

Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

Análise da qualidade do ar: um estudo de séries temporais para dados de contagem

Kelly Cristina Ramos da Silva

“Se o conhecimento pode criar problemas, não é através da ignorância que podemos solucioná-los”

Isaac Asimov (1920-1992).

São Carlos, 19 de maio de 2013.

Análise da qualidade do ar: um estudo de séries temporais para dados de contagem

Kelly Cristina Ramos da Silva

Orientador: Prof. Dr. Adriano Polpo de Campos

Dissertação de Mestrado a ser submetida ao Programa de Pós Graduação em Estatística da Universidade Federal de São Carlos (PPGEst-UFSCar), como parte dos requisitos necessários à obtenção do título de Mestre em Estatística.

**Ficha catalográfica elaborada pelo DePT da
Biblioteca Comunitária da UFSCar**

S586aq

Silva, Kelly Cristina Ramos da.

Análise da qualidade do ar : um estudo de séries temporais para dados de contagem / Kelly Cristina Ramos da Silva. -- São Carlos : UFSCar, 2013.

113 f.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2013.

1. Análise de séries temporais. 2. Análise multivariada. 3. Modelos de regressão. 4. Dados de contagem. 5. Ar - qualidade. I. Título.

CDD: 519.55 (20ª)



UNIVERSIDADE FEDERAL DE SÃO CARLOS
Centro de Ciências Exatas e de Tecnologia
Programa de Pós-Graduação em Estatística
Via Washington Luis, Km 235 - C.P.676 - CGC 45358058/0001-40
FONE: (016) 3351-8292 – Email: ppgest@ufscar.br
13565-905 - SÃO CARLOS-SP - BRASIL

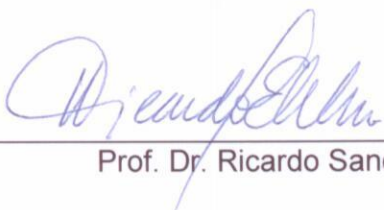
FOLHA DE APROVAÇÃO

Aluno(a) : Kelly Cristina Ramos da Silva

DISSERTAÇÃO DE MESTRADO DEFENDIDA E APROVADA EM 30/04/2013
PELA COMISSÃO JULGADORA:

Presidente 
Prof. Dr. Adriano Polpo de Campos (DEs-UFSCar/Orientador)

1º Examinador 
Prof. Dr. Carlos Alberto de Bragança Pereira (IME-USP)

2º Examinador 
Prof. Dr. Ricardo Sandes Ehlers (ICMC-USP)

Resumo

O objetivo deste trabalho foi investigar a quantidade mensal de dias desfavoráveis à dispersão de poluentes na atmosfera da região metropolitana de São Paulo (RMSP). Foram considerados dois conjuntos de dados provenientes do monitoramento da qualidade do ar da RMSP: (1) um contendo observações mensais das séries temporais do período anual e (2) outro contendo observações mensais das séries temporais do período de maio a setembro. Foram utilizadas duas classes de modelos: os Modelos Vetoriais Autorregressivos (VAR) e os Modelos Aditivos Generalizados para Localização, Escala e Forma (GAMLSS), ressaltando que as técnicas apresentadas nessa dissertação da classe VAR têm ênfase na modelagem de séries temporais estacionárias e as da classe GAMLSS têm ênfase nos modelos para dados de contagem, sendo eles: Delaporte (DEL), Binomial Negativa tipo I (NBI), Binomial Negativa tipo II (NBII), Poisson (PO), Poisson Inflacionada de Zeros (ZIP), Poisson Inversa Gaussiana (PIG) e Sichel (SI). O modelo VAR foi utilizado apenas para o conjunto de dados (1), obtendo uma boa previsão da quantidade mensal de dias desfavoráveis, apesar do ajuste ter apresentado resíduos relativamente grandes. Os GAMLSS foram utilizados em ambos conjuntos de dados, sendo que os modelos NBII e ZIP melhor se ajustaram aos conjuntos de dados (1) e (2) respectivamente. Além disso, realizou-se um estudo de simulação para compreender melhor os GAMLSS investigados. Os dados foram gerados de três diferentes distribuições Binomiais Negativas. Os resultados obtidos mostraram que, tanto os modelos NBI e NBII como o modelo PIG, ajustaram bem os dados gerados. As técnicas estatísticas utilizadas nessa dissertação foram importantes para descrever e compreender o problema da qualidade do ar.

Palavras-chave: Modelo de séries temporais multivariado, modelos de regressão univariados para dados de contagem, qualidade do ar, superdispersão.

Abstract

The aim of this study was to investigate the monthly amount of unfavourable days to pollutant dispersion in the atmosphere on the metropolitan region of São Paulo (RMSP). It was considered two data sets derived from the air quality monitoring on the RMSP: (1) monthly observations of the times series of annual period and (2) monthly observations of the times series of period form May to September. It was used two classes of models: the Vector Autoregressive models (VAR) and Generalized Additive Models for Location, Scale and Shape (GAMLSS). The techniques presented in this dissertation was focus in: VAR class had emphasis on modelling stationary time series; and GAMLSS class had emphasis on models for count data, like Delaporte (DEL), Negative Binomial type I (NBI), Negative Binomial type II (NBII), Poisson (PO), inflated Poisson Zeros (ZIP), Inverse Poisson Gaussian (PIG) and Sichel (SI). The VAR was used only for the data set (1) obtaining a good prediction of the monthly amount of unfavourable days, although the adjustment had presented relatively large residues. The GAMLSS were used in both data sets, and the NBII model had good performance to data set (1), and ZIP model for data set (2). Also, it was made a simulation study to better understanding of the GAMLSS class for count data. The data were generated from three different Negative Binomial distributions. The results shows that the models NBI, NBII, and PIG adjusted well the data generated. The statistic techniques used in this dissertation was important to describe and understand the air quality problem.

Keywords: Multivariate time series model, univariate regression models for counting data, air quality, over dispersion.

Sumário

Lista de Figuras

Lista de Tabelas

1	Introdução	p. 19
1.1	Objetivos e justificativa	p. 21
1.1.1	Objetivos Específicos	p. 21
1.2	Suporte computacional	p. 22
1.3	Descrição geral da dissertação	p. 22
2	Apresentação do problema e análise descritiva	p. 23
2.1	Qualidade do ar	p. 23
2.2	Controle estatístico de processo	p. 26
2.3	Descrição das variáveis	p. 26
2.4	Descrição dos conjuntos de dados	p. 27
2.4.1	Conjunto de dados referente ao período de janeiro a dezembro	p. 28
2.4.2	Conjunto de dados referente ao período de maio a setembro	p. 28
2.5	Análise descritiva	p. 29
2.5.1	Análise unidimensional	p. 29
2.5.2	Análise bidimensional	p. 39
2.5.3	Análise Inferencial	p. 43
3	Modelo VAR	p. 47

3.1	Introdução	p. 47
3.1.1	Definição	p. 48
3.1.2	Teste de raiz unitária	p. 49
3.1.3	Teste de cointegração	p. 50
3.1.4	Método de estimação	p. 50
3.1.5	Métodos de análise de diagnóstico	p. 51
3.2	Método de previsão	p. 52
4	GAMLSS	p. 55
4.1	Introdução	p. 55
4.2	Definição	p. 56
4.2.1	Função de ligação	p. 57
4.2.2	Termos paramétricos direcionados a séries temporais e priores de alinhamento	p. 58
4.3	Algoritmos	p. 59
4.4	Método de estimação	p. 59
4.5	Métodos para a análise de diagnóstico	p. 60
4.6	Modelos de regressão para dados de contagem	p. 60
4.7	Distribuições paramétricas de mistura	p. 61
4.7.1	Distribuição Delaporte	p. 62
4.7.2	Distribuição Binomial Negativa	p. 62
4.7.3	Distribuição Poisson Inflacionada de Zeros	p. 63
4.7.4	Distribuição Poisson Inversa Gaussiana	p. 63
4.7.5	Distribuição Sichel	p. 64
5	Aplicação	p. 65
5.1	Estimação de modelo - VAR	p. 65
5.1.1	Escolha do número de defasagens	p. 65

5.1.2	Análise do ajuste do modelo	p. 66
5.1.3	Análise de diagnóstico	p. 67
5.1.4	Análise da previsão	p. 69
5.2	Estimação do modelo - GAMLSS	p. 70
5.2.1	Seleção de modelos	p. 70
5.2.2	Análise do ajuste - conjunto de dados anual	p. 72
5.2.3	Análise do ajuste - conjunto de dados de maio a setembro	p. 73
5.2.4	Análise de diagnóstico - conjunto de dados anual.	p. 74
5.2.5	Análise de diagnóstico - conjunto de dados de maio a setembro	p. 75
6	Experimento de simulação	p. 77
6.1	Descrição dos algoritmos de geração de dados	p. 77
6.2	Análise teórica dos modelos	p. 78
6.3	Análise dos resultados da simulação	p. 79
7	Considerações finais	p. 81
7.1	Conclusões e Discussão	p. 81
7.2	Sugestões para pesquisas futuras	p. 83
	Apêndice A – Alguns conceitos básicos de regressão e inferência	p. 85
A.1	Critérios de seleção de modelo	p. 85
A.1.1	Deviance	p. 85
A.1.2	Critérios de informação	p. 85
A.2	Análise de Multicolinearidade	p. 86
A.2.1	Fator de Inflação da Variância - VIF	p. 87
A.3	Assimetria e Curtoses	p. 87
A.4	Medidas de correlação entre duas variáveis	p. 88
A.4.1	Coefficiente de correlação linear de Pearson	p. 89

A.4.2	Coeficientes de correlação de postos de Spearman e de Kendall . . .	p. 90
Apêndice B	- Ajustes sem covariáveis - Dados de jan/1999 a dez/2010	p. 93
Apêndice C	- Ajustes sem covariáveis - Dados de maio a setembro, 2001 a 2010	p. 97
Apêndice D	- Simulação	p. 99
Referências Bibliográficas		p. 103

Lista de Figuras

2.1	Número de dias desfavoráveis à dispersão de poluentes na RMSP (maio a setembro). <i>Fonte: CETESB (2001-2012).</i>	p. 25
2.2	Fenômeno natural da inversão térmica. <i>Fonte: http://ambiente.hsw.uol.com.br/inversao-termica.htm.</i>	p. 27
2.3	Decomposição da série dias desfavoráveis.	p. 30
2.4	Decomposição da série precipitação pluviométrica.	p. 30
2.5	Decomposição da série dias com precipitação.	p. 30
2.6	Decomposição da série frentes frias.	p. 31
2.7	Decomposição da série inversão térmica (0-200) metros de altitude.	p. 31
2.8	Decomposição da série inversão térmica (201-500) metros de altitude.	p. 31
2.9	Decomposição da série inversão térmica (> 500) metros de altitude.	p. 32
2.10	Decomposição da série dias com ultrapassagem do padrão para O_3	p. 32
2.11	Decomposição da série porcentagem mediana de umidade relativa do ar.	p. 33
2.12	Gráfico linear e Boxplot da série dias desfavoráveis.	p. 33
2.13	Gráfico linear e Boxplot da série precipitação pluviométrica.	p. 33
2.14	Gráfico linear e Boxplot da série dias com precipitação.	p. 34
2.15	Gráfico linear e Boxplot da série frente fria.	p. 34
2.16	Gráfico linear e Boxplot da série inversão térmica (0-200) metros de altitude.	p. 34
2.17	Gráfico linear e Boxplot da série inversão térmica (201-500) metros de altitude.	p. 34
2.18	Gráfico linear e Boxplot da série inversão térmica (> 500) metros de altitude.	p. 35
2.19	Gráfico linear e Boxplot da série dias com ultrapassagem do padrão para O_3	p. 35

2.20	Gráfico linear e Boxplot da série percentagem mediana de umidade relativa do ar.	p. 35
2.21	Médias por desvio padrão anuais para quantidade de dias desfavoráveis. . . .	p. 36
2.22	Histogramas para as variáveis do conjunto <i>Dados 1</i>	p. 38
2.23	Histogramas para as variáveis do conjunto <i>Dados 2</i>	p. 38
2.24	Diagramas de dispersão para <i>Dados 1</i>	p. 39
2.25	Diagramas de dispersão para <i>Dados 2</i>	p. 40
2.26	Efeito de precipitação, dias com precipitação e frentes frias, respectivamente, <i>versus</i> dias desfavoráveis.	p. 42
2.27	Efeito de inversão térmica (0-200), inversão térmica (201-500) e inversão térmica (>500), respectivamente, <i>versus</i> dias desfavoráveis.	p. 42
2.28	Efeito de dias com ultrapassagem do padrão para o ozônio e percentagem de umidade relativa do ar, respectivamente, <i>versus</i> dias desfavoráveis.	p. 42
5.1	Diagrama de ajuste e resíduos do modelo.	p. 67
5.2	Resíduos do modelo.	p. 68
5.3	Previsão para os dias desfavorável, de jan/1999 a dez/2009 e de jan/1999 a dez/2010.	p. 69
5.4	Previsão para os dias desfavorável, de jan/2006 a dez/2009 e de jan/2006 a dez/2010.	p. 69
5.5	Previsão para os dias desfavorável, de jan/1999 a dez/2002 e de jan/2003 a dez/2006.	p. 70
5.6	Desvios <i>versus</i> quantis normalizados (gráfico de envelope) - modelo de regressão NB.	p. 75
5.7	Resíduos <i>versus</i> valores ajustados, resíduos <i>versus</i> índices, estimativa do núcleo da densidade e quantis amostrais <i>versus</i> quantis teóricos (gráfico QQ) - modelo de regressão NB.	p. 75
5.8	Desvios <i>versus</i> quantis normalizados (gráfico de envelope) - modelo de regressão ZIP.	p. 76

5.9	Resíduos <i>versus</i> valores ajustados, resíduos <i>versus</i> índices, estimativa do núcleo da densidade e quantis amostrais <i>versus</i> quantis teóricos (gráfico QQ) - modelo de regressão ZIP.	p. 76
B.1	Ajuste PO, NBI e NBII para os dados da variável dias desfavoráveis.	p. 93
B.2	Ajuste PIG e DEL para os dados da variável dias desfavoráveis.	p. 93
B.3	Ajuste ZIP e SI para os dados da variável dias desfavoráveis.	p. 94
B.4	Desvios <i>versus</i> quantis normalizados (gráfico de envelope) - modelo sem regressão NB.	p. 95
B.5	Resíduos <i>versus</i> valores ajustados, resíduos <i>versus</i> índices, estimativa do núcleo da densidade e quantis amostrais <i>versus</i> quantis teóricos (gráfico QQ) - modelo sem regressão NB.	p. 95
C.1	Ajuste PO, NBI e NBII para os dados da variável dias desfavoráveis.	p. 97
C.2	Ajuste PIG, DEL e SI para os dados da variável dias desfavoráveis.	p. 97
C.3	Desvios <i>versus</i> quantis normalizados (gráfico de envelope) - modelo sem regressão NB.	p. 98
C.4	Resíduos <i>versus</i> valores ajustados, resíduos <i>versus</i> índices, estimativa do núcleo da densidade e quantis amostrais <i>versus</i> quantis teóricos (gráfico QQ) - modelo sem regressão NB.	p. 98

Lista de Tabelas

2.1	Estrutura do índice da qualidade do ar. <i>Fonte: http://www.fepam.rs.gov.br/qualidade/iqarpop.htm.</i>	p. 24
2.2	Medidas resumos do conjunto <i>Dados 1</i>	p. 37
2.3	Medidas resumos do conjunto <i>Dados 2</i>	p. 37
2.4	Coeficientes de correlação de Pearson, Spearman e Kendall.	p. 41
2.5	Teste de normalidade e teste de raiz unitária - <i>Dados 1</i>	p. 44
2.6	Teste de normalidade e teste de raiz unitária - <i>Dados 2</i>	p. 44
2.7	Diagnóstico de multicolinearidade - <i>Dados 1</i>	p. 45
2.8	Diagnóstico de multicolinearidade - <i>Dados 2</i>	p. 45
4.1	Funções de ligação.	p. 58
4.2	Distribuições de mistura para dados de contagem.	p. 61
4.3	Médias e variâncias das distribuições de mistura.	p. 61
5.1	Valores dos critérios para obtenção do número de defasagens.	p. 66
5.2	Ajuste do modelo PVAR.	p. 67
5.3	Testes de normalidade, de homocedasticidade e de independência dos resíduos.	p. 68
5.4	Valores dos critérios de informação - dados de jan/1999 a dez/2010.	p. 71
5.5	Valores do critério de informação - dados de maio a setembro, 2001 a 2010.	p. 71
5.6	Modelo de regressão NB estimado.	p. 73
5.7	Modelo de regressão ZIP estimado.	p. 74
5.8	Resumo dos quantis dos resíduos para o ajuste do modelo de regressão NB.	p. 75
5.9	Resumo dos quantis dos resíduos para o ajuste do modelo ZIP.	p. 76
6.1	Resultado da simulação: dados 1.	p. 80

6.2	Resultado da simulação: dados 2.	p. 80
6.3	Resultado da simulação: dados 3.	p. 80
B.1	Valores dos critérios de informação - dados de jan/1999 a dez/2010.	p. 94
B.2	Média e variância dos dias desfavoráveis e o intervalo de confiança para o coeficiente estimado.	p. 94
B.3	Modelo NB estimado.	p. 94
B.4	Resumo dos quantis dos resíduos para o ajuste do modelo sem regressão NB.	p. 95
C.1	Valores dos critérios de informação - dados de maio a setembro, 2001 a 2010.	p. 97
C.2	Média e variância dos dias desfavoráveis e do intervalo de confiança para o intercepto.	p. 98
C.3	Modelo NB estimado.	p. 98
C.4	Resumo dos quantis dos resíduos para o modelo NB.	p. 98

Lista de Abreviaturas

ADF teste Dickey-Fuller Aumentado.

AIC critério de informação Akaike.

AICc critério de informação Akaike corrigido.

BIC critério de informação Schwarz.

CETESB Companhia Ambiental do Estado de São Paulo.

Chuva quantidade mensal de precipitação pluviométrica em milímetros.

Dias com chuva quantidade mensal de dias com precipitação pluviométrica.

Dias com ultr. O_3 quantidade mensal de dias com alto nível de gás ozônio na atmosférica.

DEL distribuição de probabilidade Delaporte.

FPE Erro Preditivo Final.

GAM Modelo Aditivo Generalizado.

GAMLSS Modelos Aditivos Generalizados para Posição, Escala e Forma.

GLM Modelo Linear Generalizado.

HQ critério de informação Hamman e Quimn.

Inversão (0-200) quantidade mensal de inversão térmica com altitude (m) entre (0-200).

Inversão (201-500) quantidade mensal de inversão térmica com altitude (m) entre (201-500).

Inversão (>500) quantidade mensal de inversão térmica com altitude (m)(>500).

JB teste Jarque Bera.

Med.porc.umidade porcentagem mediana mensal de umidade relativa do ar.

NBI distribuição de probabilidade Binomial Negativa tipo I.

NBII distribuição de probabilidade Binomial Negativa tipo II.

O_3 gás ozônio.

PIG distribuição de probabilidade Poisson Inversa Gaussiana.

PO distribuição de probabilidade Poisson.

PVAR Modelo Periódico Vetorial Autorregressivo.

R software estatístico R.

RMSP Região Metropolitana de São Paulo.

SI distribuição de probabilidade Sichel.

SW teste de Shapiro-Wilk.

VAR Modelo Vetorial Autorregressivo.

VIF Fator de Inflação da Variância.

Y variável resposta.

ZIP distribuição de probabilidade Poisson Inflacionada de Zeros.

1 Introdução

A previsão da quantidade de dias por mês com alto nível de poluição atmosférica é de extrema importância para os órgãos públicos para adoção de medidas que ajudem a reduzi-la, bem como medidas de proteção à saúde e o bem estar do Homem e do meio ambiente em geral.

Para determinar a concentração de poluentes presentes na atmosfera, a Companhia Ambiental do Estado de São Paulo (CETESB) realiza um acompanhamento sistemático (um monitoramento) da qualidade do ar no Estado de São Paulo. Esse monitoramento viabiliza a elaboração de diagnósticos da qualidade do ar e, conseqüentemente, de ações governamentais para o controle das emissões de poluentes (CETESB, 2001-2012).

Em geral, as quantidades monitoradas pela CETESB relativas à qualidade do ar são variáveis aleatórias discretas, as quais estão presentes em diversos problemas reais, sendo possível encontrar variáveis aleatórias discretas assimétricas com características em comum, tais como superdispersão, inflação de valores, longas caudas e truncamento (LORD; WASHINGTON; IVAN, 2005). Em biologia, por exemplo, o número de uma determinada espécie de planta que sofreu mutação em um período experimental é uma variável aleatória discreta. Em educação, o número de educandos que melhoraram seu desempenho acadêmico em um determinado período letivo é uma variável aleatória discreta. Um fator relevante presente nesse tipo de estudo é o tempo. Para estas observações, a estrutura temporal deve ser levada em conta.

Há diferentes propostas na literatura para a modelagem de séries temporais de contagem. West, Harrison e Migon (1985) e Sims e Zha (1998) propuseram os Modelos Dinâmicos Bayesianos Generalizados para modelagem de dados de contagem com estrutura temporal. Até onde se sabe, os primeiros trabalhos em modelagem de séries temporais periódicas são Jones e Brelford (1967), Pagano (1978) e Troutman (1979), que analisaram as principais propriedades dos Modelos Univariados Periódicos Autorregressivos (PAR).

A classe de Modelos Vetoriais Autorregressivos Clássica (VAR) foi introduzida por Sims (1980). Desde então, essa classe de modelos passou a ser frequentemente utilizada para descre-

ver o comportamento dinâmico de séries temporais econômicas e financeiras, por possibilitar a incorporação da relação de cointegração¹ entre duas ou mais séries integradas de mesma ordem e pela boa previsão do comportamento futuro, a curto e a médio prazo, de séries temporais interrelacionadas (HAMILTON, 1994; CHATFIELD, 2004). Segundo Ursu e Duchesne (2008), a estrutura VAR é uma importante classe de modelos adequada para séries temporais de diferentes áreas de pesquisa, como climatologia e hidrologia.

Um segundo fator relevante é a presença de superdispersão, também conhecida como extra-variância. Em dados de contagens é usual a escolha da distribuição de probabilidade Poisson, que tem média e variância iguais e considera que as observações da variável resposta tem taxa constante de incidência entre unidades individuais. Todavia, em estudos envolvendo dados de contagem, de diferentes áreas de pesquisa, é usual ter uma elevada superdispersão nos dados quando comparado ao esperado em um modelo de Poisson.

