behavior registered in samples of size 100, that is, at 2% of contamination the results of the robust methods are competitive with a slight advantage for the AH estimates and at 5% and 10% contamination the RGW and AH estimations have the smallest values of MSD with a slight advantage for AH estimates. The Cens estimates present a competitive result, however, when the level of contamination is 10% the MSD increases. The selection of tuning constants can influence the regulation and control of possible outliers. In addition, the Cens method bound the observations by modifying the structure of the likelihood function and not the observations or weight of observations, like the other methods.

3.7.3 Simulation Under Non Parametric Gamma Model

The non parametric gamma estimations was evaluated based on the brain image data, introduced in Section 3.8. The brain data were used to simulate and control the parameters following Aeberhard et al. (2021). In summary, the brain image data have a non negative response variable representing the physiological activation level of voxels in a brain location, represented by x_1 and x_2, defining the location of each voxel and used as covariates. The set of voxel and response variables forms a sample of size 1567. The covariate data were used to generate a response for each voxel according to a GAMLSS with a gamma distribution with expectation $\boldsymbol{\mu}$ and variance $\sigma^2\mu^2$ where the systematic components is non parametric defined as $\log(\mu) = \eta_1 = s_1(x_1, x_2)$ and $\log(\sigma) = \eta_2 = s_2(x_1, x_2)$. The smooth function s_1 and s_2 has not been defined in Aeberhard et al. (2021), hence, these functions were constructed to represent the main features of the fitted surfaces on the real data. The smooth functions are available as:

$$s_1(x_1, x_2) = 0.9e^{-0.05(x_1-65)^{2/11}(x_2-32)^{2/12}}\cos(5\pi x_1 x_2)$$

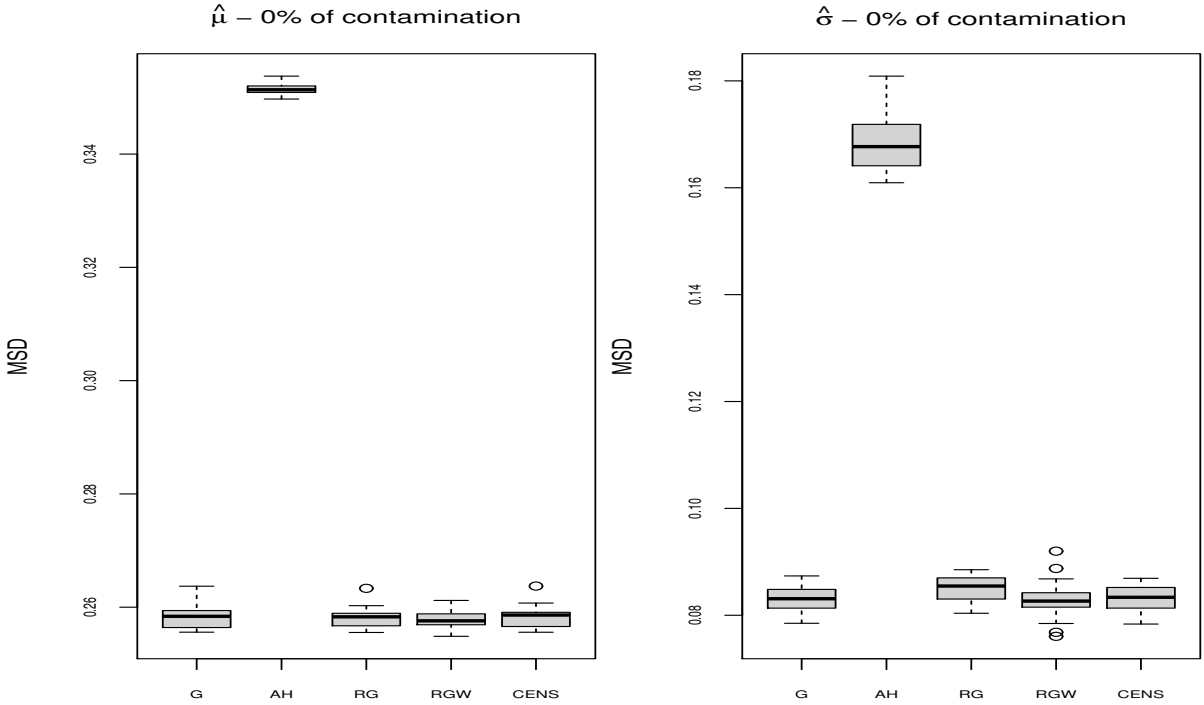
$$s_2(x_1, x_2) = -1.8e^{-0.05((x_1-65)^{2/11}(x_2-32)^{2/12})}$$

The contaminated data are generated in a similar way to Aeberhard et al. (2021), were modify a simulated data set by choosing at random 78 (5%) of the responses falling in the upper-right corner of the brain slice, for $x_1 > 70$ and $x_2 > 30$, and by adding 25 to their original value. The simulations were based on the contamination levels of 0% and 5% with 200 replications, fitting a gamma GAMLSS with log links for both parameters $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$. The bivariate Thin Plate Regression Splines with $k = 20$ bases was used to approximate the s_1 and s_2 smooth functions. In addition the Maximum likelihood based methods (REML) was used for estimating the smoothing hyper-parameters (see Rigby and Stasinopoulos (2005); Stasinopoulos et al. (2017)).

The MSD of $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\sigma}}$ for simulation without contamination, that is, 0% of contamination, shows that the AH estimation has a poor performance with high MSD values. Figure 12 shows the boxplots of MSD of $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\sigma}}$ for the contamination level of 5% and the five methods (G, RG, RGW, AH and Cens). The MSD of AH and RGW estimates shows the best results for the both parameters, with better performance for $\boldsymbol{\mu}$ estimates of AH and better performance for $\boldsymbol{\sigma}$ estimates of RGW.

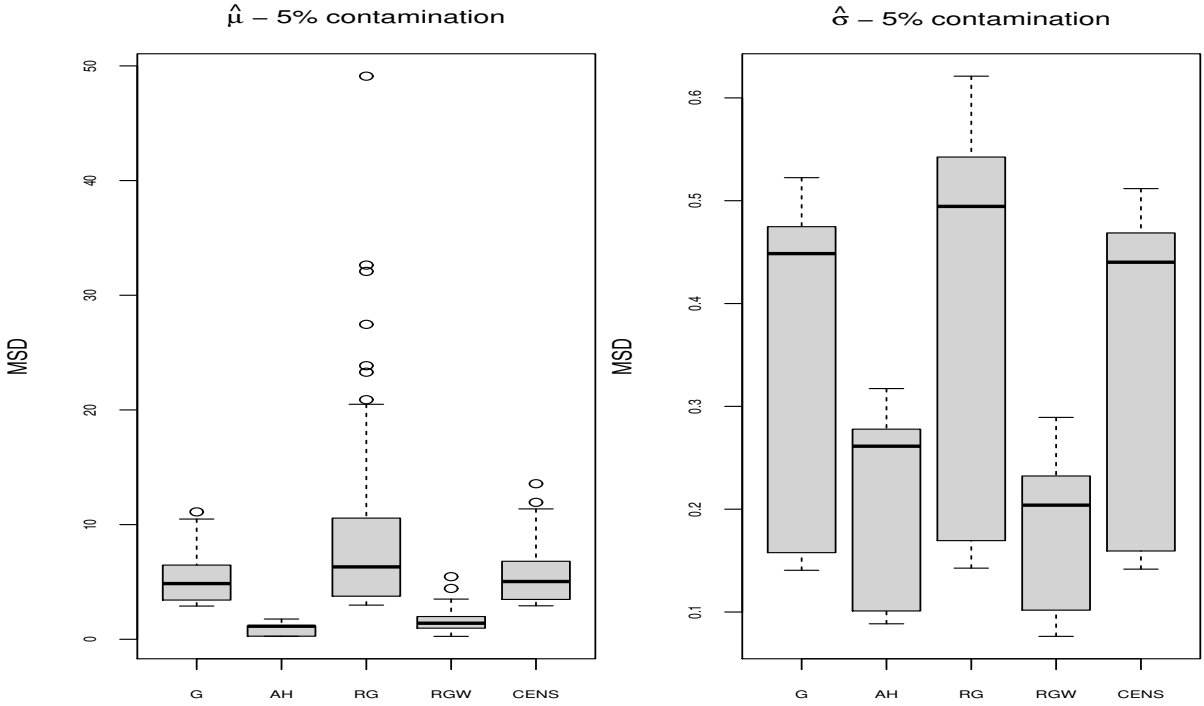
Figure 13 shows colored surfaces representing the average bias of $\hat{\mu}$ and $\hat{\sigma}$, across replications, that is, $\frac{1}{200}\sum_{j=1}^{200}(\hat{\theta}_{ij} - \theta_i)$ where $i = 1, \ldots, n$, θ_i is μ_i or σ_i, and 200 is the number of replicates. It is desirable that the values of $\hat{\mu}$ and $\hat{\sigma}$ bias be near to 0. Observing the results, the scale of color of surface bias near of 0 is purple for $\hat{\mu}$ and blue $\hat{\sigma}$. Note that the coloring patterns are not the same between the figures for both parameters. The bias for $\hat{\mu}$ (first column of Figure 13) of G, RG and Cens shows a large positive bias in the top-right corner of the brain slice, which is precisely the area that is contaminated ($x_1 > 70$ and $x_2 > 30$). The bias for $\hat{\mu}$ of AH estimates shows negative values $(-1.67, -1)$ in center of the brain, while the bias of the RGW estimates concentrates higher values in the area of the right edges of the brain image. The surface for the average bias of $\hat{\sigma}$ (second column of Figure 13) shows large positive bias in the center of the brain in all methods. The AH estimates indicate more high values, between 1.09 and 1.4 in the center of the brain image, while RGW method indicate more high values, between 0.47 and 0.78 in the center of the brain image. Thus, in this scenario, the AH and RGW methods are competitive.

Figure 11 – Boxplots of the Mean Squared Deviations of $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ estimates, simulated under non parametric gamma model with 0% of contamination.



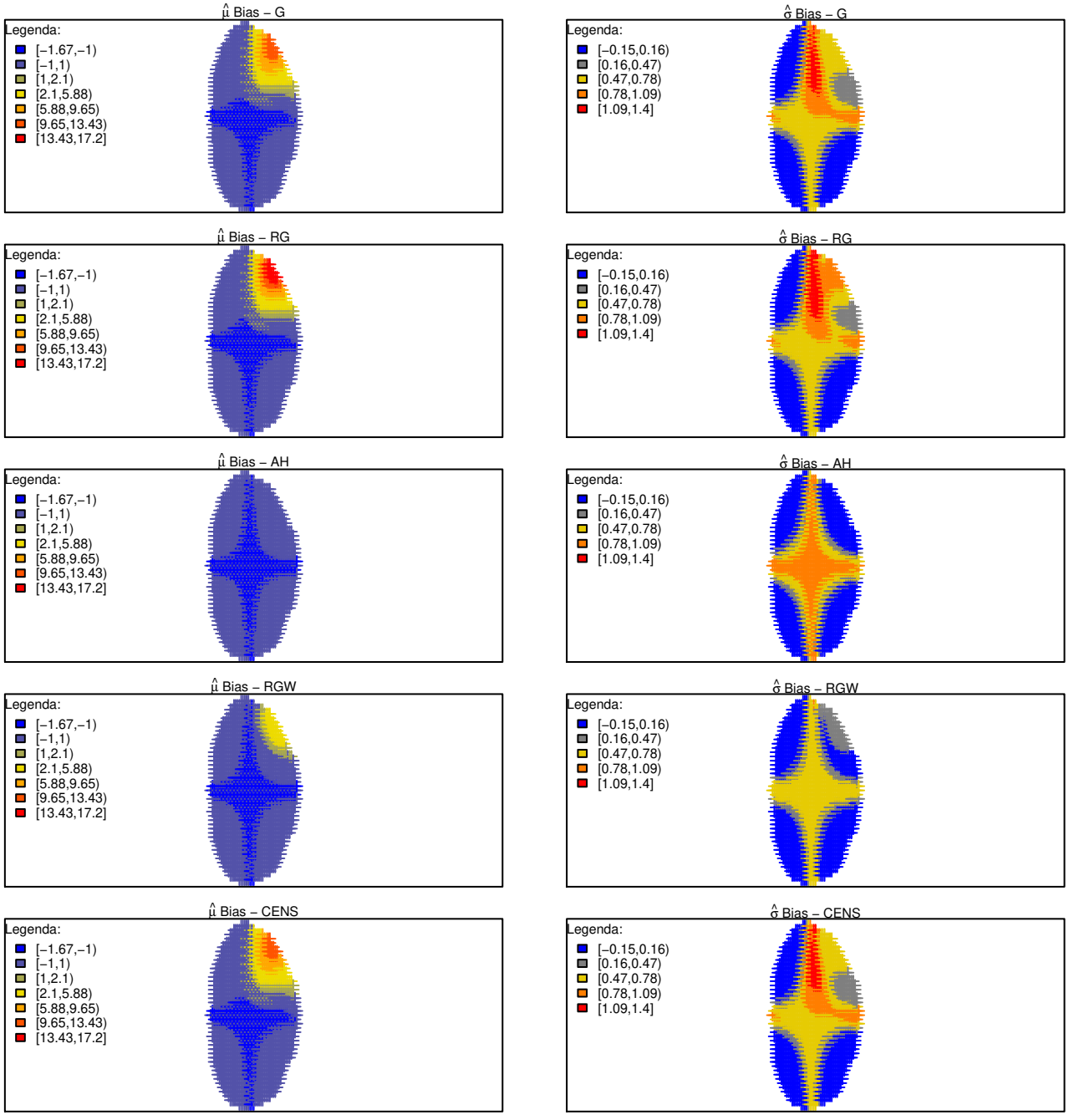
Source: The author (2021)

Figure 12 – Boxplots of the Mean Squared Deviations of $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ estimates, simulated under non parametric gamma model with 5% of contamination.



Source: The author (2021)

Figure 13 – Surfaces of the average biases for the $\hat{\mu}$ and $\hat{\sigma}$ simulated under non parametric gamma model with 5% of contamination.



Source: The author (2021)

3.8 APPLICATIONS

In this section, the performance of the proposal will be illustrated with real data of functional magnetic resonance imaging measurements - FMRI of the human brain. The FMRI, for a human brain subject to a particular experimental stimulus, was presented in Landau et al. (2004) and subsequently used in Wood (2017), are available in the R package `gamair` available on CRAN. The aim of the Wood (2017) study is to test a difference in the timing between two anatomically distinct brain regions using a measure, summarized as the median of three measurements of fundamental power quotient on each brain voxel. The variable of interest is the median of three replicate Fundamental Power Quotient values at each voxel, called `medFPQ`. This is the main measurement of brain activity, which represents the physiological response of the brain to controlled stimuli. The covariates used in model is the coordinates x_1 and x_2 identifying the location of each voxel. Following Wood (2017) and Aeberhard et al. (2021), it was modeled both the mean and variance of `medFPQ` as joint functions of the $s_1(x_1,x_2)$ and $s_2(x_1,x_2)$ to be approximated by Thin Plate Regression Spline basis functions with a smoothness penalty. Wood (2017) (p. 329) identified two outliers voxel responses ($\texttt{medFPQ} < 0.0005$), that were discarded for the subsequent analysis. Here, this two outliers voxel response will be considered.

The response variable medFPQ is a continuous variable in positive real numbers. The minimum value of the observations is 0.000003 and the maximum is 20.82. In Figure 14, we have a `medFPQ` boxplot, with 92 extreme values, equivalent to 5.86% of the observations . We can observe that the distribution is asymmetric, with a positive asymmetry coefficient equal to 5.86. In addition, the kurtosis coefficient was equal to 56.3, indicating that the distribution of these data is leptokurtic.

The estimated surfaces of μ and σ, for the five methods are given in Figure 15. The surface for σ estimates (second column in Figure 15) indicates a larger localized response variance in G estimates, localized in the left corner of the brain. This is caused by two observations in this area which are the ones in identified and excluded from the analysis in Wood (2017). The large values of $\hat{\sigma}$ may characterize the existence of outliers or the occurrence of high values of σ in location. Only the AH and RGW estimates were able to reduce the high values of σ at the top of the brain image. The largest estimated surfaces for μ of G, located in the upper-right corner of the brain, is much low when considering the RGW and AH methods, that is, the large values in μ mean brain activity implied by G estimates have been smoothed when considering robust estimation. These low values do not imply that these observations are outliers, but that they, may be, do not seem to follow the same pattern as the data given the gamma GAMLSS assumed. This behavior can be explained when we investigate the outliers identified by each method. The RG identified 35 outliers, RGW 92, AH could have 35 observations and the Cens method identified 40 outliers.

A residual analysis was carried out using the normalised quantile residuals. The main advantage of the normalised quantile residuals is that, whatever the distribution of the response variable, their true values always have a standard normal distribution given the assumption that the model is correct Dunn and Smyth (1996b). Figures 17 and 16 shows the QQ plot and worm plot of the normalised quantile residuals of the adjusted models, respectively, indicating a lack of fit caused by some change in kurtosis measures in all methods. The worm plots of G, RG, AH and Cens indicate a leptokurtic distribution with positive excess kurtosis, that is, the distribution has fatter tails. The worm plot of residuals of RGW method indicates a distribution with negative excess kurtosis.

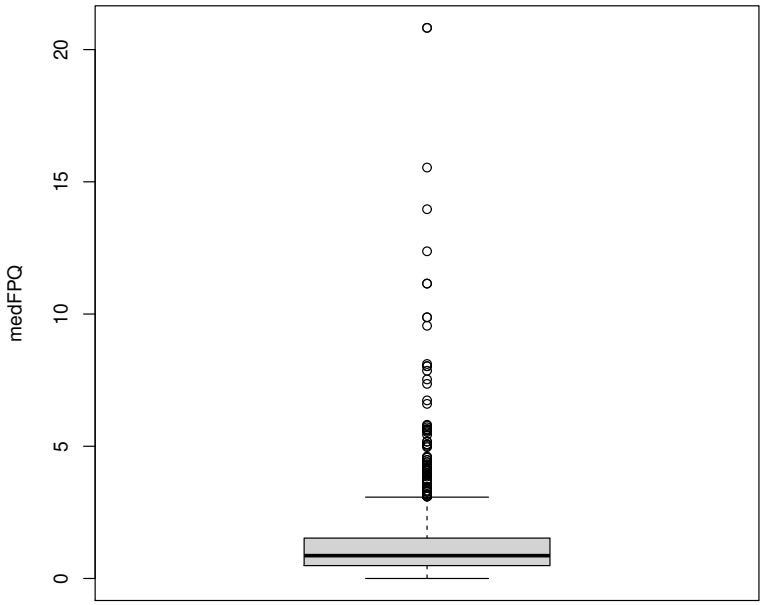
3.9 CONCLUSION

In this chapter several contributions have been introduced for robust fitting GAMLSS models. The first contribution was to evaluate the quality of the estimates produced by the robust method proposed by Rigby et al. (2019) and compare it with the Aeberhard et al. (2021) proposal. For this, based on the idea of Rigby et al. (2019), we introduce the robust fitting GAMLSS gamma model with bias correction and we carried out an extensive simulation study with several scenarios. Until now, only the correction of bias for the estimators of the beta model were presented in Rigby et al. (2019). In addition, computational codes produced are quite general since it can be employed for any likelihood. We introduced two robust estimations methods for the broad class of GAMLSS. The first modifies Rigby et al. (2019) proposal by changing the weight of the observations considered outliers (RGW); the second is based on the ideas of censored variables (cens).

Simulations based on the gamma distribution showed the RGW and Aebehard method - AH are competitive methods with the best results. When it is considered the non parametric model, the results are slightly better for the RGW method, while for parametric model without and with covariate in systematic component the results are slightly better for the AH method. Our proposed robust estimator based on censoring has of some limitations. Like any robust estimator, the proportion of contaminated data cannot be unreasonably large without the estimator starting to break at some point. The AH method presents a tuning constant selection criterion that can be considered an advantage compared to the RG, RGW and Cens methods in which the tuning constants were defined in a subjective way, which can affect the performance of the methods. In addition, the cens method bound the observations by modifying the structure of the likelihood function and not the observations or weight of observations, like the other methods. We believe that an adaptive censoring method can shows better results and it is worth to be investigate in future research.

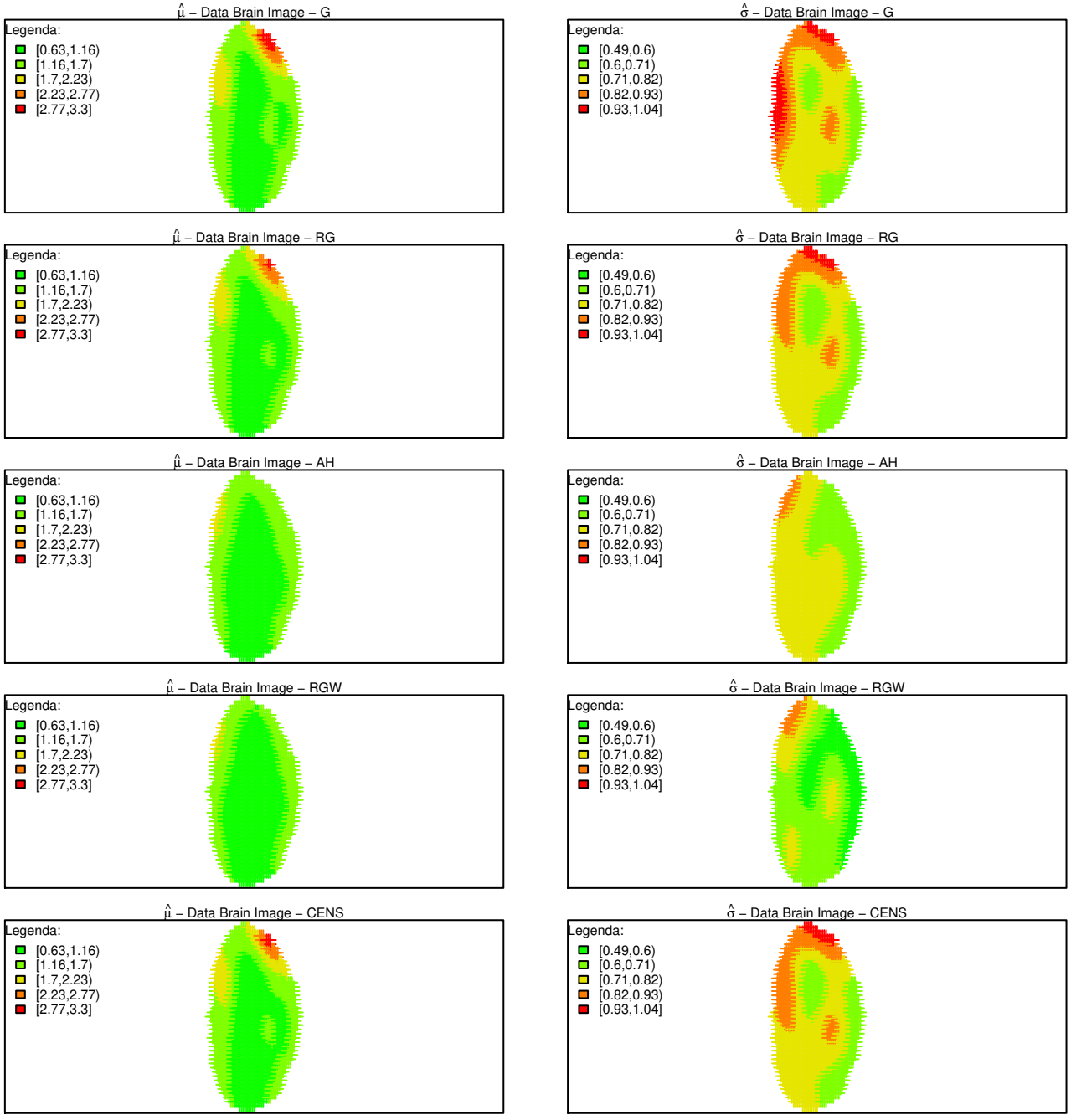
The application to the brain imaging data showed also show that our robust estimator allows the automatic detection of deviating observations. In addition, the gamma distribution was not suitable for modeling brain imaging data. Our proposal is based on a new idea of thinking about robust models. The study of this alternative idea is not complete and future studies on theoretical properties will be needed, such as the sampling distribution necessary for inference; the correction for Fisher consistency cannot be directly extended beyond continuous families of distribution due to the reliance on quantile residuals; and the challenges of selection of tuning probabilities and smoothing parameter selection are not discussed. In the next chapter, a new proposal for a robust fitting GAMLSS and a robust tuning constant selection method will be presented.

Figure 14 – Boxplot of median of Fundamental Power Quotient - `medFPQ` based on the gamma distribution.



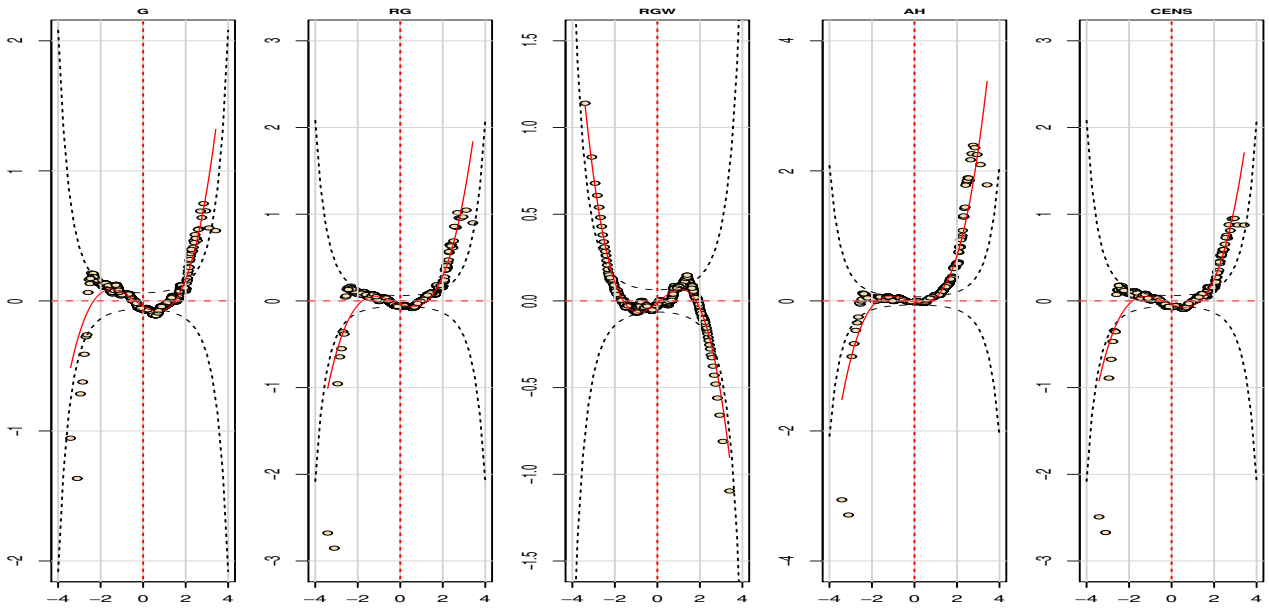
Source: The author (2021)

Figure 15 – Surfaces for the μ σ estimates of median of Fundamental Power Quotient - `medFPQ`.