A incidência de superdispersão pode, por exemplo, estar relacionada à área de estudo, ao processo de coleta dos dados, à correlação entre respostas individuais, à presença de observações discrepantes, às variáveis explicativas importantes e/ou a termos de interação omitidos (CORDEIRO; DEMÉTRIO, 2008; MCEL DUFF, 2012).

A distribuição de probabilidade Binomial Negativa tem sido amplamente utilizada para a modelagem de dados de contagem superdispersos (LORD; MANNERING, 2010; LORD; GEEDIPALLY, 2011; LEE et al., 2002; MCCULLAGH; NELDER, 1989; HILBE, 2011a). No entanto, essa distribuição pode não ser apropriada para modelar dados com longas caudas e/ou valores inflacionados. Uma alternativa são as distribuições de probabilidade Delaporte, Poisson Inversa Gaussiana, Poisson Inflacionada de Zeros e Sichel. Em recentes trabalhos, diferentes modelos foram propostos, fundamentados na estrutura dos Modelos Lineares Generalizados (GLMs). Por exemplo, o Modelo Conway-Maxwell-Poisson é uma extensão da distribuição de Poisson (CORDEIRO; RODRIGUES; CASTRO, 2012).

Um terceiro fator relevante é a inflação de valores, isto é, um excesso de um particular valor, no qual o mais comum é o excesso de zeros. Usualmente não é adequado o uso de distribuições discretas comuns para análise de dados de contagem com inflação de zeros. Desta forma, foi proposta uma classe de Modelos Inflacionados de Zero. Entre os modelos pertencentes a essa classe, está o Modelo de Poisson Inflacionado de Zero introduzido por Lambert (1992). No entanto, essa estrutura vem sendo alvo de discussões e críticas devido aos seus problemas metodológicos (LEE et al., 2002; LORD; WASHINGTON; IVAN, 2005; HUDSON;

¹As séries são cointegradas se duas ou mais séries são individualmente integradas, mas algumas das suas combinações lineares tem uma ordem baixa de integração (MARGARIDO, 2004; HATEMI, 2008).

KIM; KEATLEY, 2011; LORD; WASHINGTON; IVAN, 2007).

Para melhor compreender as distribuições de probabilidade Poisson, Binomial Negativa, Delaporte, Poisson Inflacionada de Zeros, Poisson Inversa Gaussiana e Sichel, que pertencem aos GAMLSS, realizou-se um estudo de simulação com dados gerados de três diferentes distribuições Binomiais Negativa por meio de três diferentes algoritmos.

A motivação deste trabalho surgiu do interesse em analisar a quantidade mensal de dias desfavoráveis à dispersão de poluentes na atmosfera, que é uma série temporal discreta, com excesso de zeros, monitorada pela CETESB. Utilizou-se os modelos das classes VAR com o propósito de prever o comportamento futuro do evento e os modelos GAMLSS para verificar a relação das covariáveis com a variável resposta de interesse.

1.1 Objetivos e justificativa

Busca-se neste trabalho identificar e compreender a natureza da série temporal referente à qualidade do ar na região de São Paulo. Tal compreensão dos dados, poderá auxiliar a tomada de decisões por órgãos públicos com intuito de melhorar a qualidade do ar. Além disso, pretende-se entender melhor a característica de cada um dos modelos mencionados das classes VAR e GAMLSS, tentando identificar suas qualidades e ineficiências.

1.1.1 Objetivos Específicos

- (i) Apresentar, descrever e caracterizar as estruturas VAR e GAMLSS, destacando seus métodos de estimação e análises de diagnóstico inerentes à análise de regressão.
- (ii) Identificar a natureza das séries temporais estudadas provenientes do monitoramento da qualidade do ar da RMSP mediante o emprego de técnicas estatísticas de análises descritivas e de análises inferenciais;
- (iii) Analisar as previsões do comportamento futuro da série temporal de contagem de interesse, obtidas do ajuste do modelo de previsão da classe VAR;
- (iv) Identificar e analisar o modelo que melhor se ajusta à série temporal de contagem de interesse entre os modelos para dados de contagem da classe GAMLSS propostos;
- (v) Avaliar os modelo GAMLSS propostos por meio de um estudo de simulação.

1.2 Suporte computacional

As análises estatísticas realizadas neste trabalho, apresentações gráficas, medidas resumo, testes de hipóteses e estimativas de parâmetros foram feitas no software estatístico (R Development Core Team (2009)).

Os modelos VAR estão implementados nos pacotes `MSBVAR` e `vars`, os quais contém funções para a estimação dos parâmetros dos Modelos Vetoriais Autorregressivos Clássicos (VAR) e dos Modelos Vetoriais Autorregressivos Bayesianos (BVAR), além de outras funções relacionadas aos modelos VAR e BVAR (Brandt e Appleby (2007)).

Os modelos GAMLSS estão disponíveis em seis pacotes: `gamlss`, `gamlss.cens`, `gamlss.dist`, `gamlss.mx`, `gamlss.nl`, `gamlss.tr` que possibilitam o ajuste de mais de cinquenta distribuições diferentes, tais como distribuições de misturas finitas, distribuições truncadas, modelos não lineares e distribuições para variável resposta censurada (RIGBY; STASINOPOULOS, 2007).

1.3 Descrição geral da dissertação

Esta dissertação está dividida em sete capítulos.

No Capítulo 2 é introduzido o problema da qualidade do ar, são descritos dois conjuntos de dados reais e são analisadas as medidas descritivas, os testes de hipóteses e os gráficos das variáveis investigadas.

Nos capítulos 3 e 4 são abordadas as questões conceituais que dão suporte para modelagem do evento de interesse, os métodos de estimação e de análises de diagnóstico referentes às estruturas VAR e GAMLSS.

No Capítulo 5 é apresentado os resultados e análises das aplicações dos métodos descritos nos capítulos 3 e 4 aos dados previamente apresentados no Capítulo 2.

No Capítulo 6 é descrito o estudo de simulação para os modelos estudados da classe GAMLSS.

No Capítulo 7 são apresentadas as conclusões pertinentes e possibilidades de trabalhos futuros.

2 *Apresentação do problema e análise descritiva*

2.1 **Qualidade do ar**

O problema da poluição atmosférica não é recente e nem de total responsabilidade do Homem. A própria natureza durante milhares de anos vem lançando gases e materiais particulados originários de atividades vulcânicas e tempestades, dentre outras fontes naturais de poluição. Normalmente a própria atmosfera é responsável pela dispersão dos poluentes, misturando-os eficientemente num grande volume de ar, o que contribui para que a poluição fique em níveis aceitáveis na maior parte do tempo. A dispersão dos poluentes atmosféricos depende tanto da topografia quanto das condições meteorológicas locais.

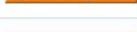

Com o crescimento do número de frotas de veículos automotores e indústrias, principalmente nos grandes centros urbanos, os níveis de poluição atmosférica nas regiões metropolitanas aumentaram significativamente (JERRETT et al., 2005; CETESB, 2001-2012). O ozônio¹, por exemplo, destaca-se atualmente como o poluente com maior número de ultrapassagens do padrão, por diversas vezes tendo ultrapassado o valor de $200 \mu\text{g}/\text{m}^3$, o que determina a má qualidade do ar, sendo que o nível máximo aceitável para esse poluente é de $160 \mu\text{g}/\text{m}^3$. Ele é um gás que se encontra em boa parte da atmosfera, mas nas camadas mais baixas da atmosfera (superfície) ele é gerado a partir da formação de bolhas de poluição emitidas da superfície que reagem com a radiação solar. Para os níveis entre $101 \mu\text{g}/\text{m}^3$ a $150 \mu\text{g}/\text{m}^3$ (moderados) do gás concentrado na atmosfera, é recomendado as pessoas a diminuição de exercícios físicos pesados ao ar livre ao amanhecer, principalmente para pessoas com doenças pulmonares, idosos e crianças. O ozônio é capaz de modificar tanto o equilíbrio ambiental de ecossistemas como a bioquímica das plantas, podendo provocar prejuízos econômicos significativos na produção

¹O ozônio é um gás com efeitos tóxicos e altamente oxidante e é formado na atmosfera da reação entre os compostos orgânicos voláteis (COVs) e óxidos de nitrogênio (NOx) em presença de luz solar para $x = \{1, 2\}$. Os NOx são lançados na atmosfera por meio de processos de combustão (veicular e industrial) e os compostos orgânicos voláteis são emitidos por meio de processos evaporativos, da queima incompleta de combustíveis automotivos e em processos industriais (CETESB, 2002)




agrícola, reduzindo a safra de forma discreta. Devido a esses motivos, tem-se discutido medidas para tentar evitar altas concentrações de ozônio na superfície, apesar de o ozônio ser o único poluente que apresenta suas maiores concentrações no período compreendido fora dos meses de maio a setembro, devido a necessidade da luz solar para poder reagir. Nesse período do ano há diminuição de ocorrência de inversões térmicas em baixas altitudes (Seção 2.3), ventos com altas velocidades e altos índices de precipitação pluviométrica, fatores que favorecem a dispersão dos poluentes atmosféricos (CETESB, 2002; LIRA et al., 2012).

A Tabela 2.1 mostra os índices e classificações dos poluente atmosféricos assim como os índices de qualidade do ar adotados pela CETESB.

Tabela 2.1: Estrutura do índice da qualidade do ar. Fonte: <http://www.fepam.rs.gov.br/qualidade/iqarpop.htm>.

ÍNDICE DA QUALIDADE DO AR (IQAr)								
Qualidade	Índice	Níveis de Cautela sobre a Saúde	PTS ($\mu\text{g}/\text{m}^3$)	PI10 ($\mu\text{g}/\text{m}^3$)	S02 ($\mu\text{g}/\text{m}^3$)	NO2 ($\mu\text{g}/\text{m}^3$)	CO (PPm)	O3 ($\mu\text{g}/\text{m}^3$)
 Boa	0-50		0-80	0-50	0-80	0-100	0-4,5	0-80
 Regular	51-100		81-240	51-150	81-365	101-320	4,6-9,0	81-160
 Inadequada	101-199	* Insalubre para Grupos Sensíveis	241-375*	151-250*	366-586* 587-800	321-1130*	9,1-12,4* 12,5-15,0	161-322* 323-400
 Má	200-299	Muito Insalubre	376-625	251-350 351-420*	801-1600	1131-2260	15,1-30	401-800
 Péssima	300-399	Perigoso	626-875	421-500	1601-2100	2261-3000	30,1-40	801-1000
 Crítica	Acima de 400	Muito Perigoso	> 876	> 500	> 2100	> 3000	> 40	> 1001

Os índices, até a classificação REGULAR, atendem aos Padrões de Qualidade do Ar, estabelecido pela Resolução CONAMA 03 de 28/06/1990.

PADRÕES E CLASSIFICAÇÃO DA QUALIDADE DO AR		
Qualidade	Índice	Padrões de Qualidade do Ar* - CONAMA
 Boa	0-50	Abaixo dos Padrões de Qualidade > 1
 Regular	51-100	Abaixo dos Padrões de Qualidade > 2
 Inadequada	101-200	Acima dos Padrões de Qualidade
 Má	201-300	Acima do Nível de Atenção
 Péssima	301-400	Acima do Nível de Alerta
 Crítica	Acima de 400	Acima do Nível de Emergência

* Resolução CONAMA nº 03 de 28/06/1990.
> 1 Atende ao padrão primário anual
> 2 Atende aos padrões primários de qualidade

A modelagem da quantidade mensal de dias desfavoráveis à dispersão de poluentes na atmosfera deve ser realizada de forma coerente, de modo que as experiências de outros pesquisadores na análise de eventos relacionados ao problema da qualidade do ar sirvam como guia na escolha de métodos e de variáveis explicativas relevantes. Camalier, Cox e Dolwick (2007) utilizaram as técnicas dos Modelos Lineares Generalizados (GLMs) para modelar a relação entre parâmetros meteorológicos e o ozônio, Medeiros e Gouveia (2005) empregaram as técnicas de

Regressão Linear e Regressão Logística para analisar se há relação entre o efeito da exposição materna à poluição do ar com o baixo peso dos bebês ao nascer e Jerrett et al. (2005) analisou o efeito provocado em indivíduos saudáveis expostos à poluição do ar da área urbana mediante o emprego de seis diferentes modelos: (i) Modelos de Regressão para Superfície²; (ii) Modelos de Aproximação³; (iii) Modelos Estatísticos de Interpolação⁴; (iv) Modelos de Dispersão Linear⁵; (v) Modelos Meteorológicos de Emissão Integrada (IME), os Modelos Meteorológicos e os Modulares Químicos⁶; e (vi) Modelos Híbridos⁷.

Na Figura 2.1 é ilustrado os totais de observações do período de maio a setembro dos anos 2002 a 2011 da variável dias desfavoráveis à dispersão de poluentes na atmosfera. Nota-se uma tendência crescente a partir de 2005, em 2008 uma queda e um retorno do crescimento em 2009. Em diversos momentos, nota-se que os valores observados estão acima da média dos últimos 10 anos. No ano de 2011, houve a ocorrência de 56 dias desfavoráveis no período, fato semelhante ao ocorrido em 2008 e 2010, com 59 dias desfavoráveis. A maioria dos dias desfavoráveis ocorreram nos meses de junho e julho, em dias com ocorrências de altas porcentagens de calmaria⁸ e ausência de precipitação pluviométrica (CETESB, 2001-2012).



Figura 2.1: Número de dias desfavoráveis à dispersão de poluentes na RMSF (maio a setembro).

Fonte: CETESB (2001-2012).

Em 1986 o Governo Federal fundou o Programa de Controle de Poluição do Ar por

²Os modelos de regressão para superfície procuram prever a concentração de poluição do ar em uma região próxima à superfície e caracterizar o seu tráfego.

³Os modelos de aproximação utilizam de medidas de proximidade do indivíduo em relação às fontes poluidoras. Esses modelos auxiliam na identificação da relação entre a poluição do ar e a ocorrência de doenças em populações expostas a fontes poluidoras.

⁴Os modelos estatístico de interpolação são baseados em técnicas estatísticas determinísticas e estocásticas.

⁵Os modelos de dispersão linear frequentemente consideram equações Gaussianas (BELLANDER et al., 2001). Esses modelos admitem suposições sobre os processos determinísticos, utilizam de dados de emissão de poluentes, de informações sobre as condições meteorológicas e sobre a topografia local para a estimação da concentração de poluição do ar.

⁶IME são conjuntamente associados para simulação dinâmica de poluição atmosférica.

⁷Os modelos híbridos combinam o monitoramento pessoal ou local com métodos frequentemente utilizados para estimar a exposição à poluição do ar.

⁸Calmaria é um fenômeno que ocorre nos dias com baixa velocidade média do vento, isto é, velocidade média de aproximadamente 1,5 m/s.

Veículos Automotores (Proconve). O Proconve estabelece um cronograma de redução gradual de emissões de poluentes para veículos leves e pesados. Já a CETESB, além do monitoramento, também procura medidas para a melhoria da qualidade do ar no estado de São Paulo, como rodízio de veículos, incentivo do uso de bicicletas como meio de transporte e desenvolvimento de novos combustíveis menos tóxicos (CETESB, 2001-2012).

2.2 Monitoramento da qualidade do ar no estado de São Paulo

O monitoramento da qualidade do ar no estado de São Paulo é realizado pela CETESB. As observações das séries temporais provenientes desse monitoramento se encontram disponíveis no endereço eletrônico <http://www.cetesb.sp.gov.br/ar/qualidade-do-ar/31-publicacoes-e-relatorios>.

Os métodos selecionados para monitorar as variáveis relacionadas à qualidade do ar dependem de diversos fatores, como (i) recursos disponíveis, (ii) níveis de poluição e (iii) tipos de poluentes.

2.3 Descrição das variáveis

Segue abaixo a descrição das variáveis utilizadas neste trabalho (CETESB, 2001-2012; AHRENS, 2009):

Variável resposta:

Desfav: variável quantitativa discreta, quantidade mensal de dias desfavoráveis à dispersão de poluentes na atmosfera, segundo critérios adotados pela CETESB.

Variáveis meteorológicas:

Chuva: variável quantitativa contínua, quantidade mensal de precipitação pluviométrica em milímetros obtida pelo pluviômetro da estação Mirante de Santana.

Dias: variável quantitativa discreta, quantidade mensal de dias com precipitação pluviométrica.

Frentes frias: variável quantitativa discreta, quantidade mensal de sistemas frontais, ou seja, quantidade de massa de ar frio que avançou sob uma massa de ar quente.

Med.umid: variável quantitativa contínua, porcentagem de umidade relativa do ar, ou seja, quantidade de água presente na parcela de ar pela quantidade de água que cabe na parcela de ar. As observações foram medidas às 15 horas, horário do dia em que, geralmente, a porcentagem de umidade relativa do ar apresenta seus valores mais baixos.

Inversão térmica: variável quantitativa discreta, quantidade de inversão térmica para três faixas de altitude em metros: IB: (0-200), IM:(201-500) e IA:(> 500). A inversão térmica é caracterizada quando a temperatura do ar aumenta com a altitude e, como o ar frio é mais denso do que o ar quente, os movimentos ascendentes não ocorrem com facilidade e a poluição na superfície não se dispersa, marcando assim uma atmosfera estável. Esse fenômeno ocorre com maior frequência durante a noite e ao amanhecer, isto é, períodos do dia em que a Terra não está bem aquecida pelo Sol. A Figura 2.2 ilustra tal fenômeno.

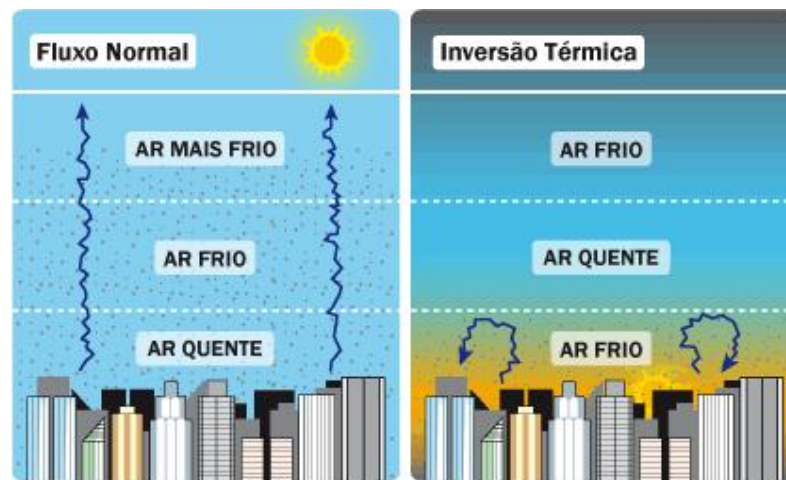


Figura 2.2: Fenômeno natural da inversão térmica. Fonte: <http://ambiente.hsw.uol.com.br/inversao-termica.htm>.

Variável poluente:

Dias O₃: variável quantitativa discreta, quantidade mensal de dias com ultrapassagens do padrão ($160 \mu\text{g}/\text{m}^3$ - uma hora) para ozônio estabelecido pela Lei Estadual 8.468, de 1976. O tempo máximo médio de amostragem do poluente é de oito horas por dia.

2.4 Descrição dos conjuntos de dados

As características dos dois conjuntos de dados usados neste trabalho são descritas nesta seção. O motivo pelo qual se analisou dois conjuntos de dados estava no interesse em incluir no estudo a variável porcentagem de umidade relativa do ar disponível diariamente nos relatórios da CETESB no período de maio a setembro. No entanto, as demais variáveis analisadas estão disponíveis mensalmente no período de janeiro a dezembro. Nesta dissertação a maioria das variáveis são meteorológicas, pelo fato das observações da variável resposta e da maioria das variáveis meteorológicas estarem disponíveis mensalmente. Não foram incluídas neste estudo as observações dos poluentes pelo fato de estarem disponíveis anualmente nos relatórios da CETESB (CETESB, 2001-2012).

2.4.1 Conjunto de dados referente ao período de janeiro a dezembro

O primeiro conjunto de dados consiste de 144 observações mensais do período de janeiro de 1999 a dezembro de 2010 de 8 das variáveis descritas na Seção 2.3 (todas as variáveis descritas, excluindo a variável porcentagem de umidade relativa do ar). Nesse conjunto de dados foram identificadas 25 observações faltantes na variável quantidade de inversões térmicas, sendo que 24 delas pertenciam à faixa de (0-200) e uma à faixa de (201-500). Para conduzir adequadamente a estimação dos dados faltantes, compreendeu-se o processo de modo a utilizar um método de imputação de dados em que proporciona-se estimativas próximas aos valores esperados. O método de imputação de dados foi o usualmente utilizado por pesquisadores da área da Meteorologia, que consiste em fazer a média aritmética do respectivo mês anterior com a do respectivo mês posterior ao mês faltante. No restante do texto este conjunto de dados é referenciado por *Dados 1*.

2.4.2 Conjunto de dados referente ao período de maio a setembro

O segundo conjunto de dados contém 50 observações mensais do período de maio a setembro de 2001 a 2010 de todas as variáveis descritas na Seção 2.3. Para incluir nesse conjunto de dados a variável porcentagem de umidade relativa do ar, foram consideradas as suas medianas mensais e o ano de 2008 foi estimado utilizando o método de imputação de dados descrito na Seção 2.4.1, ou seja, foi calculada a média aritmética das medianas dos respectivos meses de 2007 com os respectivos meses de 2009. No restante do texto este conjunto de dados é

referenciado por *Dados 2*.

2.5 Análise descritiva

Nesta seção são apresentadas as análises descritivas, gráficas e inferenciais das observações das variáveis dos dois conjuntos de dados. Inicialmente foi realizada uma análise descritiva unidimensional das séries com o intuito de caracterizar a amostra. Por meio de gráficos da decomposição das séries, são ilustrados os comportamentos das séries ao longo do tempo para identificar a presença de componente de tendência e/ou de sazonalidade. A seguir, temos os gráficos bidimensionais, os coeficientes de correlação linear de Pearson e os coeficientes de correlação de postos de Spearman e de Kendall, para verificar a existência de relação linear e/ou relação monótona entre duas variáveis (Seção A.4). Para verificar normalidade, não estacionariedade devido à presença de raiz unitária e diagnosticar a multicolinearidade foram realizados o teste de Shapiro-Wilk (SW), o teste Dickey-Fuller Aumentado (ADF) e o Fator de Inflação da Variância (VIF), respectivamente.

2.5.1 Análise unidimensional

Para análise unidimensional das séries são desenhados três diferentes gráficos. Primeiramente observa-se as figuras 2.3 a 2.11, os gráficos das séries em questão. Para todos os gráficos foram considerados o conjunto *Dados 1*, exceto para umidade relativa do ar que é disponível apenas no conjunto *Dados 2*.

Analisando-se os gráficos de decomposição das séries, figuras 2.3 a 2.11, descarta-se a hipótese de existência de componente de tendência, por elas apresentarem comportamentos aleatórios. No entanto, suspeita-se da existência de um componente sazonal associado ao ano. Essa suspeita é confirmada pelos gráficos Boxplots, em que são agrupadas as observações mensalmente, figuras 2.12 a 2.20.

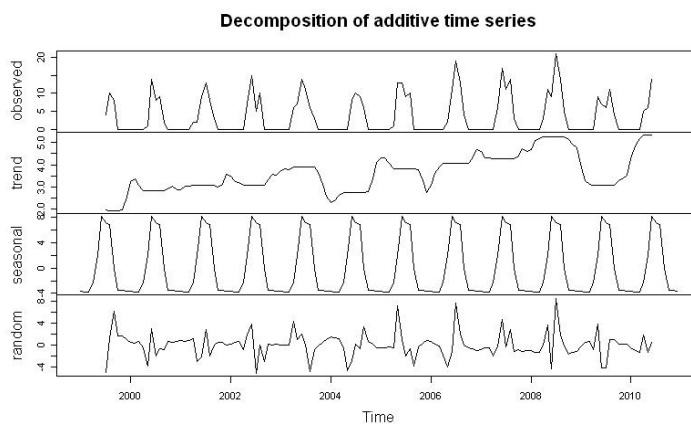


Figura 2.3: Decomposição da série dias desfavoráveis.

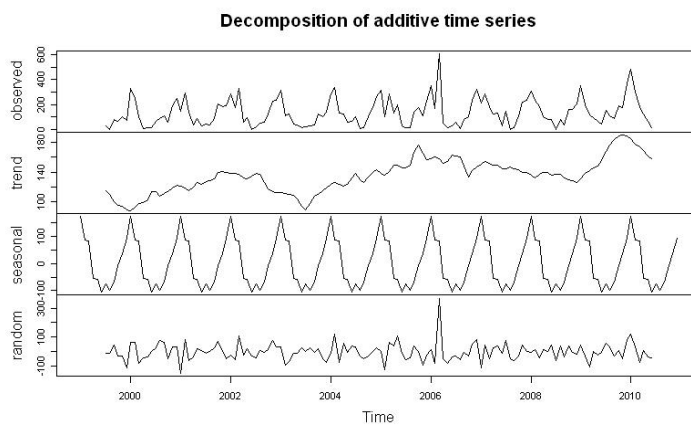


Figura 2.4: Decomposição da série precipitação pluviométrica.

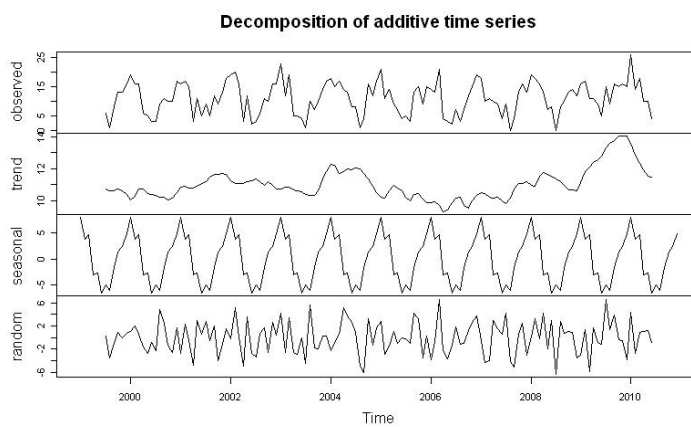


Figura 2.5: Decomposição da série dias com precipitação.

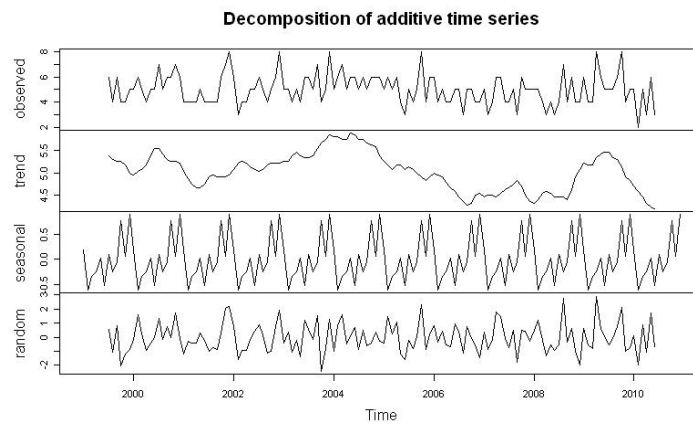


Figura 2.6: Decomposição da série frentes frias.

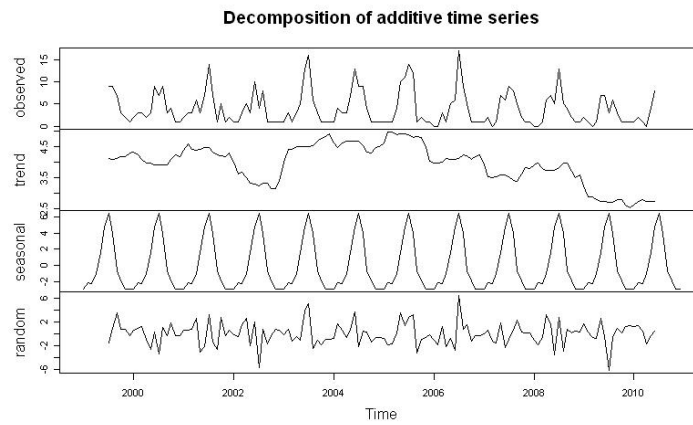


Figura 2.7: Decomposição da série inversão térmica (0-200) metros de altitude.

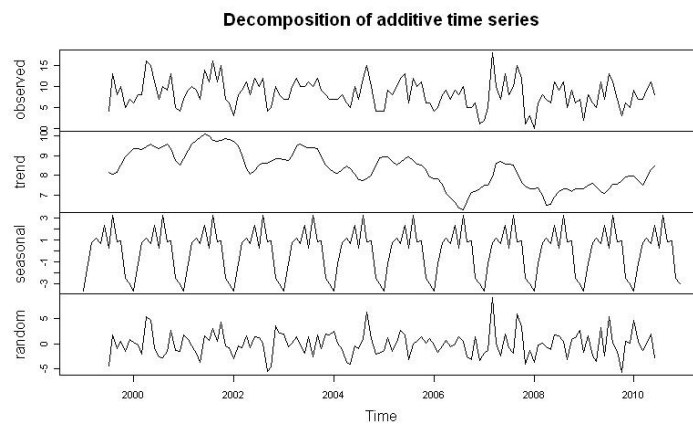


Figura 2.8: Decomposição da série inversão térmica (201-500) metros de altitude.

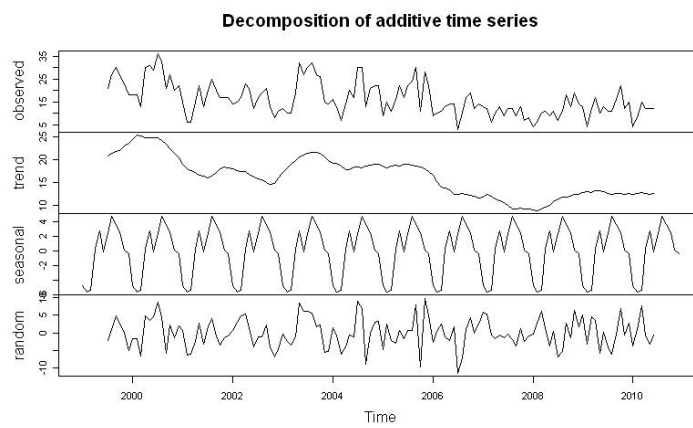


Figura 2.9: Decomposição da série inversão térmica (> 500) metros de altitude.

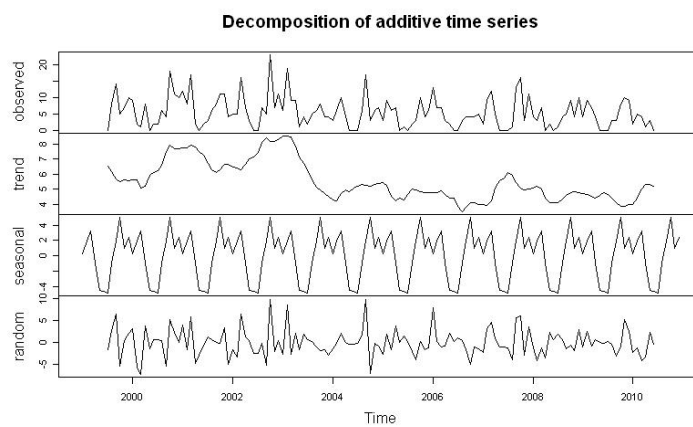


Figura 2.10: Decomposição da série dias com ultrapassagem do padrão para O_3 .

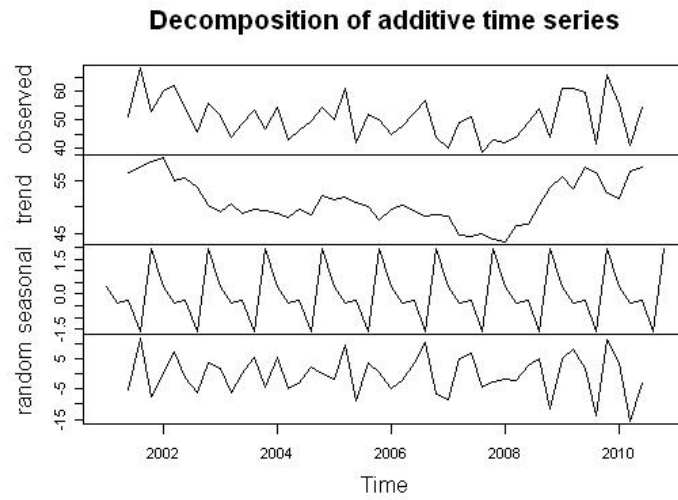


Figura 2.11: Decomposição da série porcentagem mediana de umidade relativa do ar.

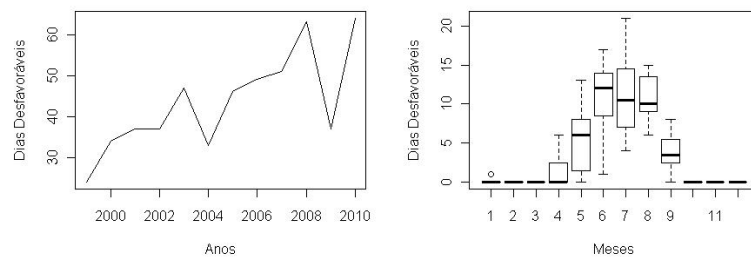


Figura 2.12: Gráfico linear e Boxplot da série dias desfavoráveis.

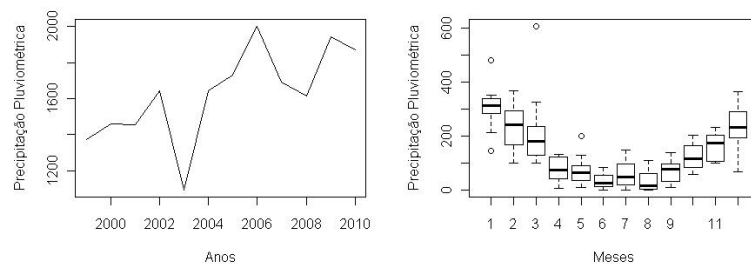


Figura 2.13: Gráfico linear e Boxplot da série precipitação pluviométrica.

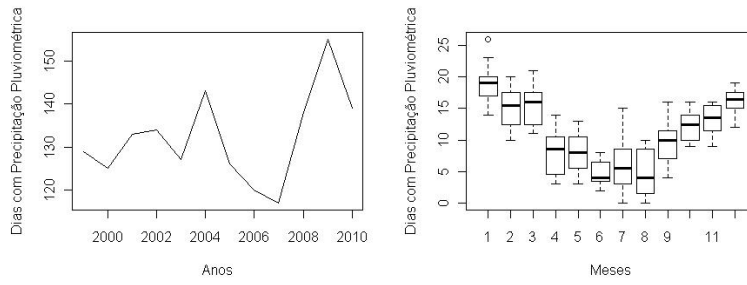


Figura 2.14: Gráfico linear e Boxplot da série dias com precipitação.

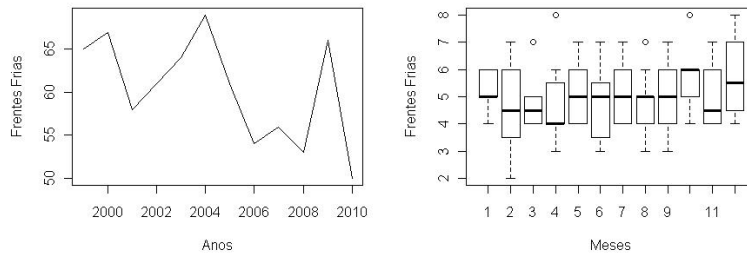


Figura 2.15: Gráfico linear e Boxplot da série frente fria.

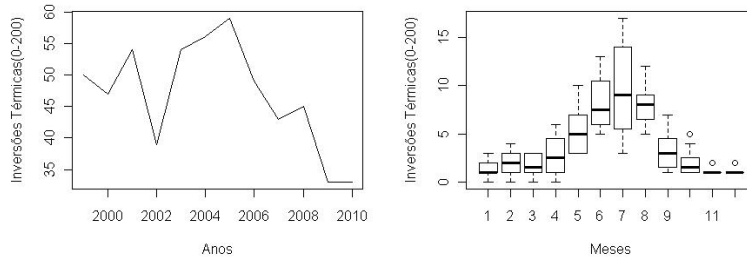


Figura 2.16: Gráfico linear e Boxplot da série inversão térmica (0-200) metros de altitude.

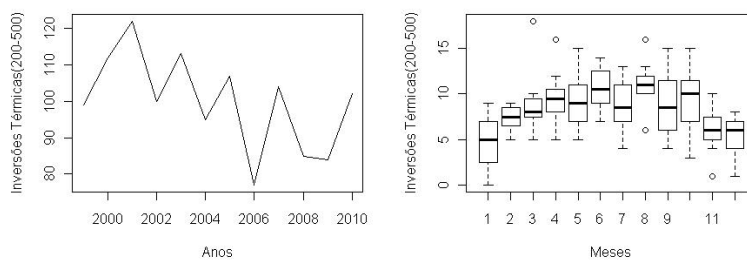


Figura 2.17: Gráfico linear e Boxplot da série inversão térmica (201-500) metros de altitude.

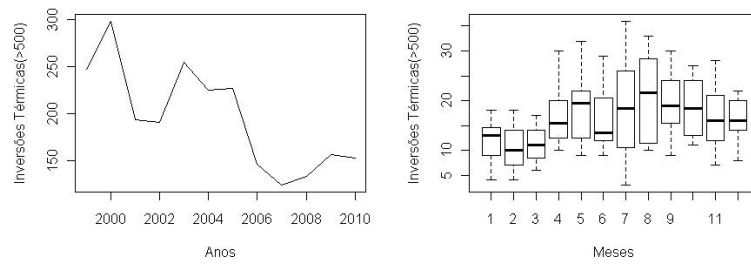


Figura 2.18: Gráfico linear e Boxplot da série inversão térmica (> 500) metros de altitude.

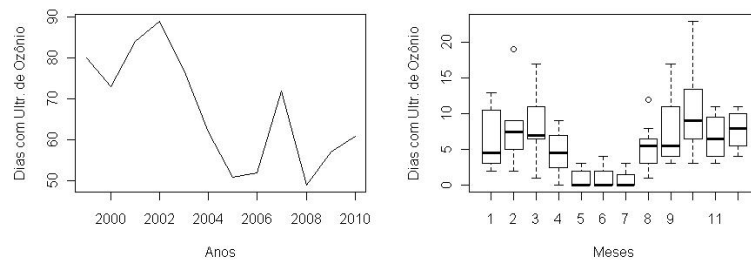


Figura 2.19: Gráfico linear e Boxplot da série dias com ultrapassagem do padrão para O_3 .

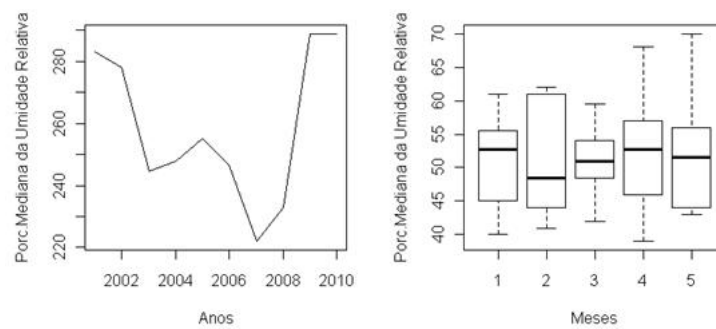


Figura 2.20: Gráfico linear e Boxplot da série porcentagem mediana de umidade relativa do ar.

As figuras 2.3 a 2.11 podem ser obtidas no R através dos comandos:

```
R > variavel.ts <- ts(dados[,k], st=1999, fr=12)
R > variavel.decomp <- decompose(variavel.ts, type = c("additive",
"multiplicative"), filter = NULL)
R > plot(variavel.decomp)
```

em que k é a k -ésima variável com $k = 1, \dots, 9$. Enquanto que as figuras 2.12 a 2.20 podem ser obtidas no R através dos comandos:

```
R > par(mfrow=c(1, 2))
R > plot(aggregate(variavel.ts), xlab="Anos", ylab="variavel.ts")
R > boxplot(variavel.ts ~ cycle(variavel.ts), xlab="Meses",
            ylab="variavel.ts")
```

O gráfico linear da variável resposta, Figura 2.12, descreve bem o problema do inverno de 2008 e 2010. Esses anos apresentaram as maiores quantidades de dias desfavoráveis à dispersão de poluentes primários dos últimos 10 anos, com a ocorrência de 59 dias desfavoráveis somente no período de maio a setembro.

Em alguns anos as séries tiveram uma variabilidade maior devido à presença de picos. Para confirmar a presença de heterocedasticidade, desenhou-se um gráfico de dispersão da média pelo desvio padrão de cada ano da variável resposta, identificando uma associação linear positiva que confirma a suposição descrita anteriormente, Figura 2.21. O código utilizado em R foi

```
R > y <- matrix(dados[, k], 12)
R > medias <- apply(y, 2, mean)
R > desvios <- apply(y, 2, sd)
R > plot(medias, desvios)
```

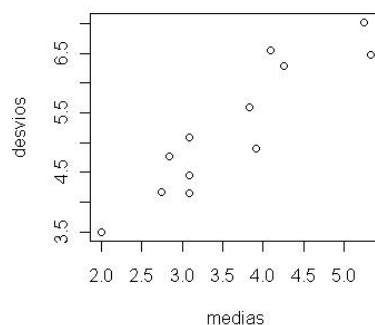


Figura 2.21: Médias por desvio padrão anuais para quantidade de dias desfavoráveis.

A frequência do evento de interesse é significativamente maior nos meses de maio a setembro comparado-se aos demais meses do ano, de maneira que a diferença entre a média do período de maio a setembro com a média do período anual é de 141% (tabelas 2.2 e 2.3).

Tabela 2.2: Medidas resumos do conjunto *Dados 1*.

	Mínimo	1Q.	Mediana	Média	3Q.	Máximo	$\hat{\sigma}$
Dias desfavoráveis	0	0	0	3.63	7	21	5.23
Chuva	0	50.02	107.70	135.59	189.78	607.90	110.70
Dias com chuva	0	7	11	11.01	16	26	5.54
Frentes frias	2	4	5	5.03	6	8	1.22
Inversão (0-200)	0	1	3	3.90	6	17	3.68
Inversão (201-500)	0	6	8	8.33	10	18	3.26
Inversão (>500)	3	11	15	16.30	21	36	7.03
Dias com ultr. O_3	0	2	5	5.60	8.25	23	4.61

Tabela 2.3: Medidas resumos do conjunto *Dados 2*.

	Mínimo	1Q.	Mediana	Média	3Q.	Máximo	$\hat{\sigma}$
Dias desfavoráveis	0	6	9	8.78	12.5	21	5
Chuva	0	16.75	42.40	55.04	84.65	199	46.32
Dias com chuva	0	4	6.5	6.70	9	16	3.90
Frentes frias	3	4	5	4.74	5	7	1.05
Inversão (0-200)	1	4	6.5	6.84	9	17	3.91
Inversão (201-500)	5	7.25	10	9.92	12	16	2.79
Inversão (>500)	3	12	16	17.3	21.75	32	7.21
Dias com ultr. O_3	0	0	2	2.88	4	17	3.71
Med.porc.umidade	39	45.25	51.25	51.76	55.88	70	7.80

Pelas figuras 2.22 e 2.23, vê-se que os dias desfavoráveis, chuva, inversão térmica com altitude entre (0-200) em metros e dias com ultrapassagem do padrão para o ozônio são assimétricas positivas. Nota-se que apenas a variável dias desfavoráveis e a variável inversão térmica com altitude entre (0-200) em metros apresentaram diferenças significativas com relação às suas respectivas distribuições para os conjuntos *Dados 1* e *Dados 2*. Com relação à variável porcentagem de umidade relativa do ar não foi possível a comparação devido a falta das suas observações para o período de jan/1999 a dez/2010.

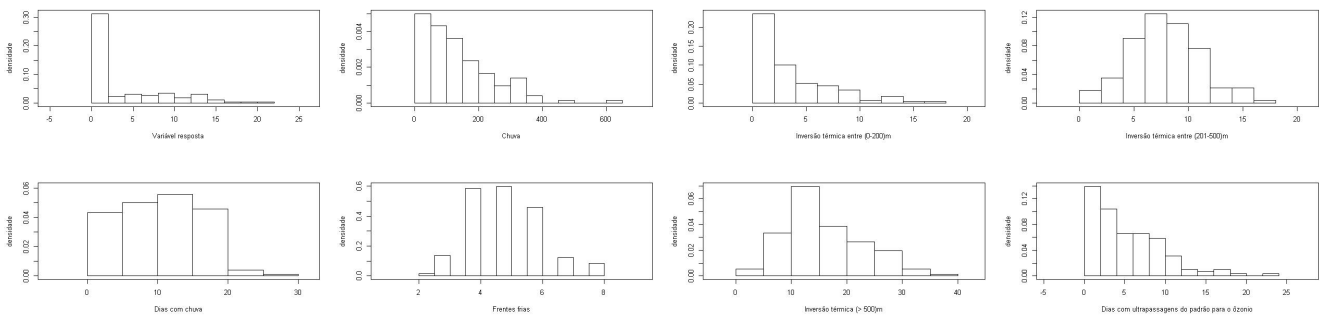


Figura 2.22: Histogramas para as variáveis do conjunto *Dados 1*.