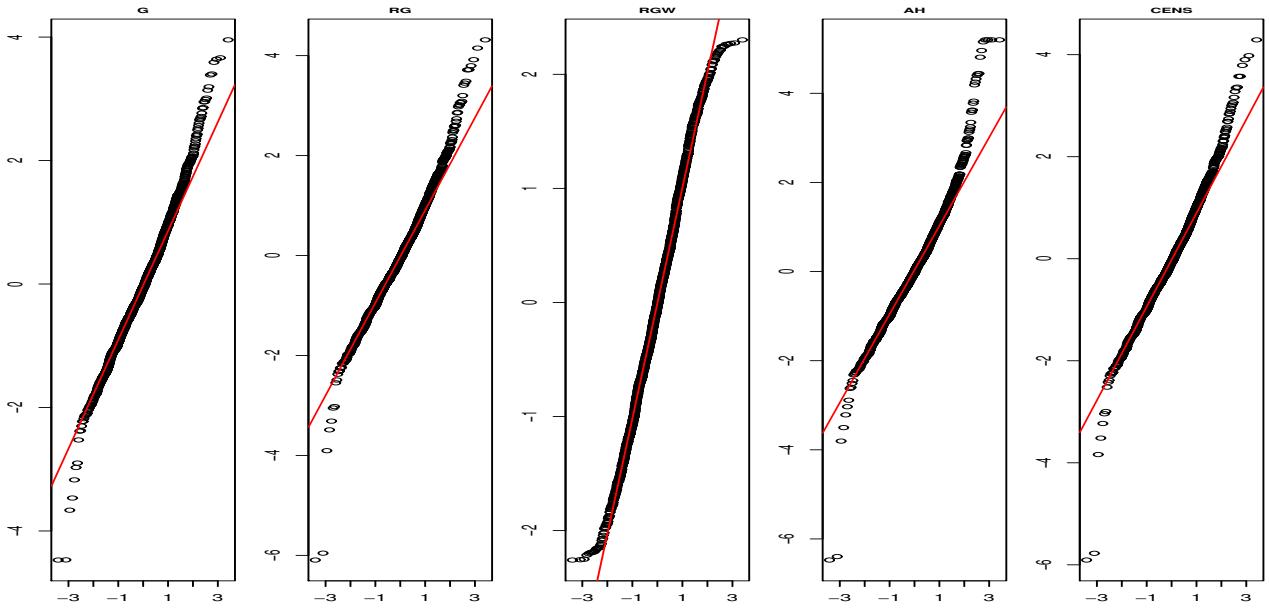
Source: The author (2021)

Figure 16 – Worm plot for normalised quantile residuals of median of Fundamental Power Quotient - `medFPQ`, based on the gamma distribution.



Source: The author (2021)

Figure 17 – QQ plot for normalised quantile residuals of median of Fundamental Power Quotient - `medFPQ`, based on the gamma distribution.



Source: The author (2021)

4 A NEW APPROACH TO ROBUSTNESS USING ADAPTIVE TRUNCATION FOR GAMLSS

4.1 INTRODUCTION

Generalized additive models for location, scale and shape (GAMLSS) are a flexible and general class of distributional regression models that have been introduced by Rigby and Stasinopoulos (2005). See also Stasinopoulos et al. (2018, 2017) for a review. GAMLSS models are very popular and widely used in the literature, due to their flexibility. These models allow the modeling of all parameters of any probability distribution, that is, they allow the use of exploratory variables to model the parameters of location, scale and shape. It's worth pointing out that, these models cover special cases generalized additive models - GAM (Hastie and Tibshirani (1990b)) and generalized linear models - GLM (Nelder and Wedderburn (1972b)).

In GAMLSS models, parametric terms (linear and non-linear) and additives are used to model p parameters, $\boldsymbol{\theta}^\top = (\theta_1, \ldots, \theta_p)$ of density probability function $f(y|\boldsymbol{\theta})$, where $\boldsymbol{y}^\top = (y_1, \ldots, y_n)$ is the vector of the response variables. The fit is carried out by maximum likelihood - ML estimation. The maximum likelihood estimators are sensitive to the presence of extreme values (outliers), meaning that the estimated can be distorted by the presence of outliers.

To solve this issues in statistical modelling and data analysis, robust methods began to emerge in the 1960s [Heritier et al. (2009)], with the aim of minimize the impact of outliers and derive methods that produce reliable parameter estimates, with associated confidence intervals and tests. Robustness is discussed in detail in Huber (1981), Hampel (1968) and Maronna et al. (2006).

Edgeworth (1887) was one of the first to propose alternatives to least square, as a robust estimator. Cantoni and Ronchetti (2001) presented robust inference for generalized linear models (GLMs) based on the notion of quasi-likelihood. Mills and Dupuis (2002) proposed a robust estimation procedure for generalized linear mixed models. Cantoni (2004) presented a robust approach to longitudinal data analysis. Alimadad and Salibian-Barrera (2011) discussed an outlier-robust fit for generalized additive models (GAMs) based on the backfitting algorithm. Croux et al. (2012b) obtained functional estimates for the mean and the dispersion in extended GAMs that are both robust and smooth.

Cadigan and Chen (2001) presents properties of robust M-estimators for poisson and negative binomial data. Cantoni and Zedini (2009) showed a robust version of the hurdle model. Aeberhard et al. (2014) presents robust inference in the negative binomial regression model. Agostinelli et al. (2014) presents robust estimators of the generalized log-gamma distribution. Valdora and Yohai (2014) proposed a family of robust estimators for generalized linear models, using an M-estimator after applying a variance stabilizing transformation to the response. These works, however, cannot be extended to the more general setting of GAMLSS and only two works proposed a robust estimation based on GAMLSS model. The first, Aeberhard et al. (2021), introduces robustness by modifying the objective function following an idea introduced by Eguchi and Kano (2001). The second work is an alternative robust estimation method for GAMLSSs is presented in book of Rigby et al. (2019). This method achieves robustness by transformation of the observed response through (normalized) quantile residuals.

This work is motivated by the small number of proposals that address robust fitting for GAMLSS

models and the existence of several real problems to which robust fitting for GAMLSS models can be applied. We can cite: Rule et al. (2004) (estimation of the glomerular filtration rate and serum creatinine), Harrell Jr (2015) (diabetes data), Conen et al. (2004) (prevalence of hyperuricemia) and Beyerlein et al. (2008) (childhood obesity). Therefore, we propose a robust fitting of a GAMLSS, where throughout the work we understand the term robust as that method is not sensitive to any arbitrary contamination in the response distribution (Hampel (1968); Huber (2004)). Our method is based on a simple adaptive truncation approach where potential outlier contaminated observations are checked and if necessary removed by truncating the distribution of the response variable. We shall show that this conceptual simple approach outperforms the standard approaches based on the influential function. The works is organized as follows. Section 4.2 introduces the GAMLSS model. Section 4.3 discusses our robust method. Section 4.4 briefly review outlier contamination and Section 4.5 reports the results of a simulation study. Two examples are discussed in Section 4.6 while concluding remarks are found in Section 4.7.

4.2 GAMLSS

Let y_i, for $i = 1,\ldots,n$, be the response variable observations independent, with probability (density) function $f(y_i|\boldsymbol{\theta}^i)$, where $\boldsymbol{\theta}^{i\top} = (\theta_1^i,\ldots,\theta_p^i)$ is a vector of p parameters that is related to the effects of explanatory variables and random effects through monotonic link function $g_k(\boldsymbol{\theta_k})$, for $k = 1,\ldots,p$. This function (g_k) is defined as additive model given by:

$$g_k(\boldsymbol{\theta}_k) = \eta_k = \mathbf{X}_k\boldsymbol{\beta}_k + \mathbf{Z}_{k1}\boldsymbol{\gamma}_{k1} + \ldots + \mathbf{Z}_{kJ_k}\boldsymbol{\gamma}_{kJ_k}, \tag{4.1}$$

where, $\boldsymbol{\theta}_k^\top = (\theta_{1k},\ldots,\theta_{nk})$, $\boldsymbol{\eta}_k^\top = (\eta_{1k},\ldots,\eta_{nk})$ are vectors of length n, $\boldsymbol{\beta}_k^\top = (\beta_{1k},\ldots,\beta_{b_kk})$ is a parameter vector of length b_k, $\boldsymbol{X}_k$ is the matrix of values of the explanatory variable, also called design matrix of order $n \times b_k$, $\mathbf{Z}_{jk}$ is the base design matrix $(n \times q_{jk})$ depending on the values of $\boldsymbol{X}_k$ and $\boldsymbol{\gamma}_{jk}$ is a q_{jk}-dimensional random vector. Model (4.1) is called the GAMLSS. The distribution of the dependent variable is not limited to the exponential family and all parameters are modelled in terms of both fixed and random effects.

The estimation of the parameters is based on the model formulated as a random-effects GAMLSS. Assume in model (4.1) that the $\boldsymbol{\gamma}_{jk}$ have independent normal distributions with $\boldsymbol{\gamma}_{jk} \sim N_{q_{jk}}(\mathbf{0}, \mathbf{G}_{jk}^-)$, where $\mathbf{G}_{jk}^-$ is the generalized inverse of a q_{jk} x q_{jk} symmetric matrix $\mathbf{G}_{jk} = \mathbf{G}_{jk}(\boldsymbol{\lambda}_{jk})$, which may depend on a vector of hyperparameters $\boldsymbol{\lambda}_{jk}$. Hence, the parameter vector $\boldsymbol{\beta}$ and the parameters of random effects $\boldsymbol{\gamma}_{jk}$, for $j = 1,2,\ldots,J_k$ and $k = 1,2,\ldots,p$ are estimated in the structure of GAMLSS, for fixed vector of smoothing hyper parameters, maximizing a function likelihood penalty penalized l_p given by:

$$l_p = l - \frac{1}{2}\sum_{k=1}^{p}\sum_{j=1}^{J_k}\boldsymbol{\gamma}_{jk}^T\mathbf{G}_{jk}\boldsymbol{\gamma}_{jk}, \tag{4.2}$$

where $l = \sum_{i=1}^{n}\log\{f(y_i|\boldsymbol{\theta}^i)\}$ is the log-likelihood function of the data given $\boldsymbol{\theta}^i$ for $i = 1,2,\ldots,n$. Rigby and Stasinopoulos (2005) proposed two algorithms to maximize the likelihood function penalized. They fit a GAMLSS for fixed or for estimates values of hyperparameters, using the CG and RS algorithms (for more details see Chapter 3 Stasinopoulos et al. (2017)).

4.3 ROBUST FITTING FOR GAMLSS

An outlier is defined as a value of y lying in the left or right tail of the distribution of Y, i. e., a value of y for which cumulative distribution function $F_Y(y|\boldsymbol{\theta})$ is very close to 0 or 1 Rigby et al. (2019). The influence function, introduced by Huber (1967), is a important definition in robustness

to outliers. Hampel (1968) defines as "the effect of an infinitesimal contamination at point x, on a estimate, standardized by the mass of the contamination". For a robust estimator of parameter $\boldsymbol{\theta}$ we want the influence function to be bounded as y moves into the left or right tail of the distribution of Y. In general, maximum likelihood estimation (MLE) leads to parameter estimates with unbounded influence function for some or all parameters. Hence, MLE are generally vulnerable to the unlimited influence of even a single outlier y.

Rigby et al. (2019) proposed a robust estimation method for GAMLSSs based in bounded the influence function to obtain the parameter estimators. The approach is similar to the idea used in Field and Smith (1994) that starting with a parametric model and modifying the usual likelihood functions to obtain robust estimates with good breakdown properties. The breaking point is defined as the amount maximum of bad specification of the probabilistic model that an estimator can resist before breakdown, that is, useless (see Huber (1996)). In this way, the proposal method achieves robustness by transforming the observed response through (normalized) quantile residuals. The bounding of y is defined as

$$y_i^* = \begin{cases} y_i, & \Phi^{-1}(\alpha) \leq r_i \leq \Phi^{-1}(1-\alpha) \\ F_Y^{-1}(\alpha; \hat{\theta}_\mathbf{i}), & r_i < \Phi^{-1}(\alpha) \\ F_Y^{-1}(1-\alpha; \hat{\theta}_\mathbf{i}), & r_i > \Phi^{-1}(1-\alpha) \end{cases} \tag{4.3}$$

where α_1 is a probability close to zero (e.g. $\alpha = 0.01$), $r_i = \Phi^{-1}[F_Y(y_i; \hat{\theta}_\mathbf{i}]$ is the normalized quantile residual (Dunn and Smyth (1996a)) and where, $\hat{\theta}_i = (\hat{\mu}_i, \hat{\sigma}_i, \hat{\nu}_i, \hat{\tau}_i)$. The value α is tuning probabilities, associated to $\Phi^{-1}(\alpha)$, that can be seen as robustness tuning constant that regulates the potential outliers and, if necessary, its removed by truncating the distribution of the response variable. This proposal has two problems: the first is that the domain of the distribution considered is modified; the second problem is that the choice of α is made before fitting the model to data, however, there is no criterion for the choice of α. In the following, we describe the use of truncated distribution to propose a new robust fitting for GAMLSS, and in Section 4.3.3, we propose a adaptive truncation, that allows automatic selection of robustness tuning constant.

4.3.1 Robust Fitting for GAMLSS Using Truncation

In statistics, a truncated distribution is a conditional distribution that results from restricting the domain of some probability distribution.

Let X be a continuous random variable with probability density function $g(x)$ and a distribution function $G(x)$, $x \in \mathbb{R}$. Given a continuous random variable Y, its probability desnity function, $f_Y(y)$, which represents the distribution of X in the interval $[a,b]$, with $-\infty < a < b < \infty$, is a truncated probability distribution.

$$f_T(y) = \begin{cases} \frac{g(y)}{G(b)-G(a)}, & a \leq y \leq b \\ 0 & o.\,c. \end{cases} \tag{4.4}$$

The robust function proposed by Rigby et al. (2019) is defined through a restriction in the domain of the probability distribution in question, based on the residuals of the non-robust modeling. The new proposal of robust fitting for GAMLSS using truncation distribution for the observations that the normalized quantile residuals, of the original fitted distribution, lying outside of bounds of tuning constant probabilities α_1 and α_2. The bound of truncation distribution, t_1 and t_2 are the quantile of the original fitted distribution, associated of tuning constants probabilities α_1 and α_2, where they are small probability in tails. Therefore, the fit a GAMLSS model using a truncated distribution relies on transforming the response variable Y:

$$y_i^* = \begin{cases} y_i, & \Phi^{-1}(\alpha_1) \leq r_i \leq \Phi^{-1}(\alpha_2) \\ F_Y^{-1}(0.5; \hat{\theta}_\mathbf{i}), & r_i < \Phi^{-1}(\alpha_1) \\ F_Y^{-1}(0.5; \hat{\theta}_\mathbf{i}), & r_i > \Phi^{-1}(\alpha_2) \end{cases} \tag{4.5}$$

where $\hat{\theta}_i = (\hat{\mu}_i, \hat{\sigma}_i, \hat{\nu}_i, \hat{\tau}_i)$.

The robust estimator for the $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\lambda}$ is defined by maximizing the rewriting penalized log-likelihood defined in Equation (4.2) as:

$$l_{pT}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = l_T - \frac{1}{2} \sum_{k=1}^{p} \sum_{j=1}^{J_k} \lambda_{kj} \gamma_{kj}^{\top} \mathbf{G}_{kj} \gamma_{kj}. \tag{4.6}$$

where $l_T(\boldsymbol{\beta}) = \sum_{i=1}^{n} log\{f_T(y_i^* | \boldsymbol{\theta}^i)\}$ is the log-likelihood function, based on truncated distribution, of the data given $\boldsymbol{\theta}^i$ for $i = 1, 2, \ldots, n$.

Next we describe the algorithms for maximization the penalised log likelihood.

4.3.2 Algorithm for the New Robust Fitting for GAMLSS Based on Truncation

Our procedure of robust fitting for GAMLSS is as follow:

Step 1 Compute the estimation $\hat{\boldsymbol{\theta}}$ on the original data by maximizing (4.2) using a CG and RS algorithm (introduced briefly in Section 3.2.1);

Step 2 For given tuning probabilities bound α_1 and α_2 and with the estimation $\hat{\boldsymbol{\theta}}$ obtained define the tuning constant t_1 and t_2 of the truncated distribution.

Step 3 Bounded the response variable y, weighting out observations that are outside of bounds, based on (4.5), generating the new y values defined as y^*.

Step 4 Generate the truncated distribution.

Step 5 Compute the estimation $\hat{\boldsymbol{\theta}}$ on the y^* data by maximizing (4.6).

Step 6 Stop if there is no change global deviance otherwise go back to step 2.

Note that this transformation is indexed by a so-called robustness tuning probabilities α_1 and α_2. In our procedure of robust fitting for GAMLSS, the choice is subjective and we need a selection criteria. In the next section we propose a adaptive criterion for choice the tuning probabilities.

4.3.3 Adaptive Truncation

The robustness tuning constant probabilities α_1 and α_2 regulates how the contribution of an observation to the objective function and regulates the level of truncation and the number of observation truncated. The choice can be made by targeting a certain loss of estimation efficiency with respect to the MLE at the assumed model. Here, we proposed a criterion based on goodness of fit at the assumed model, that relies on the Anscombe-Glynn test [Thode (2002)] of kurtosis for normal samples or Anderson-Darling test of normality [Anscombe and Glynn (1983)]. Our automatic procedure of choice of tuning probabilities as follow:

Step 1 Define a set of candidates of tuning probabilities;

Step 2 For a given tuning probabilities α_1 and α_2 use the algorithm defined in the previous section of the New Robust Fitting for GAMLSS Based on Truncation.

Step 3 Apply the normality test on the normalized quantile residual, based on the Anscombe-Glynn or Anderson-Darling test with a predefined level of significance.

Step 4 The process looking for the best tuning probabilities, change α_1 and α_2 and repeat 1-2 until the hypothesis of normality is not rejected.

4.4 OUTLIERS CONTAMINATION

The generation of contaminated data is a fundamental part of the simulation process that seeks to analyze the efficiency of robust estimators. In the literature, few proposals for robust methods for generalized additive models and generalized linear models define criteria for contamination of simulated data. In this section, we describe the contamination criteria used in robust estimation proposals for generalized linear models and generalized additive models.

The term robustness is usually used to designate that a given method of statistical analysis is not sensitive to minor violations of the assumptions. The most typical situation refers to potential deviations from the form of the assumed probability distribution, but it can also be associated with other types of requirements or assumptions, such as independence, same distribution or randomization procedure Huber (1996).

The contamination of the sample with outliers is a relatively common form of violation of the postulated probability distribution. The tail of the distribution can become long, inflating the standard deviation estimate, depending on the intensity of contamination.

Simulation studies, which evaluate the robust propositions, use different forms of contamination of the probability distribution involved. Rigby et al. (2019) in chapter 12, section 12.2.4 simulate a random sample of size 490 from a $BEo(5,5)$ distribution and contaminate it with random samples of two Uniform distribution, $U(0,0.1)$ and $U(0.9,1)$ each sample with size 5.

Aeberhard et al. (2021) simulated a non parametric GAMLSS based on the gamma distribution and Poisson distribution. In the simulation studies with gamma distribution with parameters μ and σ, the work use the smooth function with two fixed covariates x_1 and x_2. The contaminated data are generated modifying a clean simulated data set by choosing at random 5% of the responses, in a specific region of the covariates, and by adding 10 to their original value. In the simulation studies with Poisson distribution with parameters μ, the work use the smooth function with covariates $X \sim Uniform(0,1)$ distribution. The contaminated data are obtained by randomly selecting 5% of the original responses and changing them to the nearest integer yu_1^{u2} , where u_1 is drawn from the $Uniform(2,5)$ distribution and where u_2 is randomly set to either $\{-1,1\}$.

Alimadad and Salibian-Barrera (2011) proposed a robust fit for generalized additive models. Simulate studies considered poisson and binomial responses, with outliers either at the beginning or at the end of the range of the covariate used in the experiment in the following manner. Let $(y_1^*,x_1),\ldots,(y_n^*,x_n)$ be the data, and consider the observations y_j^* with $x_{(k_1)} \leq x_j \leq x_{(k_2)}$, for fixed numbers k_1 and k_2 , where $x_{(m)}$ is the mth order statistic. Then

$$y_j = (1-z_j)y_j^* + z_j w_j,$$

where $z_j \sim B(1,\delta)$, and $w_j = 10$ or $w_j \sim Poisson(30)$ depending on whether y_j has a binomial or Poisson distribution. Hence, the number of outliers (and their position in the Poisson case) in each sample is random.

Bayes et al. (2012) proposed a new regression model for proportions considering the beta rectangular distribution. The work considering two simulation studies, one without covariates and the other considering covariates. In the scenario without covariates, first, a dataset with sample size n was simulated from the Beta distribution with parameters μ and parameter of dispersion σ. Next, 5% of sample of size n are replaced by values generated from the uniform distribution $(q,1)$, where q corresponds to the 0.999 quantile of the simulated beta distribution — that is, outliers in the right tail of the distribution by considering values generated from another distribution. In the scenario with covariates, the following model were considered: $Y \sim Beta(\mu,\sigma)$ and $logit(\mu) = \beta_0 + \beta_1 X$, where $i = 1,,n$ and $X \sim Unifrom(-3,3)$. In this scenario, 2% of the random sample generated from beta distribution (y) were replaced by their contaminated values $y = y \pm \Delta$. Were consider four perturbation patterns:

i A decrease of Δ units of the response values to higher values of x,

ii An increase of Δ units of the response values to lower values of x,

iii A decrease and increase of Δ units of the response values for higher and lower values of x, respectively

iv A decrease of Δ units of the response values for central values of x.

Croux et al. (2012a) simulated 1000 datasets of size $n = 250$ coming from a Poisson-like distribution with mean function $\mu(x_1, x_2)$ and dispersion function $\gamma(x_1, x_2)$. The data are simulated in three different settings, contaminating a growing percentage $(0\%, 3\%, \textit{and } 5\%)$ of observations, uniformly located in $0.1 < x_1 < 0.2$ and $0.8 < x_2 < 0.9$, with y observations drawn from a discrete $U(25, 28)$ distribution. Fu et al. (2020) propose a robust regression and consider a linear model $(y = \beta_0 + z\beta_1 + \sigma\varepsilon)$ with a variety of error distribution types, for instance, Normal erros, $N(0, 1)$, and σ takes a value of 1,3 and 4. Wong et al. (2014) studies M-type estimators for fitting robust Generalized Additive Models in the presence of outliers. Two types of distribution were considered: the binomial and poisson distribution, and two types of univariates smooth function. The contaminated data were generated in the following manner. For binomial distribution, y is set to 0 if the original value of y is 1, and vice versa. For Poisson data, y is set to the nearest integer to yu_{12}^{u} , where u_1 is generated from $Uniform(2, 5)$ and u_2 is drawn randomly from $(1, 1)$. Overall, these works do not present a fixed criterion for contamination. The next section describes the simulation studies as well as the contamination process used.

4.5 SIMULATIONS

In this section, we investigate the sensitivity of the our robust fitting for GAMLSS proposal in the presence of outliers by considering three simulations studies. The first study simulates a parametric GAMLSS model without covariates in systematic components for the parameters. The second study is a parametric GAMLSS model with one covariate in systematic component of all parameters. The third study used a non parametric GAMLSS model. In all three studies, was based on gamma and beta distributions comparing the following methods

1. non robust fitting for GAMLSS Rigby and Stasinopoulos (2005), called G;
2. robust fitting for GAMLSS proposed in Rigby et al. (2019), called RG
3. robust fitting for GAMLSS method propose by Aeberhard et al. (2021), called AH.
4. Adaptive Truncation GAMLSS based on Anderson Darling Test with significance level 0.01. The method is called AD01.
5. Adaptive Truncation GAMLSS based on Anderson Darling Test with 0.05. The method is called AD05.
6. Adaptive Truncation GAMLSS based on Anderson Darling Test with significance level 0.1 The method is called AD1.
7. Adaptive Truncation GAMLSS based on Anscombe Test with significance level 0.01. The method is called Ansc01.
8. Adaptive Truncation GAMLSS based on Anscombe test with significance level 0.05. The method is called Ansc05.
9. Adaptive Truncation GAMLSS based on Anscombe test with significance level 0.1. The method is called Ansc1.

The evaluation of the performance of the estimates was carried out by investigating the: Mean Square Error - $MSE(\hat{\theta},\theta) = \frac{1}{K}\sum_{i=1}^{K}(\hat{\theta}_i - \theta)^2$ where $\hat{\theta}_i$ is i-th Monte Carlo estimate K is he number of replications; and the Mean Square Deviations for each sample defined as $MSD = \frac{1}{n}\sum_{j=1}^{n}(\hat{\theta}_j - \theta_j)^2$ where $j = 1,\ldots,n$.

All computations are performed using the R packages GJRM Marra and Radice (2020) and gamlss Stasinopoulos et al. (2017) in R Team et al. (2020). The choice of the value of the robustness tuning constant c of Aeberhard method, was based on median downweighting proportion (MDP) using 0.95 efficiency, propose by Aeberhard et al. (2021). The tuning probability α regulates the contribution observation based on the normalized quantile residual. The tuning probabilites for *RG* and *RGW* was defined as $\alpha = 0.01$. The truncation robust used the adaptive method for select the tuning probabilities.

4.5.1 Simulations Under Beta Distribution

In this section we use the beta distribution to evaluated the proposal. The beta distribution is very flexible in situations where the Y, dependent variable, is continuous and restricted to the some finite intervals, because its density function can take different forms depending on the values of the parameters that compose it. The probability density function of the beta distribution, denoted by $BE(\mu,\sigma)$, is given by:

$$f_Y(y|\mu,\sigma) = \frac{1}{B(\tau,\beta)} y^{\upsilon-1}(1-y)^{\upsilon-1}.$$

for $0 < y < 1$, where

$$\tau = \mu(1-\sigma^2)/\sigma^2$$

$$\upsilon = (1-\mu)(1-\sigma^2)/\sigma$$

$\tau > 0$, $\upsilon > 0$ and hence $0 < \mu < 1$ and $0 < \sigma < 1$. In this parameterization $Y \sim BE(\mu,\sigma)$, the mean of Y is $E(Y) = \mu$ and the variance is$Var(Y) = \sigma^2\mu(1-\mu)$

The aim of this section is to investigate the performance of the methods in samples of size 100, contaminated with 0%, 2%, 5% and 10% of the observations. The investigation was based on 4 contamination scenarios of the variable response, considering the location of the contamination in the tail: left tail, right tail, both tails and at model. In general, the process use a random sample with 100 observations is generated from $BE(\mu,\sigma)$ and 2%, 5% and 10% of observation are selected at random and replaced according to contamination position. When contamination is carried out in both tails, the percentage of contamination of 5% is replaced by 6%, due to the fact that the sample size is even. The contamination processes are defined below.

1. contamination on the left - observation are selected at random and replaced by sample from uniform distribution in interval $(0,u_1)$, where u_1 is the quantile of the $BE(\mu,\sigma)$ with percentile 0.001.

2. contamination on the right - observation are selected at random and replaced by sample from uniform distribution in $(u_2,1)$, where u_2 is the quantile of the $BE(\mu,\sigma)$ with percentile 0.999.

3. contamination on the left and right - Observations are selected randomly and half of them are applied to process 1 and the other half to process 2.

4. without contamination - A random sample with 100 observations is generated from $BE(\mu,\sigma)$ without contamination.

In the following, details on the models used will be presented.

4.5.1.1 Parametric GAMLSS Based On Beta Distribution with Constants Systematic Components for μ and σ.

In this subsection the simulation studies consider the parametric GAMLSS model under beta distribution with $\mu = 0.7$ and $\sigma = 0.05$, without covariates, that is, $Y \sim BE(\mu, \sigma)$. The model can be defined with the systematic components defined as

$$\eta_1 = logit(\mu) = \beta_{11}$$

and

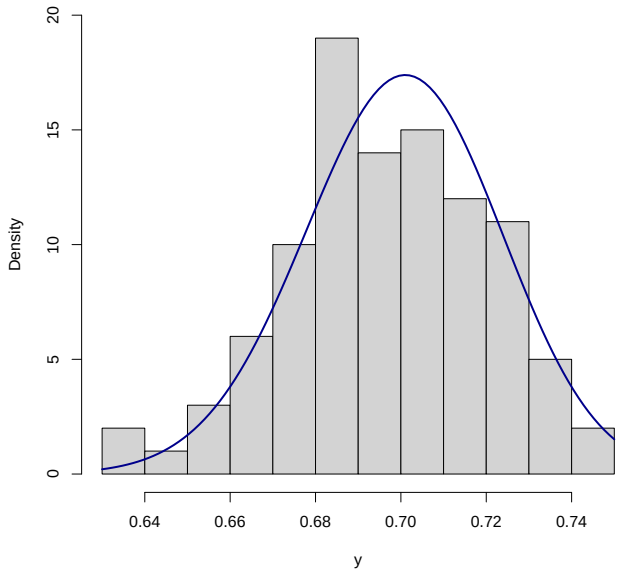
$$\eta_2 = logit(\sigma) = \beta_{12}.$$

The simulations study was based on 200 replicates, from the $BE(\mu = 0.7, \sigma = 0.05)$ distribution, with sample size 100, considering contamination of 0%, 2%, 5% and 10% of the observations. Figure 18 shows the density of $BE(0.7, 0.05)$ used in simulations and the histogram of one sample of size 100.

Figure 19 displays boxplots of $\hat{\mu}$ (left column of figure) and $\hat{\sigma}$ (right column of figure) for all methods, assumed Beta GAMLSS model without covariates in systematic component and without contamination. Only the AH (Aeberhard method) σ estimates show poor performance. The other methods shows good quality in the estimation of the parameters in comparison with the G method (non robust fitting for GAMLSS). Figure 20 displays boxplots of $\hat{\mu}$ (left column of figure) and $\hat{\sigma}$ (right column of figure) for all methods under left contamination. Each line of figure represents one level of contamination. The G estimation (non robust fitting for GAMLSS) shows poor performance in all levels of contamination, underestimating μ and overestimating σ, as expected. Among the robust methods, the AH underestimates the values of μ and σ for levels of 2% (first line) and 5%(second line). Under the level of 10% (third line) AH estimates are controlled showing good performance, while RG estimates are underestimated for μ overestimated for σ. All variations of our proposal showed good performance in all levels of contamination. Figure 21 displays boxplots of $\hat{\mu}$ and $\hat{\sigma}$ for all methods under right contamination. The G estimation overestimating μ and σ in all levels. Among the robust methods, the AH overestimate the values of μ and underestimates the values of σ for all levels of contamination. All variations of our proposal showed similar and good performance in all levels of contamination. Lastly, the results for the both tail contamination is displays in Figure 22. In this scenario, the AH estimations underestimates $\hat{\sigma}$ at the levels of 2% and 5%. Table 9 shows the results of MSE of the estimations of μ and σ. The truncation robust fitting for GAMLSS based on the Anscombe test, presents the lowest values of MSE at all levels of contamination and distribution tails.

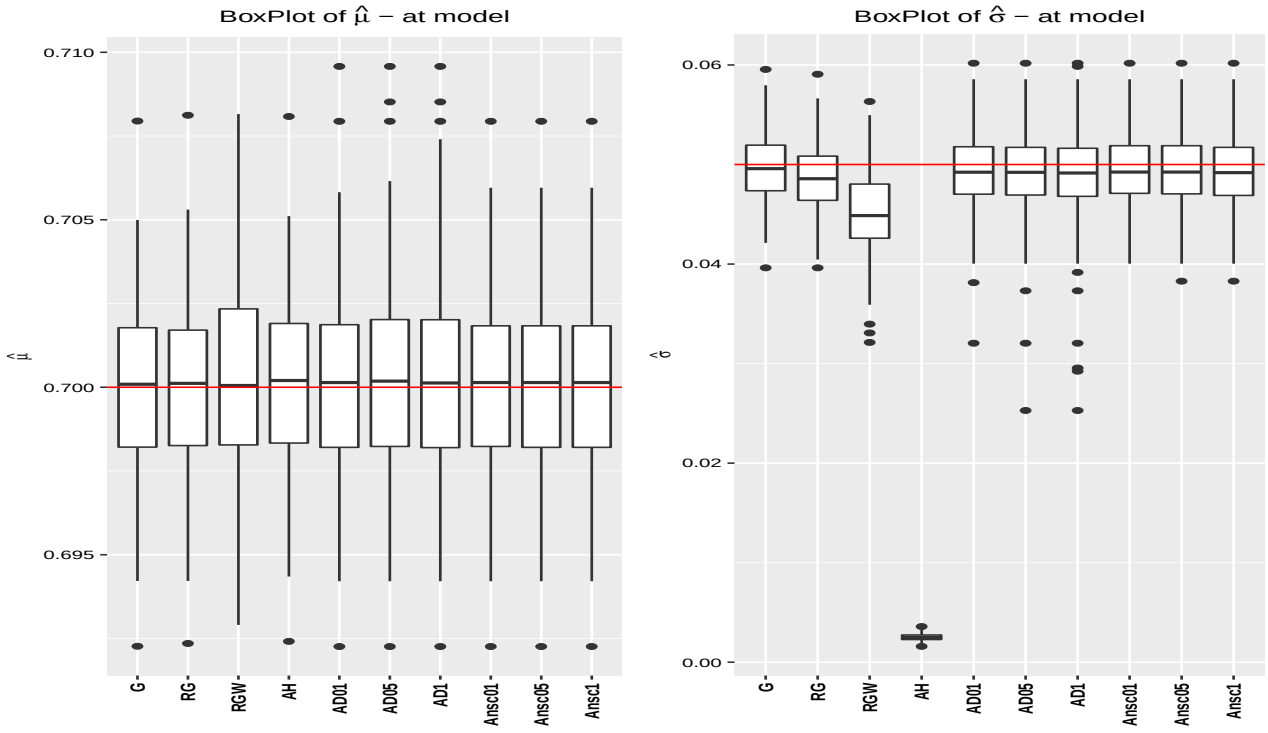
Overall, these simulation results yields one main conclusion. Our proposed robust method performs better to the existing alternatives in the robust fitting for GAMLSS.

Figure 18 – Histogram of a random sample and probability density function of a GAMLSS model based on a beta distribution with parameters $\mu = 0.7$ and $\sigma = 0.05$.



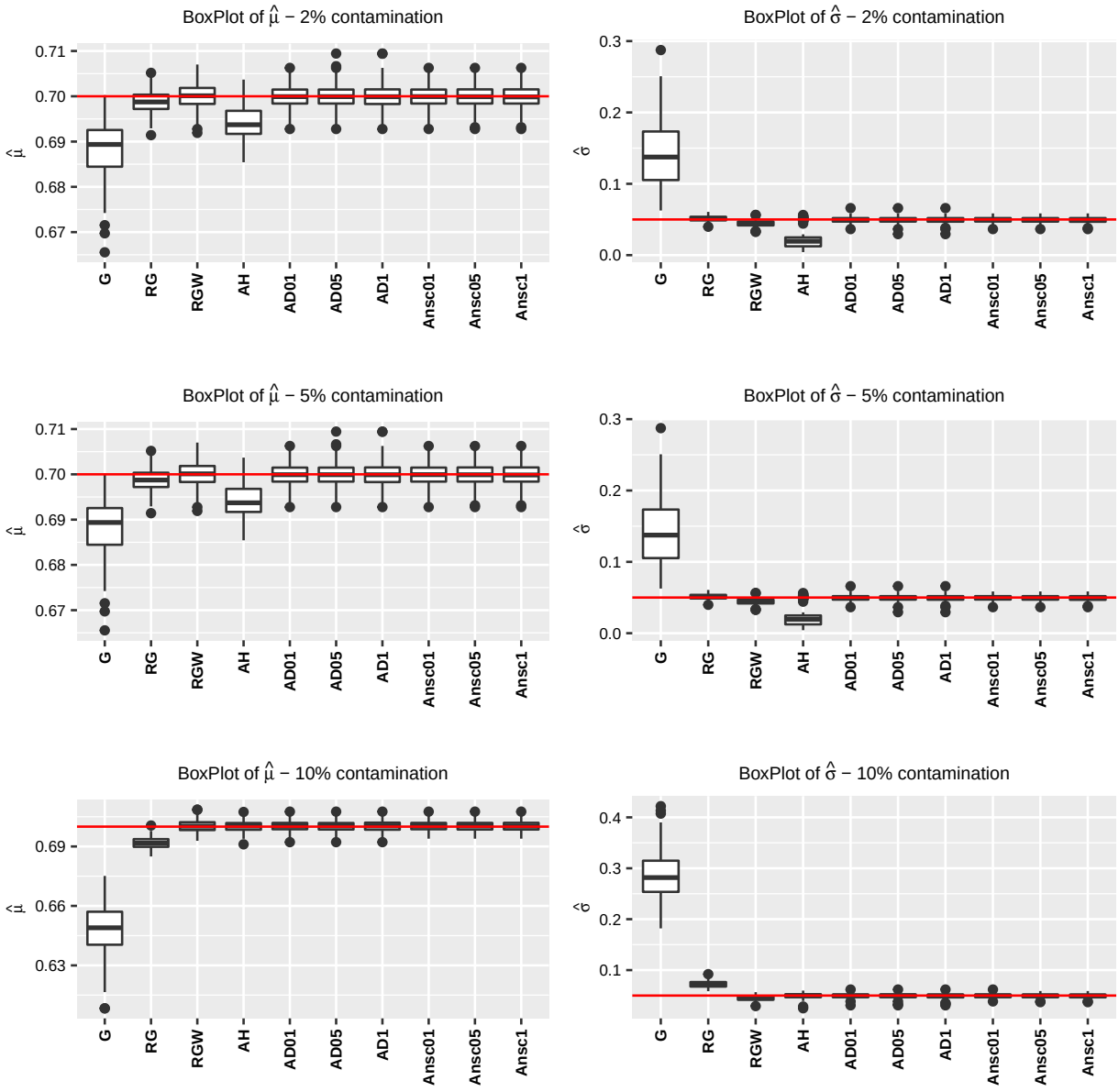
Source: The author (2021)

Figure 19 – Boxplots of $\hat{\mu}$ and $\hat{\sigma}$ simulated at model (without contamination), based on $Beta(\mu, \sigma)$ without covariates in systematic component and sample size 100. The red line is the parameter value ($\mu = 0.7$ and $\sigma = 0.05$).



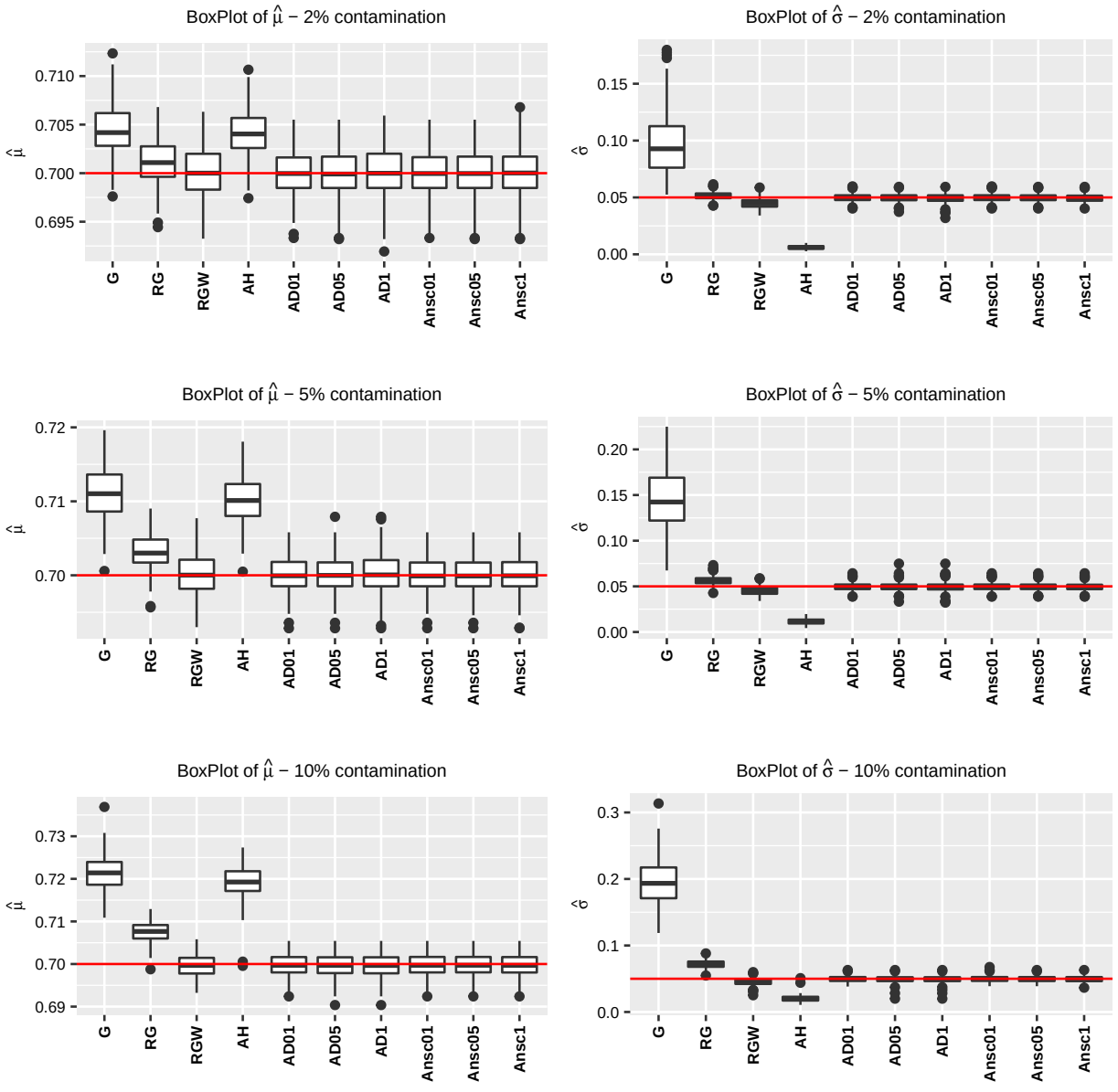
Source: The author (2021)

Figure 20 – Boxplots of $\hat{\mu}$ (first column of the figure) and $\hat{\sigma}$ (second column of the figure) simulated under left contamination (2% - first line of the figure , 5% - second line of the figure and 10% - third line of the figure), based on $Beta(\mu,\sigma)$ without covariates in systematic component and sample size 100. The red line is the parameter value ($\mu = 0.7$ and $\sigma = 0.05$).



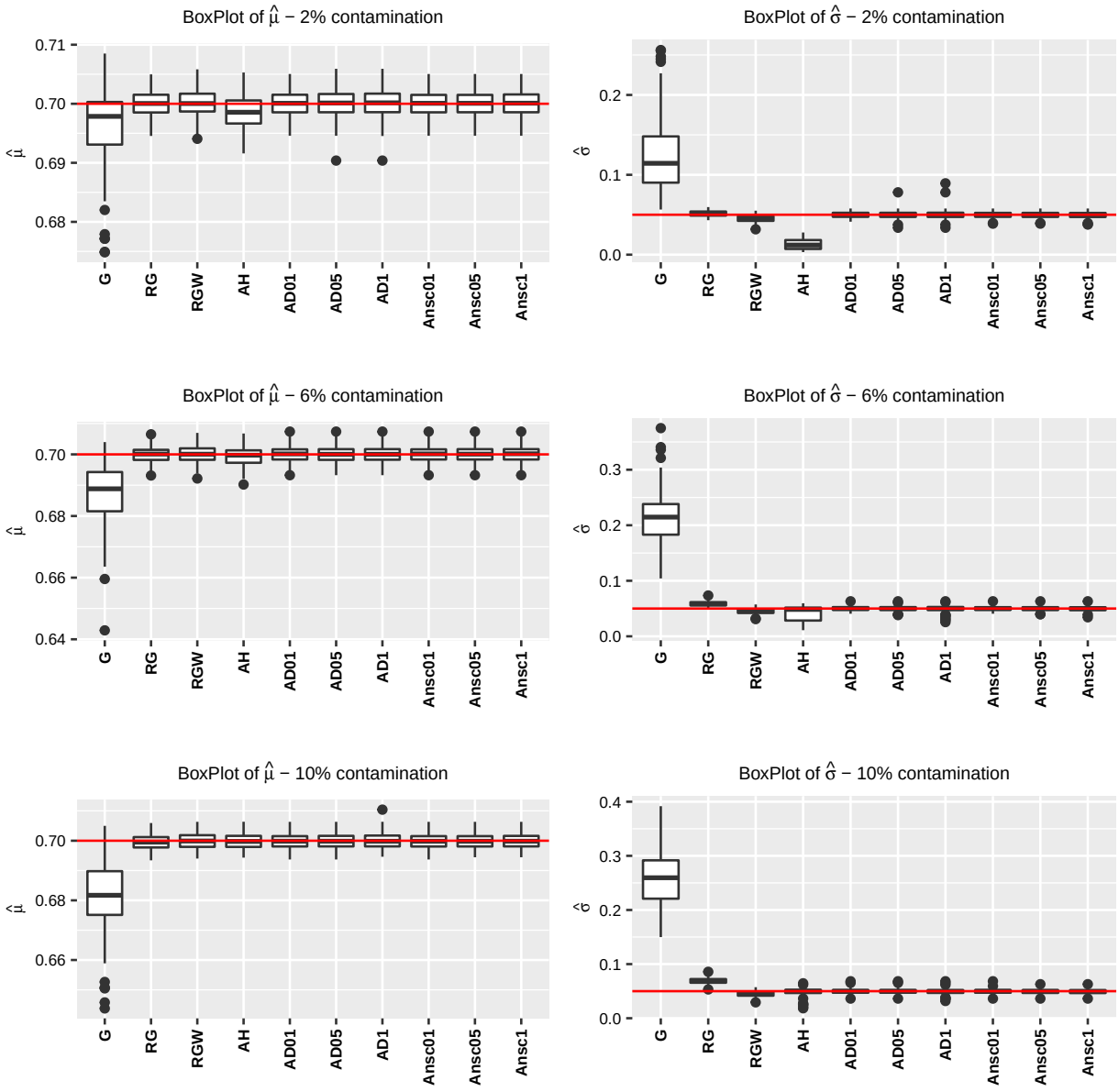
Source: The author (2021)

Figure 21 – Boxplots of $\hat{\mu}$ (first column of the figure) and $\hat{\sigma}$ (second column of the figure) simulated under right contamination (2% - first line of the figure , 5% - second line of the figure and 10% - third line of the figure), based on $Beta(\mu, \sigma)$ without covariates in systematic component and sample size 100. The red line is the parameter value ($\mu = 0.7$ and $\sigma = 0.05$).



Source: The author (2021)

Figure 22 – Boxplots of $\hat{\mu}$ (first column of the figure) and $\hat{\sigma}$ (second column of the figure) simulated under left and right contamination (2% - first line of the figure , 5% - second line of the figure and 10% - third line of the figure), based on $Beta(\mu, \sigma)$ without covariates in systematic component and sample size 100. The red line is the parameter value ($\mu = 0.7$ and $\sigma = 0.05$)



Source: The author (2021)

Table 9 – Mean squared errors of $\hat{\mu}$ and $\hat{\sigma}$ with 0%, 2%, 5% and 10% of contamination in the left tail, right tail and both tails, based on $Beta(\mu, \sigma)$ distribution with constant systematic component.

					without contamination						
Level	$\hat{\theta}$	G	RG	RGW	AH	AD01	AD05	AD1	Ansc01	Ansc05	Ansc1
0%	$\hat{\mu}$	**0.000006**	0.000006	0.000008	0.000006	0.000007	0.000007	0.000008	0.000006	0.000006	0.000006
	$\hat{\sigma}$	**0.000011**	0.000013	0.000042	0.002257	0.000015	0.000018	0.000023	0.000013	0.000013	0.000013
					Right contamination						
Level	$\hat{\theta}$	G	RG	RGW	AH	AD01	AD05	AD1	Ansc01	Ansc05	Ansc1
2%	$\hat{\mu}$	0.000027	0.000007	0.000007	0.000023	**0.000006**	**0.000006**	0.000007	**0.000006**	**0.000006**	**0.000006**
	$\hat{\sigma}$	0.002948	0.000013	0.000042	0.001932	**0.000012**	0.000013	0.000017	**0.000012**	**0.000012**	**0.000012**
5%	$\hat{\mu}$	0.000137	0.000015	0.000007	0.000112	**0.000005**	0.000006	0.000007	**0.000005**	**0.000005**	**0.000005**
	$\hat{\sigma}$	0.009983	0.000065	0.000044	0.001483	0.000018	0.000023	0.000027	**0.000018**	**0.000018**	**0.000018**
10%	$\hat{\mu}$	0.000472	0.000063	0.000007	0.000383	**0.000006**	0.000007	0.000007	**0.000006**	**0.000006**	**0.000006**
	$\hat{\sigma}$	0.022170	0.000521	0.000051	0.000909	0.000018	0.000024	0.000026	0.000018	**0.000017**	0.000019
					Left contamination						
Level	$\hat{\theta}$	G	RG	RGW	AH	AD01	AD05	AD1	Ansc01	Ansc05	Ansc1
2%	$\hat{\mu}$	0.000176	0.000007	0.000007	0.000048	**0.000005**	0.000006	0.000006	**0.000005**	**0.000005**	**0.000005**
	$\hat{\sigma}$	0.010392	0.000013	0.000047	0.000985	0.000013	0.000015	0.000016	**0.000011**	0.000012	0.000014
5%	$\hat{\mu}$	0.00102	0.00002	0.0000056	**0.00004**	0.000005	0.000005	0.000006	0.0000057	0.000005	0.000005
	$\hat{\sigma}$	0.03175	0.00006	0.000047	0.00015	0.00001	0.00001	0.00004	0.00001	0.00001	0.00001
10%	$\hat{\mu}$	0.002934	0.000076	0.000008	0.000007	**0.000006**	**0.000006**	0.000007	**0.000006**	**0.000006**	**0.000006**
	$\hat{\sigma}$	0.057655	0.000524	0.000049	0.000021	0.000017	0.000019	0.000022	**0.000015**	0.000016	0.000017
					left and right contamination						
Level	$\hat{\theta}$	G	RG	RGW	AH	AD01	AD05	AD1	Ansc01	Ansc05	Ansc1
2%	$\hat{\mu}$	0.000047	0.000005	0.000005	0.000010	0.000005	0.000005	0.000005	0.000005	0.000005	0.000005
	$\hat{\sigma}$	0.007238	0.000014	0.000042	0.001415	**0.000012**	0.000018	0.000027	0.000013	0.000013	0.000015
5%	$\hat{\mu}$	0.000228	**0.000007**	0.000009	0.000009	**0.000007**	**0.000007**	**0.000007**	**0.000007**	**0.000007**	**0.000007**
	$\hat{\sigma}$	0.028493	0.000089	0.000046	0.000241	**0.000014**	0.000016	0.000026	**0.000014**	0.000015	0.000016
10%	$\hat{\mu}$	0.000460	0.000006	0.000007	0.000006	0.000006	0.000006	0.000007	0.000006	0.000006	0.000006
	$\hat{\sigma}$	0.046628	0.000386	0.000047	0.000041	**0.000016**	0.000017	0.000021	**0.000016**	0.000016	0.000017

G - non robust fitting for GAMLSS, RG - robust fitting for GAMLSS by Rigby et al. (2019), RGW - robust fitting for GAMLSS modified, AH - Aeberhar method, AD01 - truncation robust fitting for GAMLSS with Anderson Darlin test and significance level 1%, AD05 - truncation robust fitting for GAMLSS with Anderson Darlin test and significance level 5%, AD1 - truncation robust fitting for GAMLSS with Anderson Darlin test with significance level 10%, Ansc01 - truncation robust fitting for GAMLSS with Anscombe test and significance level 1%, Ansc05 - truncation robust fitting for GAMLSS with Anscombe test and significance level 5%, Ansc1 - truncation robust fitting for GAMLSS with Anscombe test and significance level 10%

4.5.1.2 Parametric GAMLSS Based on Beta Distribution with Linear Systematic Components for μ and σ

In this subsection, was conduct a specific study based on $Y_i \sim Beta(\mu_i, \sigma_i)$ with systematic components defined as

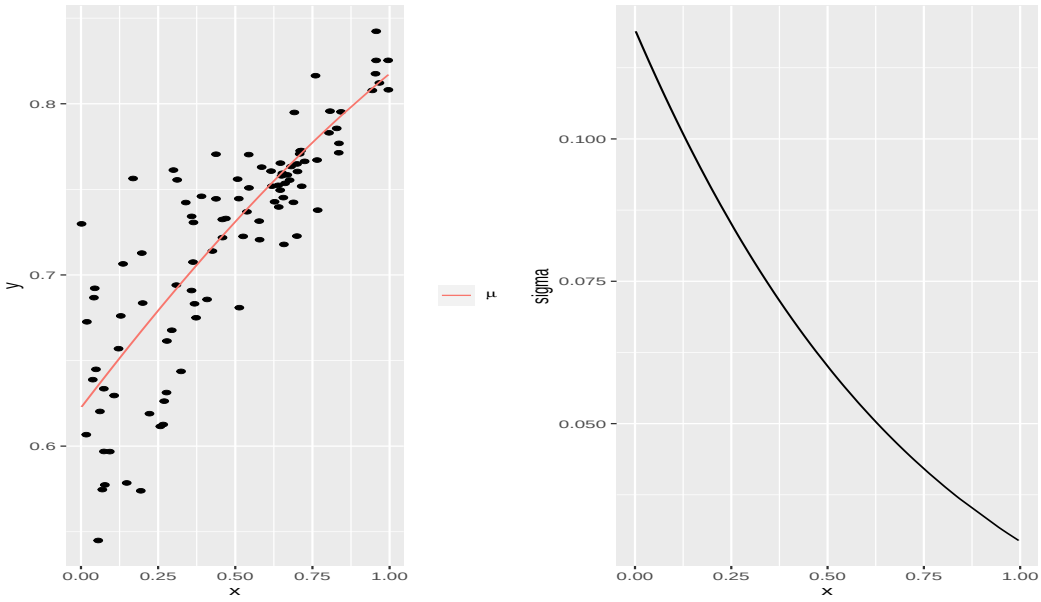
$$\eta_{i1} = logit(\mu_i) = \beta_{11} + B_{21}X_i,$$

and

$$\eta_{i2} = logit(\sigma_i) = \beta_{12} + \beta_{22}X_i,$$

where, $i = 1, \ldots, n$, $\beta_{11} = 0.5$, $\beta_{21} = 1$, $\beta_{12} = -2$, $\beta_{22} = -1.5$ and $X_i \sim Uniform(0,1)$ is a fixed covariate. The simulations studies was based on 200 replicates, from the beta distribution with sample sizes 100 and the contamination was carried out in 0%, 2%, 5% and 10% of the observations. Figure 23 shows the μ and σ used in simulations and the histogram of one sample generated based on model.

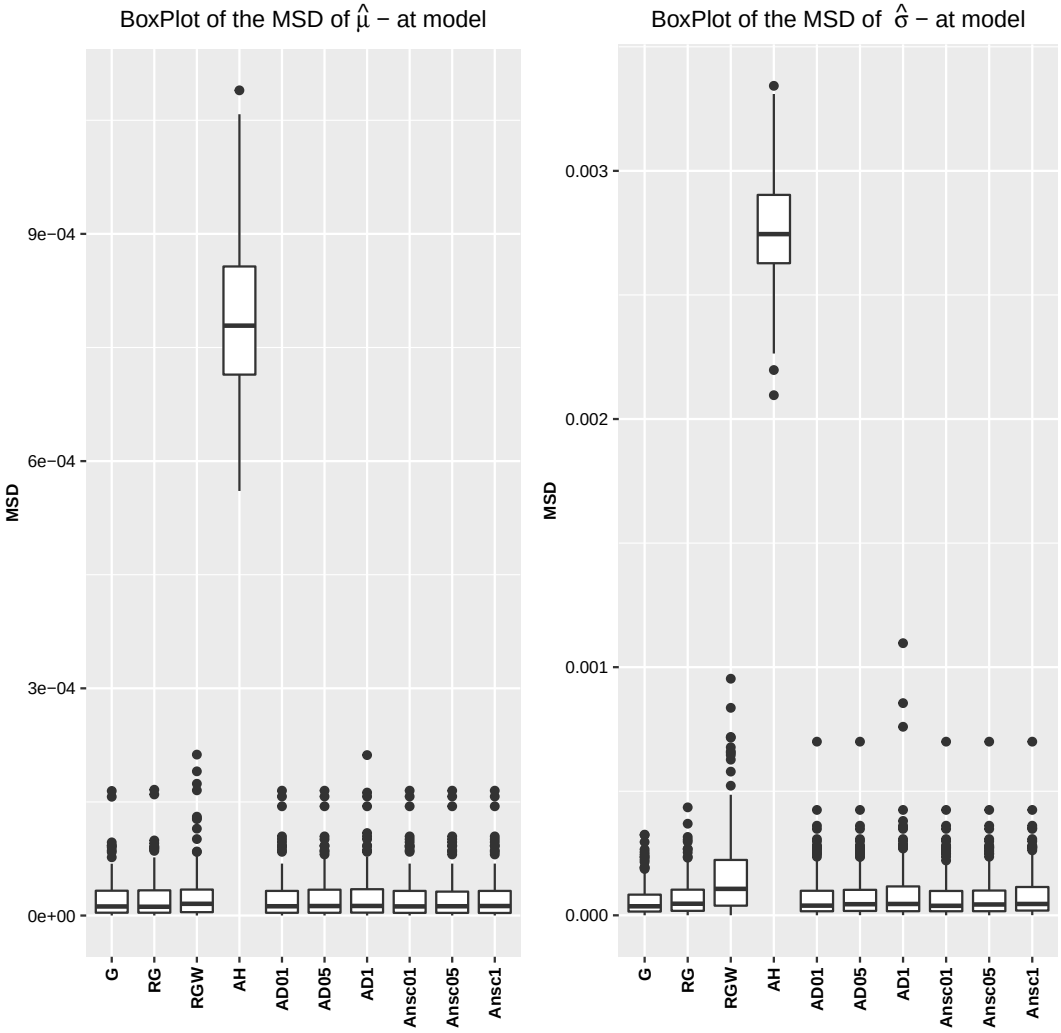
Figure 23 – μ and σ parameters used in the simulations of GAMLLSS model under $Beta(\mu, \sigma)$ with linear systematic components and a scatter plot of a random sample generated based on the proposed model without contamination.