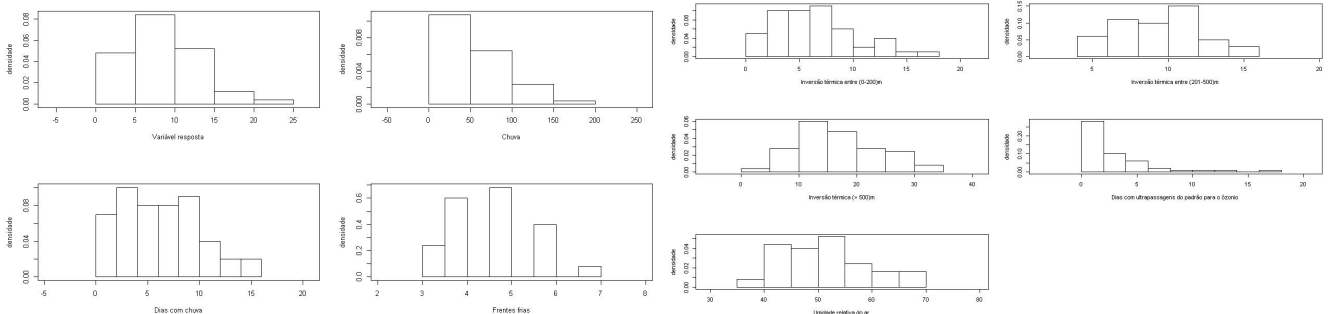


Figura 2.23: Histogramas para as variáveis do conjunto *Dados 2*.

O código R para obter as figuras 2.22 e 2.23 foi

```
R > dest <- density(dados$variável, na.rm = TRUE)
R > int <- hist(dados$variável, plot = FALSE)
R > hist(dados$variável, xlim = range(int$breaks, dest$x),
        ylim = range(int$density, dest$y), xlab = "variável",
        ylab = "densidade", freq = FALSE)
R > box()
```

2.5.2 Análise bidimensional

Para análise bidimensional das séries são interpretados os gráficos de dispersão e os coeficientes de correlação linear de Pearson, de postos de Spearman e de Kendall. Nota-se que há moderada relação linear entre a dias desfavoráveis *versus* dias com precipitação e entre a dias desfavoráveis *versus* inversão térmica com altitude entre (0-200) em metros (figuras 2.24 e 2.25 e Tabela 2.4).

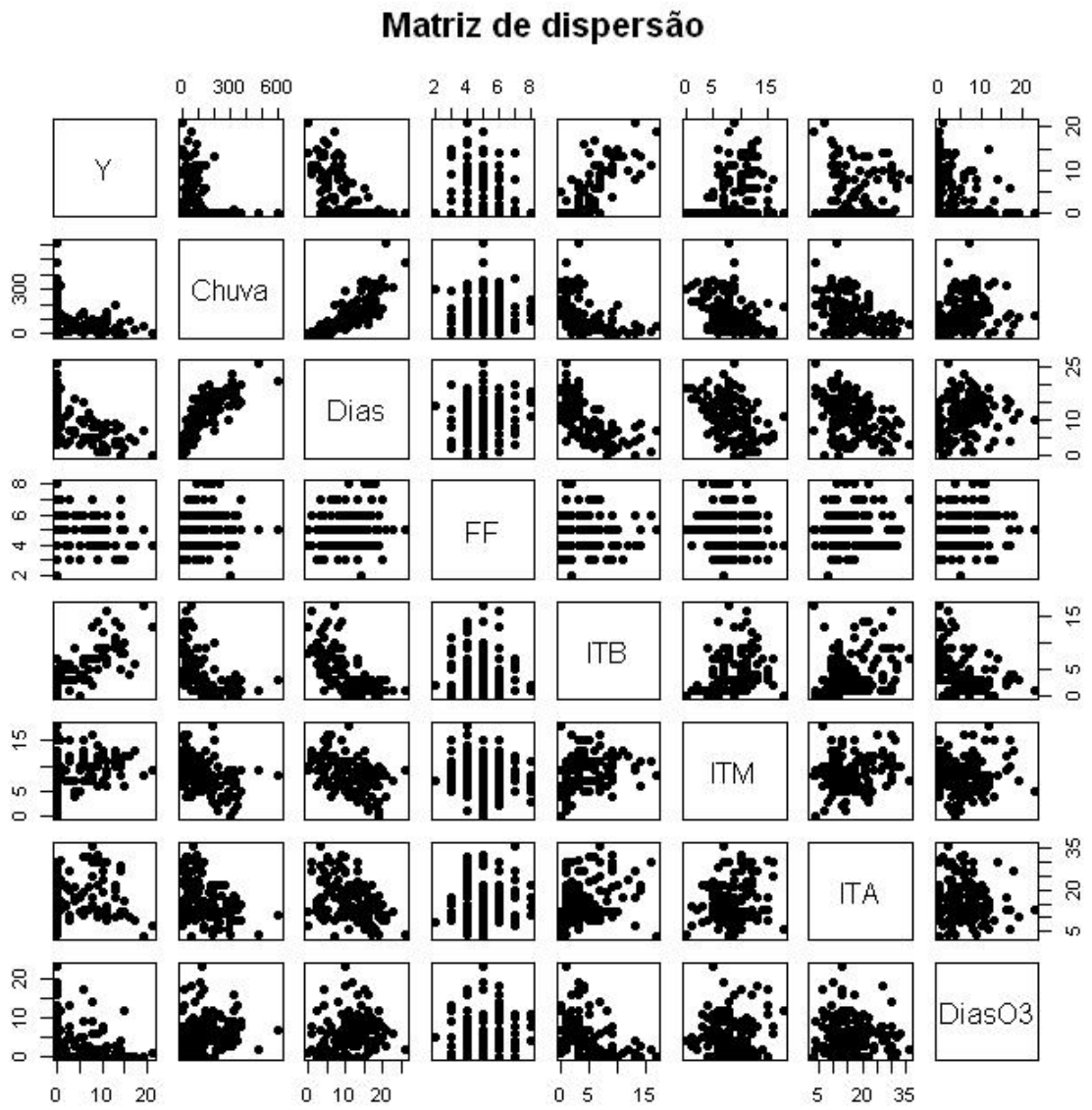


Figura 2.24: Diagramas de dispersão para *Dados 1*.

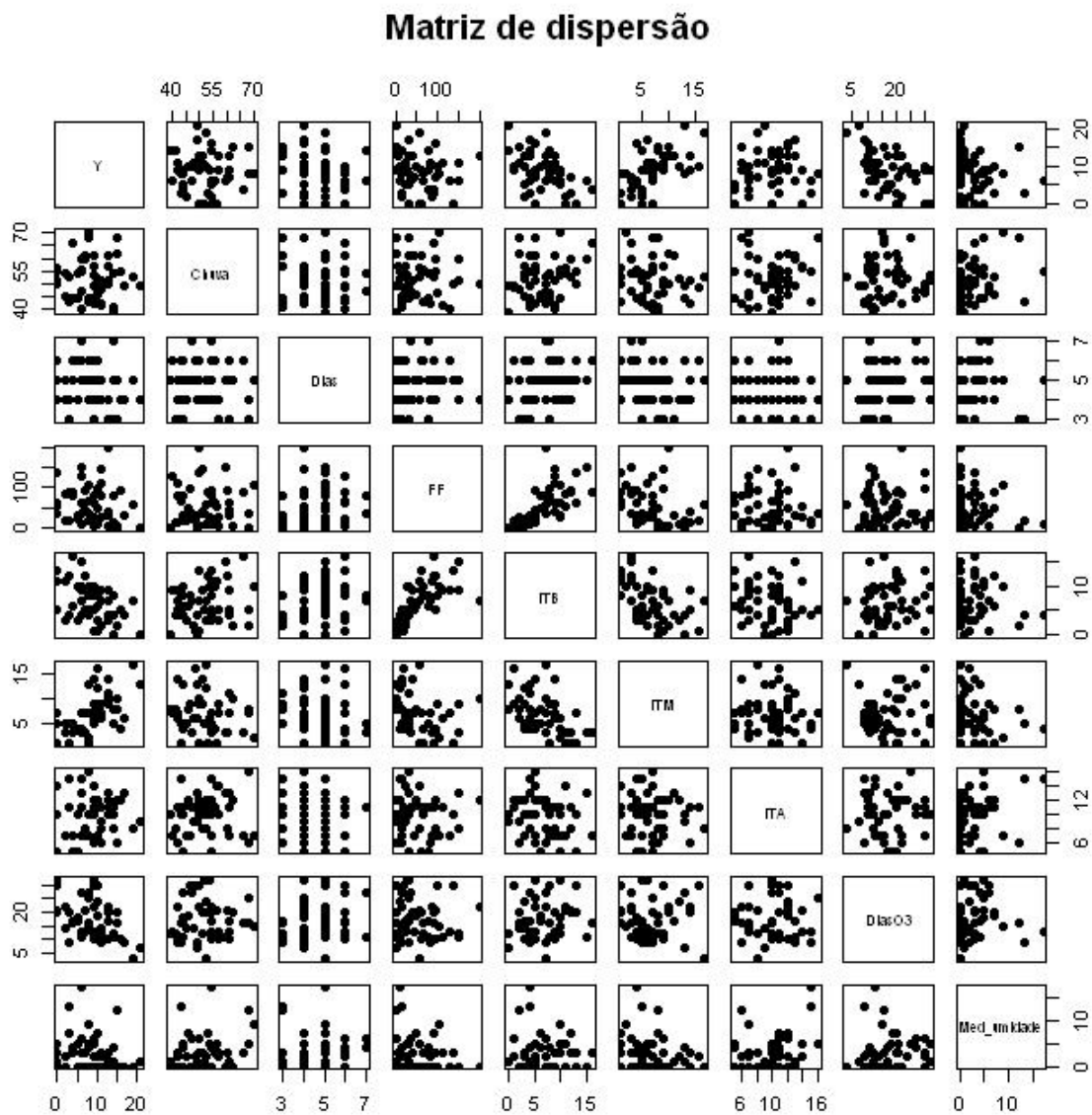


Figura 2.25: Diagramas de dispersão para *Dados 2*.

O código R para obter as figuras 2.24 e 2.25 foi

```
R > nomes <- c("Y", "Chuva", "Dias", "FF", "ITB",
              "ITM", "ITA", "Dias03", "Med_umidade")
R > pairs(dados, pch=16, labels = nomes, main="Matriz de dispersao")
```

Tabela 2.4: Coeficientes de correlação de Pearson, Spearman e Kendall.

	Pearson	p-valor	Spearman	p-valor	Kendall	p-valor
Y versus Chuva	-0.53	<0.01	-0.68	<0.01	-0.52	<0.01
Y versus Dias com chuva	-0.69	<0.01	-0.74	<0.01	-0.58	<0.01
Y versus Frentes frias	-0.20	0.02	-0.15	0.065	-0.13	0.059
Y versus Inversão (0-200)	0.82	<0.01	0.81	<0.01	0.69	<0.01
Y versus Inversão (201-500)	0.36	<0.01	0.44	<0.01	0.34	<0.01
Y versus Inversão (>500)	0.08	0.33	0.12	0.15	0.09	0.14
Y versus Dias com ultr. O ₃	-0.42	<0.01	-0.51	<0.01	-0.39	<0.01
Y versus Med.porc.umidade	-0.008	0.95	-0.05	0.72	-0.03	0.79

O código R para obter os coeficientes de correlação de Pearson, Spearman e Kendall é

```
R > cat("\n Y x variável")
R > cor.test(dados[,i], dados[,k], method = "pearson")
R > cor.test(dados[,i], dados[,k], method = "kendall")
R > cor.test(dados[,i], dados[,k], method = "spearman")
```

em que, $i, k = 1, \dots, 9$.

Para a construção dos gráficos de efeito nas Figuras 2.26 a 2.28 foi considerado o ajuste do modelo Binomial Negativo (mais detalhes são apresentados no Capítulo 5), o conjunto *Dados 1* e a série umidade relativa do ar do conjunto *Dados 2*. Em particular, nota-se que as variáveis: precipitação pluviométrica, dias com precipitação, frentes frias, inversão térmica(> 500) e dias com ultrapassagem do padrão para o ozônio apresentam uma associação inversamente proporcional com relação à variável dias desfavoráveis. Uma associação diretamente proporcional pode ser vista entre a variável dias desfavoráveis e a variável inversão térmica com altitude entre (0-200) em metros, e entre a variável dias desfavoráveis e a variável inversão térmica com altitude entre (201-500) em metros.

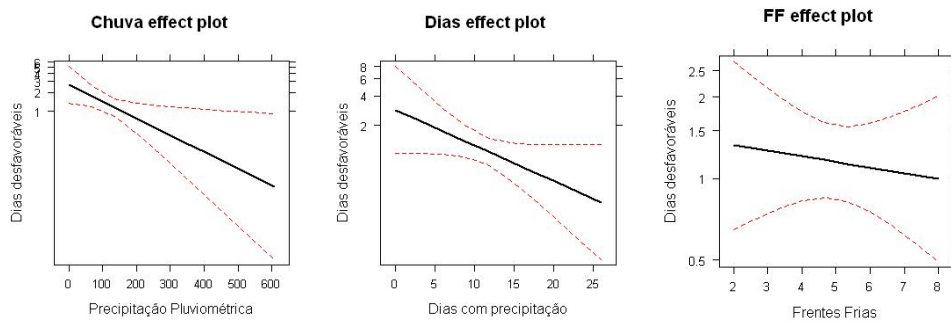


Figura 2.26: Efeito de precipitação, dias com precipitação e frentes frias, respectivamente, *versus* dias desfavoráveis.

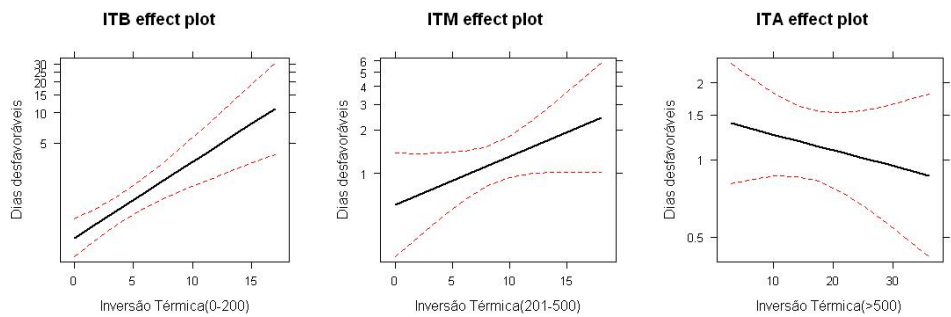


Figura 2.27: Efeito de inversão térmica (0-200), inversão térmica (201-500) e inversão térmica (>500), respectivamente, *versus* dias desfavoráveis.

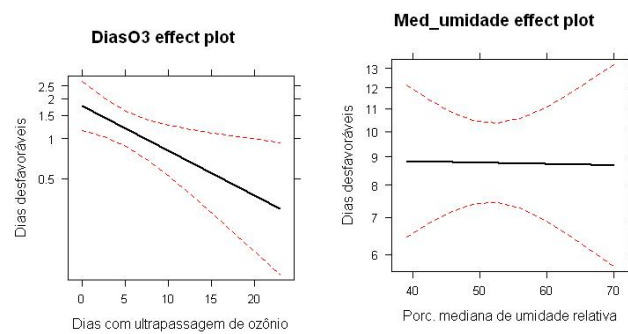


Figura 2.28: Efeito de dias com ultrapassagem do padrão para o ozônio e porcentagem de umidade relativa do ar, respectivamente, *versus* dias desfavoráveis.

O código R para obter as figuras 2.26 a 2.28 foi

```
R > library(dismo)
R > library(effects)
R > library(pscl)
```

```
R > modelo <- glm.nb(Y ~ X1 + X2 + ... + Xp, link = "log",
                    data=dados)
R > plot(effect("Xi", modelo, ylevels=list(Y = 0:20)), multiline=TRUE,
        xlab= "Xi", ylab= "Y", rug=FALSE)
```

em que dados é o conjunto de dados, Y é a variável resposta e X_i é a i -ésima covariável, para $i = 1, \dots, p$.

O ozônio apresenta seus maiores índices quando há maior frequência de luz solar (primavera e verão). Porém, na primavera e no verão os fenômenos que auxiliam na dispersão de poluentes ocorrem com maior frequência, o que justifica a associação inversamente proporcional entre as variáveis dias desfavoráveis e dias com ultrapassagem do padrão para o ozônio, Figura 2.28.

A variável resposta e a porcentagem de umidade relativa do ar possuem alguma associação com a variável precipitação pluviométrica, mas não estão associadas entre si, Figura 2.28 (AHRENS, 2009).

2.5.3 Análise Inferencial

Para análise inferencial das séries realizou-se o teste de Shapiro-Wilk (SW), o teste Dickey-Fuller Aumentado (ADF) e calculou-se o Fator de Inflação da Variância (VIF).

Segundo o SW, podemos concluir que há evidências para rejeitarmos a hipótese de normalidade das séries ($p - valor \leq 0.03$), com exceção da série inversão térmica com altitude entre (201-500) em metros (valor-p = 0.23). Segundo o ADF, podemos concluir que não há evidências para rejeitar a hipótese de estacionariedade das séries ($p - valor < 0.01$), Tabela 2.5.

Nota-se que em ambos os conjuntos de dados pelo menos uma das séries analisadas não segue uma distribuição normal. No caso quando todas as séries analisadas são individualmente normais, seguimos com um teste de normalidade multivariada. Neste caso, podemos afirmar que nos conjuntos de dados *Dados 1* e *Dados 2* não há normalidade multivariada.

A estacionariedade das séries foi testada com o propósito de utilizar o modelo VAR para séries estacionárias.

Tabela 2.5: Teste de normalidade e teste de raiz unitária - *Dados 1.*

	Teste Shapiro-Wilk	p-valor	Teste ADF	p-valor
Dias desfavoráveis	0.73	<0.01	8.19	0.01
Chuva	0.91	<0.01	8.56	0.01
Dias com chuva	0.98	0.03	8.84	0.01
Frentes frias	0.93	<0.01	4.34	0.01
Inversão (0-200)	0.83	<0.01	8.01	0.01
Inversão (201-500)	0.99	0.23	6.12	0.01
Inversão (>500)	0.96	<0.01	5.4	0.01
Dias com ultr. O_3	0.92	<0.01	5.98	0.01

No caso dos dados de maio a setembro, podemos concluir que não há evidências para rejeitarmos a hipótese de normalidade das séries ($0.07 \leq p - valor \leq 0.97$), com exceção da série dias com ultrapassagem do padrão para o ozônio ($valor - p = 0.02$). Para quatro delas, sendo: dias com precipitação, frentes frias, inversão térmica(> 500) e porcentagem mediana de umidade relativa do ar, podemos concluir que há evidências para a rejeitar hipótese de estacionariedade das séries ($0.11 \leq p - valor \leq 0.9$), Tabela 2.6.

Tabela 2.6: Teste de normalidade e teste de raiz unitária - *Dados 2.*

	Teste Shapiro-Wilk	p-valor	Teste ADF	p-valor
Dias desfavoráveis	0.07	0.97	-4.2	0.01
Chuva	0.15	0.2	-3.9	0.02
Dias com chuva	0.13	0.4	-3.2	0.11
Frentes frias	0.18	0.07	-3.0	0.15
Inversão (0-200)	0.14	0.43	-3.9	0.02
Inversão (201-500)	0.13	0.35	-4.2	0.01
Inversão (>500)	0.12	0.42	-2.4	0.43
Dias com ultr. O_3	0.22	0.02	-4.6	0.01
Med.porc.umidade	0.08	0.89	-1.2	0.9

O código R para obter os testes SW, ADF é

```
R > shapiro.test(variável)
R > library(tseries)
R > adf.test(variável)
```

Analisando os valores do VIF, podemos concluir que não há problema de multicolinearidade nos conjuntos *Dados 1* e *Dados 2*, isto é, os valores obtidos do VIF são inferiores a dez (tabelas 2.7 e 2.8).

Tabela 2.7: Diagnóstico de multicolinearidade - *Dados 1*.

	Tolerância	VIF
Intercepto	.	0
Chuva	0.32	3.15
Dias com chuva	0.30	3.35
Frentes frias	0.80	1.25
Inversão (0-200)	0.63	1.60
Inversão (201-500)	0.70	1.43
Inversão (>500)	0.77	1.30
Dias com ultr. O_3	0.68	1.47

Tabela 2.8: Diagnóstico de multicolinearidade - *Dados 2*.

	Tolerância	VIF
Intercepto	.	0
Med.porc.umidade	0.87	1.15
Frentes frias	0.84	1.19
Chuva	0.45	2.24
Dias com chuva	0.40	2.49
Inversão (0-200)	0.75	1.33
Inversão (201-500)	0.88	1.14
Inversão (>500)	0.85	1.17
Dias com ultr. O_3	0.77	1.3

O código R para obter o VIF é

```
R > library(car)
R > vif(modelo)
```

As análises realizadas neste capítulo viabilizam a identificação das características dos dados, de modo que se procure um modelo que melhor se ajuste aos dados. Por exemplo, uma condição necessária para a utilização do modelo VAR é a estacionariedade das séries temporais,

a qual essa hipótese supostamente foi atendida para o *Dados 1*. A estabilidade é uma importante característica desse processo.

A maioria das séries estudadas não seguem uma distribuição normal, além de serem na maioria de contagens. O modelo VAR é indicado para o ajuste de multivariadas séries temporais contínuas estacionárias que segue uma distribuição normal multivariada.

Nesta dissertação foi analisado o desempenho do modelo VAR irrestrito para as multivariadas séries temporais com as características descritas acima. No próximo capítulo é apresentado uma visão geral da teoria da classe VAR.

3 *Modelo VAR*

3.1 Introdução

Os Modelos Vetoriais Autorregressivos Clássicos (VAR) foram introduzidos por Sims (1980) como uma classe de modelos alternativa aos Modelos Macroeconômicos Estruturais que, em sua maioria, eram formados por uma grande quantidade de equações que apresentavam restrições teóricas difíceis de serem testadas e previsões imprecisas.

Os VAR tiveram boa aceitação tanto pela comunidade acadêmica como pelo Banco Central, por possuírem funcionamento simples, previsões avaliadas como bem sucedidas e respeitarem o processo de geração das séries. Porém, esses modelos apresentam algumas limitações, como o grande número de parâmetros a serem estimados. O número de parâmetros cresce de forma quadrática à medida que cada variável é incluída no modelo, podendo ocorrer multicolinearidade e perda de graus de liberdade. As estimativas dos parâmetros do modelo podem ser difíceis de serem interpretadas, devido à existência de interações complexas entre as variáveis. Além disso, seu ajuste pode fornecer um modelo não parcimonioso (DIEBOLD, 1998).