Source: The author (2021)

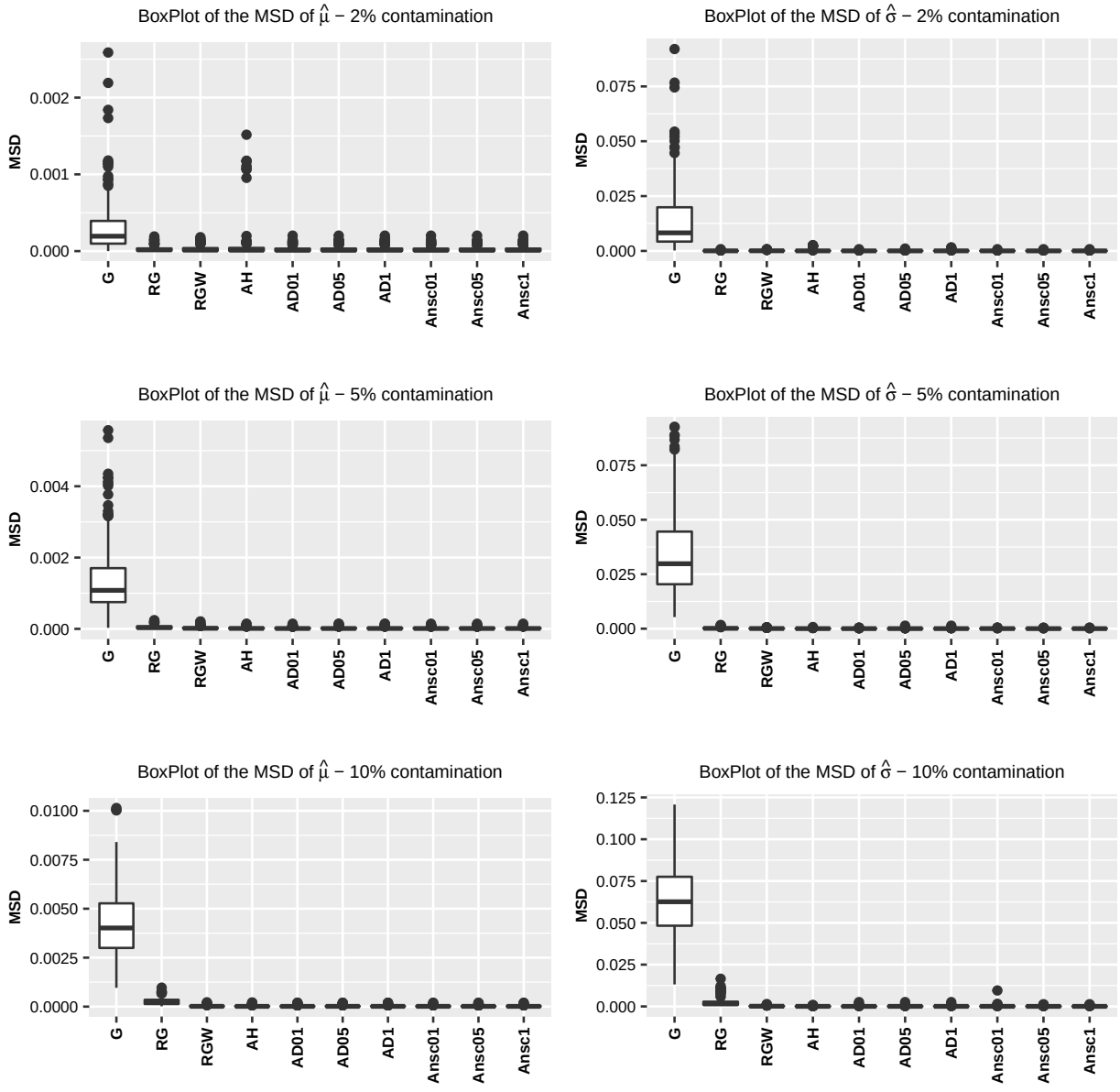
Figure 24 displays boxplots of the MSD of $\hat{\mu}$ (left column of figure) and $\hat{\sigma}$ (right column of figure) for all methods, based on the model without contamination. Only the AH method shows largest MSD for the μ and σ estimates, however, the values are small. The other methods shows similar values of MSD in comparison with the G method (non robust fitting for GAMLSS). Under left contamination (Figure 25), the robust methods has similar performance,standing out negatively AH estimations with some outliers in the boxplot of the MSD of $\hat{\mu}$ under 2% of contamination. Under right contamination (Figure 26), the AH estimates has the largest MSDs despite the small values at the level of 2% (first line). At the levels of 5% and 10% the results are similar. Under left and right contamination (Figure 27), again the AH estimations shows poor performance for $\hat{\mu}$ in levels of 2% with has the largest MSDs and tends to vary more than the others. For all the others robust methods the estimations shows good performance. Therefore, our proposed robust method performs better to the existing alternatives in the robust fitting for GAMLSS.

Figure 24 – Boxplots of the MSD of $\hat{\mu}$ and $\hat{\sigma}$, simulated at model (without contamination), based on $Beta(\mu, \sigma)$ with linear systematic component and sample size 100.



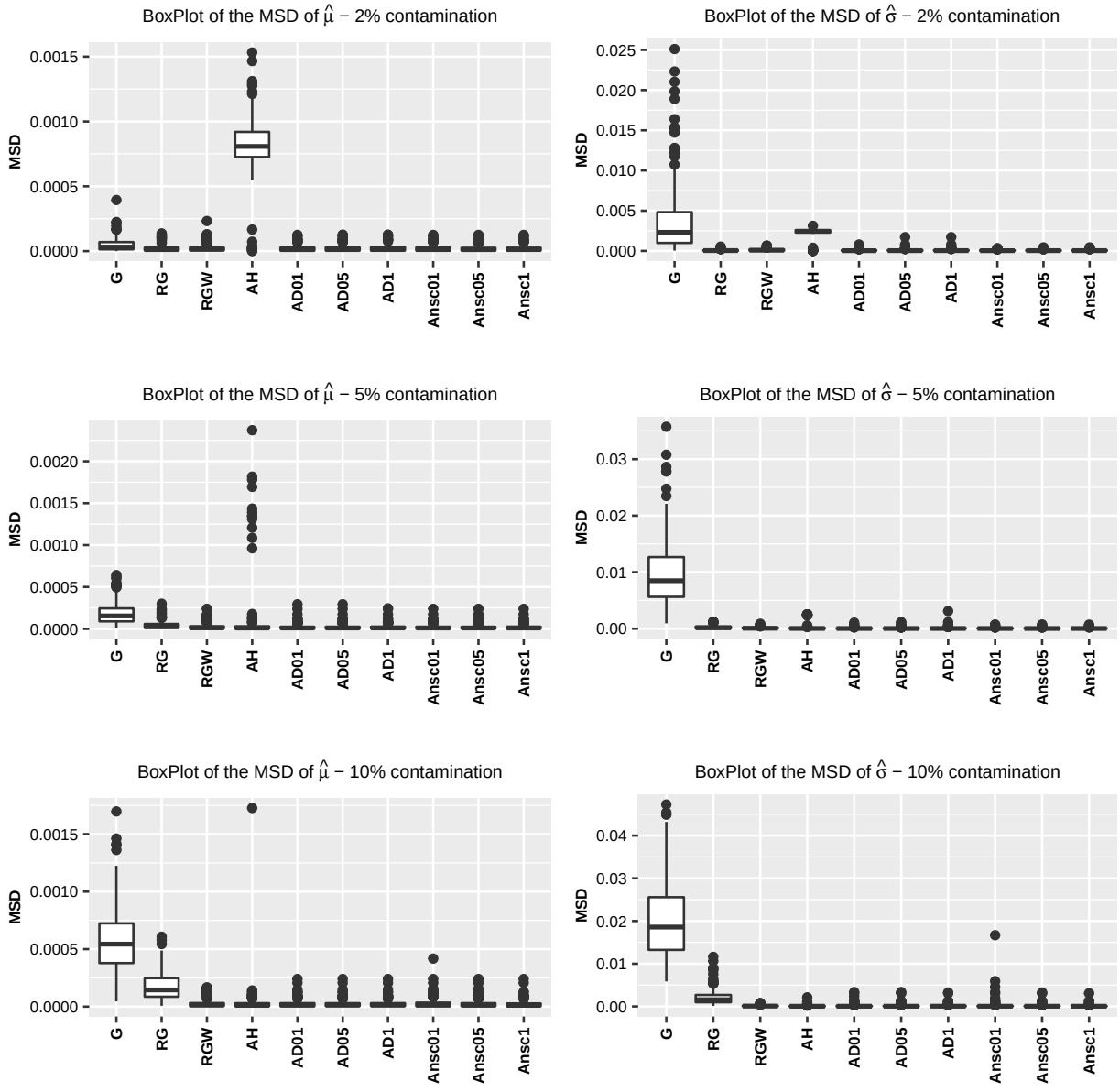
Source: The author (2021)

Figure 25 – Boxplots of the MSD of $\hat{\mu}$ (first column of the figure) and $\hat{\sigma}$ (second column of the figure), simulated under left contamination (2% - first line of the figure , 5% - second line of the figure and 10% - third line of the figure) and based on $Beta(\mu, \sigma)$ with linear systematic component and sample size 100.



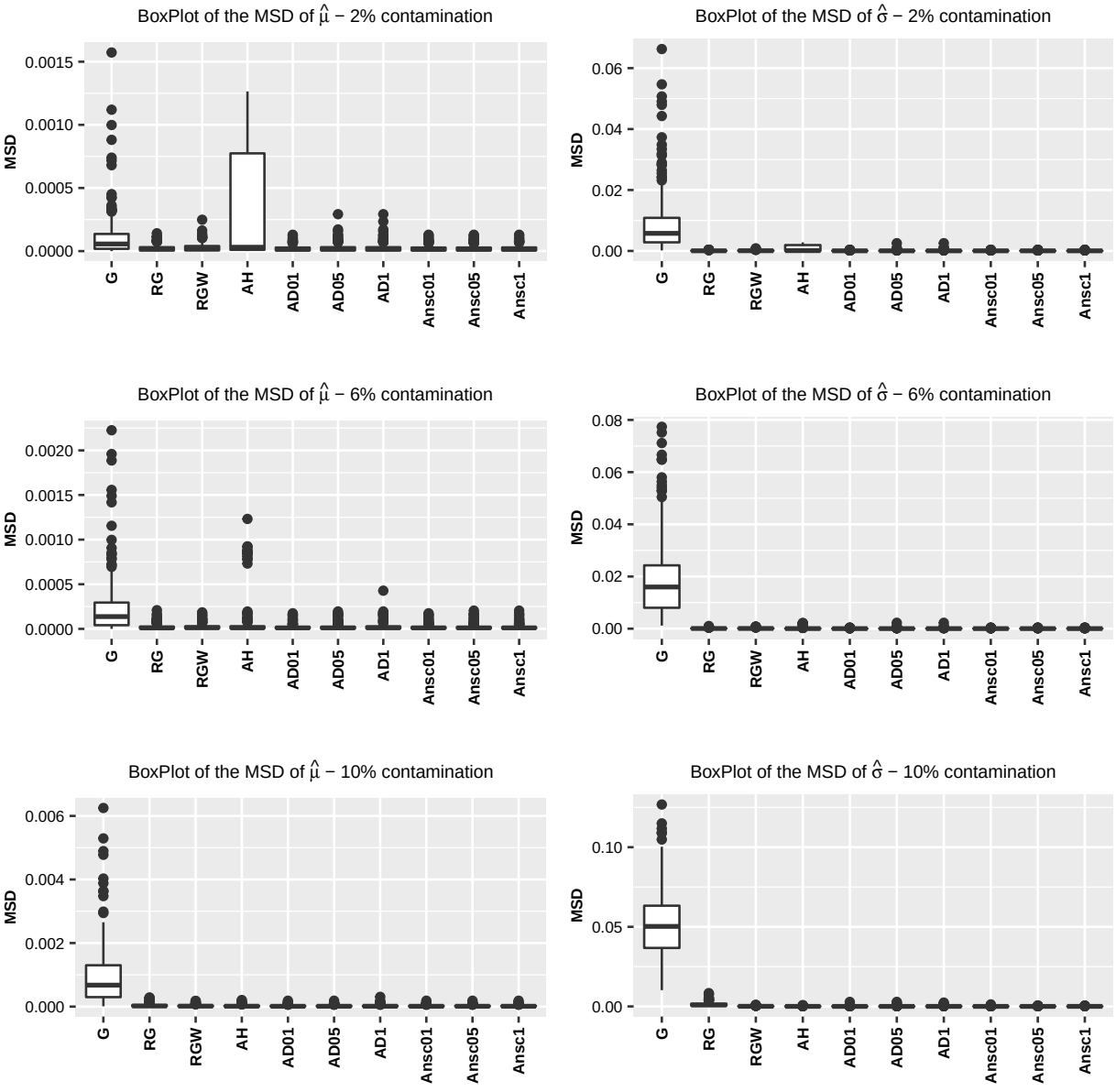
Source: The author (2021)

Figure 26 – Boxplots of the MSD of $\hat{\mu}$ (first column of the figure) and $\hat{\sigma}$ (second column of the figure), simulated under right contamination (2% - first line of the figure , 5% - second line of the figure and 10% - third line of the figure) and based on $Beta(\mu, \sigma)$ with linear systematic component and sample size 100.



Source: The author (2021)

Figure 27 – Boxplots of the MSD of $\hat{\mu}$ (first column of the figure) and $\hat{\sigma}$ (second column of the figure), simulated under left and right contamination (2% - first line of the figure , 5% - second line of the figure and 10% - third line of the figure) and based on $Beta(\mu, \sigma)$ with liner systematic component and sample size 100.



Source: The author (2021)

4.5.1.3 Non Parametric GAMLSS Based on Beta Model using Nonparametric Systematic Component with P-splines

The nonparametric beta model are evaluate based on the systematic components defined as

$$logit(\mu) = \eta_1 = s_1(X).$$

and

$$logit(\sigma) = \eta_2 = s_2(X).$$

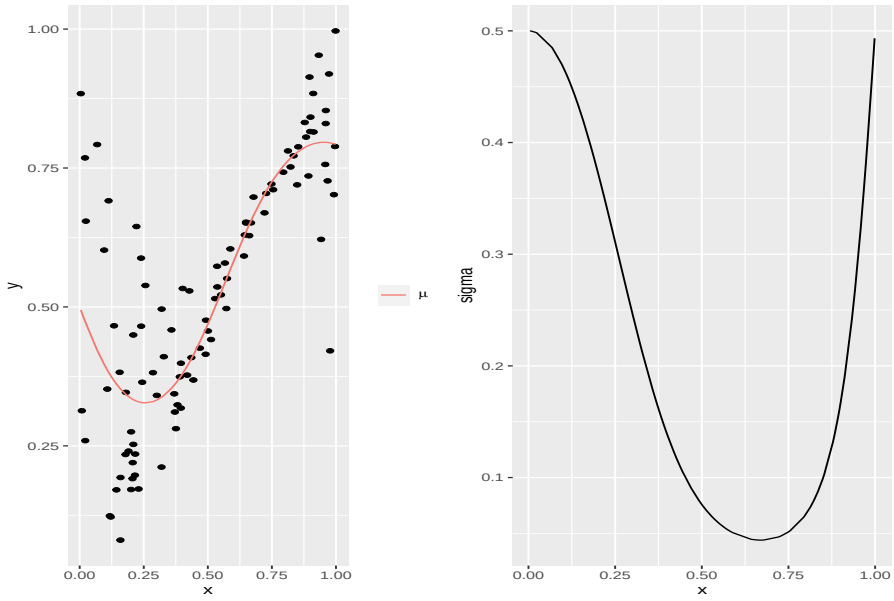
The smooth functions are

$$s_1(X) = -0.5 * X^2 + 5 \times \cos(X \times pi) \times exp(-X) \times -X$$

$$s_2(X) = -2(X^2 + 2X)sin(X\pi).$$

where $X \sim U(0,1)$ is a fixed covariate. The simulations studies were based on 200 replicates, from the beta distribution with sample sizes 100 and the contamination was carried out in 2%, 5% and 10% of the observations. Figure 28 shows the μ and σ used in simulations and the histogram of one sample generated based on model.

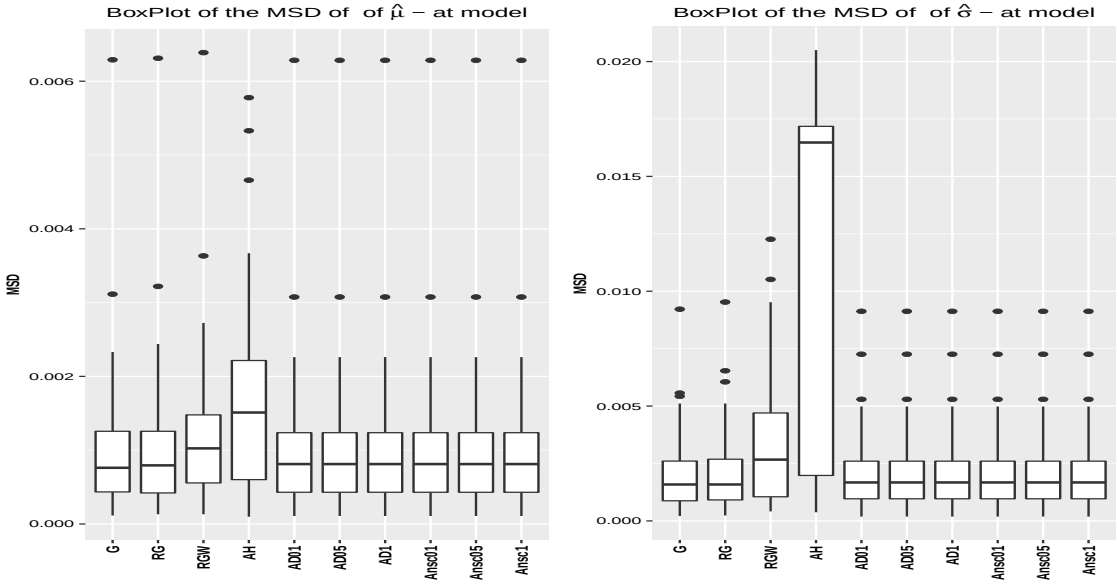
Figure 28 – μ and σ parameters used in the simulations of GAMLLSS model under $Beta(\mu, \sigma)$ with linear systematic components and a scatter plot of a random sample generated based on the proposed model without contamination.



Source: The author (2021)

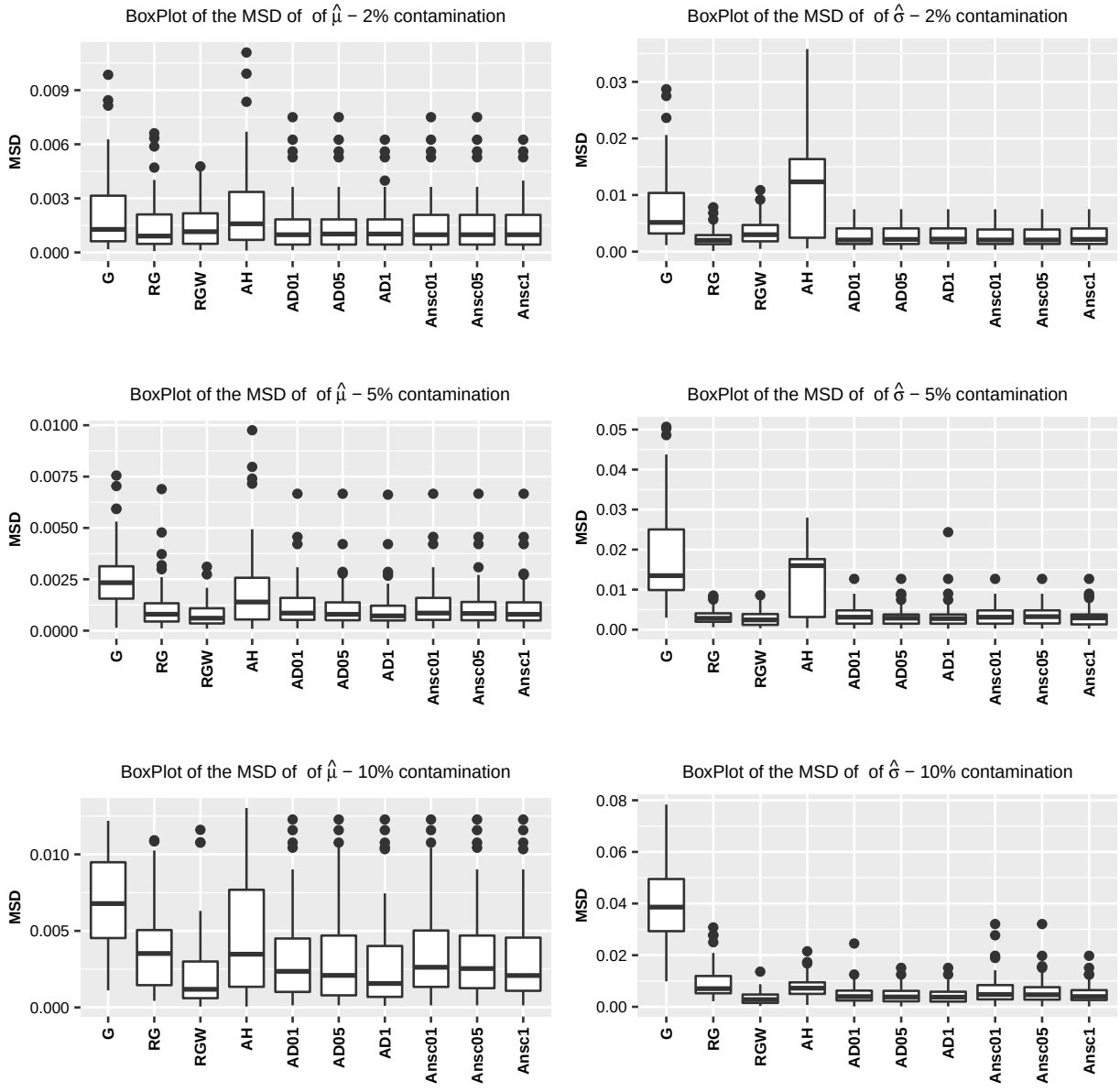
Figure 29 displays boxplots of the MSD of $\hat{\mu}$ and $\hat{\sigma}$ for all methods simulated without contamination. The AH estimation shows poor performance for the σ estimates, presenting largest MSDs and tends to vary more than the others. Our proposal performs similarly to the G method (non robust fitting for GAMLSS). Under left contamination (Figure 30), our robust methods has similar performance, while the AH has the largest MSDs. Under right contamination (Figure 31), the non parametric model is less sensitive to contamination in 2% of observations, showing a similar results. Only the error of AH estimates show greater variability. At the level of 5% contamination (second line of figure) the results are similar, with largest MSDs for AH. At the level of 10% the RGW, AD05 and AD1 has the lowest MSDs. Finally, Figure 32 displays boxplots of the MSD of $\hat{\mu}$ and $\hat{\sigma}$ under left and right contamination. Among robust methods, the AH has the largest MSDs. At the levels of 5% and 10% the results are similar.

Figure 29 – Boxplots of the MSD of $\hat{\mu}$ and $\hat{\sigma}$, simulated at model (without contamination), based on $Beta(\mu, \sigma)$ with non parametric systematic component and sample size 100.



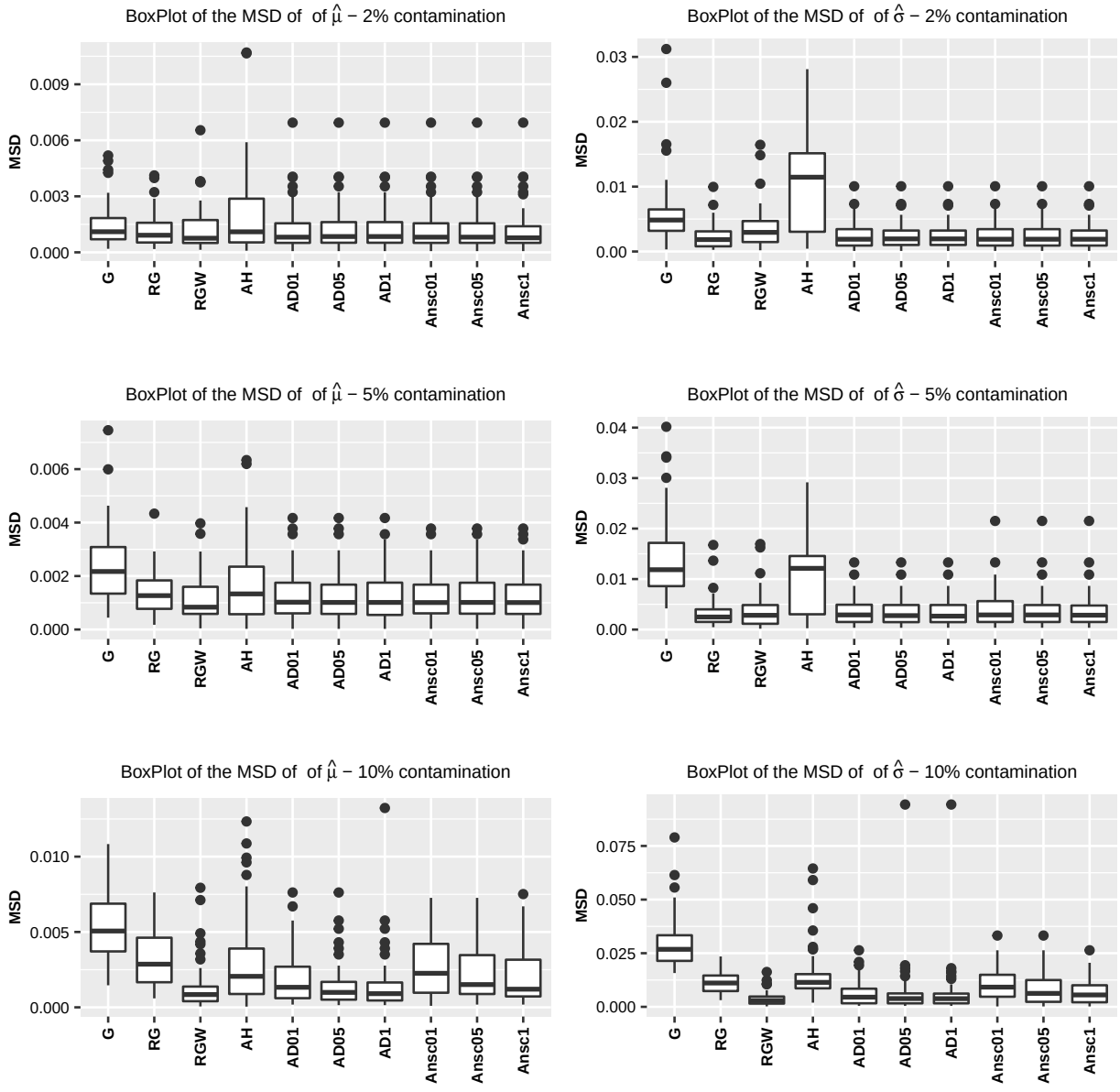
Source: The author (2021)

Figure 30 – Boxplots of the MSD of $\hat{\mu}$ (first column of the figure) and $\hat{\sigma}$ (second column of the figure), simulated under left contamination (2% - first line of the figure , 5% - second line of the figure and 10% - third line of the figure), based on $Beta(\mu, \sigma)$ with non parametric systematic component and sample size 100.