Para investigações empíricas na área de macroeconomia os modelos VAR estão entre os modelos mais utilizados. Em outras áreas, como climatologia, hidrologia, meteorologia, engenharia elétrica, o uso dessa classe de modelos não é frequente apesar de haver uma estrutura adequada para modelar séries temporais multivariadas (URSU; DUCHESNE, 2008; CAVALCANTI, 2010).

A estrutura VAR é uma generalização dos Modelos Autorregressivos de Médias Móveis (ARMA) para séries multivariadas e consiste de sistemas de equações simultâneas. Essa estrutura é capaz de capturar a existência de interrelações entre variáveis a partir de restrições que permitam identificar o componente exógeno¹ de cada variável, tornando-se viável a estimação do efeito de uma mudança na economia ou choque de uma variável sobre as demais. Além disso, são impostas poucas restrições à sua estrutura, basicamente a seleção das defasagens.

¹Variável(eis) explicativa(s) ou exógena(s) são compreendidas como variável(eis) independente(s) determinada(s) por ocorrências exteriores à teoria de interesse e que apesar de externas ao sistema em causa, o influenciam.

Na prática, a escolha das defasagens é feita com base em testes estatísticos. Os parâmetros do modelo são usualmente estimados usando-se os métodos de máxima verossimilhança ou de mínimos quadrados ordinários. No caso da presença do componente sazonal nas séries, adiciona-se à estrutura do modelo as variáveis dicotômicas (*dummy*) sazonais e o modelo é denominado Modelo Periódico Vetorial Autorregressivo (PVAR) (TROUTMAN, 1979; PFAFF, 2008).

Atualmente, com os avanços na capacidade computacional, o ajuste de modelos multivariados tornaram-se mais fáceis. Porém, no processo de construção, ainda há muito mais dificuldades nos modelos multivariados se comparados aos modelos univariados. Aparentemente os modelos multivariados são mais vulneráveis à perda de especificação e possuem mais parâmetros a serem estimados do que os modelos univariados. Além disso, a modelagem da dependência serial dentro de cada série e interdependência entre as séries pode ser complexa. Nos modelos multivariados é necessária uma quantidade suficiente de informação para entender o contexto e identificar todas as variáveis explicativas relevantes (CHATFIELD, 2004).

Frequentemente em macroeconomia existe o interesse em estimar explicitamente o efeito de uma variável exógena sobre as variáveis endógenas ², ou seja, modelagem de interdependência contemporânea entre as variáveis endógenas, na qual se emprega o modelo estruturado da classe VAR denominado SVAR, sendo que o VAR é um caso particular do SVAR na ausência de efeito contemporâneo da variável não observada sobre as variáveis endógenas. Para o ajuste do SVAR podem ser necessárias restrições sobre aos parâmetros referentes aos valores passados e/ou aos parâmetros referentes ao processo de resíduos do modelo.

No ajuste de modelo SVAR, isto é, em análise estrutural, admite que a estrutura causal dos dados sobre investigação, e os resultados impactos causais de choques inesperados para variáveis específicas são usualmente resumidos com função de resposta impulsionada (IRF) e decomposição da variância dos erros de previsão, as quais são baseados na decomposição de médias móveis de Wold (PFAFF, 2008).

Análises utilizando de inferência causal podem ser também feitas a partir do ajuste da classe VAR. Nesta dissertação, não foi empregada tal inferência devido uma discussão apresentada na literatura referente a sua interpretação (??).

²Variáveis observadas que dependem do sistema e o condiciona.

3.1.1 Definição

Seja $\mathbf{Y} = (\mathbf{Y}_t, \forall t \in \mathbb{N})$ um vetor aleatório periódico autorregressivo de um processo estocástico de dimensão $(d \times 1)$. De forma geral, uma parametrização para o Modelo Periódico Vetorial Autorregressivo (PVAR) é

$$\mathbf{Y}_{ns+v} = \sum_{k=1}^{p(v)} \phi_k(v) \mathbf{Y}_{ns+v-k} + \boldsymbol{\varepsilon}_{ns+v}, \quad (3.1)$$

em que \mathbf{Y}_{ns+v} é uma realização durante a v -ésima sazonalidade no ano n , com $v = 1, \dots, s$, com s e v fixos. Os coeficientes do modelo autorregressivo de ordem $p(v)$ durante a sazonalidade v são definidos por $\phi_k(v) = (\phi_{k,ij}(v))$, $i, j = 1, \dots, d$, $k = 1, \dots, p(v)$, d é o número de variáveis endógenas e $p(v) = p$ é o número de defasagens.

O processo de erros periódico de dimensão $(d \times 1)$ referente ao Modelo (3.1) é definido por $\boldsymbol{\varepsilon} = \{\boldsymbol{\varepsilon}_t, \forall t \in \mathbb{N}\}$ com vetor de médias iguais a zero e uma matriz de variâncias e covariâncias predefinida não singular, denotados por $E(\boldsymbol{\varepsilon}_t) = \mathbf{0}$ e $E(\boldsymbol{\varepsilon}_{ns+v} \boldsymbol{\varepsilon}_{ns+v}^T) = \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}(v) = (\sigma_{\varepsilon,ij}(v))$, respectivamente, para $i, j = 1, \dots, d$ (URSU; DUCHESNE, 2008).

3.1.2 Teste de raiz unitária

Considere a estrutura geral do modelo VAR, definido por

$$\mathbf{Y}_t = \phi_1 \mathbf{Y}_{t-1} + \phi_2 \mathbf{Y}_{t-2} + \dots + \phi_p \mathbf{Y}_{t-p} + \boldsymbol{\varepsilon}_t, \quad (3.2)$$

em que \mathbf{Y}_t é o vetor de variáveis aleatórias no instante t , ϕ_i é o i -ésimo vetor de parâmetros para $i = 1, \dots, p$.

No contexto de séries temporais, o Modelo (3.2) pode ser reescrito em função do operador defasagem B , como segue

$$(I_d - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p) y_t = \boldsymbol{\varepsilon}_t. \quad (3.3)$$

A operação à esquerda pode ser fatorada, como segue

$$(I_d - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p) = (I_d - \lambda_1 B)(I_d - \lambda_2 B) \dots (I_d - \lambda_p B), \quad (3.4)$$

em que $(\lambda_1, \dots, \lambda_p)$ são os autovalores do polinômio característico.

A matriz \mathbf{F} de parâmetros é definida,

$$\mathbf{F} = \begin{pmatrix} \phi_1 & \phi_2 & \dots & \phi_p \\ I_d & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & I_d & 0 \end{pmatrix}$$

Um processo VAR pode gerar séries temporais estacionárias, isto é, a média, a variância e a estrutura de covariância do processo são invariantes no tempo. A estacionariedade das séries pode ser verificada avaliando-se o determinante do polinômio característico para $|B| \leq 1$, definido por

$$\det(I_d - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p) \neq 0. \quad (3.5)$$

As séries são estacionárias se os seus autovalores estão dentro do círculo de raio unitário. Se a solução do polinômio tem uma raiz unitária para $B = 1$, então algumas ou todas as variáveis no processo VAR(p) são integradas de ordem um, podendo assim existir cointegração entre as variáveis (HAMILTON, 1994; PFAFF, 2008; LÜTKEPOHL, 2006). No caso de séries não estacionárias, é recomendado o uso do Modelo Vetorial de Correção de Erros (VECM) não definido nesta dissertação.

Teste de Dickey-Fuller Aumentado - ADF

Considere o Modelo (3.1), reescrito como $\Delta \mathbf{Y}_t = \phi_k^*(v) \mathbf{Y}_{ts+v-k} + \boldsymbol{\varepsilon}_{ts+v}$, em que $\phi_k^* = \sum_{k=1}^p \phi_k - 1$. O teste de hipóteses ADF tem a finalidade de testar se a solução do polinômio característico tem uma raiz unitária para o $B = 1$. As hipóteses do teste são definidas como, (MORETTIN; TOLOI, 2006)

$$H_0 : \phi_k^* = 0 \quad \text{e} \quad H_1 : \phi_k^* < 0.$$

3.1.3 Teste de cointegração

Os testes de hipótese de cointegração são empregados em séries temporais não estacionárias integradas de mesma ordem³, em que se testa a hipótese de presença de relação de longo prazo entre as séries. Por exemplo, o teste de Johansen e Juselius (1990), que consiste na determinação do número de vetores de cointegração existentes entre as séries. O método torna-se necessário para determinar a ordem da defasagem do evento de interesse, pois esse procedimento tem como

³Séries integradas de mesma ordem são séries que necessitam do mesmo número de diferenciações para se tornarem estacionárias.

base a hipótese de que, ao introduzir um número suficiente de defasagens no modelo, é possível obter uma estrutura de resíduos bem comportada (MARGARIDO, 2004).

Os testes de hipótese utilizados para identificar a presença de cointegração em séries não estacionárias são ferramentas poderosas para amostras de séries econômicas. No entanto, essas ferramentas não são totalmente aceitas, sendo alvo de críticas (DAMGHANI et al., 2012; HATEMI, 2008).

3.1.4 Método de estimação

As estimativas dos parâmetros do Modelo (3.1) podem ser obtidas pelo método dos mínimos quadrados ordinários (OLS). OLS é uma técnica de otimização matemática que consiste em minimizar a soma dos quadrados das diferenças entre os valores preditos, \hat{y} , e o valores observados, y , tais diferenças são denominadas resíduos. Esse método permite que os parâmetros de cada equação do modelo sejam estimados sem perda de sua relativa eficiência na generalização.

Seja a matriz de dimensão $(d \times dp)$ de coeficientes estimados definida por

$$\hat{\mathbf{\Pi}} = \begin{pmatrix} \hat{\phi}_{111} & \hat{\phi}_{112} & \cdots & \hat{\phi}_{11p(v)} & \cdots & \hat{\phi}_{1d1} & \hat{\phi}_{1d2} & \cdots & \hat{\phi}_{1dp(v)} \\ \vdots & \vdots & \vdots & \vdots & & & & & \\ \hat{\phi}_{d11} & \hat{\phi}_{d12} & \cdots & \hat{\phi}_{dp(v)} & \cdots & \hat{\phi}_{dd1} & \hat{\phi}_{dd2} & \cdots & \hat{\phi}_{ddp(v)} \end{pmatrix},$$

em que $\hat{\phi}_{ijk}$ é o parâmetro da i -ésima equação, j -ésima variável e k -ésima defasagem e $vec(\hat{\mathbf{\Pi}})^t = (\hat{\phi}_1, \dots, \hat{\phi}_d)$ é o operador que empilha as colunas da matriz de dimensão $(d \times dp)$, $i, j = 1, \dots, d, k = 1, \dots, p(v)$.

Considerando o Modelo (3.1), tem-se que as estimativas da matriz de covariância assintótica estacionário e ergódico e de $vec(\hat{\mathbf{\Pi}})$, que são consistentes e com distribuição assintótica dada por uma distribuição normal (HAMILTON, 1994), são

$$\hat{Cov}(vec(\hat{\mathbf{\Pi}})) = \hat{\Sigma} \oplus (\mathbf{Z}^T \mathbf{Z})^{-1} \quad (3.6)$$

e

$$\hat{\Sigma} = \frac{1}{n-dp} \sum_{t=1}^n \hat{\epsilon}_{ns+v} \hat{\epsilon}_{ns+v}^T, \quad (3.7)$$

em que $\epsilon_{ns+v} = Y_{ns+v} - \hat{\mathbf{\Pi}} \mathbf{Z}_t$ é o resíduo multivariado de mínimos quadrados no tempo t , para $\mathbf{Z}_t^T = (1, Y_{t-1}^T, \dots, Y_{t-p}^T)$.

Algumas das limitações do método OLS são descritas: (1) os resíduos devem ser indepen-

dentes e aleatórios; (2) o modelo deve ser linear nos parâmetros, ou seja, as variáveis devem apresentar uma relação linear entre si.

Um modelo é linear quando pode ser escrito na forma matricial, como segue,

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (3.8)$$

em que \mathbf{Y} é o vetor ($n \times 1$) de observações da variável resposta, X é a matriz ($n \times p$) de observações das p variáveis explicativas, $\boldsymbol{\beta}$ é o vetor de parâmetros do modelo e $\boldsymbol{\varepsilon}$ é o vetor ($n \times 1$) de resíduos aleatórios.

3.1.5 Métodos de análise de diagnóstico

Para análise de diagnóstico são apresentados três testes de hipótese: o teste Jarque-Bera (JB), o teste Portmanteau e o modelo autorregressivo condicional heterocedasticidade (ARCH).

O teste de hipótese de Portmanteau é utilizado para testar a existência de conjuntos de resíduos com autocorrelação diferentes de zero. Uma formulação do teste é dada por

$$Q_h = T \sum_{j=1}^h tr(\hat{C}_j^T \hat{C}_0^{-1} \hat{C}_j \hat{C}_0^{-1}), \quad (3.9)$$

em que $\hat{C}_i = \frac{\sum_{t=i+1}^T \hat{\varepsilon}_t \hat{\varepsilon}_{t-i}^T}{T}$, $tr(\hat{C}_j^T \hat{C}_0^{-1} \hat{C}_j \hat{C}_0^{-1})$ é o traço da matriz $\hat{C}_j^T \hat{C}_0^{-1} \hat{C}_j \hat{C}_0^{-1}$ e T é o instante de tempo.

A estatística do teste tem distribuição aproximada $\chi_{(d^2h-n^*)}^2$, em que n^* é o número de coeficientes excluídos do termo determinístico do VAR. As distribuições limitadas são válidas para $h \rightarrow \infty$.

O modelo ARCH é utilizado para modelar a variância e é aplicado frequentemente no estudo de séries temporais financeiras que exibem períodos de oscilação seguidos por um período de relativa calma. Uma parametrização do modelo ARCH é dada por

$$vech(\hat{\varepsilon}_t \hat{\varepsilon}_t^T) = \beta_0 + A_1 vech(\hat{\varepsilon}_t \hat{\varepsilon}_{t-1}^T) + \dots + A_p vech(\hat{\varepsilon}_t \hat{\varepsilon}_{t-p}^T) + \mathbf{v}_t, \quad (3.10)$$

$$VARCH(p) = \frac{1}{2} T d(d+1) R_m^2 \quad (3.11)$$

e

$$R_m^2 = 1 - \frac{2}{d(d+1)} tr(\hat{\Omega} \hat{\Omega}_0^{-1}), \quad (3.12)$$

em que A_i é uma matriz de coeficientes, \mathbf{v}_t é o processo de erros esféricos, d é o número de

variáveis, $\hat{\Omega}$ é a matriz de covariância e $vech$ é o operador para matrizes simétricas que empilha a diagonal principal em colunas. A estatística $VARCH(p)$ tem distribuição $\chi^2\left(\frac{pd^2(d+1)^2}{4}\right)$.

A estatística Jarque-Bera (JB) é baseada no terceiro e no quarto momento e tem distribuição $\chi^2_{(2d)}$. O teste univariado JB é definido por

$$JB = \frac{n}{6} \left(A(\mathbf{Y}_t)^2 + \frac{(K(\mathbf{Y}_t) - 3)^2}{4} \right), \quad (3.13)$$

em que n é o número de observações ou graus de liberdade, $A(\mathbf{Y}_t)$ e $K(\mathbf{Y}_t)$ são os coeficientes de assimetria e curtoses respectivamente (Apêndice A.3).

O estimador de mínimos quadrados vai coincidir com o estimador de máxima verossimilhança, quando o processo de resíduos do modelo for normalmente distribuído, admitindo que os valores iniciais dados ao método numérico são adequados, ou seja, quando a forma do estimador não é fechada, utiliza-se um método numérico, como o algoritmo Newton-Rapson para se obter estimativas para os parâmetros do modelo (HAMILTON, 1994; PFAFF, 2008).

3.2 Método de previsão

As estimativas de previsões futuras de séries para horizontais $h \geq 1$ podem ser obtidas recursivamente, após ajuste do Modelo (3.1), dado que os resíduos do modelo são não correlacionados. O vetor de previsões estimadas é definido por

$$\hat{\mathbf{y}}_{t+h/t} = \hat{\phi}_1 \mathbf{y}_{t+h-1/t} + \dots + \hat{\phi}_p \mathbf{y}_{t+h-p/t}. \quad (3.14)$$

O intervalo de confiança estimado para a série futura prevista é definido por

$$(\hat{\mathbf{y}}_{d,t+h/t} - c_{1-\gamma/2} \hat{\sigma}_d(h), \hat{\mathbf{y}}_{d,t+h/t} + c_{1-\gamma/2} \hat{\sigma}_d(h)), \quad (3.15)$$

em que $c_{1-\gamma/2}$ é o $1 - \gamma/2$ percentil da distribuição Normal e $\hat{\sigma}_d(h)$ é a estimativa do desvio padrão da d -ésima variável h passos a frente.

O intervalo de confiança é inferido da matriz de covariâncias empírica dos erros de previsão, a qual à sua decomposição é baseada na matriz ortogonal estimada de coeficientes da variável resposta $\hat{\Psi}_n$. A estimativa da variância dos erros de previsão é definida por

$$\hat{\sigma}_d^2(h) = \sum_{n=0}^{h-1} (\hat{\Psi}_{d,1,n}^2 + \dots + \hat{\Psi}_{dd,n}^2). \quad (3.16)$$

Neste capítulos nos restringimos à definição das ferramentas básicas para o ajuste de um modelo VAR irrestrito. O interesse dessa dissertação está em analisar o desempenho das previsões do modelo VAR estável para séries temporais de contagens.

No próximo capítulos, será dada uma visão geral da classe de modelos de regressão univariada denominada modelos generalizados para locação, forma e escala (GAMLSS) e em particular dos modelos para dados de contagens: Delaporte, Binomial Negativa, Poisson Inversa Gaussiana, Poisson Inflacionada de Zeros e Sichel.

4 *GAMLSS*

4.1 Introdução

A presença da Estatística é cada vez mais intensa na sociedade. Estudos nos mais variados ramos de atividade frequentemente utilizam alguma técnica de estatística para inferir a respeito do evento de interesse. No entanto, muitas das técnicas de análise estatística possuem fortes suposições que, se violadas, podem gerar resultados duvidosos e irrealistas. Por este motivo, novas técnicas estatísticas mais flexíveis e menos restritivas tem sido desenvolvidas.

Dentre as técnicas de modelagem de regressão univariada, Rigby e Stasinopoulos (2007, 2005), Akantziliotou, Rigby e Stasinopoulos (2002) introduziram os Modelos Aditivos Generalizados para Localização, Escala e Forma (GAMLSS). Estes são modelos estatísticos de regressão (semi)paramétricos, pelo fato de permitirem o ajuste dos parâmetros da distribuição da variável resposta, assim como a inclusão de funções de suavização não paramétricas.

Nos GAMLSS admite-se que a distribuição para a variável resposta é uma distribuição de uma família geral, podendo ser uma distribuição com alta assimetria, com alta curtose, entre outros. Nessa estrutura, a parte sistemática do modelo é ampliada para permitir a modelagem de todos os parâmetros da distribuição da variável resposta como função paramétrica linear, não linear, função não paramétrica aditiva de variáveis explicativas e efeitos aleatórios.

A classe GAMLSS foi especialmente desenvolvida com o propósito de superar algumas das limitações associadas aos Modelos Lineares Generalizados (GLM) e aos Modelos Aditivos Generalizados (GAM) introduzidos por Nelder e Wedderburn (1972) e Hastie e Tibshirani (1990), respectivamente, que são uma das classes de modelos mais importantes na literatura. As limitações mais discutidas associadas às classes GLM e GAM são (i) que a distribuição da variável resposta deve pertencer à família exponencial e (ii) que apenas o parâmetro de localização (a média) da distribuição é modelado explicitamente a partir das variáveis explicativas.

Uma limitação da estrutura GAMLSS, por exemplo, é o fato de que seu método de estimação utiliza a primeira e a segunda derivada da função de verossimilhança maximizada,

sendo que há distribuições em que tais derivadas são difíceis de serem obtidas. Nesse caso, os algoritmos numéricos disponíveis no pacote `gamlss` implementados no software R utilizam valores aproximados para as derivadas (RIGBY; STASINOPOULOS, 2006, 1996a).

4.2 Definição

Sejam $\mathbf{y}^T = (y_1, \dots, y_n)$ um vetor de n observações independentes da variável resposta, Y com função de (densidade ou massa) de probabilidade $f(y_i | \boldsymbol{\theta}^i)$ condicionada a um vetor de p parâmetros $\boldsymbol{\theta}^{iT} = (\theta_1, \dots, \theta_p)$ relacionado às variáveis explicativas e aos efeitos aleatórios. Se as variáveis explicativas forem estocásticas ou dependerem de seus valores passados, compreende-se que $f(y_i | \boldsymbol{\theta}^i)$ é condicionada a esses valores. No entanto, y_i é condicionalmente independente do vetor de parâmetros $\boldsymbol{\theta}^i$, para todo $i = 1, \dots, n$ (RIGBY; STASINOPOULOS, 2005; FLORENCIO, 2010).

Para associar cada parâmetro θ_k da distribuição de Y às variáveis explicativas e aos efeitos aleatórios é considerada uma função de ligação $g_k(\cdot)$ (Seção 4.2.1).

A equação do modelo aditivo é definida por

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \mathbf{X}_k \boldsymbol{\beta}_k + \sum_{j=1}^{J_k} \mathbf{Z}_{jk} \boldsymbol{\gamma}_{jk}, \quad (4.1)$$

em que $\boldsymbol{\beta}_k^T = (\beta_{1k}, \dots, \beta_{J'_k})$ é um vetor de parâmetros de dimensão J'_k , \mathbf{X}_k e \mathbf{Z}_{jk} são matrizes de planejamento conhecidas com dimensões $(n \times J'_k)$ e $(n \times q_{jk})$, respectivamente, $\boldsymbol{\gamma}_{jk}$ é um vetor de variáveis aleatórias de dimensão q_{jk} com distribuição q_{jk} -variada com vetor de médias iguais a zero e a matriz de variâncias e covariâncias é denotada por $\mathbf{G}_{jk}^{-1}(\lambda_{jk})$ de dimensão $(q_{jk} \times q_{jk})$, que depende de parâmetros de suavização λ_{jk} . O vetor $\boldsymbol{\gamma}_{jk}$ é denotado por $\boldsymbol{\gamma}_{jk} \sim N_{q_{jk}}(\mathbf{0}, \mathbf{G}_{jk}^{-1}(\lambda_{jk}))$, em que $k = 1, \dots, p$.

O preditor linear $\boldsymbol{\eta}_k$ é constituído de uma componente paramétrica $\mathbf{X}_k \boldsymbol{\beta}_k$ e de uma componente aditiva $\mathbf{Z}_{jk} \boldsymbol{\gamma}_{jk}$ para $k = 1, \dots, p$ e $j = 1, \dots, J_k$, sendo que a Equação (4.1) facilita a incorporação de diferentes tipos de combinações de termos aditivos e de efeitos aleatórios, além de ser apropriada no uso de algoritmos de retroajuste. Por exemplo, o método iterativo de Newton-Raphson e o método iterativo de *Scoring* de Fisher.