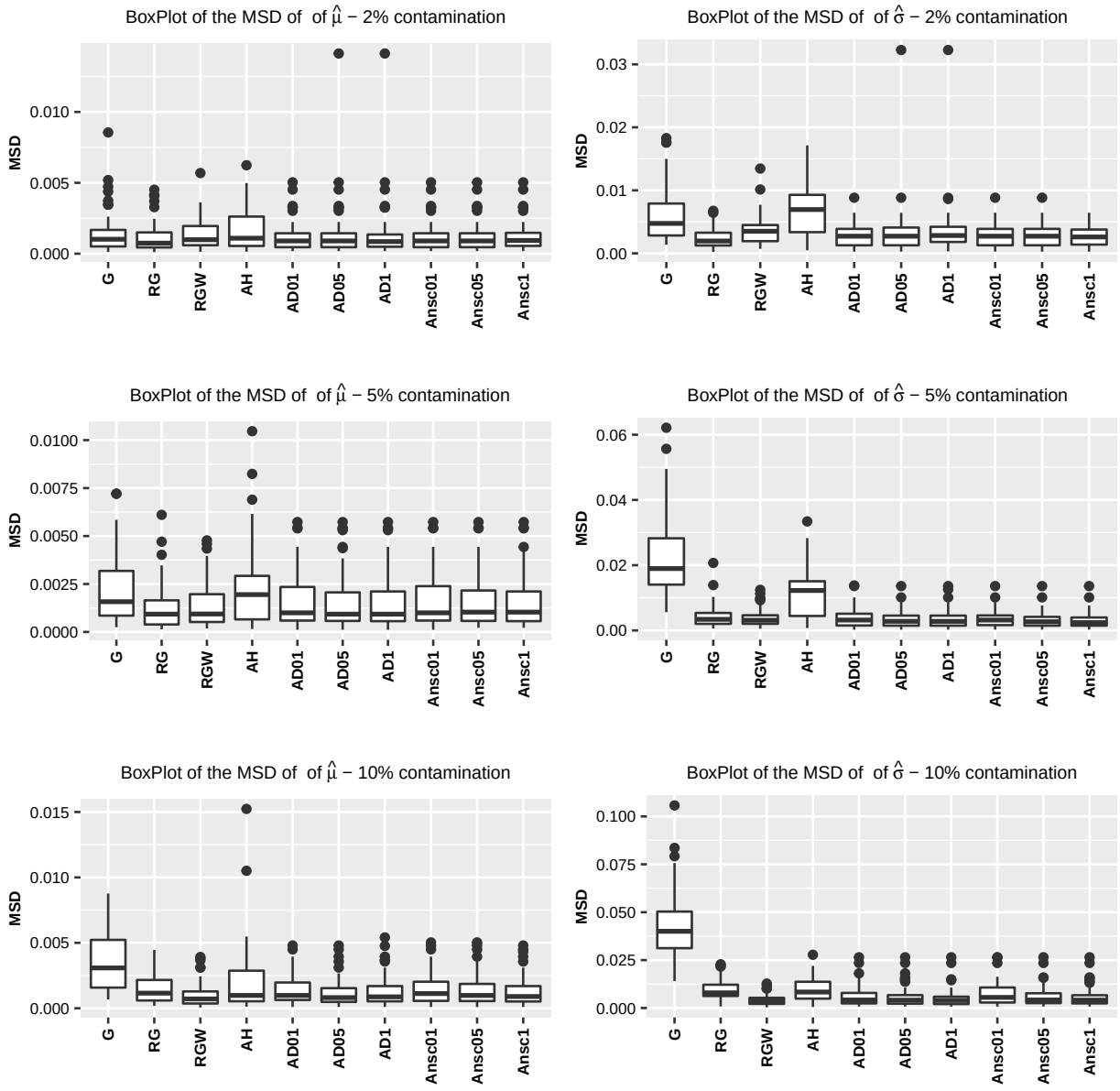
Source: The author (2021)

Figure 31 – Boxplots of the MSD of $\hat{\mu}$ (first column of the figure) and $\hat{\sigma}$ (second column of the figure), simulated under right contamination (2% - first line of the figure , 5% - second line of the figure and 10% - third line of the figure), based on $Beta(\mu, \sigma)$ with non parametric systematic component and sample size 100.



Source: The author (2021)

Figure 32 – Boxplots of the MSD of $\hat{\mu}$ (first column of the figure) and $\hat{\sigma}$ (second column of the figure), simulated under left and right contamination (2% - first line of the figure , 5% - second line of the figure and 10% - third line of the figure), based on $Beta(\mu, \sigma)$ with non parametric systematic component and sample size 100.



Source: The author (2021)

4.5.2 Simulations Under Gamma Distribution

In this section we use the gamma distribution to evaluate the proposal. The gamma distribution is appropriate for positively skewed data and its density function, denoted by $GA(\mu,\sigma)$, is given by

$$f_y(y|\mu,\sigma) = \frac{y^{1/\sigma^2-1}}{(\sigma^2\mu)^{1/\sigma^2}\Gamma(1/\sigma^2)}$$

for $y > 0$, where $\mu > 0$ and $\sigma > 0$. Here $E(Y) = \mu$ and $Var(Y) = \sigma^2\mu^2$.

The investigation used samples of size 100, based on 4 contamination scenarios of the variable response, considering the location of the contamination: left tail, right tail and both tails. A random sample with 100 observations is generated from $GA(\mu,\sigma)$ and 2%, 5% and 10% of the observations are randomly selected and replaced according to contamination position. When contamination is carried out on the left and right tail, the percentage of contamination of 5% is replaced by 6%. The contamination processes are defined below.

1. contamination on the left - observation are selected at random and replaced by the quantile of the $GA(\mu,\sigma)$ distribution with the 0.0001 percentile;
2. contamination on the right - observations are selected randomly and a constant k defined as the quantile of the distribution $G(\mu,\sigma)$ with a percentile of 0.9999 is added;
3. contamination on the left and right - Observations are selected randomly and half of them are applied to process 1 and the other half to process 2;
4. without contamination - A random sample with 100 observations is generated from $GA(\mu,\sigma)$ without contamination.

In the next subsections, details on the models used will be presented.

4.5.2.1 Parametric GAMLSS Based On Gamma Distribution without Covariates Systematic Components for μ and σ.

In this subsection the simulation studies consider the parametric GAMLSS model under gamma distribution with $\mu = 10$ and $\sigma = 0.5$ without covariates, that is, $Y \sim GA(10, 0.5)$. Hence, the model can be defined with the systematic components defined as

$$\eta_1 = \log(\mu) = \beta_{11},$$

and

$$\eta_2 = \log(\sigma) = \beta_{12}.$$

The simulations studies was based on 200 replicates, from the $GA(\mu = 10, \sigma = 0.5)$ distribution, with sample size 100, considering contamination of 0%, 2%, 5% and 10% of the observations. Figure 33 shows the density of distribution used in simulations and the histogram of sample of size 100.

Figure 34 displays boxplots of $\hat{\mu}$ (left column of figure) and $\hat{\sigma}$ (right column of figure) for all methods assumed gamma GAMLSS model without contamination and covariates in systematic component. Only the RGW estimates shows poor performance, underestimating the parameters. Our proposal shows good quality in the estimation in comparison with the G method (non robust fitting for GAMLSS), with similar results. Figure 35 displays boxplots of $\hat{\mu}$ and $\hat{\sigma}$ for all methods under left contamination. At the level of 2% contamination, the μ estimates shows good performance for all methods with slightly worse performance for G estimates. The σ estimates shows better performance for our proposal. At the level of 10% of contamination, the truncated robust GAMLLS, using the Anderson Darling test with any significance level, tends to present less variability and proximity to

the real values of the model. The AH estimations shows poor performance with high variability for σ estimates at level 5% and overestimating σ at level 10%. The methods RG, AH and Ansc01 were not able to adequately estimate the σ values at the level of 10%. Under right contamination (Figure 36), the G estimations overestimating μ and σ in all levels. Among the robust methods, the RG overestimated the values of μ and σ for all levels of contamination, the RGW estimations underestimates the values of σ in all levels of contamination. All variations of our proposal showed good performance in all levels of contamination with similar results. Ansc01 estimates tend to show greater variability. Lastly, the results for the both tail contamination is displays in Figure 37. In this scenario, our proposal present the best results with less variability and similar results for all variations. The AH estimations of σ at level 5% and 10% shows high variability with a tendency to overestimate. Table 10 shows the MSE of $\hat{\mu}$ and $\hat{\sigma}$. The truncation robust based on Anscombe test so does not show better results at 10% contamination levels.

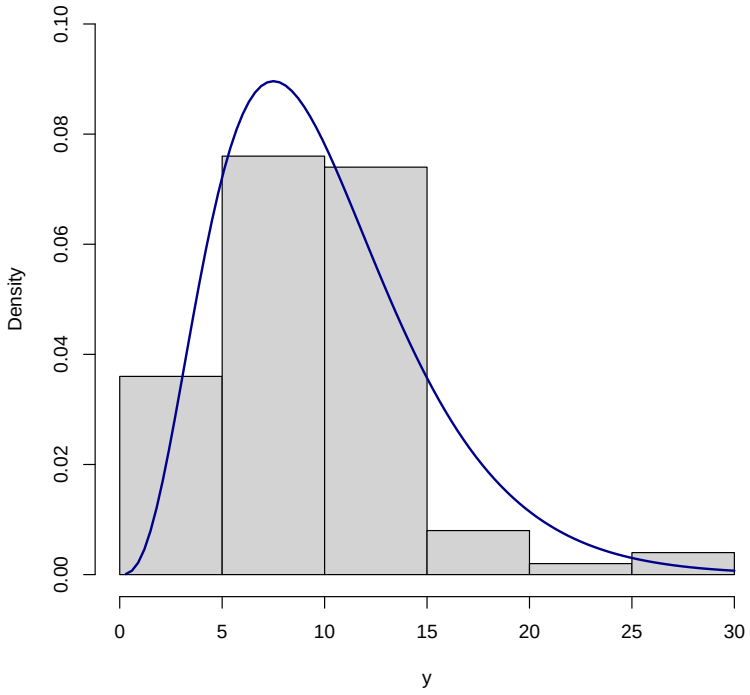
Overall, these simulation results yields one main conclusion. Our proposed robust method performs better than existing alternatives in the robust fitting for GAMLSS, and the truncation robust fitting for GAMLSS, based on Anscombe test, is slightly best performance for contamination levels of 2% and 5%, while the truncation robust fitting for GAMLSS, based on Anderson Darling test, is better for the conatamination level of 10%

Table 10 – Mean square error for $\hat{\mu}$ and $\hat{\sigma}$ with 0%, 2%, 5% and 10% of contamination in the left tail, right tail and both tails, of $GA(\mu, \sigma)$ with constant systematic component.

Without contamination											
Level	$\hat{\theta}$	G	RG	RGW	AH	AD01	AD05	AD1	Ansc01	Ansc05	Ansc1
0%	$\hat{\mu}$	**0.000006**	0.000006	0.000008	0.000006	0.000007	0.000007	0.000008	0.000006	0.000006	0.000006
	$\hat{\sigma}$	**0.000011**	0.000013	0.000042	0.002257	0.000015	0.000018	0.000023	0.000013	0.000013	0.000013
Right contamination											
Level	$\hat{\theta}$	G	RG	RGW	AH	AD01	AD05	AD1	Ansc01	Ansc05	Ansc1
2%	$\hat{\mu}$	1.080033	0.361641	0.252016	0.231639	0.241528	0.260609	0.276917	0.232586	0.231526	**0.228580**
	$\hat{\sigma}$	0.007348	0.001595	0.004670	0.001594	0.001509	0.002116	0.002589	**0.001494**	0.001660	0.001701
5%	$\hat{\mu}$	5.978484	1.611620	0.339583	0.307492	0.293229	0.343079	0.396373	0.289707	**0.288429**	0.293412
	$\hat{\sigma}$	0.031593	0.005848	0.004064	0.001410	0.001493	0.001943	0.002818	**0.001371**	0.001396	0.001464
10%	$\hat{\mu}$	22.329934	14.243905	0.315961	**0.287171**	0.311421	0.311654	0.336634	5.827071	0.293656	0.309258
	$\hat{\sigma}$	0.072660	0.047567	0.004842	**0.001815**	0.002429	0.002943	0.003238	0.020127	0.002325	0.002315
Left contamination											
Level	$\hat{\theta}$	G	RG	RGW	AH	AD01	AD05	AD1	Ansc01	Ansc05	Ansc1
0%	$\hat{\mu}$	0.279989	0.271174	0.367058	0.286428	0.268324	0.274299	0.279722	0.264916	0.260221	0.268902
	$\hat{\sigma}$	0.006445	0.002070	0.004586	0.002582	0.001344	0.001931	0.002459	**0.001326**	0.001343	0.001554
5%	$\hat{\mu}$	0.458018	0.402517	0.386445	0.311100	0.292978	0.319177	0.320348	0.302044	**0.292071**	0.292063
	$\hat{\sigma}$	0.030220	0.009062	0.005025	0.020678	**0.001673**	0.002606	0.002723	0.002549	0.001703	0.001769
10%	$\hat{\mu}$	1.133534	1.007512	0.442105	0.523201	**0.297604**	0.361759	0.376285	0.979308	0.456819	0.334813
	$\hat{\sigma}$	0.093362	0.084681	0.016044	0.104466	**0.002187**	0.002829	0.003092	0.089168	0.024137	0.007426
left and right contamination											
Level	$\hat{\theta}$	G	RG	RGW	AH	AD01	AD05	AD1	Ansc01	Ansc05	Ansc1
2%	$\hat{\mu}$	0.329630	0.236114	0.330853	0.247510	0.250449	0.260435	0.274652	**0.235935**	0.237472	0.256391
	$\hat{\sigma}$	0.007275	0.001629	0.005240	0.001619	0.002224	0.002474	0.002911	**0.001235**	0.001290	0.001558
5%	$\hat{\mu}$	0.359355	0.097202	0.093907	0.066363	0.076502	0.076619	0.087286	**0.063558**	0.065757	0.068647
	$\hat{\sigma}$	0.011915	0.002521	0.001189	0.001214	0.000480	0.000388	0.000368	0.000345	0.000352	**0.000341**
10%	$\hat{\mu}$	3.634607	1.176805	0.410479	0.372144	0.342864	0.370330	0.389817	0.362483	**0.342256**	0.340621
	$\hat{\sigma}$	0.103202	0.043173	0.004439	0.029092	0.001697	0.002118	0.002500	0.002717	**0.001654**	0.002059

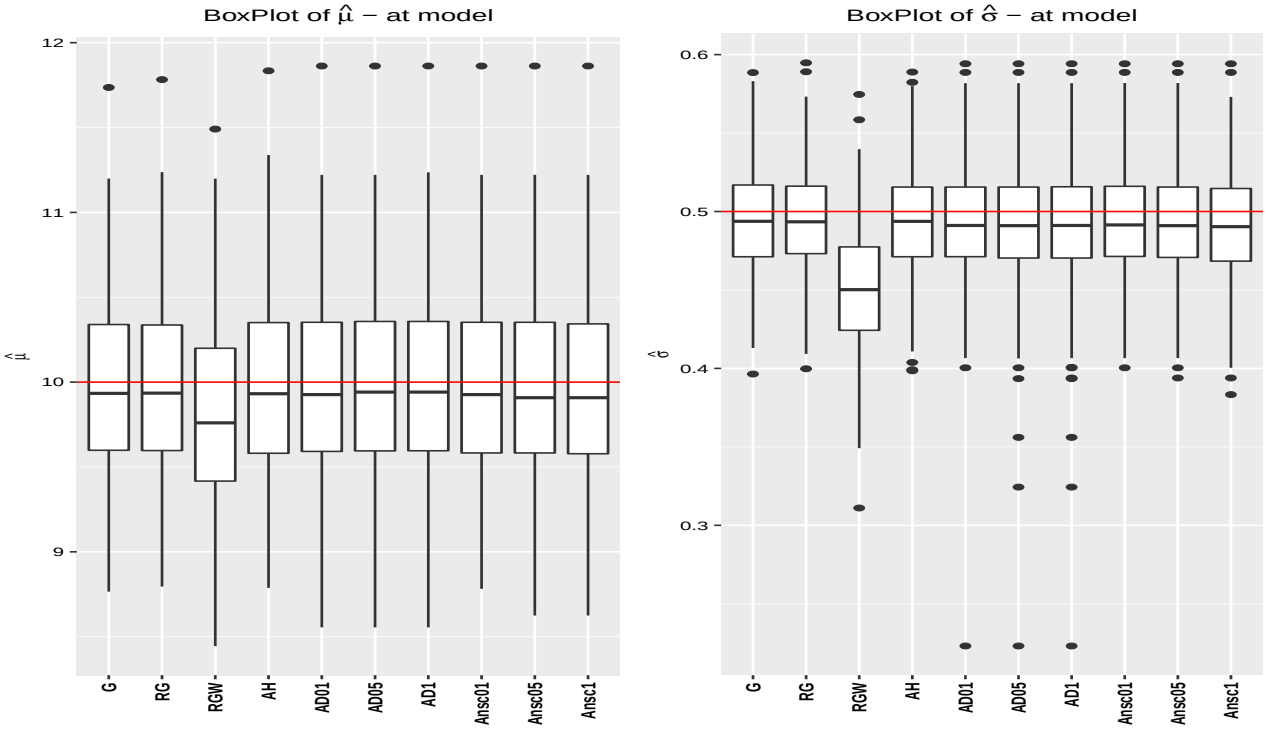
G - non robust fitting for GAMLSS, RG - robust fitting for GAMLSS by Rigby et al. (2019), RGW - robust fitting for GAMLSS modified, AH - Aeberhar method, AD01 - truncation robust fitting for GAMLSS with Anderson Darlin test and significance level 1%, AD05 - truncation robust fitting for GAMLSS with Anderson Darlin test and significance level 5%, AD1 - truncation robust fitting for GAMLSS with Anderson Darlin test with significance level 10%, Ansc01 - truncation robust fitting for GAMLSS with Anscombe test and significance level 1%, Ansc05 - truncation robust fitting for GAMLSS with Anscombe test and significance level 5%, Ansc1 - truncation robust fitting for GAMLSS with Anscombe test and significance level 10%

Figure 33 – Histogram of a random sample and probability density function of a GAMLSS model based on a Gama distribution with parameters $\mu = 10$ and $\sigma = 0.5$.



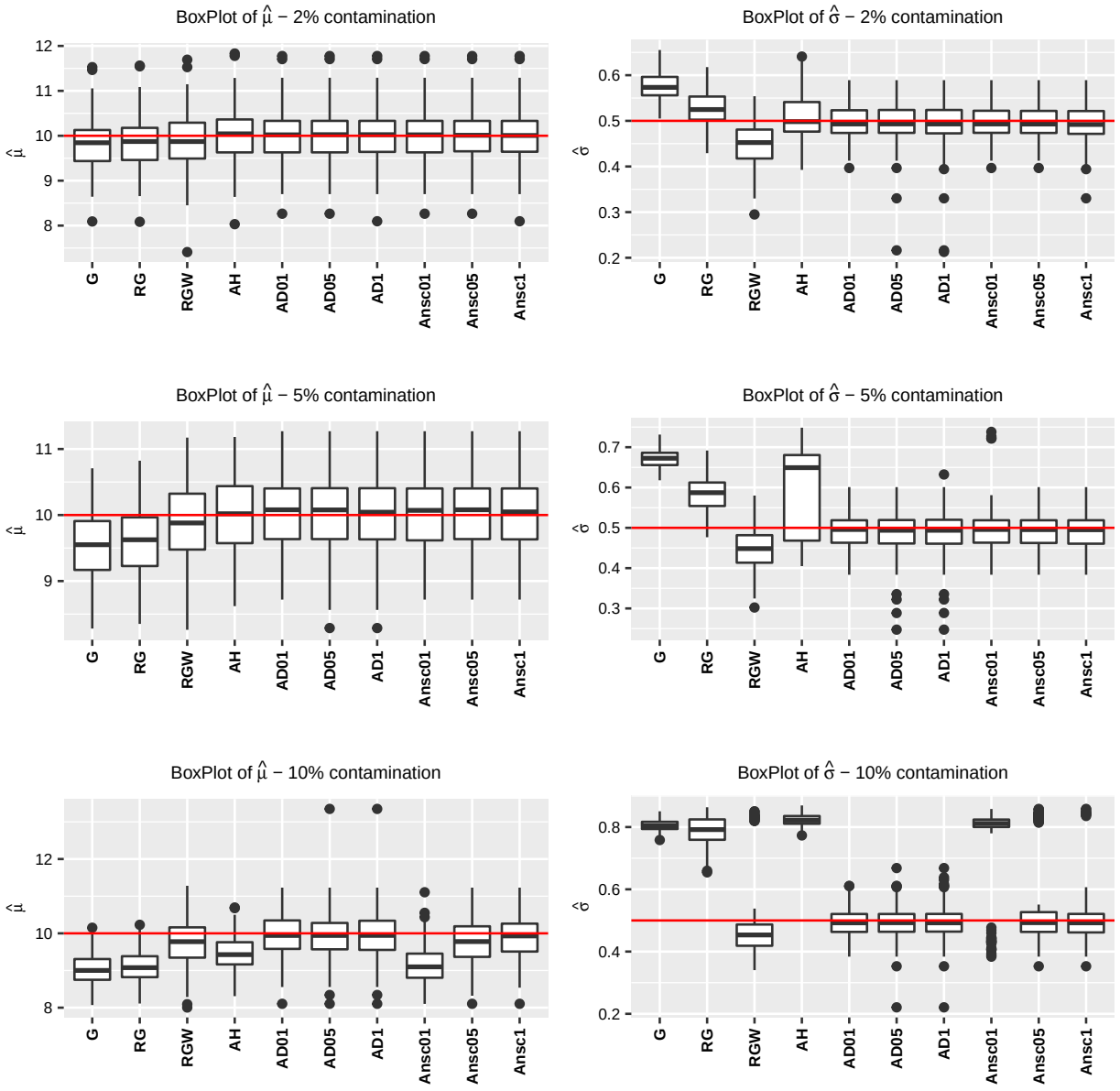
Source: The author (2021)

Figure 34 – Boxplots of $\hat{\mu}$ and $\hat{\sigma}$ simulated at model (without contamination), based on $GA(\mu,\sigma)$ without covariates in systematic component and sample size 100. The red line is the parameter value ($\mu = 10$ and $\sigma = 0.5$).



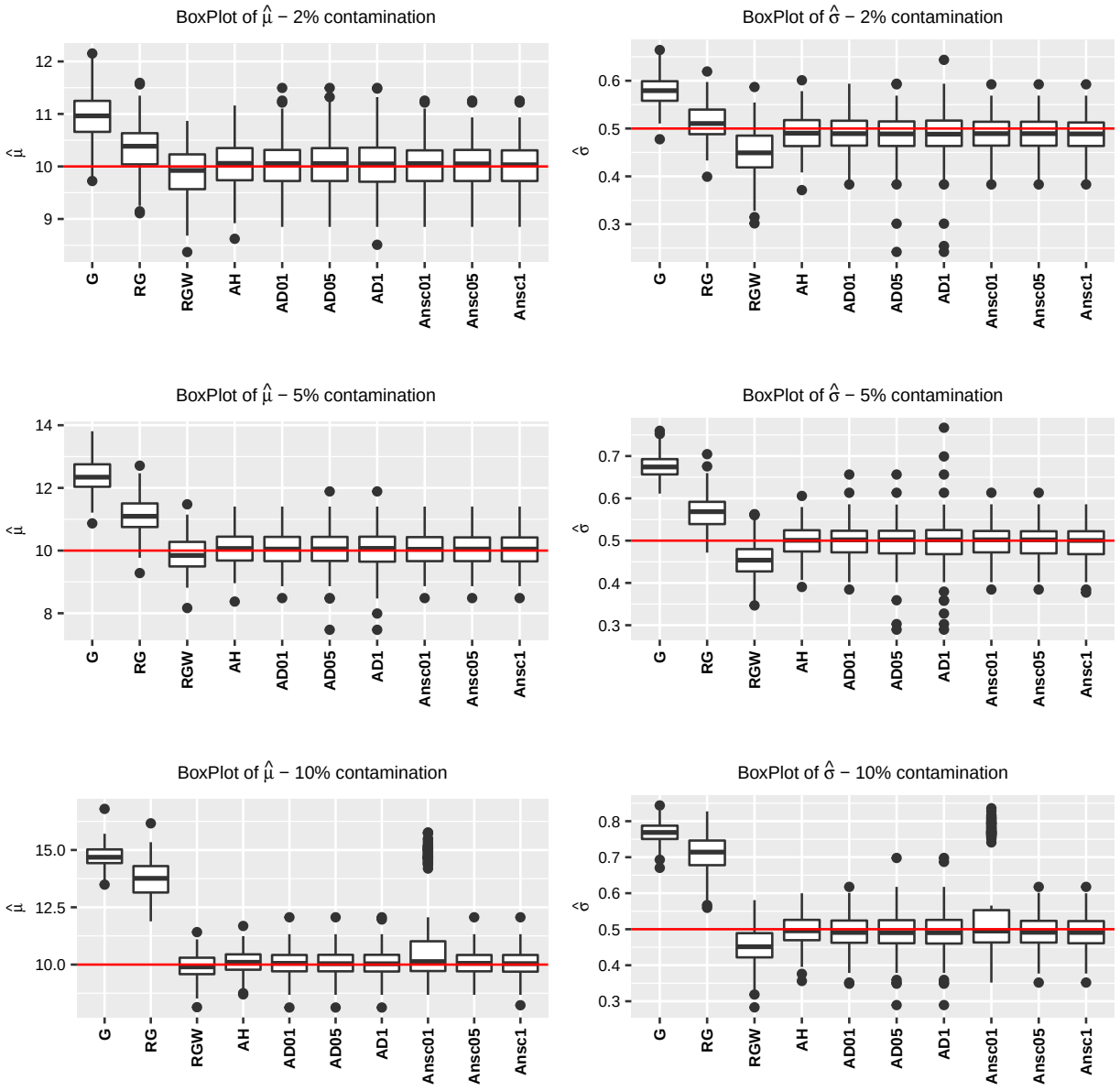
Source: The author (2021)

Figure 35 – Boxplots of $\hat{\mu}$ (first column of the figure) and $\hat{\sigma}$ (second column of the figure) simulated under left contamination (2% - first line of the figure , 5% - second line of the figure and 10% - third line of the figure), based on $GA(\mu, \sigma)$ without covariates in systematic component and sample size 100. The red line is the parameter value ($\mu = 10$ and $\sigma = 0.5$).



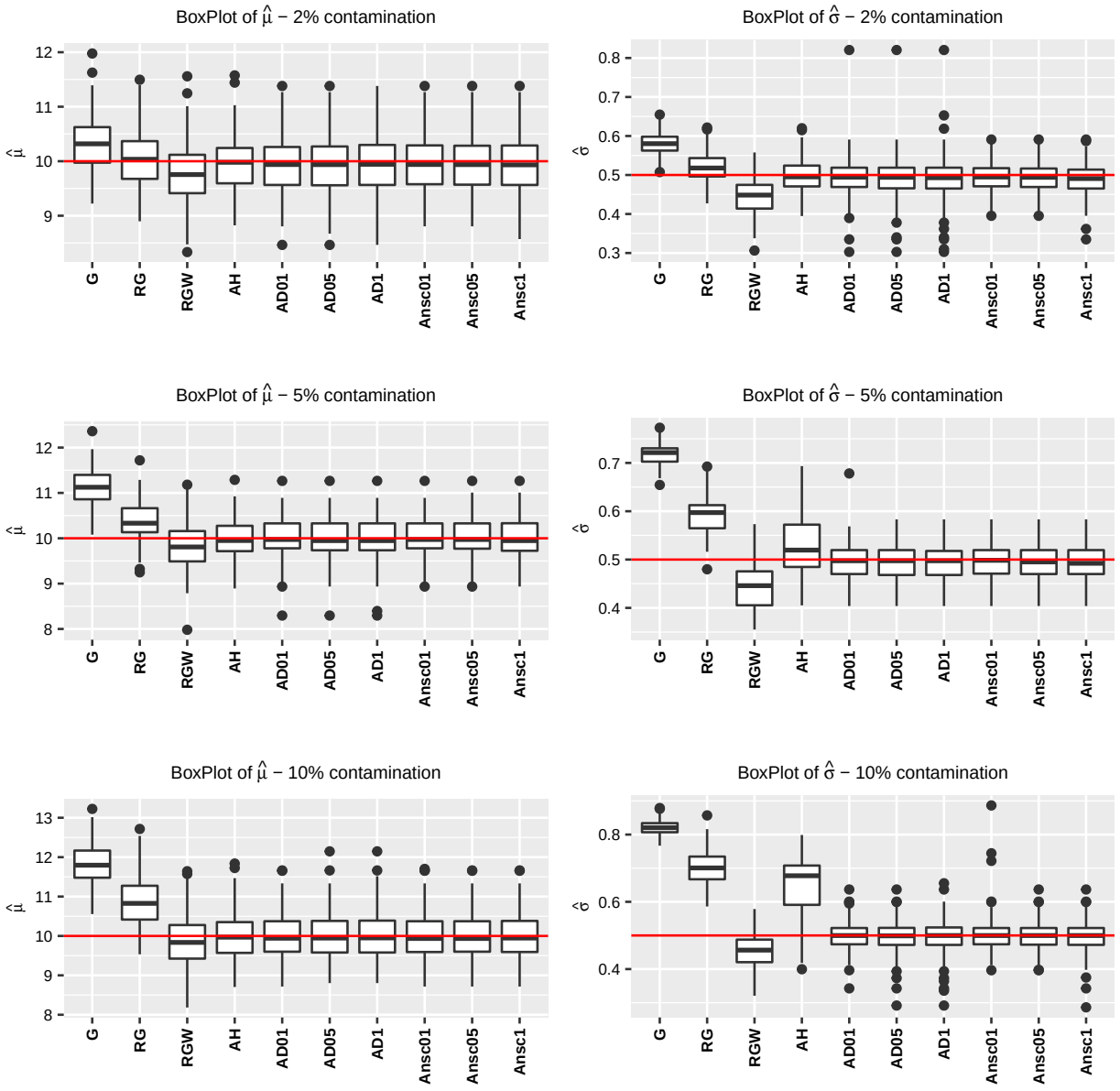
Source: The author (2021)

Figure 36 – Boxplots of $\hat{\mu}$ (first column of the figure) and $\hat{\sigma}$ (second column of the figure) simulated under right contamination (2% - first line of the figure , 5% - second line of the figure and 10% - third line of the figure), based on $GA(\mu, \sigma)$ without covariates in systematic component and sample size 100. The red line is the parameter value ($\mu = 10$) and $\sigma = 0.5$).