Um caso especial da Equação (4.1) ocorre quando $\mathbf{Z}_{jk} = \mathbf{I}_n$ e $\boldsymbol{\gamma}_{jk} = q_{jk}(\mathbf{x}_{jk})$, resultando em um modelo contendo termos paramétricos, não paramétricos e efeitos aleatórios, em que \mathbf{I}_n é a matriz identidade de dimensão $(n \times n)$ e as combinações de j e k são específicas. Assim, a

equação do modelo semiparamétrico aditivo linear é definida por

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \mathbf{X}_k \boldsymbol{\beta}_k + \sum_{j=1}^{J_k} q_{jk}(\mathbf{x}_{jk}), \quad (4.2)$$

em que $q_{jk}(\mathbf{x}_{jk})$ é uma função não conhecida avaliada nas observações da variável explicativa X_{jk} para todo $j = 1, \dots, J_k$.

A classe GAMLSS consiste de uma combinação de termos paramétricos lineares ou não lineares, de modo que a Função paramétrica (4.1) pode ser estendida para permitir termos paramétricos não lineares. A equação do modelo semiparamétrico aditivo não linear é definida por

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = h_k(\mathbf{X}_k \boldsymbol{\beta}_k) + \sum_{j=1}^{J_k} q_{jk}(\mathbf{x}_{jk}), \quad (4.3)$$

em que h_k é uma função não linear.

Se não existirem termos aditivos associados aos parâmetros da distribuição ($J_k = 0$), então são definidas as equações do modelo paramétrico linear simples e do modelo paramétrico não linear simples, respectivamente, por

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \mathbf{X}_k \boldsymbol{\beta}_k \quad (4.4)$$

e

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = h_k(\mathbf{X}_k \boldsymbol{\beta}_k). \quad (4.5)$$

A estrutura GAMLSS pode ser aplicada aos parâmetros de qualquer distribuição populacional. No entanto, até onde se sabe, os pacotes disponíveis no software R referentes à estrutura GAMLSS se restringem a distribuições das famílias com no máximo quatro parâmetros, denotados usualmente por $\boldsymbol{\theta}^{iT} = (\mu, \sigma, \nu, \tau)$ sendo que, os primeiros dois parâmetros, μ e σ , na Equação (4.1), são caracterizados como parâmetro de localização e parâmetro de escala, respectivamente. Os demais parâmetros, se existirem, são caracterizados como parâmetros de forma (RIGBY; STASINOPOULOS, 2005, 2007).

4.2.1 Função de ligação

A escolha de uma função de ligação deve ser resultado de um exame intensivo dos dados e deve ser compatível com a distribuição proposta aos dados. A função de ligação é uma função conhecida, bijetora, contínua e diferenciável pelo menos até a segunda ordem, que relaciona o parâmetro com a componente sistemática, $\boldsymbol{\eta}_k = \mathbf{X}_k \boldsymbol{\beta}_k$. No caso de observações de contagem,

a função de ligação canônica é a log natural, pelo fato de restringir os valores esperados ao intervalo $(0, \infty)$ (HILBE, 2011b).

Segue abaixo, a relação de funções de ligação utilizadas nos ajustes dos modelos de contagem da subclasse GAMLSS paramétrica.

Tabela 4.1: Funções de ligação.

Distribuição	$g_1(\mu)$	$g_2(\sigma)$	$g_3(\nu)$
Delaporte	log	log	logito
Binomial Negativa tipo I	log	log	-
Binomial Negativa tipo II	log	log	-
Poisson	log	-	-
Poisson Inflacionada de zeros	log	logito	-
Poisson Inversa gaussiana	log	log	-
Sichel	log	log	identidade

No software R, a função `show.link()` mostra as funções de ligação avaliadas para as distribuições paramétricas de cada família disponível na função `gamlss()`.

4.2.2 Termos paramétricos direcionados a séries temporais e priori de alinhamento

Sejam as estatísticas de ordem $x_{(1)} < \dots < x_{(n)}$ observações equidistantes da variável explicativa X , a qual X corresponde a uma unidade de tempo, como dias, meses ou anos. Considere-se também o i -ésimo passeio aleatório definido por

$$h[x_{(i)}] = h[x_{(i-1)}] + \varepsilon_i \quad \text{e} \quad h[x_{(i)}] = 2h[x_{(i-1)}] - h[x_{(i-2)}] + \varepsilon_i,$$

em que, ε_i são erros independentes distribuídos normalmente com parâmetros zero e λ^{-1} e $i > 2$, denotado por $\varepsilon_i \sim N(0, \lambda^{-1})$ e com priores uniformes difundidas para os passeios aleatórios de primeira e de segunda ordem. Esses são alguns dos termos que podem ser incorporados na estrutura GAMLSS.

Um caso mais geral ocorre quando a variável aleatória X é contínua, suas estatísticas de ordem não são equidistantes e a distribuição a priori para $h(\mathbf{x})$ é denotada como uma função de suavização, em que $\mathbf{h} = h(\mathbf{x})$ é uma função desconhecida com distribuição normal multivariada com vetor de médias iguais a zero e uma matriz de variâncias e covariâncias denotada por $\lambda^{-1} \mathbf{K}^{-1}$. O vetor \mathbf{h} é denotado por $\mathbf{h} \sim N_n(\mathbf{0}, \lambda^{-1} \mathbf{K}^{-1})$, em que \mathbf{K} é uma matriz conhecida

penalizada que depende apenas do vetor de observações \mathbf{x} , conhecido, da(s) variável(eis) explicativa(s) (RIGBY; STASINOPOULOS, 2005; FAHRMEIR; TUTZ, 2001; HARVEY, 1989).

4.3 Método de estimação

Considere o Modelo (4.1) e $\boldsymbol{\gamma}_{jk}$ com distribuição normal a priori, denotado por $\boldsymbol{\gamma}_{jk} \sim N_{q_{jk}}(\mathbf{0}, \mathbf{G}_{jk}^{-1})$, em que \mathbf{G}_{jk} é uma matriz simétrica generalizada inversível ($q_{jk} \times q_{jk}$) que depende do vetor de hiperparâmetros λ_{jk} . Se \mathbf{G}_{jk} não tiver matriz inversa, então o λ_{jk} é compreendida a priori como uma função de densidade imprópria proporcional a $\exp(-\frac{1}{2}\boldsymbol{\gamma}_{jk}^T \mathbf{G}_{jk} \boldsymbol{\gamma}_{jk})$ (RIGBY; STASINOPOULOS, 2005; FLORENCIO, 2010).

Os parâmetros de suavização, $\boldsymbol{\lambda}'_{jk}$ s, os vetores de parâmetros, $\boldsymbol{\beta}'_k$ s, os parâmetros de efeitos aleatórios, $\boldsymbol{\gamma}'_{jk}$ s, são estimados da função de verossimilhança penalizada maximizada, definida por

$$\ell_p = \ell - \frac{1}{2} \sum_{k=1}^p \sum_{j=1}^{J_k} \lambda_{jk} \boldsymbol{\gamma}_{jk}^T \mathbf{G}_{jk} \lambda_{jk}, \quad (4.6)$$

em que $\ell = \sum_{i=1}^n \log f(y_i | \boldsymbol{\theta}^i)$ é a função de verossimilhança maximizada condicionada ao vetor de parâmetros da distribuição $\boldsymbol{\theta}^i$.

No caso, quando não há termos aditivos no modelo, os parâmetros $\boldsymbol{\beta}'_k$ s são estimados da função de verossimilhança maximizada ℓ , isto é, $\ell_p = \ell$.

As estimativas de máxima verossimilhança de $\boldsymbol{\beta}$, $\boldsymbol{\lambda}$ e $\boldsymbol{\gamma}$ são os valores $\boldsymbol{\beta}'_k$ s, $\boldsymbol{\lambda}'_{jk}$ s e $\boldsymbol{\gamma}'_{jk}$ s que maximizam a função de verossimilhança, ℓ_p .

O estimador de máxima verossimilhança (EMV) sob as condições de regularidade é um estimador aproximadamente não viciado nos casos assintóticos, converge em probabilidade para o verdadeiro valor populacional, isto é, é um estimador consistente. É também eficiente, invariante e segue assintoticamente uma distribuição normal com média o valor do parâmetro populacional e com variância o inverso da informação de Fisher do parâmetro.

Para algumas distribuições de probabilidade complexas e quando temos termos aditivos no modelo o método de máxima verossimilhança não apresenta uma forma explícita para o EMV exigindo assim o uso de métodos numéricos, como o algoritmo de Newton-Rapson. O algoritmo de Newton-Rapson pode ser difícil de ser implementado, pois não utiliza de expressões analíticas para estimar os parâmetros do modelo e além de necessitar de valores iniciais adequados para à sua convergência (CORDEIRO; DEMÉTRIO, 2008).

4.4 Métodos para a análise de diagnóstico

Os métodos para a análise de diagnósticos, as comparações de modelos e a detecção de pontos discrepantes para os modelos de contagem paramétricos da classe GAMLSS têm sido discutidos.

Os quantis dos resíduos aleatórios normalizados de Dunn e Smyth (1996) e o gráfico de envelope (*worn plot*) de Buuren e Fredriks (2001) são utilizados para verificar a adequabilidade do modelo e as inadequações nas observações, respectivamente. Na estrutura GAMLSS, os resíduos são definidos

$$\hat{r}_i = \phi^{-1}(u_i), \quad (4.7)$$

em que ϕ^{-1} é a inversa da função de distribuição acumulada da normal padrão e $u_i = (F(y_i - 1 | \boldsymbol{\theta}^i), F(y_i | \boldsymbol{\theta}^i))$, se y_i é uma observação de contagem, então $F(y | \boldsymbol{\theta}^i)$ é a função de distribuição acumulada da variável resposta.

4.5 Algoritmos

Para maximizar a função de verossimilhança (penalizada), $\ell(\ell_p)$, dois algoritmos numéricos denominados por RS e CG estão implementados no software R na função `gamlss()` (COLE; GREEN, 1992; RIGBY; STASINOPOULOS, 1996a).

Para obter as estimativas dos parâmetros dos modelos da família de dispersão é indicado o uso do algoritmo RS, que é geralmente mais estável e mais rápido se comparado ao algoritmo CG quando se trabalha com um número suficientemente grande de observações. O RS não utiliza o valor esperado das derivadas cruzadas no ajuste do parâmetro de locação (da média) e do parâmetro de dispersão (de escala) da Equação (4.1). Para muitas funções (densidade ou massa) de probabilidade os parâmetros são informações ortogonais¹. Em particular, para as distribuições Binomial Negativa, Gama, Inversa Gaussiana, Logística e Normal, seus parâmetros são informações ortogonais completas.

O algoritmo CG é indicado para distribuições com alta correlação nas estimativas dos parâmetros e na falta de convergência do algoritmo RS. O CG usa o valor esperado ou aproximado da primeira e segunda derivadas cruzadas da função de verossimilhança maximizada. Uma terceira opção é o algoritmo *mixed* que consiste em uma mistura dos algoritmos CG e RS e pode ser utilizado quando há efeitos aleatórios no ajuste do modelo.

¹Quando os valores esperados das derivadas cruzadas da função de verossimilhança maximizada são iguais a zero.

As principais vantagens dos algoritmos são: (i) permitir diferentes modelos de diagnóstico para cada parâmetro da distribuição, isto é, permitir o procedimento de ajuste modular; (ii) facilitar a adição de termos aditivos extra e de distribuições extras, e (iii), em geral, são rápidos e estáveis (RIGBY; STASINOPOULOS, 2005; FLORENCIO, 2010).

4.6 Modelos de regressão para dados de contagem

4.6.1 Modelo Poisson

Considere uma série temporal Y_t e um processo de covariáveis $Z_{t-p}^T = (1, X_t, Y_{t-p})$, em que X_t é uma componente de tendência ou uma componente sazonal e Y_{t-p} é a série no instante $t - p$, $\forall p \in \mathbb{N}$ (KEDEM; FOKIANOS, 2002).

A distribuição de probabilidade da Poisson com parâmetro μ_t para séries temporais de contagem é definida por

$$f(y) = \frac{\exp(-\mu_t)\mu_t^{y_t}}{y_t!}, \quad (4.8)$$

em que se supõe que a média e a variância da variável resposta são iguais, denotado por $E(Y_t) = \text{Var}(Y_t) = \mu_t$.

4.6.2 Algumas distribuições paramétricas baseadas no modelo de Poisson

Nesta seção, são apresentadas cinco distribuições discretas de mistura: Delaporte, Binomial Negativa, Poisson Inflacionada de Zeros, Poisson Inversa Gaussiana e Sichel. As distribuições investigadas neste trabalho pertencem à família dos modelos para dados de contagem com no máximo três parâmetros.

As distribuições de mistura na estrutura GAMLSS e suas respectivas médias e variâncias são apresentadas nas tabelas 4.2 e 4.3.

Tabela 4.2: Distribuições de mistura para dados de contagem.

Distribuição	Parâmetros	Distribuição de mistura
Delaporte	$(\mu; \sigma; \nu)$	Gama modificada, SG $(1, \sigma, \nu)$
Binomial Negativa tipo I	$(\mu; \sigma)$	Gama, GA $(1, \sigma)$
Binomial Negativa tipo II	$(\mu; \sigma)$	Gama, GA $(1, \frac{\sigma}{\mu})$
Poisson	(μ)	-
Poisson Inflacionada de Zeros	$(\mu; \sigma)$	Binomial, BI $(1, 1 - \sigma)$
Poisson Inversa Gaussiana	$(\mu; \sigma)$	Inversa Gaussiana, IG $(1, \sigma)$
Sichel	$(\mu; \sigma; \nu)$	Generalizada IG, GIG $(1, \sigma, \nu)$

Tabela 4.3: Médias e variâncias das distribuições de mistura.

Distribuição	Média	Variância
Delaporte	μ	$\mu + \sigma(1 - \nu)^2 \mu^2$
Binomial Negativa tipo I	μ	$\mu + \sigma \mu$
Binomial Negativa tipo II	μ	$\mu + \sigma \mu^2$
Poisson	μ	μ
Poisson Inflacionada de Zeros	$(1 - \sigma)\mu$	$(1 - \sigma)\mu + \sigma(1 - \sigma)\mu^2$
Poisson Inversa Gaussiana	μ	$\mu + \sigma \mu^2$
Sichel	μ	$\mu + h(\sigma, \nu)\mu^2$

4.6.3 Distribuição Delaporte

A distribuição Delaporte (DEL) foi introduzida por Delaporte (1959). Os primeiros trabalhos realizados com o emprego da distribuição Delaporte foram na modelagem do número de reivindicações de seguro de motor portfólio (WILLMOTA; SUNDTB, 1989). Atualmente, sua maior utilidade é na área de ciência atuarial.

A DEL é uma composição da distribuição de Poisson com a distribuição Gama modificada, denotada por $\text{Poisson}(\lambda + \text{Gama}(\alpha, \beta))$, em que λ é o parâmetro da Poisson e α e β são os parâmetros da distribuição Gama. Se α e β forem ambos iguais a zero temos uma distribuição de Poisson. Se λ for igual a zero temos uma distribuição Binomial Negativa.

Uma parametrização da função de distribuição de probabilidade da DEL é definida por

$$f(\alpha, \beta, \lambda) = \sum_{i=0}^k \frac{\Gamma(\alpha + i) \beta^i \lambda^{k-i} \exp^{-\lambda}}{\Gamma(\alpha) i! (1 + \beta)^{\alpha+i} (k-i)!}, \quad (4.9)$$

em que $\lambda > 0$, $\alpha > 0$, $\beta > 0$ e $k \in 0, 1, 2, \dots, \infty$.

4.6.4 Distribuição Binomial Negativa

Nesta seção são descritas duas estruturas da distribuição de probabilidade Binomial Negativa: a generalizada e a canônica.

A distribuição de probabilidade Binomial Negativa generalizada (NB-P) foi introduzida por Greene (1994). Nessa distribuição a variância é uma generalização da variância da distribuição de probabilidade Binomial Negativa, definida por $\mu + \frac{\mu^p}{\nu}$, em que μ é a variância da distribuição Poisson e $\frac{\mu^p}{\nu}$ é a variância da distribuição Gama, ν é o parâmetro de dispersão e p é um parâmetro a ser estimado (MCCULLAGH; NELDER, 1989; HILBE, 2011a).

Em particular, são definidas as funções de probabilidade das distribuições Binomial Negativa tipo I (NBI) e Binomial Negativa tipo II (NBII), respectivamente,

$$f(y; \nu, \mu) = \binom{\nu + y - 1}{\nu - 1} \left(\frac{\nu/\mu}{\nu/\mu + 1} \right)^y \left(\frac{1}{\nu/\mu + 1} \right)^\nu, \quad (4.10)$$

em que $\nu > 1$, $\mu > 0$ e

$$f(y; \nu, \mu) = \binom{\nu\mu + y - 1}{\nu\mu} \left(\frac{\nu}{\nu + 1} \right)^y \left(\frac{1}{\nu + 1} \right)^{\nu\mu}, \quad (4.11)$$

em que $y = 1, \dots, \nu$, $\nu > 0$ e $\mu > 0$.

De modo que a distribuição Binomial Negativa tipo I é denotada por $NBI(Y; \nu, \frac{\nu}{\mu})$, com parâmetros de forma ν e de escala $\frac{\nu}{\mu}$ e a distribuição Binomial Negativa tipo II é denotada por $NBII(Y; \nu\mu, \nu)$, com parâmetros de forma $\nu\mu$ e de escala ν .

A distribuição de probabilidade Binomial Negativa Canônica (NB-C) é baseada na função de probabilidade da Binomial Negativa com probabilidade $\frac{1}{(1+\nu\mu)}$ e função de ligação $\frac{1}{(\exp(-\eta)-1)^\nu}$, em que η é o preditor linear. No software R, os parâmetros do modelo NB-C podem ser estimados a partir da função `m1.nbc` do pacote `COUNT`. Ressalta-se que a construção de um modelo de regressão Binomial Negativa exato deve ser baseado na parametrização do modelo NB-C (HILBE, 2011a). Uma formulação para a função de máxima verossimilhança maximizada do modelo NB-C é dada

$$\ell_{NB-C} = \sum_{i=1}^n y_i(\eta) + \frac{1}{\nu} \ln(1 - \exp(\eta)) + \ln\Gamma(y_i + \frac{1}{\nu}) - \ln\Gamma(y_i + 1) - \ln\Gamma(\frac{1}{\nu}). \quad (4.12)$$

4.6.5 Distribuição Poisson Inflacionada de Zeros

Na classe de modelos inflacionados de zeros supõe-se que os dados são gerados de um processo dual: um estágio de zero e um estágio de não zero. Usualmente a probabilidade da ocorrência do estágio zero é estimada por meio de um modelo de regressão binário ou logístico e a probabilidade da ocorrência do estágio não zero é estimada por meio de um modelo de regressão de contagem, sendo que os pesos atribuídos para cada distribuição devem ser complementares.

No caso, da distribuição Poisson inflacionada de zeros (ZIP) temos uma mistura entre um componente do modelo de contagem e um componente do modelo binário, por exemplo: a função de probabilidade da ZIP pode ser composta da mistura entre a distribuição Logística binária e a distribuição de Poisson, (HILBE, 2011a).

Uma parametrização da função de probabilidade da ZIP é dada por

$$P(Y = k) = \begin{cases} w + (1 - w)\exp(-\mu), & \text{se } k = 0 \\ (1 - w)\exp(-\mu)\frac{\mu^y}{y!}, & \text{se } k > 0 \end{cases},$$

em que w é a probabilidade de mistura e $0 \leq w \leq 1$. Neste caso, a contagem de observações diferentes de zero tem distribuição de Poisson com parâmetro μ (MORGAN; PALMER; RIDOUT, 2007).

Em (LORD; WASHINGTON; IVAN, 2007) é discutido a respeito da lógica dessa classe de modelos, as quais algumas questões são levantadas, por exemplo: (i) como se deve proceder se as características específicas que classificam os dois estágios não são identificadas? e (ii) para analisar os dois estágios simultaneamente, poderíamos utilizar apenas um modelo binário ao em vez de analisar os estágios de forma independente, considerando as condições de divisão apropriadas entre os estágios? Além disso, o autor destaca que pelo que se conhece pouco se discute com relação aos problemas lógicos dos modelos inflacionados de zeros na modelagem de dados de seguros.

4.6.6 Distribuição Poisson Inversa Gaussiana

A distribuição de probabilidade Poisson Inversa Gaussiana (PIG) consiste em uma mistura entre a distribuição de Poisson e a distribuição Inversa Gaussiana com parâmetros μ e σ , denotada por $PIG(\mu, \sigma)$. A PIG é uma apropriada distribuição para dados com alta assimetria positiva (ATKINSON, 1982).

Uma parametrização da função da PIG é definida por

$$f_Y(y; \mu, \sigma) = \left(\frac{2\alpha}{\pi}\right)^{\frac{1}{2}} \frac{\mu^y \exp\left(\frac{1}{\sigma}\right) K_{y-\frac{1}{2}}(\alpha)}{(\alpha\sigma)^y y!}, \quad (4.13)$$

em que $K_\lambda(t) = \frac{1}{2} \int_0^\infty x^{\lambda-1} \exp\left(-\frac{1}{2}t(x+x^{-1})\right) dx$ é a função de Bessel modificada de terceiro tipo, $\alpha^2 = \frac{1}{\sigma^2} + \frac{2\mu}{\sigma}$, $y \geq 0$, $\mu > 0$ e $\sigma > 0$.

4.6.7 Distribuição Sichel

A distribuição Sichel foi introduzida por Sichel (1975) com o objetivo de modelar contagens de palavras. Ela consiste em uma mistura entre a distribuição de Poisson e a distribuição Generalizada Inversa Gaussiana (GIG). A distribuição Sichel pertence à família de distribuições para dados de contagem com três parâmetros com alta assimetria positiva e é denotada por $Y|Z \sim PO(\mu Z)$, em que $Z \sim GIG(1, \sigma, \nu)$.

A PIG é um caso particular da SI quando o parâmetro de forma ν é igual a $-\frac{1}{2}$. Uma parametrização da distribuição Sichel, além de mais detalhes a respeito, estão descritos na Seção (6.2).

No próximo capítulos são apresentados os procedimentos e resultados da aplicação das técnicas descritas nos capítulos 3 e 4 aos *Dados 1* e *Dados 2*.