Source: The author (2021)

Figure 37 – Boxplots of $\hat{\mu}$ (first column of the figure) and $\hat{\sigma}$ (second column of the figure) simulated under left and right contamination (2% - first line of the figure , 5% - second line of the figure and 10% - third line of the figure), based on $GA(\mu, \sigma)$ without covariates in systematic component and sample size 100. The red line is the parameter value ($\mu = 10$ and $\sigma = 0.5$).



Source: The author (2021)

4.5.2.2 Parametric GAMLSS Based on Gamma Distribution with Covariates in Systematic Components For μ and σ.

In this subsection, was conduct a specific study based on $Y_i \sim Gamma(\mu_i, \sigma_i)$ with systematic components defined as

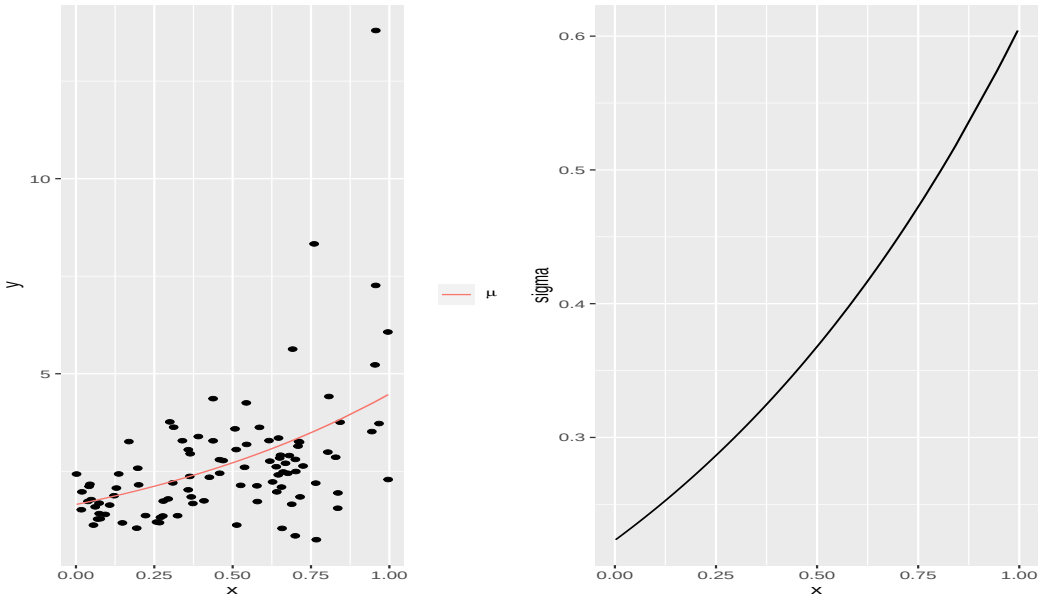
$$\eta_1 = \log(\mu_i) = \beta_{11} + \beta_{21} X_i$$

and

$$\eta_2 = \log(\sigma_i) = B_{12} + B_{22} X_i,$$

where $i = 1, \ldots, n$, $\beta_{11} = 0.5$, $\beta_{21} = 1$, $\beta_{12} = -1.5$, $\beta_{22} = 1$ and $X_i \sim Uniform(0,1)$ is a fixed covariate. The simulations studies was based on 200 replicates, from the gamma distribution with sample sizes 100 and the contamination was carried out in 2%, 5% and 10% of the observations. Figure 38 shows the μ and σ parameters used in the simulations of GAMLLSS model under $GA(\mu, \sigma)$ with linear systematic components and a scatter plot of a random sample generated based on the proposed model without contamination.

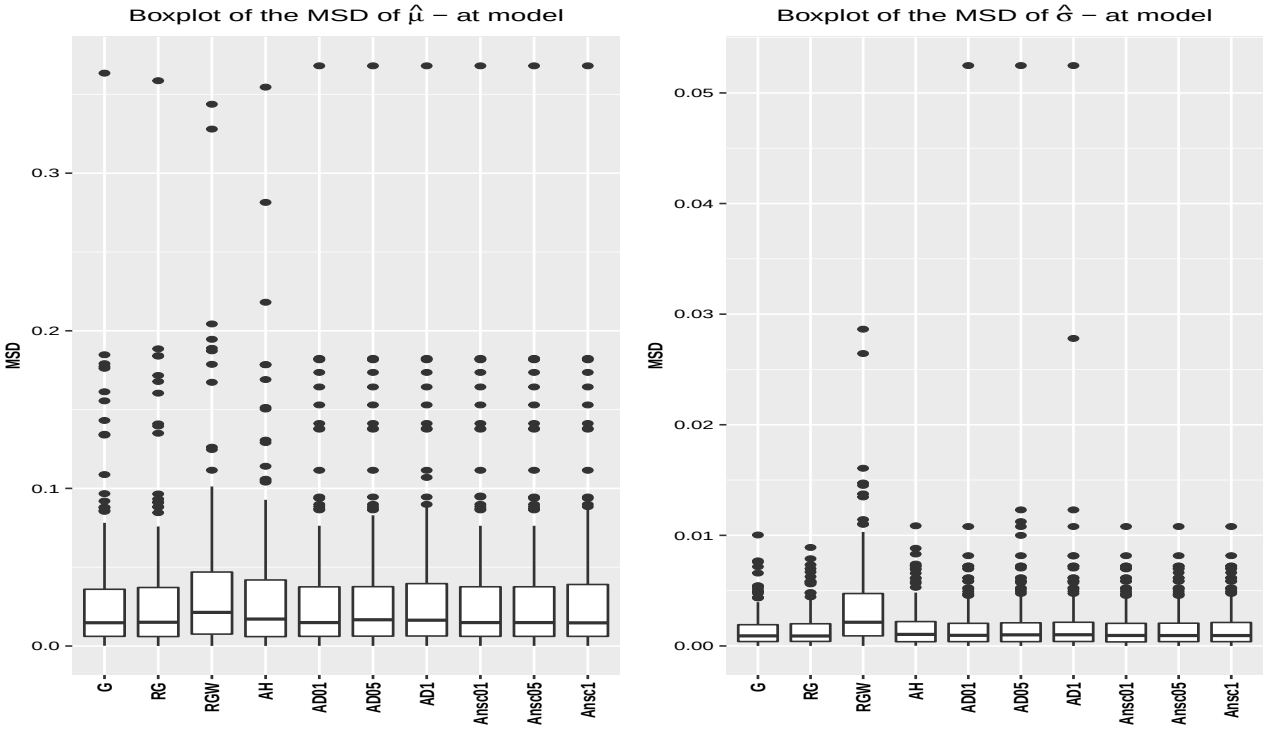
Figure 38 – μ and σ parameters used in the simulations of GAMLLSS model under $GA(\mu, \sigma)$ with linear systematic components and a scatter plot of a random sample generated based on the proposed model without contamination.



Source: The author (2021)

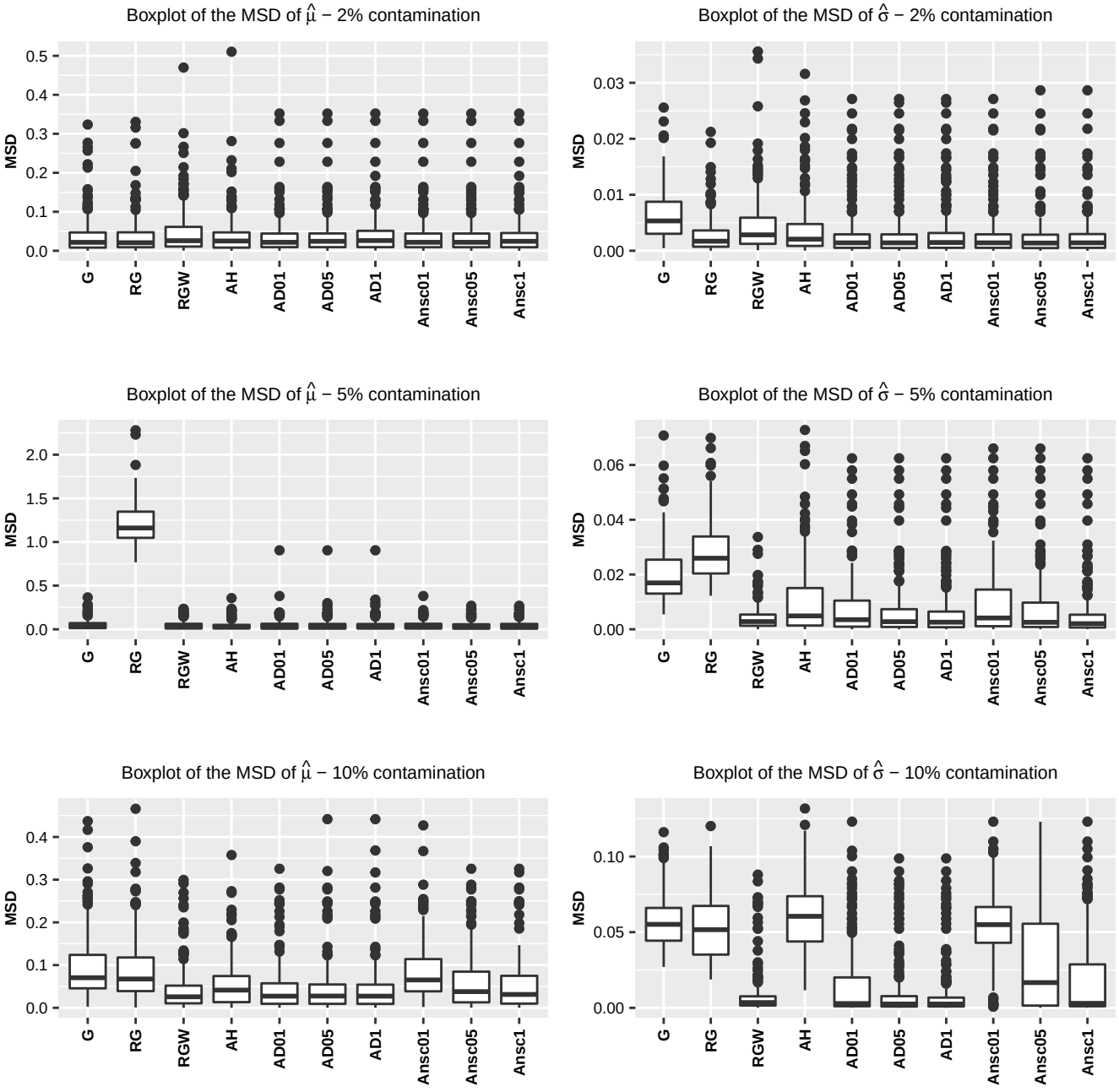
Figure 39 displays boxplots of the MSD of $\hat{\mu}$ (left column of figure) and $\hat{\sigma}$ (left column of figure) for all methods simulated without contamination. Only the RGW estimations shows high variability of MSD for the σ estimates. The other methods shows similar values of MSD in comparison with the G method (non robust fitting for GAMLSS). Under left contamination (Figure 40), the μ estimations of G method it does not present severe distortions in its estimates due to the fact that the gamma distribution has an influence function limited on the left. The σ estimates shows the best results for the Truncation proposal based on Anderson Darling test. Under right contamination (Figure 41), the performance of MSDs of $\hat{\mu}$ are similar at level of 2% of contamination among the robust methods. The MSDs of $\hat{\sigma}$ has high variability. At levels 5% and 10% of contamination, the AH estimates are slightly performance. All variations of truncation robust shows extreme values. Under left and right contamination (Figure 42), the estimations shows good performance at level 2% and 5% with similar results. At level 10% the RGW estimation shows slightly better performance for both parameters.

Figure 39 – Boxplots of the MSD of $\hat{\mu}$ and $\hat{\sigma}$, simulated at model (without contamination), based on gamma model with covariates in systematic component and sample size 100.



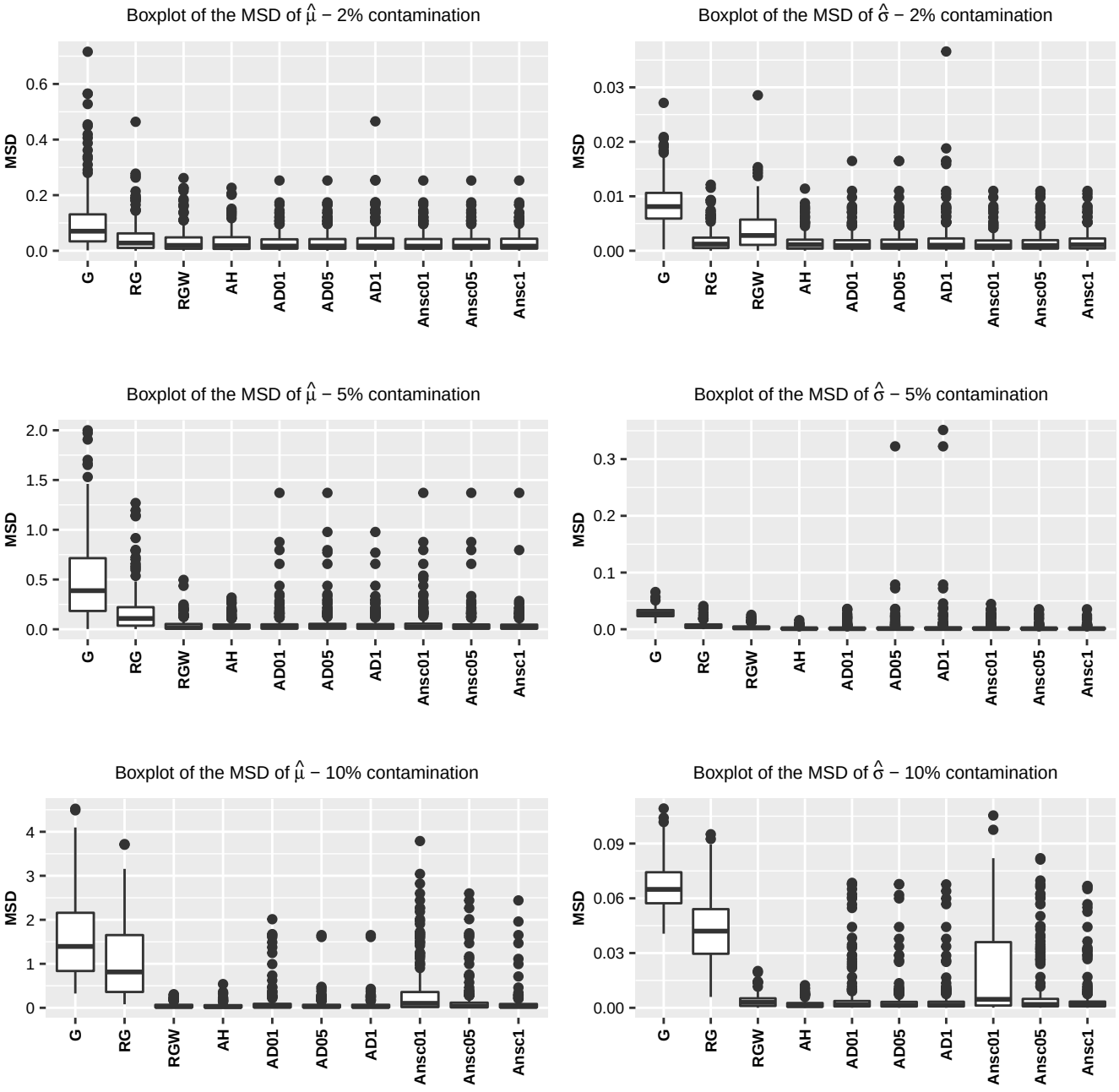
Source: The author (2021)

Figure 40 – Boxplots of the MSD of $\hat{\mu}$ (first column of the figure) and $\hat{\sigma}$ (second column of the figure), simulated under left contamination (2% - first line of the figure , 5% - second line of the figure and 10% - third line of the figure) and based on gamma model with covariates in systematic component and sample size 100.



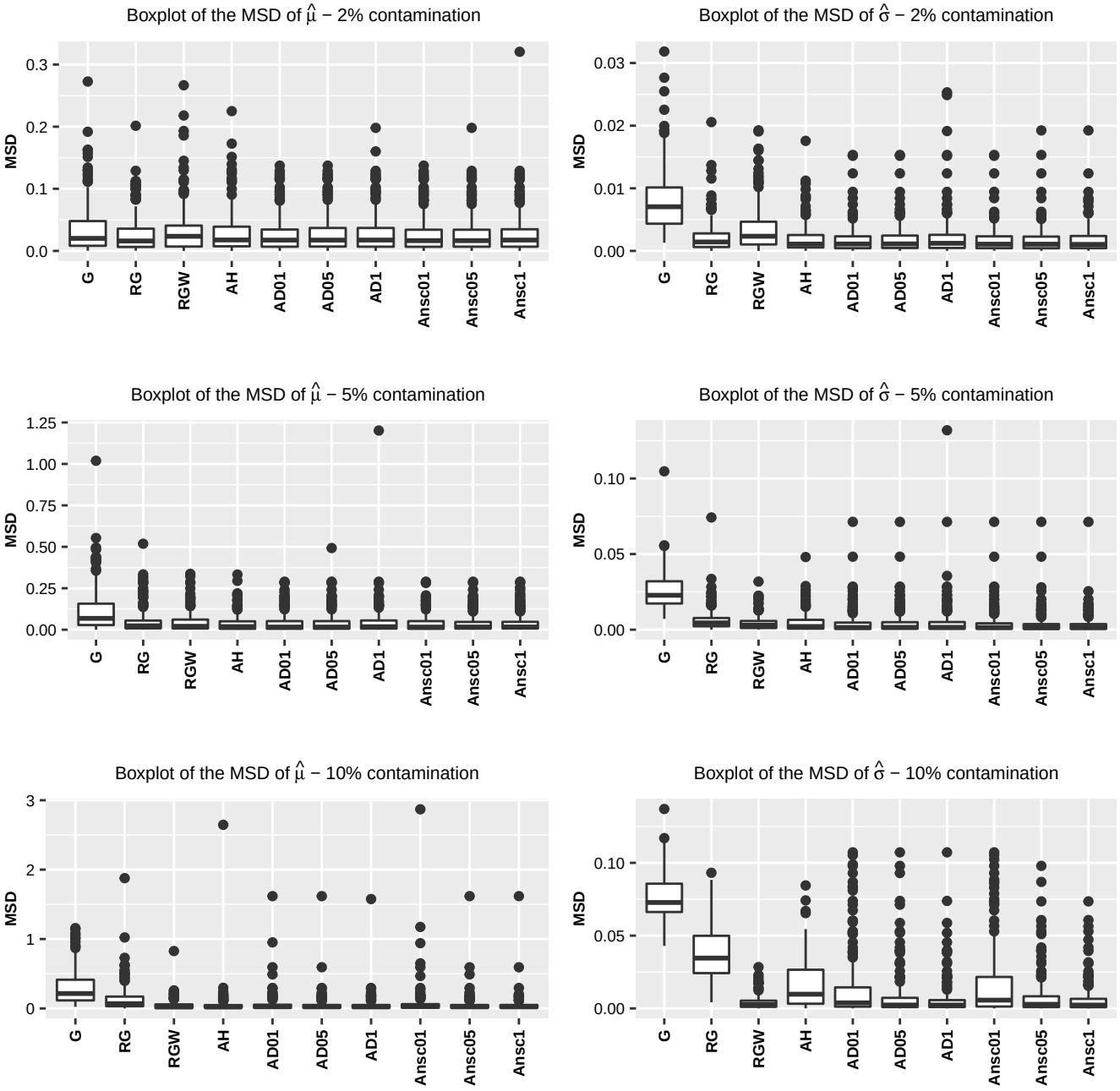
Source: The author (2021)

Figure 41 – Boxplots of the MSD of $\hat{\mu}$ (first column of the figure) and $\hat{\sigma}$ (second column of the figure), simulated under right contamination (2% - first line of the figure , 5% - second line of the figure and 10% - third line of the figure) and based on gamma model with linear systematic component and sample size 100.



Source: The author (2021)

Figure 42 – Boxplots of the MSD of $\hat{\mu}$ (first column of the figure) and $\hat{\sigma}$ (second column of the figure), simulated under left and right contamination (2% - first line of the figure , 5% - second line of the figure and 10% - third line of the figure) and based on Gama model with liner systematic component and sample size 100.



Source: The author (2021)

4.5.2.3 Non Parametric GAMLSS based on Gamma Model Using NonParametric Systematic Component with P-splines.

The non parametric gamma model were evaluate based on the systematic components defined as

$$\log(\mu) = \eta_1 = s_1(X)$$

and

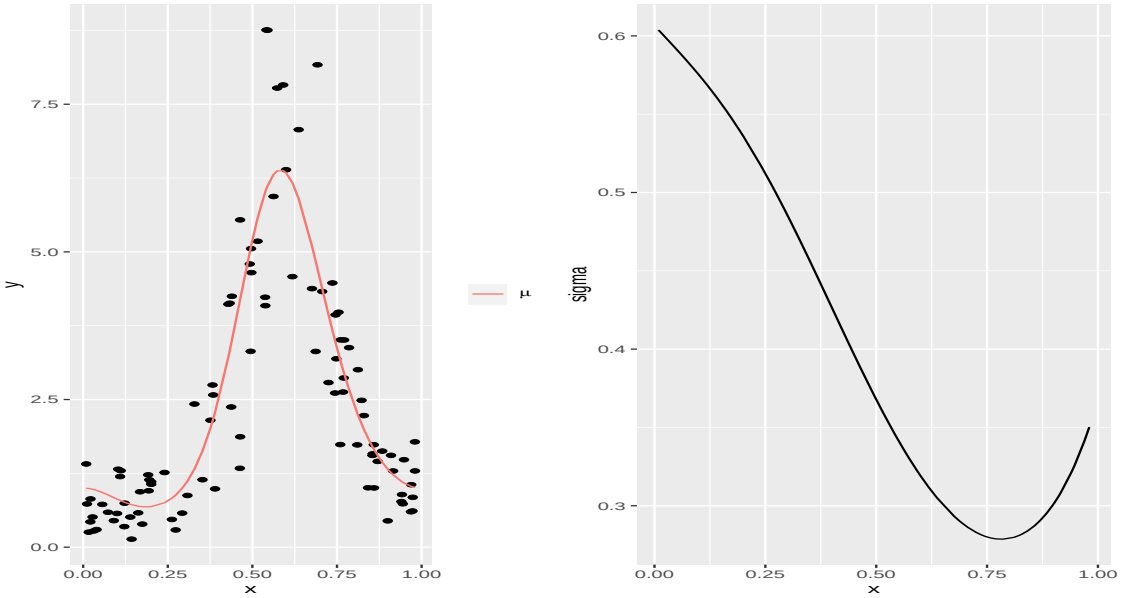
$$\log(\sigma) = \eta_2 = s_2(X).$$

The smooth functions are

$$s_1(X) = 2Xe^X \sin(2\pi(1-X)^2),$$
$$s_2(X) = -(X+1)^1/2 - X^2 \sin(X\pi),$$

where $X \sim U(0,1)$ is a fixed covariate. The simulations study was based on 200 replicates, from the gamma distribution with sample sizes 100 and the contamination was carried out in 2%, 5% and 10% of the observations. Figure 43 shows the μ and σ parameters used in the simulations of GAMLLSS model under gamma distribution with non parametric systematic components and a scatter plot of a random sample generated based on the proposed model without contamination.

Figure 43 – μ and σ parameters used in the simulations of GAMLLSS model under gamma distribution with non parametric systematic components and a scatter plot of a random sample generated based on the proposed model without contamination.

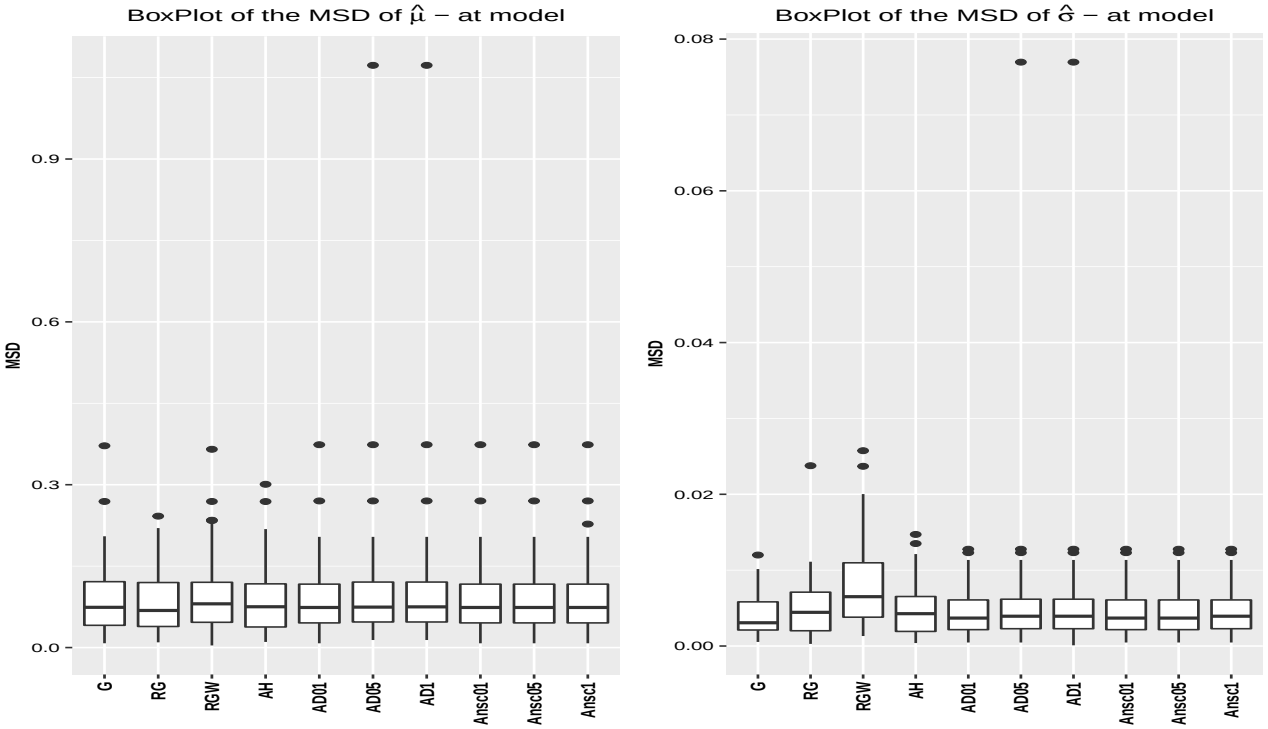


Source: The author (2021)

Figure 44 displays boxplots of the MSD of $\hat{\mu}$ and $\hat{\sigma}$ for all methods. The boxplots of the MSDs of μ performs similarly in comparison with the G method. Only the MSDs of RGW estimation shows larger variability. Under left contamination (Figure 45), the MSDs of $\hat{\mu}$ of robust methods has similar performance with 2% contamination, while for the MSD of $\hat{\sigma}$ the RG and truncation estimation has a slight better performance. At level 5% the MSDs for the $\hat{\sigma}$ for RGW has a slight better performance with competitive results for AD01 and Ansc1. The AH estimations has the largest MSDs of $\hat{\sigma}$. With 10% of contamination the boxplots of the MSD of $\hat{\mu}$ has the RGW and Ansc1 with the best performances, while the MSD of $\hat{\sigma}$ the best performance are RGW with competitive results for AD05 and AD1 methods. The AH estimation has poor performance with the larger MSD. Under right contamination (Figure 46), the AH estimations shows the best performance. Finally, Figure 47 displays boxplots of the MSDs of $\hat{\mu}$ and $\hat{\sigma}$ under left and right contamination. At level 2% the boxplot

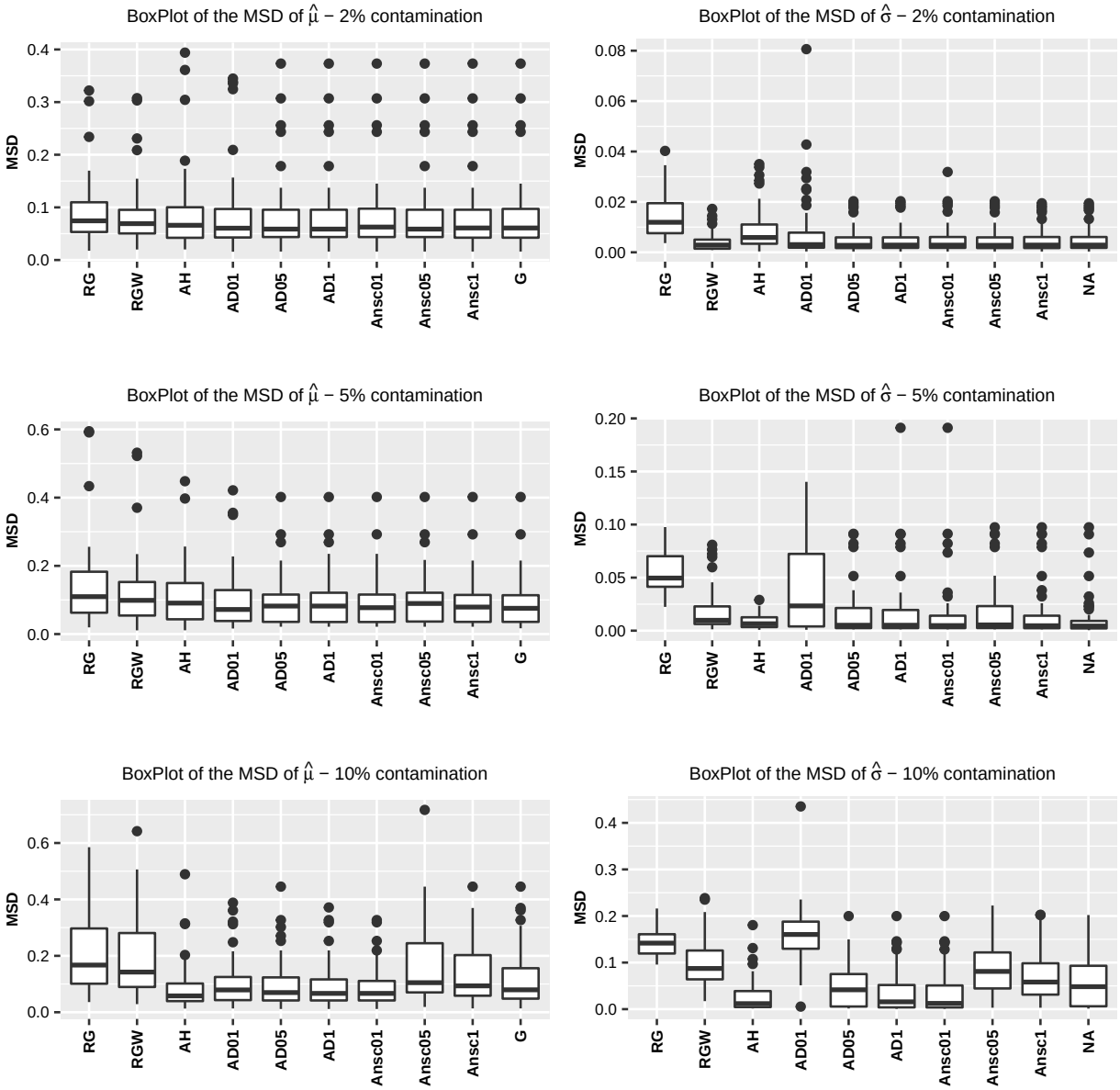
of MSDs of $\hat{\mu}$ and $\hat{\sigma}$ results are similar and the AH method shows greater variability. At level 5% the boxplot of the MSD of $\hat{\mu}$ and $\hat{\sigma}$ for the AH method has the best performance. Lastly, at the levels of 10% the results indicate that MSD of $\hat{\mu}$ of AH method has the lowest values, with competitive results for AD1 method. For the MSD of $\hat{\sigma}$ the AH method shows the smallest values.

Figure 44 – Boxplots of the MSD of $\hat{\mu}$ and $\hat{\sigma}$, simulated at model (without contamination), based on $GA(\mu, \sigma)$ with non parametric systematic component and sample size 100.



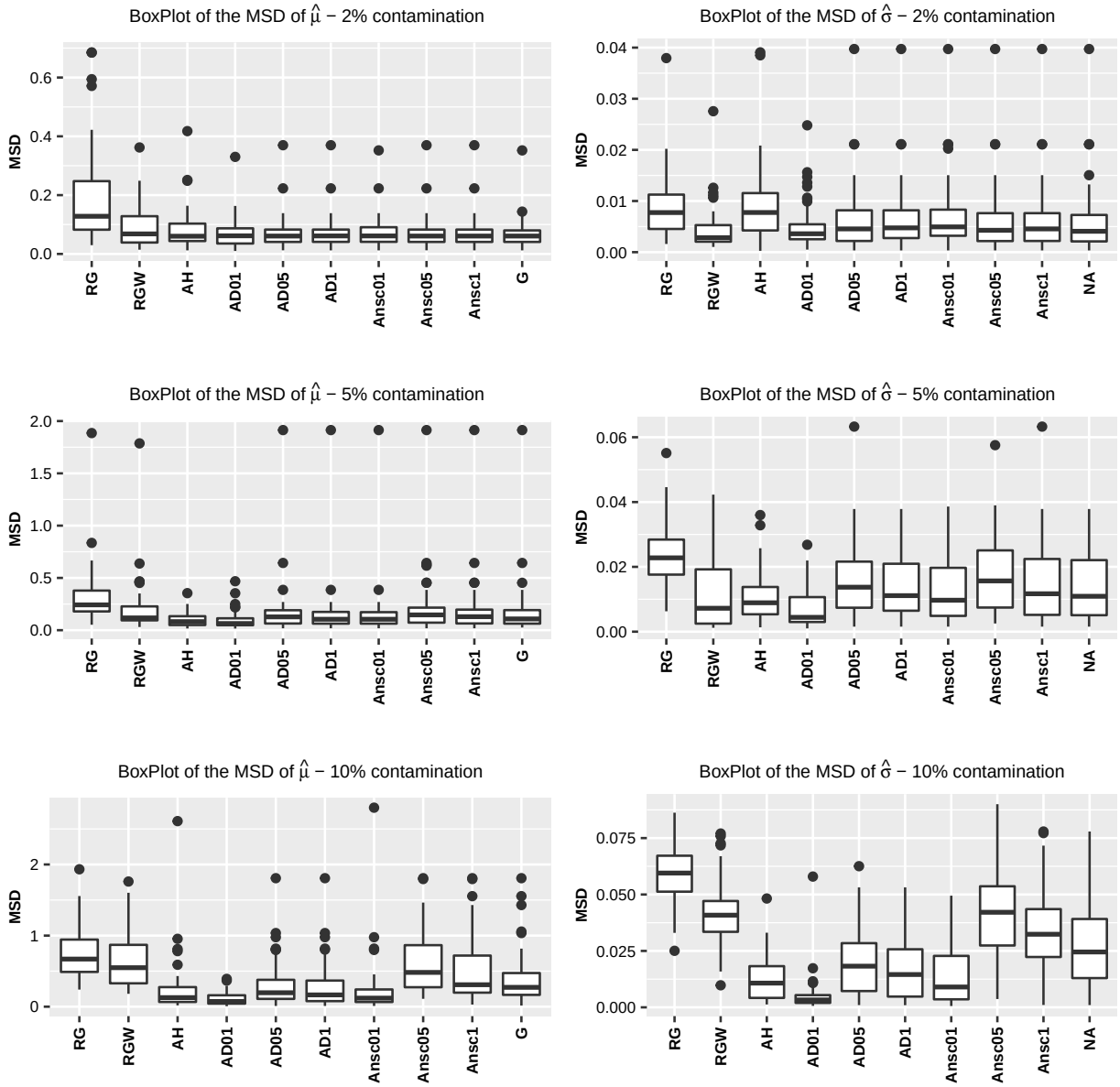
Source: The author (2021)

Figure 45 – Boxplots of the MSD of $\hat{\mu}$ (first column of the figure) and $\hat{\sigma}$ (second column of the figure), simulated under left contamination (2% - first line of the figure , 5% - second line of the figure and 10% - third line of the figure), based on $GA(\mu, \sigma)$ with non parametric systematic component and sample size 100.



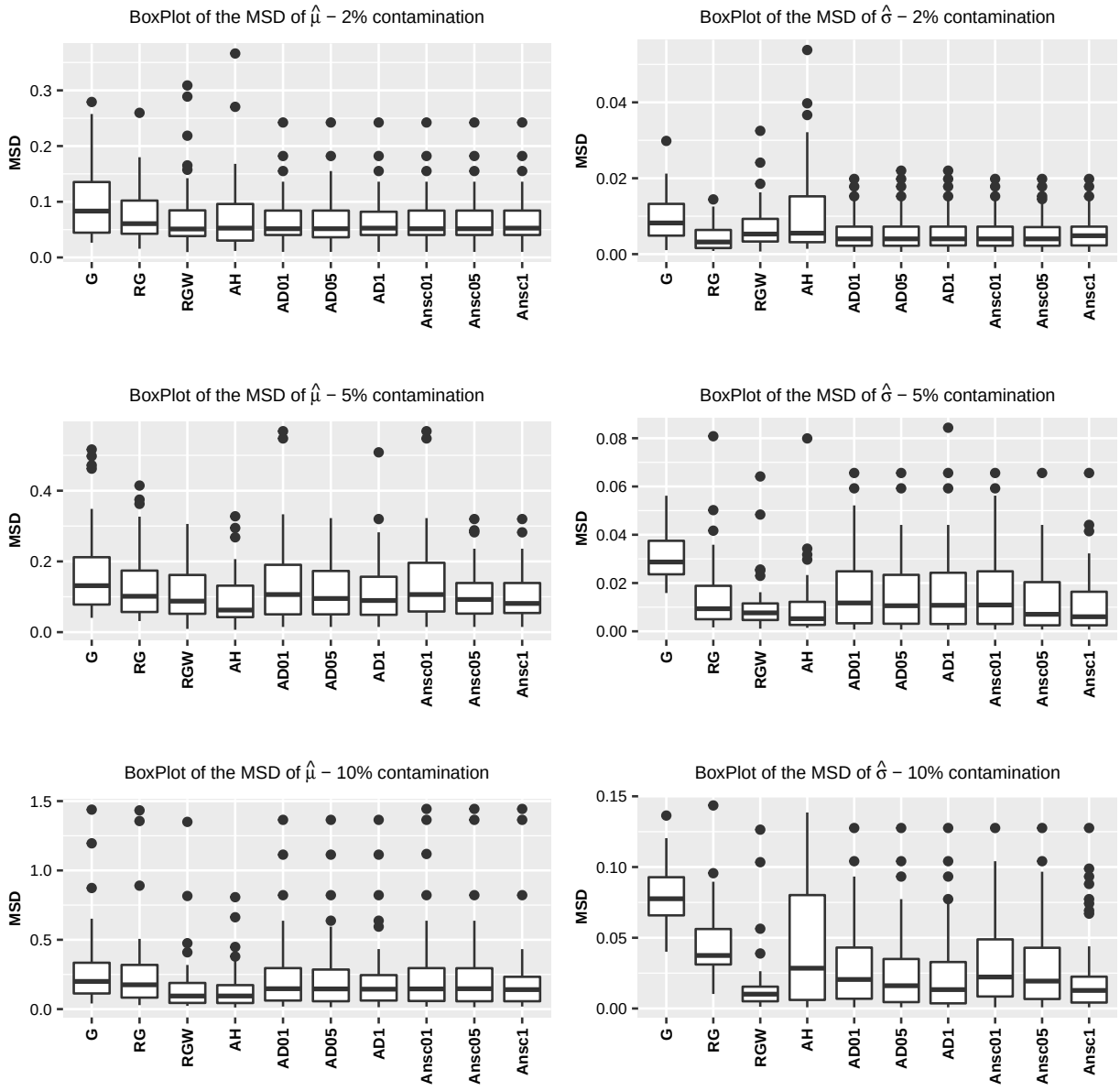
Source: The author (2021)

Figure 46 – Boxplots of the MSD of $\hat{\mu}$ (first column of the figure) and $\hat{\sigma}$ (second column of the figure), simulated under left contamination (2% - first line of the figure , 5% - second line of the figure and 10% - third line of the figure), based on $GA(\mu, \sigma)$ with non parametric systematic component and sample size 100.



Source: The author (2021) x

Figure 47 – Boxplots of the MSD of $\hat{\mu}$ (first column of the figure) and $\hat{\sigma}$ (second column of the figure), simulated under left and right contamination (2% - first line of the figure , 5% - second line of the figure and 10% - third line of the figure), based on $GA(\mu, \sigma)$ with non parametric systematic component and sample size 100.



Source: The author (2021)

4.6 APPLICATIONS

In order, to illustrate the advantages from the use of our proposal of robust fitting for GAMLSS, based on truncation distribution, we consider two real data sets. The first investigate the vulnerability of child to poverty in Ceara state, Brazil. The second study investigated the outbreak of H1N1 flu by looking at the weekly counts of influenza-like-illness (ILI) doctor visits in United States. Influenza-like-illness is a medical diagnosis of possible influenza or other illness causing a set of common symptoms. These include fever, shivering, chills, malaise, dry cough, loss of appetite, body aches, and nausea, typically in connection with a sudden onset of illness.

4.6.1 Extreme Child Poverty

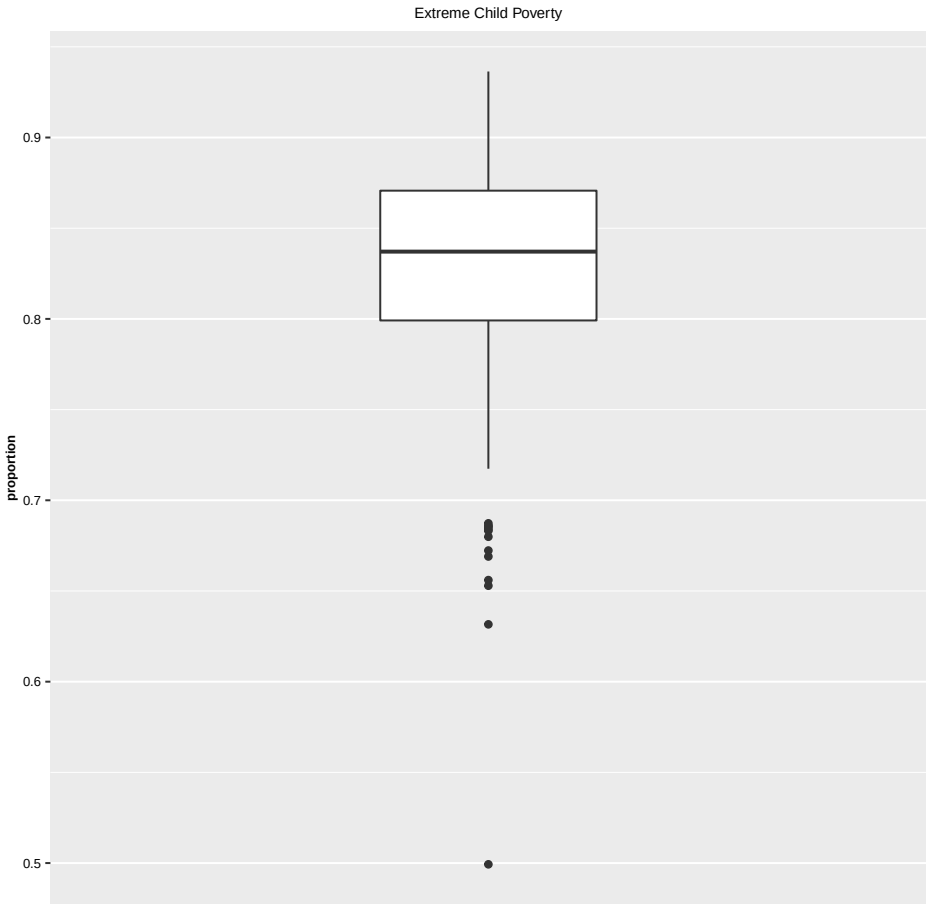
Brazil is historically characterized by having a high number of individuals in a state of extreme poverty. According to a ONU study (2010), five Brazilian cities are among the twenty most unequal in the world, with the largest income differences between rich and poor in the country and with some of the notable regions in the world for presenting pockets of poverty. In this context, it is essential to know and understand the behavior of the levels of poverty, in order to identify policy proposals to combat poverty, that can revert their levels more quickly.

In this subsection, we consider the proportion of children (0–14 year olds) vulnerable to poverty, in the municipalities of the state of Ceará in Brazil, in 2010. The data can be obtained from IBGE (2012). We are interested in modeling the proportion of children vulnerable to poverty - PCVP. Here, a child is considered vulnerable to poverty if the per capita household income is at most BRL 255, in 2010. The PCVP data set comprises 184 observations. The boxplot of the proportion of children vulnerable to poverty (Figure 48) identified thirteen extreme observations. The PCPV data has values in the range 0.4993 and 0.9364, with a mean of 0.8266 and standard deviation of 0.066.

The PCVP data is modeled here using a beta distribution, due to the data are restricted over some finite interval, see Gupta and Nadarajah (2004). The method presented here are: truncation robust fitting for GAMLSS, based on Anderson Darling test with significance level 1% - AD, Ansccomb test with significance level 1% - Ansc, Aeberhard method - AH, non robust fitting for GAMLSS - G, robust fitting for GAMLSS - RG and weight robust fitting for GAMLSS - RGW. The results of truncation robust fitting for GAMLSS based on the anderson darling test with a significance level of 5% and 10% are the same for level 1%, as well as, the results of Anscome with a significance level of 5% and 10% are the same for level 1%.

The final result on the estimation is presented in Table 11. This table shows the values of statistics for model comparison in order to evaluate the ability of methods to fit the data. According to this table, the AD01 and RGW estimation have the smallest confidence interval amplitudes and the lowest values of AIC. Figure 49 presents the residual analysis for models. This figure allow shows that the distribution of the residuals for RGW and AD01 is not far from the normal distribution, which indicates that this model is appropriate for the PCVP data. The estimated and the observed histograms of the PCVP data are presented in Figure 50 for all methods, which confirms that the AD01 and RGW estimations provides a better fit for these data, with a slightly better performance for the AD01 estimations. Assuming the AD01 beta model as a final model, we can see in Table 11 that the median of proportion of children vulnerable to poverty is close to 1 with low dispersion.

Figure 48 – Boxplot of the proportion of children vulnerable to poverty - PCVP in Ceará, Brazil in 2010.



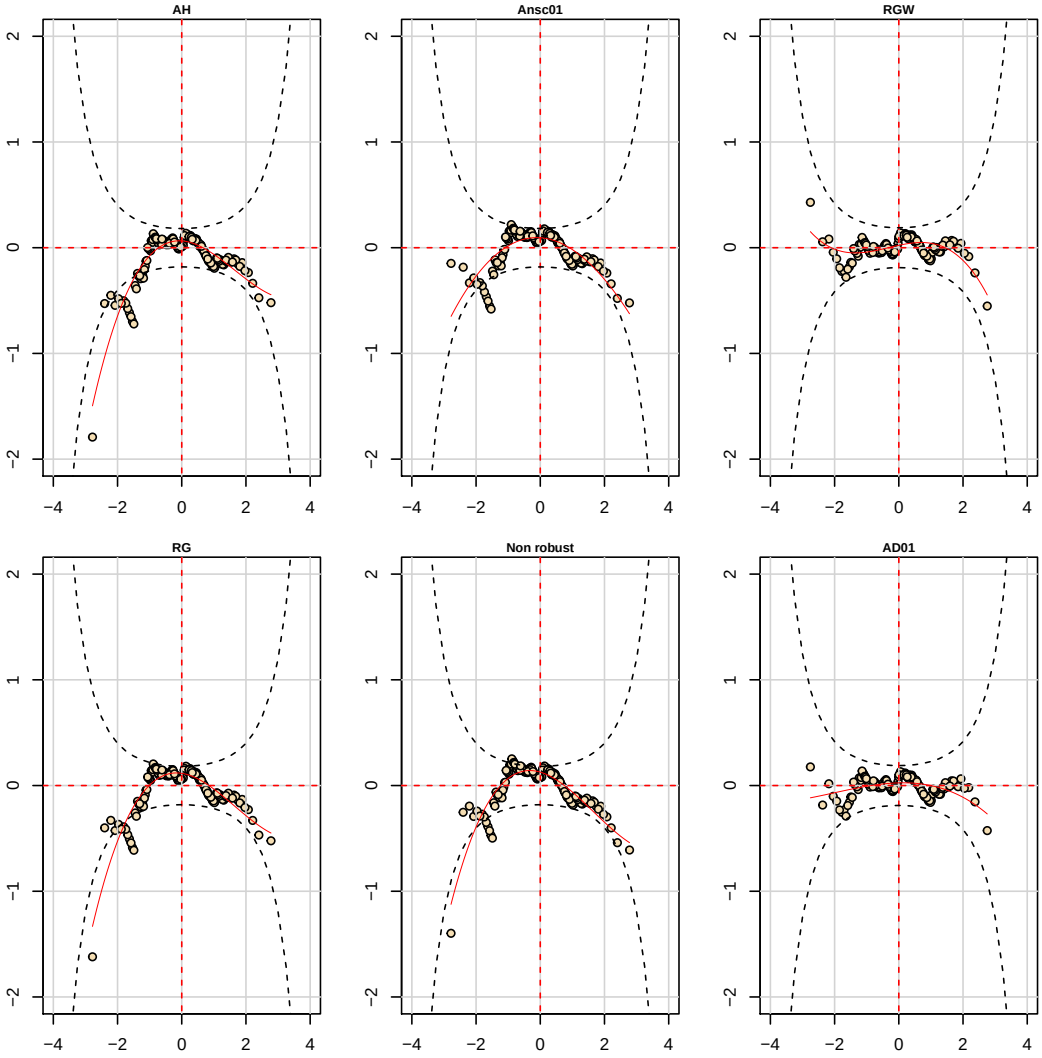
Source: The author (2021)

Table 11 – Estimates and 95% Wald confidence intervals for the parameters of the beta models and different methods.

Model	$\hat{\mu}$	$\hat{\sigma}$	AIC
AD01	0.84 (0.831, 0.845)	0.13 (0.120, 0.148)	-572.28
Ansc01	0.83 (0.819, 0.836)	0.16 (0.142, 0.173)	-524.85
AH	0.83 (0.821, 0.840)	0.15 (0.133, 0.173)	-107.67
RGW	0.84 (0.832, 0.846)	0.12 (0,113, 0.139)	-566.88
RG	0.83 (0.819, 0.836)	0.16 (0.142, 0.171)	-507.93
G	0.83 (0.817, 0.835)	0.16 (0.148, 0.180)	-508.84

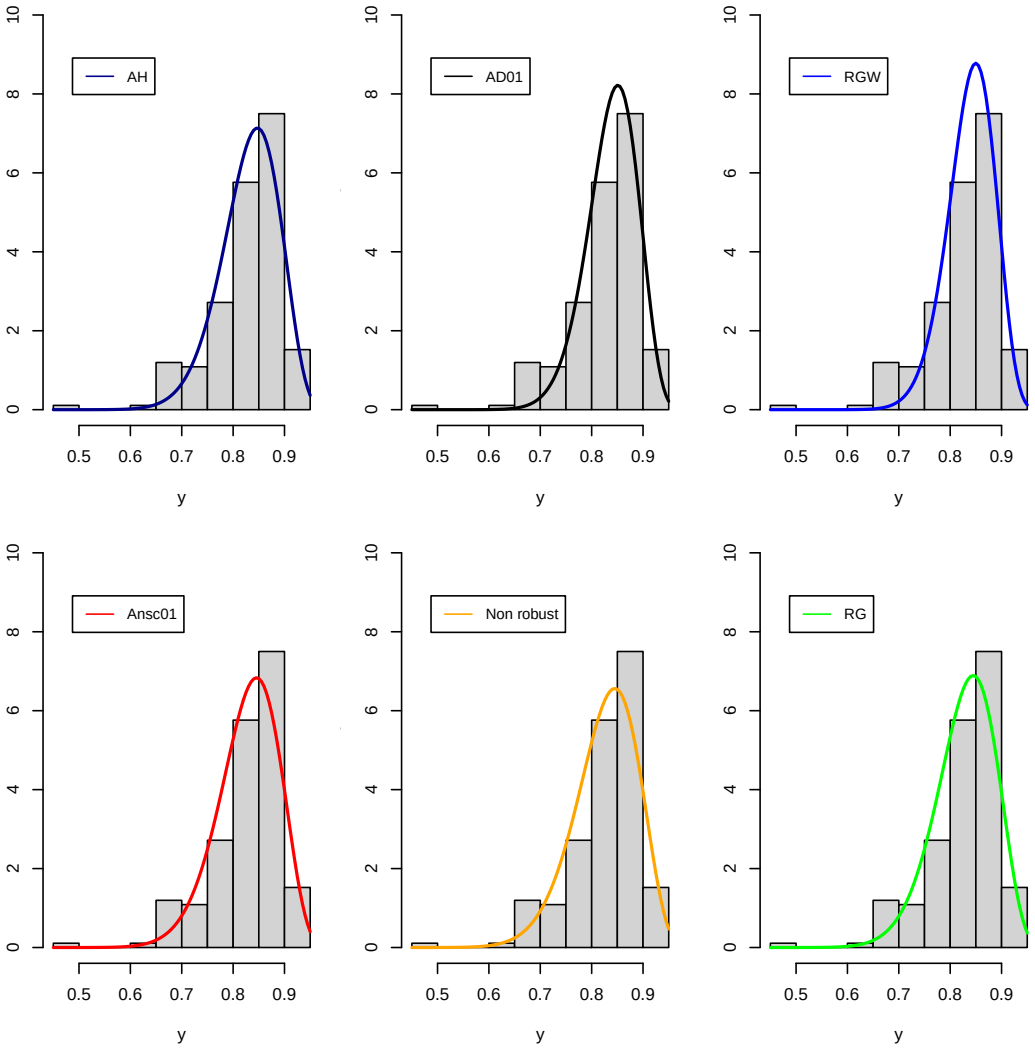
Source: The author (2021)

Figure 49 – Worm plot for residuals of models for truncation robust fitting for GAMLSS, based on Anderson Darling test with significance level 1% - AD, Ansccomb test with significance level 1% - Ansc, Aeberhard method - AH, non robust fitting for GAMLSS, robust fitting for GAMLSS - RG and weight robust fitting for GAMLSS - RGW.



Source: The author (2021)

Figure 50 – Estimated density and the observed histograms of the proportion of children vulnerable to poverty - PCVP data.