5 *Aplicação*

Neste capítulo, são apresentados os resultados da aplicação das técnicas descritas nos capítulos 3 e 4, ressaltando que as técnicas apresentadas no Capítulo 3 não foram empregadas ao conjunto *Dados 2*, devido à presença de raiz unitária em algumas de suas séries, como mostrado na Tabela 2.6. Já as técnicas apresentadas no Capítulo 4 foram empregadas aos conjuntos *Dados 1* e *Dados 2*.

5.1 **Estimação de modelo - VAR**

Nesta seção são mostrados os resultados obtidos mediante o emprego do modelo VAR. A seção foi dividida em quatro etapas. A primeira etapa consiste na seleção do número de defasagens p , que foi usado posteriormente no ajuste do modelo. A segunda etapa consiste no ajuste do modelo e na seleção das covariáveis. A terceira etapa consiste na análise de diagnóstico. A quarta e última etapa consiste na análise do comportamento previsto das observações futuras da variável resposta.

5.1.1 **Escolha do número de defasagens**

Para dar início à modelagem do problema proposto, foram calculados os valores de quatro critérios de informação. Os menores valores apresentados dos critérios se referem ao número ótimo de defasagens a ser considerado no ajuste do modelo, ou seja, será considerado que a observação de cada variável no instante de tempo t depende da sua observação no instante de tempo $t - 1$ (Tabela 5.1).

Tabela 5.1: Valores dos critérios para obtenção do número de defasagens.

p	AIC	HQ	BIC	FPE
1	2.37	2.44	2.54	1.94
2	2.38	2.50	2.69	2.12
3	2.38	2.56	2.82	2.23
4	2.37	2.61	2.96	2.23

```
R > library("vars")
R > library("urca")
R > library("MSBVAR")
R > VARselect(dados, lag.max = 8, type = "both")
```

5.1.2 Análise do ajuste do modelo

O modelo consiste em um VAR periódico, isto é, um PVAR devido à adição de uma componente sazonal com ciclo de doze meses no ajuste do modelo.

Analisando o gráfico da série original e o gráfico da série prevista pelo modelo PVAR (linha tracejada), conclui-se que o ajuste do modelo em geral é satisfatório. Entretanto, os picos de maior variabilidade da série original não foram previstos pelo modelo (Figura 5.1).

Para selecionar as covariáveis relevantes ao modelo, foi realizado um procedimento manual, que consiste no ajuste do modelo com todas as covariáveis (modelo saturado). Em seguida analisou-se o p -valor de cada covariável e retirou-se do modelo a covariável com p -valor $> \alpha$, em que α é previamente fixado. Esse procedimento foi realizado até que todas as covariáveis presentes no modelo fossem estatisticamente significativas a um nível de significância $\alpha = 0.05$, ou seja, permaneceram no modelo apenas as covariáveis que ao final do processo apresentaram p -valor < 0.05 .

Em particular a quantidade mensal esperada de dias desfavoráveis à dispersão de poluentes na atmosfera é 3.55 no instante t , se considerarmos que no instante $t - 1$ ocorreu apenas um dia desfavorável e uma inversão térmica com altitude entre (0-200) metros, Equação (5.2).

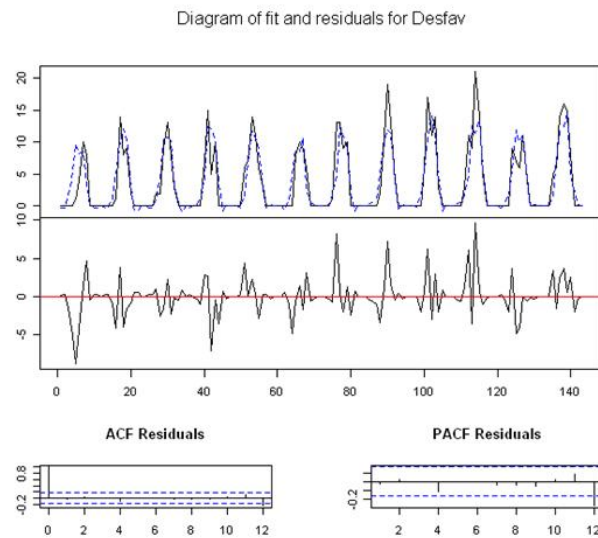


Figura 5.1: Diagrama de ajuste e resíduos do modelo.

Tabela 5.2: Ajuste do modelo PVAR.

	Intercepto	Y_{t1}	Inversão(0-200) $_{t1}$	
Y	3.5	0.36	-0.31	p-valor <0.01
Inversão(0-200)	2.38	2.50	2.69	

```
R > model <- VAR(dados, p = j, type = c("const", "trend",
    "both", "none"), season = s, exogen = NULL,
    lag.max = NULL,
    ic = c("AIC", "HQ", "SC", "FPE"))
R > plot(model, names = "variável")
```

em que j é o número de defasagens e s é o período de sazonalidade da série.

5.1.3 Análise de diagnóstico

Para a análise de diagnóstico do modelo ajustado PVAR são interpretados o gráfico dos valores esperados dos resíduos, a função de densidade dos resíduos, os gráficos de autocorrelações *versus* defasagens (correlograma) e os três testes de hipótese: o teste Portmanteau, o teste Jarque Bera e o modelo ARCH.

Com relação às hipóteses de não correlação (p -valor = 0.14) e de homoscedasticidade dos resíduos (p -valor = 1), podemos concluir que não há evidências para rejeitar essas hipóteses.

No entanto, com relação às hipóteses de normalidade, de ausência de curtoses e simetria (p -valor < 0.001), podemos concluir que há evidências para rejeitar essas hipóteses (Tabela 5.3).

Tabela 5.3: Testes de normalidade, de homocedasticidade e de independência dos resíduos.

	χ^2	valor-p
Portmanteau	1007.6	0.14
JB	516.53	< 0.001
Skewness	77.95	< 0.001
Curtoses	438.6	< 0.001
ARCH	4968	1

```
R > ser <- serial.test(model, lags.pt = 16, type = "PT.asymptotic")
R > ser$serial
R > norm <- normality.test(model)
R > norm$jb.mul
R > arch <- arch.test(model, lags.multi = 5)
R > arch$arch.mul
```

Nos gráficos de diagnóstico dos resíduos do modelo PVAR(1) observa-se que a *fac* (função de autocorrelação) tem uma correlação não nula afastada (na defasagem doze) que no entanto não apresenta grandes problemas. A suposição de normalidade dos resíduos foi rejeitada de modo que as estimativas dos parâmetros do modelo podem ser instáveis (Figura 5.2). Os resíduos do modelo PVAR(1) são relativamente grandes se encontram em torno de (-10,10).

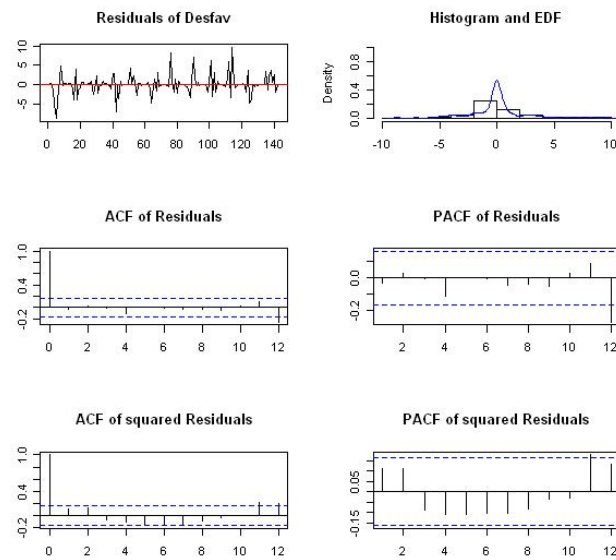


Figura 5.2: Resíduos do modelo.

```
R > plot(arch, names = "variável")
```

5.1.4 Análise da previsão

Para análise do comportamento estimado futuro da variável resposta, foram considerados seis diferentes períodos de tempo.

Inicialmente se considerou as observações dos anos 1999 a 2009 para estimar o comportamento do ano de 2010. As previsões obtidas para 2010 são subestimadas, devido o modelo VAR utilizar de um método recursivo e temos que os anos 1999 a 2005 tiveram menos de 15 ocorrências do evento por mês. Já para os anos de 2006 a 2008 a frequência do evento aumentou mostrando picos entre 15 a 23 dias por mês. No entanto quando se considerou os anos de 2006 a 2009, nota-se uma melhor aproximação do estimado com o previsto ver (figuras 5.3 a 5.5) e (CETESB, 2001-2012), destaca-se que usualmente a maior frequência do evento ocorre nos meses de junho a agosto.

No geral, ao se comparar a série original com a série prevista tem-se que o modelo apresenta boas previsões para os próximos doze meses, isto é um ciclo, independente do número de valores passados e do ano a qual se deseja estimar.

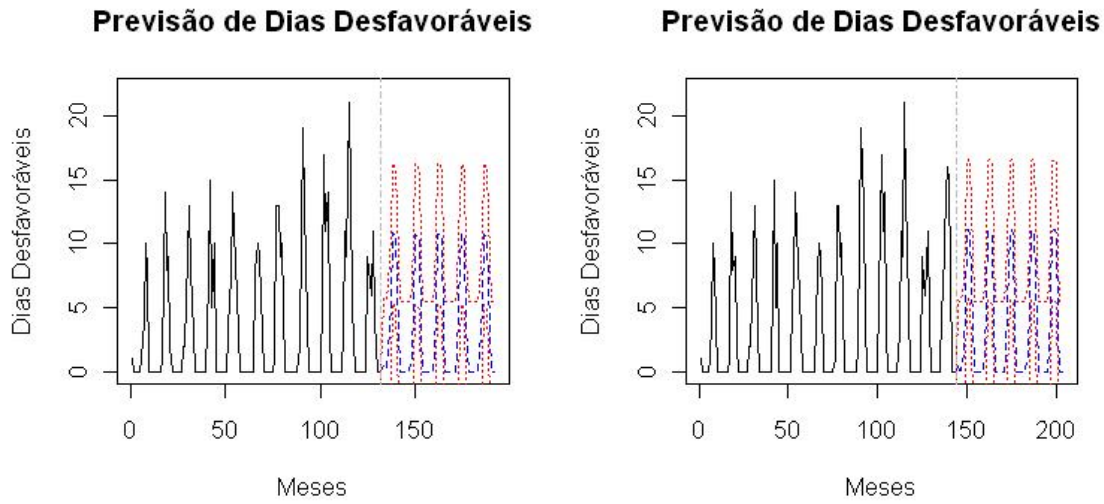


Figura 5.3: Previsão para os dias desfavorável, de jan/1999 a dez/2009 e de jan/1999 a dez/2010.

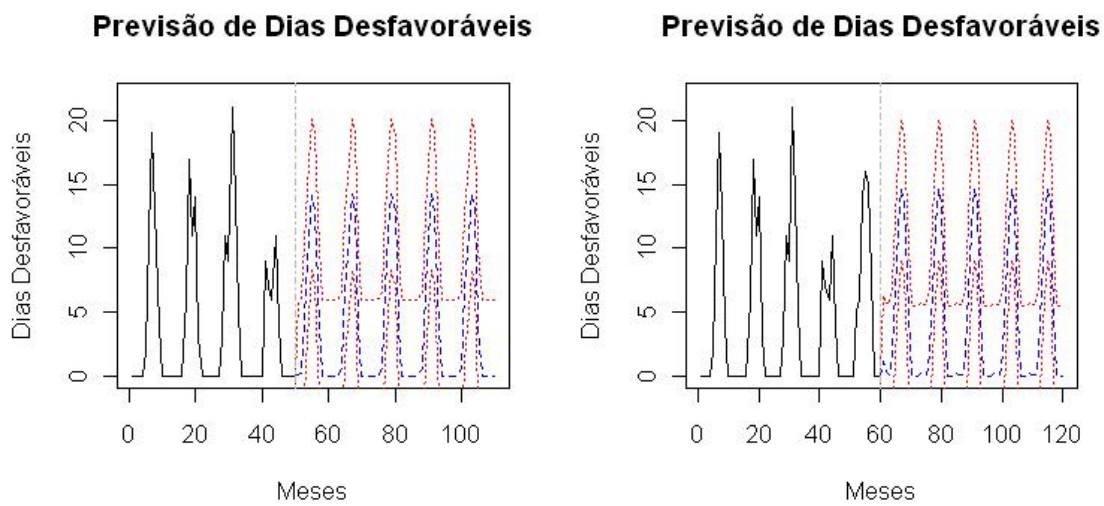


Figura 5.4: Previsão para os dias desfavorável, de jan/2006 a dez/2009 e de jan/2006 a dez/2010.

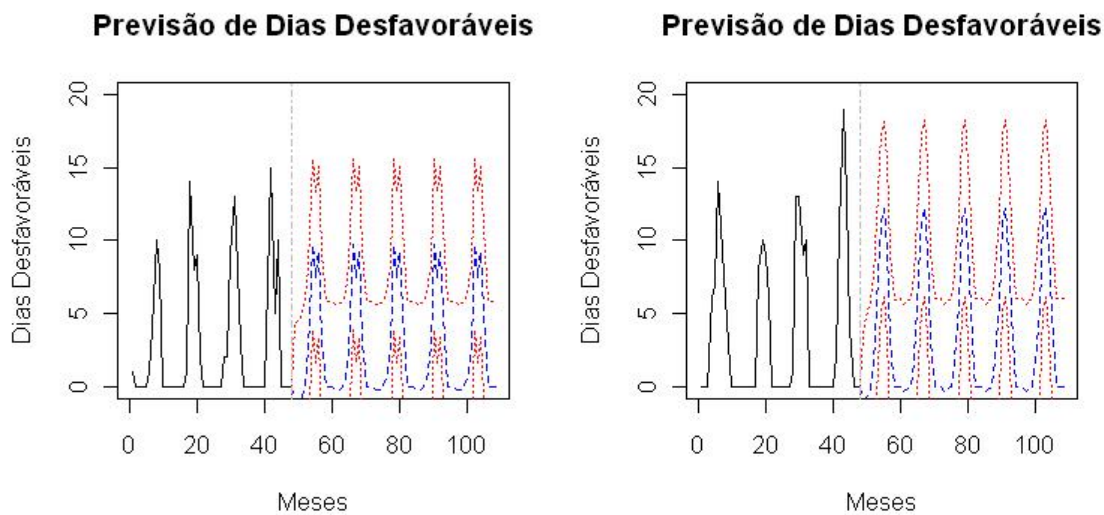


Figura 5.5: Previsão para os dias desfavorável, de jan/1999 a dez/2002 e de jan/2003 a dez/2006.

```
R > predictions <- predict(model, n.ahead = 60, ci = 0.95)
R > plot(predictions, names="variável resposta", xlab="Meses",
         ylab="variável resposta", main= "Previsão da variavel",
         ylim=c(0,22))
```

5.2 Estimação do modelo - GAMLSS

Nesta seção são apresentados os resultados obtidos mediante o uso dos modelos da classe GAMLSS aos dois conjuntos de dados investigados. A seção foi dividida em três etapas. A primeira etapa consiste nos ajustes dos sete modelos e avaliação da qualidade dos ajustes. A segunda etapa consiste na seleção das covariáveis e nas análises das estimativas dos parâmetros dos modelos. A terceira e última etapa consiste nas análises de diagnóstico.

5.2.1 Seleção de modelos

Nesta seção são utilizados quatro critérios de informação: AIC, BIC, AICc e Deviance (Seção A.1). Inicialmente foi realizada uma análise prévia do ajuste da variável resposta sem covariáveis e foi observado que, dentre as sete distribuições analisadas, a que melhor se ajusta às observações da variável resposta tanto do período anual quanto do período de maio a setembro é a distribuição Binomial Negativa log linear, a qual apresentou menor valor nos critérios (tabelas B.1 e C.1).

Com relação aos ajustes dos sete modelos de regressão a variável resposta, a distribuição Binomial Negativa log linear melhor se ajustou às observações anuais. Para os dados de maio a setembro o modelo de regressão Poisson Inflacionado de Zeros log linear apresentou menor valor nos critérios (tabelas 5.4 e 5.5). Nesse caso, não ocorreu a convergência dos parâmetros do modelo Sichel, possivelmente devido ao tamanho e ao número de variáveis do conjunto *Dados 2*.

Tabela 5.4: Valores dos critérios de informação - dados de jan/1999 a dez/2010.

	DEL	NBI	NBII	PO	ZIP	PIG	SI
AIC	32.6	38.1	0.0	129.5	64.7	33.9	35.8
BIC	35.5	38.1	0.0	126.5	64.7	33.9	38.7
AICc	32.9	38.1	0.0	129.2	64.7	33.9	36.1
Deviance	480.7	488.2	450.1	581.6	514.8	484.0	483.9

Tabela 5.5: Valores do critério de informação - dados de maio a setembro, 2001 a 2010.

	DEL	NBI	NBII	PO	ZIP	PIG
AIC	11.5	9.5	6.7	7.5	0.0	9.5
BIC	13.4	9.5	6.7	5.6	0.0	9.5
AICc	11.8	9.5	6.7	7.2	0.0	9.5
Deviance	264.2	264.2	261.4	264.2	254.7	264.2

```

R > library(gamlss)
R > con <- gamlss.control(trace = FALSE)
R > M1 <- gamlss(Y ~ X1 + X2 + ... + Xp, data=dados, family = PO,
               link = c(log, log), method = RS(), control = con)
R > M2 <- gamlss(Y ~ X1 + X2 + ... + Xp, data=dados, family = NBI,
               link = c(log, log), method = RS(), control = con)
R > M3 <- gamlss(Y ~ X1 + X2 + ... + Xp, data=dados, family = NBII,
               link = c(log, log), method = RS(), control = con)
R > M4 <- gamlss(Y ~ X1 + X2 + ... + Xp, data=dados, family = PIG,
               link = c(log, log), method = RS(), control = con)
R > M5 <- gamlss(Y ~ X1 + X2 + ... + Xp, data=dados, family = SICHEL,
               link = c(log, log, identity), method = RS(),
               control = con)
R > M6 <- gamlss(Y ~ X1 + X2 + ... + Xp, data=dados, family = DEL,
               link = c(log, log, logit), method = RS(),
               control = con)
R > M7 <- gamlss(Y ~ X1 + X2 + ... + Xp, data=dados, family = ZIP,
               link = c(log, logit) , method = RS(),
               control = con)

R > library(bbmle)
R > AIC <- ICTab (M1, M2,\ldots, Mj, type = c("AIC"), weights = TRUE,
               delta = TRUE, sort = TRUE, nobs= n)
R > BIC <- ICTab (M1, M2,\ldots, Mj, type = c("BIC"), weights = TRUE,
               delta = TRUE, sort = TRUE, nobs= n)
R > AICc <- ICTab (M1, M2,\ldots, Mj, type = c("AICc"),
               weights = TRUE, delta=TRUE, sort=TRUE, nobs= n)

```

em que M_j é o j -ésimo modelo, j é o número de modelos comparados e n é o número de observações ou número total de linhas do conjunto de dados.

5.2.2 Análise do ajuste - conjunto de dados anual

Nesta seção são apresentadas e analisadas as estimativas do modelo de regressão Binomial Negativa tipo II log linear. Para a seleção das covariáveis foram realizados os mesmos procedimentos descritos na Seção 5.1.2.

Se considerarmos que em um mês não ocorreram dias com precipitação pluviométrica e dias com ultrapassagem do padrão do ozônio e ocorreram uma inversão térmica com altitude entre (0-200) e uma entre (201-500) em metros é esperado um acréscimo de 22% na média estimada da variável resposta. Por outro lado, se considerarmos que em um mês ocorreram um dia com precipitação pluviométrica e um dia com ultrapassagem do padrão do ozônio e sem a ocorrência de inversão térmica com altitude entre (0-200) e entre (201-500) em metros é esperado um decréscimo de 20% na média estimada da variável resposta (Tabela 5.6).

O valor do parâmetro de dispersão para o ajuste sem covariáveis é aproximadamente 3.18 vezes maior se comparado ao ajuste com covariáveis (tabelas B.3 e 5.6).

A relação entre a variável resposta e a quantidade de dias com ultrapassagem do padrão de ozônio é inversamente proporcional, esta relação esta vinculada com o período do ano. Períodos do ano com maior quantidade de radiação solar se têm maiores quantidades de ozônio devido ser um poluente a qual é fruto de uma reação fotoquímica. No entanto, tem-se menores quantidades de dias desfavoráveis, porque nestes períodos do ano a superfície está mais aquecida, de modo que as ocorrências de inversões térmicas com baixas altitude são menores, além dos fenômenos que auxiliam na dispersão dos poluentes ocorrem com maior frequência, como a precipitação pluviométrica.

Na terceira coluna das tabelas (5.6 e 5.7) a estimativa para a média da variável resposta é dada, considerando um parâmetro significativo e os demais nulos.

$$\log(\hat{\mu}) = 1.1 - 0.12Dias.chuva + 0.12IB + 0.08IM - 0.1Dias.ozonio. \quad (5.1)$$

ou

$$\hat{\mu} = \exp(1.1 - 0.12Dias.chuva + 0.12IB + 0.08IM - 0.1Dias.ozonio). \quad (5.2)$$

Tabela 5.6: Modelo de regressão NB estimado.

Parâmetro	Variável	Estimativa	Exp(estimativa)	Erro	p-valor
$\hat{\mu}$	Intercepto	1.1	3	0.53	0.04
	Dias com chuva	-0.12	0.89	0.03	< 0.001
	Inversão (0-200)	0.12	1.13	0.02	< 0.001
	Inversão (201-500)	0.08	1.09	0.04	0.03
	Dias com ultr. O_3	-0.1	0.9	0.03	0.002
$\hat{\sigma}$		1.36	3.89	0.2	< 0.001

```
R > library(gamlss)
R > con <- gamlss.control(trace = FALSE)
R > M3 <- gamlss(Y ~ X1 + X2 + ... + Xp, data=dados, family = NBII,
               link = c(log, log) , method = RS(), control = con)
```

em que Y é a variável resposta, X_p é a p -ésima covariável, p é o número de covariáveis no modelo e dados é o conjunto de dados.

5.2.3 Análise do ajuste - conjunto de dados de maio a setembro

Nesta seção são apresentadas e analisadas as estimativas do modelo de regressão Poisson Inflacionada de Zeros log linear. Para a seleção das covariáveis foram realizados os mesmos procedimentos descritos na Seção 5.1.2.

Se ocorrer uma inversão térmica com altitude entre (0-200) em metros e não ocorrer dias com precipitação pluviométrica e nem ocorrer inversão térmica com altitude \geq 500 em um mês é esperado um acréscimo de 6,4% na média estimada da variável resposta. Caso contrário, neste caso é esperado um decréscimo de 4,7% na média estimada da variável resposta (Tabela 5.7).