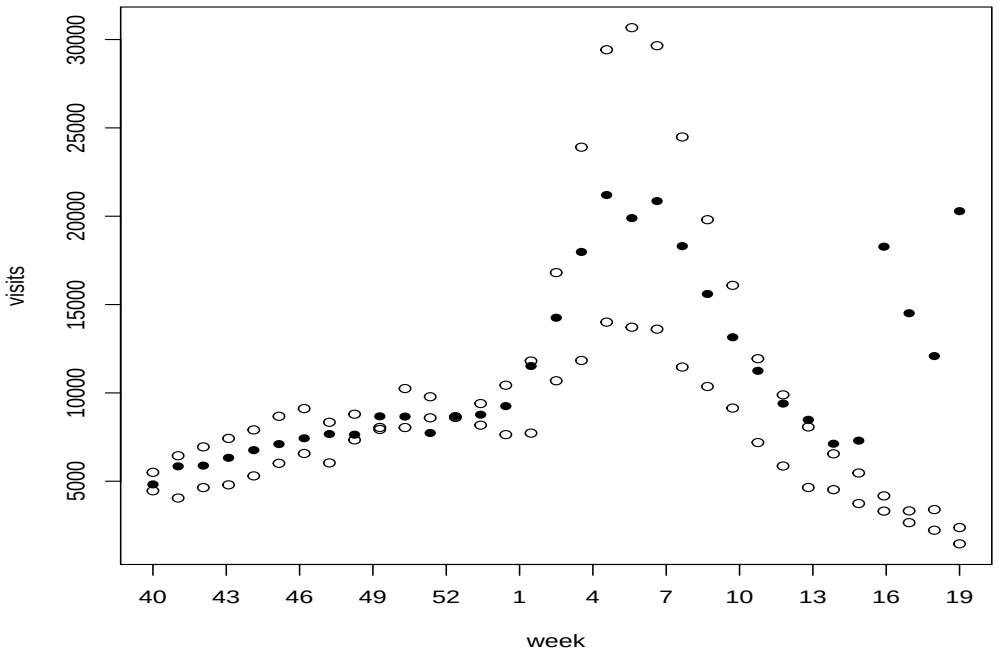


Source: The author (2021)

4.6.2 Influenza-Like Illness - ILI

Alimadad and Salibian-Barrera (2011) propose an outlier-robust fit for Generalized Additive Models with applications to disease outbreak detection. They illustrate the use of this approach on the detection of the outbreak of H1N1 flu by looking at the weekly counts of influenza-like-illness (ILI) doctor visits that was reported through the U.S. They consider data for the 2006-2007, 2007–2008, and 2008–2009 seasons and fit a GAM model with a logarithmic link function, but they do not report which probability distribution was used. The data is available on- line (US Centre for Disease Control, `http://www.cdc.gov/flu/weekly/fluactivity.htm`). Each seasons consists of weekly counts from week 40 through week 20 of the following calendar year. Figure 51 shows the weekly number of influenza-like-illness visits in the United States for the 2006–2008 flu seasons. A large number of cases were registered in weeks 17 to 20 indicated in the figure with solid circles. This behavior was due to the worldwide epidemic outbreak of H1N1 that started in the Spring of 2009. We applied the robust fitting for GAMLSS methods to fit the data based on the time (week number) as covariate and the penalized splines to create an approximating function that attempts to capture patterns in the data.

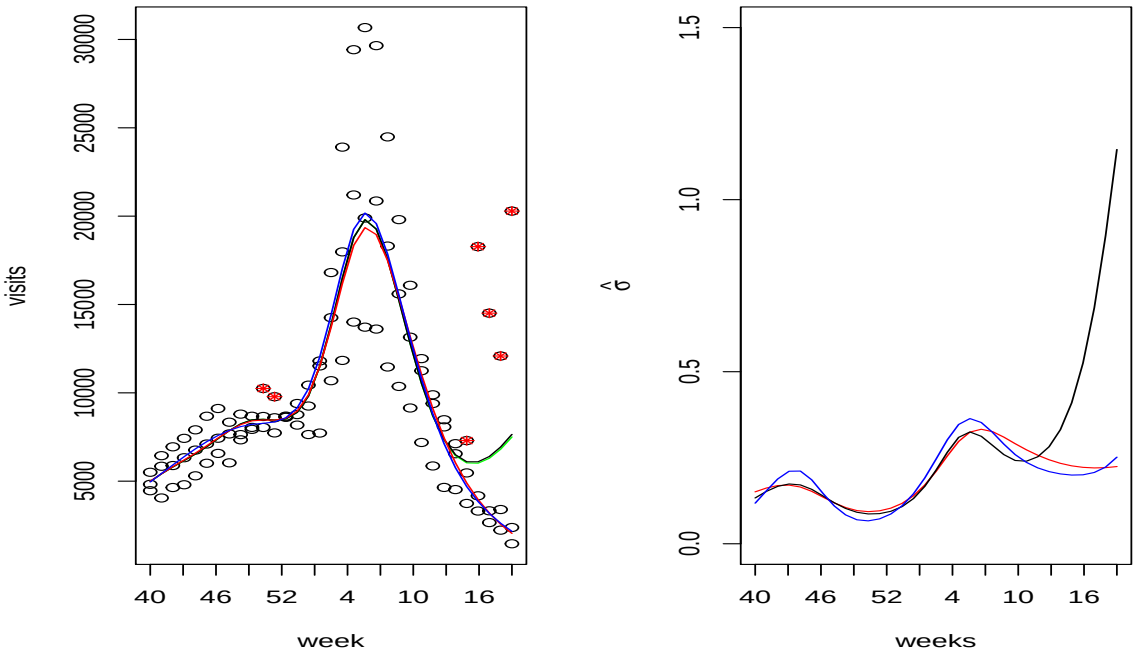
Figure 51 – Weekly counts of influenza-like-illness outpatient visits in the United considering data for the 2006 - 2008. The 2008 season is indicated with solid circles.



Source: The author (2021)

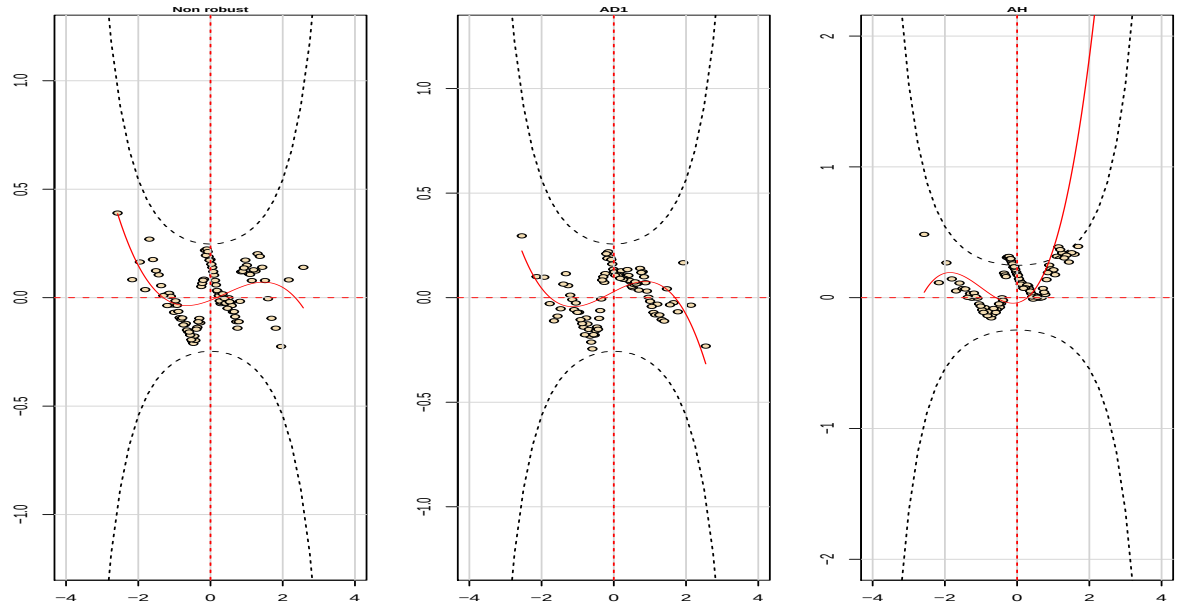
All methods evaluated in the previous sections were adjusted. The resulting fit is shown in Figure 51. It is possible to see that the robust fit is not affected by the atypical large counts of the last weeks of the 2008 season in μ estimates (left panel). The σ estimates shown the same pattern reducing the values of the estimates in the week with high numbers of visits. The red points indicated that the truncated robust was able to identify the outliers. Figure 53 features the worm plot of the residuals of the models. We can see that method AD1 obtained a slightly better fit, with all points within the confidence bands. Besides not presenting any standards that indicate lack of fit. Therefore, the truncated method presents the best results.

Figure 52 – Non robust fitting for GAMLSS, Aeberhard method, truncation robust fitting for GAMLSS fits to μ (left panel) and σ (right panel) to the weekly number of influenza-like-illness visits in the United States for the 2006 - 2008 flu seasons. In panel the black line denotes the non robust fit, the blue line corresponds to the truncation robust fitting for GAMLSS fit associated to Anderson Darling test significance level of 10%, the green line corresponds to the robust fitting for GAMLSS proposed by Rigby et al. (2019), while the Aebehard method fit is indicated with a red line. Right panel contains estimate of σ to each model.



Source: The author (2021)

Figure 53 – Worm plots for fitted models Truncated robust fitting for GAMLSS basend on Anderson Darling test with $\alpha = 10\%$, non robust fitting for GAMLSS - G and Aeberhard method - AH.



Source: The author (2021)

4.7 CONCLUSIONS

We introduced a robust estimation method for the class of GAMLSS, based on a new strategy of handling outliers. The use of truncated distributions in GAMLSS models as a strategy to deal with outliers, allows that all the computational and theoretical advantages and aspects of the theory of GAMLSS models to be used. Our implementation in software R is stable and quite general since it can be employed for any probability distribution. The distributions that are allowed for a GAMLSS are wide, they just need to be parametric. Currently, in the *gamlss* package Stasinopoulos and Rigby (2007) of R, there are 100 distributions implemented, including discrete, continuous and mixed. All distributions implement in the package can be truncated freely. It is also possible to implement a new distribution. In addition, we introduce a simple and effective selection criterion for tuning probabilities. We believe this criterion has broad applicability in the implementation of robust methods in many contexts, because its involves selecting the best truncation distribution for the response variable subject to contamination. However, perhaps the insertion of an efficiency measure along the lines of the selection criteria, as presented by Aeberhard et al. (2021), improves the proposal.

Simulations showed that our robust estimator has the best performance compared to existing approaches to robust fitting for GAMLSS models in most established scenarios and generated estimates. In addiction, our proposal has low computational cost in comparison with the Aeberhard method. For example, the processing time of our proposal for modeling extreme child poverty data represented 5% of the time required for processing the Aeberhard method.

Our application to the Influenza-Like Illness - ILI and extreme child poverty showed that our robust estimator allows for the automatic detection of deviating observations through truncation. Therefore, it is a competitive alternative to other methods.

Future work includes implementations for discrete distributions, study other measures to assess the performance of estimators and works on robust selection of smoothing parameters.

5 CONCLUDING REMARKS

This thesis presents two independent themes with different background. The first theme presents a new method for detecting spatial clusters. The main contribution of the proposal is the alternative of application to any family of distributions and the ability to apply continuous variables duly associated with a population at risk. The results showed that the ELS method was efficient for clusters with larger number of regions, being able to reduce Type I error. False alarm rates should be taken into consideration when using scan methods. As shown in the analysis, the presence of zero inflation is associated with more than 60% type I error probability using the KS method. In terms of Public Health management, this probability indicates that at least 60% of cluster identify will be false alarms. The mean is a non-robust statistic that is affected by the presence of outliers and this fact linked to the choice of the Poisson model inflated by zeros may have influenced the quality of the results. Thus, using the proportion or a median with parameters of interest in modeling empirical likelihood can generate more consistent results. An improvement for future work could be to compare with other non parametric methods, and to consider methods to deal with the non-circular form of the cluster and empirical likelihood approach.

The second theme presents several contributions and two proposals to robust estimation for generalized additive models of location, scale and shape - GAMLSS, which focus on contamination situations in the tails of distributions. The main motivation is the scarcity of robust methods for GAMLSS models.

Based on the idea of Rigby et al. (2019), we introduce the robust fitting for gamma model with bias correction and we carried out an extensive simulation study with several scenarios. Until now, only the correction of bias for the estimators of the beta model were presented. Furthermore, we modified Rigby et al. (2019) proposal to eliminate outliers. The first proposal seeks transformations in order to limit the influence function associated with the probability distribution of interest, modifying the logarithm structure of the likelihood function, using concepts of censorship. The second proposal is based on a simple adaptive truncation, where observations identified as possible outliers are verified and, if necessary, removed by truncation of the response variable distribution. We also presents an adaptive proposal for defining the tuning constant, necessary for estimating the model.

The robust estimator based on censoring has of some limitations. Like any robust estimator, the proportion of contaminated data cannot be unreasonably large without the estimator starting to break at some point. The choice of tuning constant is arbitrary. We believe that adaptive censoring method can show better results and it is worth to be investigate in future research.

Simulations showed that the robust truncated adaptive method has the best performance compared to existing approaches to robust GAMLSS models, in most established scenarios. The robust truncated adaptive method has low computational cost in comparison with the Aeberhard method. For example, the processing time of our proposal for modeling extreme child poverty data represented 5% of the time required for processing the Aeberhard method.

Our proposals is based on a new idea of thinking about robust models. The study of this alternative idea is not complete and future studies on theoretical properties will be needed, such as the sampling distribution necessary for inference; the correction for Fisher consistency cannot be directly extended beyond continuous families of distribution due to the reliance on quantile residuals; and the challenges of selection of tuning probabilities and smoothing parameter selection are not discussed.

Future work includes implementations for discrete distributions, study other measures to assess the performance of estimators and works on robust selection of smoothing parameters.

REFERENCES

Aamodt, G., Samuelsen, S. O., and Skrondal, A. (2006). A simulation study of three methods for detecting disease clusters. *International journal of health geographics*, 5(1):1–11.

Aeberhard, W. H., Cantoni, E., Marra, G., and Radice, R. (2021). Robust fitting for generalized additive models for location, scale and shape. *Statistics and Computing*, 31(1):1–16.

Aeberhard, W. H., Cantoni, E., and S., H. (2014). Robust inference in the negative binomial regression model with an application to falls data. *Biometrics*, 70(4):920–931.

Agostinelli, C., Marazzi, A., and Yohai, V. J. (2014). Robust estimators of the generalized log-gamma distribution. *Technometrics*, 56(1):92–101.

Alimadad, A. and Salibian-Barrera, M. (2011). An outlier-robust fit for generalized additive models with applications to disease outbreak detection. *Journal of the American Statistical Association*, 106(494):719–731.

Anscombe, F. J. and Glynn, W. J. (1983). Distribution of the kurtosis statistic b 2 for normal samples. *Biometrika*, 70(1):227–234.

Bayes, C. L., Bazán, J. L., García, C., et al. (2012). A new robust regression model for proportions. *Bayesian Analysis*, 7(4):841–866.

Besag, J. and Newell, J. (1991). The detection of clusters in rare diseases. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 154(1):143–155.

Beyerlein, A., Fahrmeir, L., Mansmann, U., and Toschke, A. M. (2008). Alternative regression models to assess increase in childhood bmi. *BMC medical research methodology*, 8(1):1–9.

Breen, R. (1996). *Regression Models: Censored, Sample Selected, or Truncated Data*. Sage.

Cadigan, N. G. and Chen, J. (2001). Properties of robust m-estimators for poisson and negative binomial dat. *Journal of Statistical Computation and Simulatio*, 70(3):273–288.

Cantoni, E. (2004). A robust approach to longitudinal data analysis. *Canadian Journal of Statistic*, 32(2):162–180.

Cantoni, E. and Ronchetti, E. (2001). Robust inference for generalized linear models. *Journal of the American Statistical Association*, 96(455):1022–1030.

Cantoni, E. and Zedini, A. (2009). A robust version of the hurdle model. Technical Report 07, Faculté des sciences économiques et sociales, Université de Genève.

Cançado, A., da Silva, C., and Silva, M. (2011). A zero-inflated poisson-based spatial scan statistic. *Emerging Health Threats Journal*, 4.

Chung, K., Yang, D.-H., and Bell, R. (2004). Health and gis: toward spatial statistical analyses. *Journal of Medical Systems*, 28(4):349–360.

Colosimo, E. A. and Giolo, S. R. (2006). *Análise de Sobrevivência aplicada*. Editora Blucher.

Conen, D., Wietlisbach, V., Bovet, P., Shamlaye, C., Riesen, W., Paccaud, F., and Burnier, M. (2004). Prevalence of hyperuricemia and relation of serum uric acid with cardiovascular risk factors in a developing country. *BMC public health*, 4(1):1–9.

Costa, M. A. and Assunção, R. M. (2005). A fair comparison between the spatial scan and the besag–newell disease clustering tests. *Environmental and Ecological Statistics*, 12(3):301–319.

Costa, M. A., Assunção, R. M., and Kulldorff, M. (2012). Constrained spanning tree algorithms for irregularly-shaped spatial clustering. *Computational Statistics & Data Analysis*, 56(6):1771–1783.

Croux, C., Gijbels, I., and Prosdocimi, I. (2012a). Robust estimation of mean and dispersion functions in extended generalized additive models. *Biometrics*, 68(1):31–44.

Croux, C., Gijbels, I., and Prosdocimi, I. (2012b). Robust of mean and dispersions functions in extendd generalized additive models. *Biometrics*, 68(1):31–44.

Cuzick, J. and Edwards, R. (1990). Spatial clustering for inhomogeneous populations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 52(1):73–96.

De Bastiani, F., Rigby, R. A., Stasinopoulous, D. M., Cysneiros, A. H., and Uribe-Opazo, M. A. (2018). Gaussian Markov random field spatial models in gamlss. *Journal of Applied Statistics*, 45(1):168–186.

De Castro, M., Cancho, V. G., and Rodrigues, J. (2010). A hands-on approach for fitting long-term survival models under the gamlss framework. *Computer Methods and Programs in Biomedicine*, 97(2):168–177.

de Lima, M. S., Duczmal, L. H., Neto, J. C., and Pinto, L. P. (2015). Spatial scan statistics for models with overdispersion and inflated zeros. *Statistica Sinica*, 25:225–241.

de Saúde do Estado de São Paulo, S. (2019). Boletim epidemiológico sarampo n 14. Technical report, Governo do Estado de São Paulo.

Duczmal, L. H., Moreira, G. J., Burgarelli, D., Takahashi, R. H., Magalhães, F. C., and Bodevan, E. C. (2011). Voronoi distance based prospective space-time scans for point data sets: a dengue fever cluster analysis in a southeast Brazilian town. *International Journal of Health Geographics*, 10(1):1–14.

Dunn, P. K. and Smyth, G. K. (1996a). Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, 5(3):236–244.

Dunn, P. K. and Smyth, G. K. (1996b). Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, 5(3):236–244.

Dwass, M. (1957). Modified randomization tests for nonparametric hypotheses. *The Annals of Mathematical Statistics*, 28(1):181–187.

Edgeworth, F. Y. (1887). A new method of reducing observations relating to several quantities. *Phil. Mag. (Fifth Series*, 24:222–223.

Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Eguchi, S. and Kano, Y. (2001). Robustifing maximum likelihood estimation by psi-divergence. *ISM Research Memorandam*, 802:762–763.

Elliott, P. and Wartenberg, D. (2004). Spatial epidemiology: current approaches and future challenges. *Environmental Health Perspectives*, 112(9):998–1006.

Farcomeni, A. and Ventura, L. (2012). An overview of robust methods in medical research. *Statistical Methods in Medical Research*, 21(2):111–133.

Field, C. and Smith, B. (1994). Robust estimation: A weighted maximum likelihood approach. *International Statistical Review/Revue Internationale de Statistique*, 62(3):405–424.

Fritz, C. E., Schuurman, N., Robertson, C., and Lear, S. (2013). A scoping review of spatial cluster analysis techniques for point-event data. *Geospatial Health*, 7(2):183–198.

Fu, L., Wang, Y.-G., and Cai, F. (2020). A working likelihood approach for robust regression. *Statistical Methods in Medical Research*, 29(12):3641–3652.

Glasbey, C. and Khondoker, M. (2009). Efficiency of functional regression estimators for combining multiple laser scans of cdna microarrays. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 51(1):45–55.

Gómez-Rubio, V. and López-Quílez, A. (2010). Statistical methods for the geographical analysis of rare diseases. *Advances in Experimental Medicine and Biology*, 686:151–171.

Gupta, A. K. and Nadarajah, S. (2004). *Handbook of Beta Distribution and its Applications*. CRC press.

Hampel, F. (1968). *Contributions to the theory of robust estimation*. PhD thesis, University of California, Berkeley.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (2011). *Robust Statistics: The Approach Based on Influence Functions*, volume 196. Wiley.

Harrell Jr, F. E. (2015). *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*, volume 2. Springer.

Hastie, T. and Tibshirani, R. (1990a). Exploring the nature of covariate effects in the proportional hazards model. *Biometrics*, 46(4):1005–1016.

Hastie, T. J. and Tibshirani, R. J. (1990b). *Generalized Additive Models*. Chapman and Hall, London.

Heritier, S., Cantoni, E., Copt, S., and Victoria-Feser, M.-P. (2009). *Robust Methods in Biostatistics*. John Wiley & Sons.

Hossain, A., Rigby, R., Stasinopoulos, M., and Enea, M. (2016). Centile estimation for a proportion response variable. *Statistics in medicine*, 35(6):895–904.

Huber, P. J. (1965). A robust version of the probability ratio test. *The Annals of Mathematical Statistics*, 36(6):1753–1758.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 221–233. University of California Press.

Huber, P. J. (1981). *Robust Statistics*. John Wiley and Sons.

Huber, P. J. (1996). *Robust Statistical Procedures*. Siam.

Huber, P. J. (2004). *Robust Statistics*. John Wiley & Sons.

IBGE (2012). Censo brasileiro de 2010. *Rio de Janeiro: Instituto Brasileiro de Geografia E Estatísica - IBGE*. Available in: http://www.atlasbrasil.org.br/acervo/biblioteca.

Jung, I. (2009). A generalized linear models approach to spatial scan statistics for covariate adjustment. *Statistics in Medicine*, 28(7):1131–1143.

Klein, J. P. and Moeschberger, M. L. (2006). *Survival Analysis: Techniques for Censored and Truncated Data*. Springer.

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics-Theory and Methods*, 26(6):1481–1496.

Kulldorff, M. and Nagarwalla, N. (1995). Spatial disease clusters: detection and inference. *Statistics in Medicine*, 14(8):799–810.

Kulldorff, M., Tango, T., and Park, P. J. (2003). Power comparisons for disease clustering tests. *Computational Statistics & Data Analysis*, 42(4):665–684.

Landau, S., Ellison-Wright, I., and Bullmore, E. (2004). Tests for a difference in timing of physiological response between two brain regions measured by using functional magnetic resonance imaging. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 53(1):63–82.

Loh, J. M. and Zhu, Z. (2007). Accounting for spatial correlation in the scan statistic. *The Annals of Applied Statistics*, 1(2):560–584.

Longman, I. (1956). Note on a method for computing infinite integrals of oscillatory functions. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 52, pages 764–768. Cambridge University Press.

Maronna, R. A., Martin, R. D., and Yohai, V. J. (2006). *Robust Statistics*. Wiley Series in Probability and Statistics. John Wiley and Sons Ltd.

Marra, G. and Radice, R. (2020). GJRM: Generalised joint regression modelling. *url http://CRAN. R-project. org/package= GRJM. R package version 0.2-2*.

Marra, G., Radice, R., Bärnighausen, T., Wood, S. N., and McGovern, M. E. (2017). A simultaneous equation approach to estimating hiv prevalence with nonignorable missing responses. *Journal of the American Statistical Association*, 112(518):484–496.

Mills, J. E. ad Field, C. A. and Dupuis, D. J. (2002). Marginally specified generalized linear mixed models: A robust approach. *Biomemtrics*, 58(4):727–734.

Moore, D. A. and Carpenter, T. E. (1999). Spatial analytical methods and geographic information systems: use in health research and epidemiology. *Epidemiologic Reviews*, 21(2):143–161.

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

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

Nocedal, J. and Wright, S. (2006). *Numerical Optimization*. Springer.

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2):237–249.

Ozonoff, A., Webster, T., Vieira, V., Weinberg, J., Ozonoff, D., and Aschengrau, A. (2005). Cluster detection methods applied to the upper cape cod cancer data. *Environmental Health*, 4(1):1–9.

Páez, A. and Scott, D. M. (2004). Spatial statistics for urban analysis: a review of techniques with examples. *GeoJournal*, 61(1):53–67.

Piessens, R., de Doncker-Kapenga, E., Überhuber, C. W., and Kahaner, D. K. (2012). *QUADPACK: A subroutine package for automatic integration*, volume 1. Springer Science & Business Media.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramires, T. G., Nakamura, L. R., Righetto, A. J., Pescim, R. R., Mazucheli, J., and Cordeiro, G. M. (2019). A new semiparametric Weibull cure rate model: fitting different behaviors within GAMLSS. *Journal of Applied Statistics*, 46(15):2744–2760.

Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(3):507–554.

Rigby, R. A. and Stasinopoulos, D. M. (2013). Automatic smoothing parameter selection in GAMLSS with an application to centile estimation. *Statistical Methods in Medical Research*, 23(4):318–332.

Rigby, R. A., Stasinopoulos, M. D., Heller, G. Z., and De Bastiani, F. (2019). *Distributions for Modeling Location, Scale, and Shape: Using GAMLSS in R*. CRC Press.

Rudge, J. and Gilchrist, R. (2005). Excess winter morbidity among older people at risk of cold homes: a population-based study in a london borough. *Journal of Public Health*, 27(4):353–358.

Rule, A. D., Larson, T. S., Bergstralh, E. J., Slezak, J. M., Jacobsen, S. J., and Cosio, F. G. (2004). Using serum creatinine to estimate glomerular filtration rate: accuracy in good health and in chronic kidney disease. *Annals of internal medicine*, 141(12):929–937.

Shanks, D. (1955). Non-linear transformations of divergent and slowly convergent sequences. *Journal of Mathematics and Physics*, 34(1-4):1–42.

Smith, A., Hofner, B., Lamb, J. S., Osenkowski, J., Allison, T., Sadoti, G., McWilliams, S. R., and Paton, P. (2019). Modeling spatio temporal abundance of mobile wildlife in highly variable environments using boosted GAMLSS hurdle models. *Ecology and Evolution*, 9(5):2346–2364.

Stasinopoulos, D. M. and Rigby, R. A. (2007). Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software*, 23(7):1–46.

Stasinopoulos, M. D., Rigby, R. A., and Bastiani, F. D. (2018). Gamlss: a distributional regression approach. *Statistical Modelling*, 18(3-4):248–273.

Stasinopoulos, M. D., Rigby, R. A., Heller, G. Z., Voudouris, V., and De Bastiani, F. (2017). *Flexible Regression and Smoothing: Using GAMLSS in R*. CRC Press.

Tango, T. (1995). A class of tests for detecting 'general'and 'focused'clustering of rare diseases. *Statistics in Medicine*, 14(21-22):2323–2334.

Tango, T. (2000). A test for spatial disease clustering adjusted for multiple testing. *Statistics in Medicine*, 19(2):191–204.

Tango, T. and Takahashi, K. (2005). A flexibly shaped spatial scan statistic for detecting clusters. *International journal of health geographics*, 4(1):1–15.

Team, R. C. et al. (2020). R: A language and environment for statistical computing.

Thode, H. C. (2002). *Testing for Normality*. CRC press.

Valdora, M. S. and Yohai, V. J. (2014). Robust estimators for generalized linear models. *Journal of Statistical Planning and Inference*, 146(3):31–48.

Veronesi, R. and Focaccia, R. (2015). *Tratado de Infectologia*. Atheneu.

WHO, M. G. R. S. G. (2006). *WHO Child Growth Standards: Length/height-for-age, weight-for-age, weight-for-length, weight-for-height and body mass index-for-age: Methods and development*. Geneva: World Health Organization.

WHO, M. G. R. S. G. (2007). *WHO Child Growth Standards: Head circumference-for-age, arm circumference-for-age, triceps circumference-for-age and subscapular skinford-for-age: Methods and development*. Geneva: World Health Organization.

WHO, M. G. R. S. G. (2009). *WHO Child Growth Standards: Growth velocity based on weight, length and head circumference: Methods and development*. Geneva: World Health Organization.

Wong, R. K., Yao, F., and Lee, T. C. (2014). Robust estimation for generalized additive models. *Journal of Computational and Graphical Statistics*, 23(1):270–289.

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R*. CRC press.

Wu, C. and Yan, Y. (2012). Empirical likelihood inference for two-sample problems. *Statistics and Its Interface*, 5(3):345–354.

Wynn, P. (1956). On a device for computing the e m (s n) transformation. *Mathematical Tables and Other Aids to Computation*, 10(54):91–96.

Yao, Z., Tang, J., and Zhan, F. (2011). Detection of arbitrarily-shaped clusters using a neighbor-expanding approach: a case study on murine typhus in south texas. *International Journal of Health Geographics*, 10(1):1–17.

Zhang, T. and Lin, G. (2009). Spatial scan statistics in loglinear models. *Computational Statistics & Data Analysis*, 53(8):2851–2858.

Zhang, T., Zhang, Z., and Lin, G. (2012). Spatial scan statistics with overdispersion. *Statistics in Medicine*, 31(8):762–774.