O valor esperado de dias desfavoráveis do período de maio a setembro é aproximadamente 2.7 vezes maior se comparado ao valor esperado de dias desfavoráveis anual.

$$\log(\hat{\mu}) = 2.4 - 0.03Dias.chuva + 0.06IB - 0.02IA. \quad (5.3)$$

ou

$$\hat{\mu} = \exp(2.4 - 0.03Dias.chuva + 0.06IB - 0.02IA). \quad (5.4)$$

Tabela 5.7: Modelo de regressão ZIP estimado.

Parâmetro	Variável	Estimativa	Exp(estimativa)	Erro	p-valor
$\hat{\mu}$	Intercepto	2.4	11	0.19	< 0.001
	Dias com chuva	-0.03	0.97	0.01	0.04
	Inversão (0-200)	0.06	1.06	0.01	< 0.001
	Inversão (>500)	-0.02	0.98	0.007	0.001
$\hat{\sigma}$		-2.55	$\left(\frac{\exp(estimativa)}{1-\exp(estimativa)}\right)$ 0.08	0.6	< 0.001

```
R > library(gamlss)
R > con <- gamlss.control(trace = FALSE)
R > M < gamlss(Y ~ X1 + X2 + ... + Xp, data=dados, family = ZIP,
              link = c(log, logit), method = RS(), control = con)
```

em que Y é a variável resposta, X_p é a p -ésima covariável, p é o número de covariáveis no modelo e $dados$ é o conjunto de dados.

5.2.4 Análise de diagnóstico - conjunto de dados anual.

Nesta seção são analisados os gráficos de envelope, os gráficos QQ, os gráficos de resíduos *versus* valores ajustados, os gráficos de resíduos *versus* índices, os gráficos das estimativas dos núcleos das densidades, os testes de hipótese de Filliben, os valores esperados e as dispersões dos resíduos.

No gráfico de envelope (Figura 5.6) observa-se que não há grandes afastamentos dos resíduos do zero e nenhum ponto se encontra fora da região de confiança; nos gráficos de resíduos *versus* valores ajustados e resíduos *versus* índices, nota-se que os pontos estão distribuídos aleatoriamente. Além disso não há pontos afastados do intervalo $(-2, 2)$, descartando-se assim a possibilidade de qualquer comportamento e da presença de pontos aberrantes. No gráfico da estimativa do núcleo da densidade, nota-se que os resíduos aparentam estar distribuídos simetricamente em torno do zero e no gráfico QQ os pontos estão distribuídos em uma reta (Figura

5.7). Contudo, os resíduos se comportam bem para o modelo com estrutura de regressão. No entanto o mesmo não se pode afirmar para o modelo sem covariáveis (figuras B.4 e B.5).

Analisando-se as medidas resumo podemos concluir que as médias estimadas dos resíduos estão próximas do zero com dispersão estimada próxima do um e com relação à assimetria e à curtoses, as estatísticas de teste não apresentaram grandes afastamentos do zero e do três, respectivamente. As estatísticas do teste de Filliben estão próximas do um, então há evidências que os resíduos são normalmente distribuídos (tabelas B.4 e 5.8).

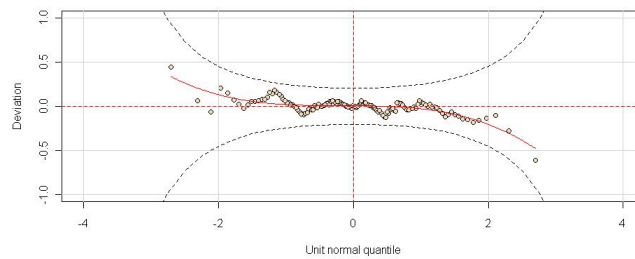


Figura 5.6: Desvios *versus* quantis normalizados (gráfico de envelope) - modelo de regressão NB.

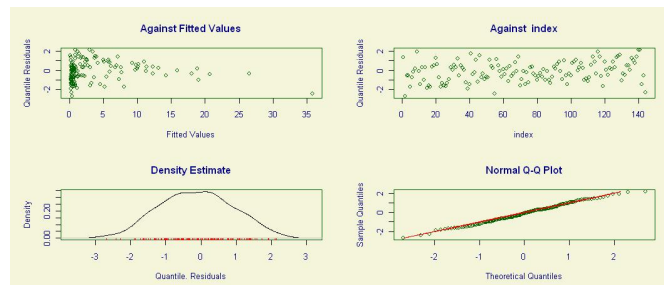


Figura 5.7: Resíduos *versus* valores ajustados, resíduos *versus* índices, estimativa do núcleo da densidade e quantis amostrais *versus* quantis teóricos (gráfico QQ) - modelo de regressão NB.

Tabela 5.8: Resumo dos quantis dos resíduos para o ajuste do modelo de regressão NB.

$\hat{\mu}$	$\hat{\sigma}^2$	$\hat{A}(Y)$	$\hat{K}(Y)$	Teste de Filliben
-0.12	1.04	0.006	2.4	1

```
R > wp(M)
```

```
R > plot(M)
```


5.2.5 Análise de diagnóstico - conjunto de dados de maio a setembro

Nesta seção são analisados os gráficos de envelope, os gráficos QQ, os gráficos de resíduos *versus* valores ajustados, os gráficos de resíduos *versus* índices, os gráficos das estimativas do núcleo da densidade, os testes de hipótese de Filliben, os valores esperados e as dispersões dos resíduos.

As conclusões aqui são análogas às apresentadas na Seção 5.2.4, com exceção do gráfico QQ que apresenta pontos na cauda inferior afastados da reta e o gráfico da estimativa do núcleo da densidade, que apresenta uma deformação na cauda inferior (figuras C.3, C.4, 5.8 e 5.9 e Tabela 5.9).

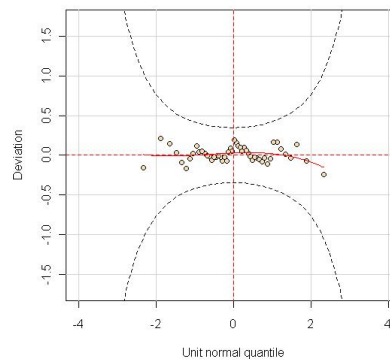


Figura 5.8: Desvios *versus* quantis normalizados (gráfico de envelope) - modelo de regressão ZIP.

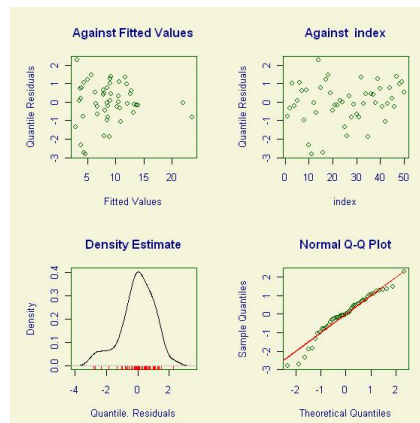


Figura 5.9: Resíduos *versus* valores ajustados, resíduos *versus* índices, estimativa do núcleo da densidade e quantis amostrais *versus* quantis teóricos (gráfico QQ) - modelo de regressão ZIP.

Tabela 5.9: Resumo dos quantis dos resíduos para o ajuste do modelo ZIP.

$\hat{\mu}$	$\hat{\sigma}^2$	$\hat{A}(Y)$	$\hat{K}(Y)$	Teste de Filliben
-0.02	1.17	-0.6	3.3	1

```
R > wp(M)
```

```
R > plot(M)
```

5.3 Comentários finais

O processo de busca pelo modelo que melhor se adéqua aos dados: análise dos dados, seleção de modelos, estimativa de parâmetros, ajuste de tendências temporais, avaliação do ajuste e previsão apresentados neste trabalho se remetem principalmente a análise de multivariadas séries temporais.

A análise dos resíduos do modelo é uma etapa essencial para verificar se o modelo bem representa os dados. No ajuste do modelo PVAR(1) temos altos resíduos, isto é, resíduos entre (-10,10), mas são esperados resíduos entre (-2,2), indicando mau ajuste. No entanto, as previsões obtidas desse ajuste são boas, ou seja, são próximas aos valores reais.

O modelo Binomial Negativa é um modelo muito bem sucedido na literatura para modelagem de dados de contagem superdispersos e entre os modelos para dados de contagem estudados foi selecionado inicialmente de acordo com os critérios de seleção para a modelagem dos *Dados 1*. A suposição de adequação foi confirmada na análise dos resíduos, na qual se obteve resíduos pequenos e bem comportados. Já o ajuste do modelo Poisson Inflacionado de Zeros aos *Dados 2* os resíduos apresentam uma leve assimetria a direita, isto é, apresenta valores mais extremos na cauda inferior.

No próximo capítulos é apresentado o estudo de simulação, na qual procurou-se investigar a relação entre os modelos para dados de contagem e identificar a qualidade do ajuste desses modelos a dados gerados de três distribuições Binomial Negativa.

6 *Experimento de simulação*

Neste capítulo é descrito o estudo de simulação. Para o ajuste dos modelos foram utilizadas funções que se encontram implementadas no pacote `gamlss` do software R. A função com a qual se obteve os resultados finais do estudo de simulação foi construída em linguagem C (Apêndice D).

6.1 Descrição dos algoritmos de geração de dados

Nesta seção são descritos os três procedimentos utilizados para a construção dos códigos de geração de dados da distribuição Binomial Negativa. Os códigos de geração de dados são baseados na geração de números pseudo aleatórios, por exemplo: na função `runif()`, `rbinom()` e `rpois()`.

- (i) O primeiro algoritmo está disponível no endereço eletrônico <http://stats.stackexchange.com/questions/9767/generating-over-dispersed-counts-data-with-serial-correlation>. O algoritmo consiste na geração de dados de uma variável aleatória Y que segue uma distribuição de Poisson com parâmetro λ , em que λ também é uma variável aleatória que segue uma distribuição D . Se a distribuição D for uma distribuição Gama, então Y segue uma distribuição Binomial Negativa com parâmetros de locação μ e de escala σ , denotada por $Y \sim NB(\mu, \sigma)$. A correlação serial foi incorporada ao considerar que o parâmetro $\log \lambda$ segue um modelo autorregressivo com uma defasagem, denotado por $\log \lambda \sim AR(1)$ para um λ a priori normalmente distribuído.
- (ii) O segundo algoritmo foi obtido em Hilbe (2011a). Em geral, os conceitos adotados na construção desse algoritmo são os mesmos adotados no algoritmo (i). No entanto, nesta geração de dados não se considerou a correlação serial e o parâmetro λ é obtido de um modelo de regressão linear com duas variáveis explicativas geradas de uma distribuição Uniforme no intervalo de zero a um.

- (iii) O terceiro e último código consiste em uma função implementada no software R, `rnbinom()`, que gera dados de uma distribuição Binomial Negativa Canônica.

6.2 Análise teórica dos modelos

Nesta seção são descritas as relações teóricas, caso existam, entre os modelos estudados da classe GAMLSS. Como descrito na Seção 6.1, é esperado que os dados sejam gerados de uma distribuição Binomial Negativa, então se espera que sejam dados superdispersos. No entanto, em alguns casos esses dados podem apresentar uma quantidade de zeros e/ou uma quantidade de valores extremos maior do que é esperado em uma distribuição Binomial Negativa. Apesar dessa distribuição ser capaz de capturar bem a taxa de superdispersão, existem casos de dados com excesso de zeros e/ou longas caudas, que ela não se ajusta bem. Por esta razão é que foram consideradas as distribuições Delaporte, Poisson Inversa Gaussiana e Sichel que são mais flexíveis quando comparadas às distribuições Binomial Negativa e Poisson Inflacionada de Zeros.

As distribuições Delaporte, Poisson Inversa Gaussiana e Sichel, assim como a distribuição Binomial Negativa consistem, de certa forma, em uma mistura ou em uma composição da distribuição de Poisson com a distribuição Generalizada Inversa Gaussiana. Ressaltando que na definição original da distribuição Binomial Negativa é pressuposto que o número de ocorrências de um determinado evento em sucessivos e iguais períodos de tempo é estacionário, independente e tem distribuição de Poisson com média constante por unidade de tempo λ , em que λ é uma variável aleatória com distribuição Gama. A distribuição Gama pode ser derivada da distribuição Generalizada Inversa Gaussiana. Uma parametrização da GIG é definida por

$$f(\lambda) = \frac{\left[\frac{2(1-\theta)^{\frac{1}{2}}}{\alpha\theta}\right]^{\gamma} \lambda^{\gamma-1} \exp^{-\left(\left[\frac{1}{\theta}\right]-1\right)\lambda - \left(\frac{\alpha^2\theta}{4\lambda}\right)}}{2K_{\gamma}[\alpha(1-\theta)^{\frac{1}{2}}]}, \quad (6.1)$$

em que $-\infty < \gamma < \infty$, $0 \leq \theta \leq 1$, $\alpha \geq 0$, $K_{\gamma}(z)$ é a função modificada de Bessel do segundo tipo de ordem γ com argumento z , e λ é a curtoses da distribuição.

A distribuição Generalizada Inversa Gaussiana é um modelo encaixado, ou seja, outras distribuições são derivadas como casos particulares dela fixado o domínio de variação de seus parâmetros. Segue abaixo o intervalo de variação dos parâmetro, na qual obtêm algumas das distribuições estudadas.

- Se $\alpha > 0$ e $-\infty < \gamma < \infty$. Então, $\lambda \sim GIG(\theta, \alpha, \gamma)$;

- Se $\alpha \rightarrow 0$ e $\gamma > 0$. Então, $\lambda \sim \text{Gama}(\alpha, \gamma)$;
- Se $0 < \gamma < 1$ tem-se que $\lambda \sim \text{Gama modificada}(\theta, \alpha, \gamma)$;
- Se $\alpha > 0$ e $\gamma = \frac{-1}{2}$. Então, $\lambda \sim \text{Inversa Gaussiana}(\sigma, \gamma)$;
- Se $\gamma < 0$ e $\theta \rightarrow 1$ temos um grande número de distribuições contínuas de grande interesse prático.

De acordo com (SICHEL, 1971), após reconstruir a teoria da distribuição Generalizada Inversa Gaussiana e identificar a flexibilidade dessa distribuição e comparar com a teoria original da distribuição Binomial Negativa, sugeriu que o parâmetro da distribuição Poisson fosse modelado por uma distribuição GIG ao invés da distribuição Gama. A mistura entre as distribuições foi denotada distribuição Sichel e uma parametrização para essa distribuição é definida por

$$\phi(r|t) = \int_0^{\infty} p(r|t\lambda)f(\lambda)d\lambda = \frac{(1 - \theta_t)^{\frac{\gamma}{2}} (\frac{\alpha_t \theta_t}{2})^r K_{r+\gamma}(\alpha_t)}{K_{\gamma}(\alpha_t (1 - \theta_t)^{\frac{1}{2}}) r!}, \quad (6.2)$$

$$\alpha_t = \alpha [1 + (t - 1)\theta]^{\frac{1}{2}}, \quad (6.3)$$

$$\theta_t = \frac{t\theta}{1 + (t - 1)\theta}. \quad (6.4)$$

Analogamente, se considerarmos os mesmos domínios de variação para os parâmetros adotados para a distribuição GIG, obtêm-se diferentes distribuições discretas, a partir da distribuição Sichel.

- Se $\gamma > 0$ e $\alpha \rightarrow 0$ tem-se uma distribuição Binomial Negativa;
- Se $0 < \gamma < 1$ tem-se uma distribuição Delaporte;
- Se $\alpha > 0$ e $\gamma = \frac{-1}{2}$ tem-se uma distribuição Poisson Inversa Gaussiana;
- Se $\gamma < 0$ e $\theta \rightarrow 1$ temos um grande número de distribuições discretas de grande interesse prático.

Ressaltando que a Função (6.2) se aproxima da distribuição Binomial Negativa se $\lambda > 0$ e $\alpha \rightarrow 0$. Se $\lambda < 0$ a Função (6.2) difere totalmente da distribuição Binomial Negativa. Os modelos dados em (6.2) a (6.4) são válidos apenas quando se trabalha com fenômenos estacionários.

Um caso particular obtido da Função (6.2) fixados $t = 1$ e $\gamma = -\frac{1}{2}$, é o Modelo (6.5) que é muito discutido na literatura, pelo fato de ser um modelo que na prática é muito bem sucedido quando a distribuição Binomial Negativa não se ajusta bem aos dados (SICHEL, 1982), a qual foi denominado distribuição Poisson Inversa Gaussiana (PIG). Uma parametrização para essa distribuição é definida por

$$\phi(r) = \frac{\left(\frac{2\alpha}{\pi}\right)^{\frac{1}{2}} \exp(\alpha(1-\theta)^{\frac{1}{2}}) \left(\frac{\alpha\theta}{2}\right)^r K_{r-\frac{1}{2}}(\alpha)}{r!}. \quad (6.5)$$

No caso, do modelo Poisson Inflacionado de Zeros à sua metodologia segue uma lógica diferente dos demais modelos para dados de contagens analisados. Nesse modelo supõe que os dados são gerados de um processo dual e a modelagem consiste no ajuste de dois estágios: um estágio zero e um estágio não zero, em que os pesos atribuídos para a ocorrência de cada estágio são complementares, Seção (??).

6.3 Análise dos resultados da simulação

Nesta seção são descritos os procedimentos e são apresentados os resultados do experimento de simulação. Inicialmente foram geradas amostras da distribuição Binomial Negativa com seis diferentes tamanhos, $n = (10, 20, 30, 50, 100 \text{ e } 500)$, sendo que cada amostra de tamanho n foi replicada 1000 vezes. Esses procedimentos foram realizados usando três diferentes algoritmos para geração de observações aleatórias de uma distribuição Binomial Negativa. Para cada amostra de cada um dos três algoritmos de geração de dados foram comparados os ajustes dos sete modelos utilizando o critério de informação AIC e ao final das 1000 iterações de cada experimento obteve-se a quantidade de vezes que cada modelo bem se ajustou aos dados.

O algoritmo de geração dos dados da distribuição Binomial Negativa não influenciou nos resultados, sendo que para os três experimentos prevaleceram as distribuições Binomial Negativa tipo I e tipo II e a distribuição de Poisson Inversa Gaussiana como modelos que bem se ajustaram aos dados simulados (tabelas 6.1 a 6.3). Porém, para os dados gerados dos algoritmos (ii) e (iii) não foi possível obter as estimativas dos parâmetros da distribuição Sichel, porque à

sua distribuição marginal, a distribuição GIG, é difícil de convergir.

Tabela 6.1: Resultado da simulação: dados 1.

N	mPO	mNBI	mNBII	mPIG	mDEL	mZIP	mSichel
10	0	570	11	403	15	1	0
20	0	208	0	777	15	0	0
30	0	55	566	320	59	0	0
50	0	105	5	661	229	0	0
100	0	137	0	811	51	0	1
500	0	0	0	732	0	0	268

Tabela 6.2: Resultado da simulação: dados 2.

N	mPO	mNBI	mNBII	mPIG	mDEL	mZIP
10	0	116	417	173	02	291
20	0	129	225	601	41	04
30	0	529	37	394	19	21
50	0	536	92	317	06	49
100	0	614	36	331	05	13
500	0	419	0	529	52	0

Tabela 6.3: Resultado da simulação: dados 3.

N	mPO	mNBI	mNBII	mPIG	mDEL	mZIP
10	2	122	413	403	22	38
20	0	155	448	351	35	11
30	0	144	467	351	26	12
50	0	535	31	426	08	49
100	0	0	0	874	126	0
500	0	144	677	177	02	0

7 *Considerações finais*

7.1 **Conclusões e Discussão**

Nesta dissertação foram abordadas as principais características e propriedades da classe de modelos para análise multivariada de séries temporais (VAR) proposta por Sims (1980) e da classe de modelos de regressão univariado (GAMLSS) proposta por Rigby e Stasinopoulos (2005). De modo geral, foram apresentados os métodos para estimação dos parâmetros e de diagnóstico de cada classe, buscando enfatizar as técnicas e os procedimentos apropriados para a modelagem de séries temporais de contagem.

Foram realizadas análises descritivas e inferenciais das séries temporais, bem como uma breve revisão dos conceitos básicos e fundamentais de regressão e de inferência estatística, verificando as principais suposições dos modelos antes de sua aplicação aos dados.

Esta pesquisa surgiu do interesse em estimar a variável quantidade mensal de dias desfavoráveis à dispersão de poluentes na atmosfera, que é uma série temporal de contagem e, analisando suas principais características, procuraram-se modelos que bem descrevam o seu comportamento, além de procurar identificar as variáveis que lhe influenciam. Buscaram-se na classe de modelos de séries temporais os modelos VAR devido à sua boa capacidade de previsão do comportamento de séries futuras, além de serem modelos os quais permitem analisar as interrelações entre múltiplas séries. Em seguida, se considerou o fato de estarmos tratando com observações de contagem, então se buscou os modelos para dados de contagem e se optou pela classe GAMLSS, por ela ser uma classe de modelos mais flexível se comparada aos GLMs.

Comparou-se os conjuntos *Dados 1* e *Dados 2* e pode-se afirmar que há significativa diferença do comportamento do evento estudado com relação ao período do ano. Mostra-se, além disso, a importância de se analisar não somente o evento em si, mas também as demais variáveis que estão fortemente relacionadas a ele, informações já afirmadas nos relatórios da CETESB (2001-2012).

A abordagem do modelo de previsão VAR para as séries estacionárias e periódicas analisa-

das, contudo, se mostrou uma proeminente ferramenta quando se deseja obter o comportamento previsto de uma série temporal para um período máximo de doze meses à frente, ressaltando que é relativamente grande a margem de erro de previsão apresentada por esse modelo para séries temporais de contagem. Destaca-se na literatura o problema de identificação apresentado nessa classe de modelos que pode ser identificado pelo teste de causalidade de Granger. No entanto, em (CAVALCANTI, 2010) chama-se a atenção a proliferação de erros na interpretação desse teste.

Com relação aos modelos estimados da classe GAMLSS: NBII e ZIP pode-se concluir que ambos são proeminentes modelos para a modelagem do evento estudado. As covariáveis consideradas como significativas foram, no geral, a quantidade de dias com precipitação e a quantidade de inversões térmicas com altitude entre (0-200) em metros. No entanto, uma variável que poderia ser incluída aos modelos é a variável quantidade mensal de frentes frias, na falta da variável quantidade mensal de dias com precipitação pluviométrica, pelo fato da maioria das ocorrências de precipitação pluviométrica e de ventos com alta velocidade serem consequências diretas da passagem de uma frente fria.

De acordo com Sichel (1975), uma sugestão para a modelagem do parâmetro do modelo de Poisson λ é a distribuição GIG, por possuir grande relevância do âmbito prático, devido à sua flexibilidade de modo que podemos a partir dela obter uma grande família de distribuições contínuas. A mistura das distribuições de Poisson e GIG é denominada distribuição Sichel, analogamente a partir da mistura pode obter uma grande família de distribuições discretas. Destaca-se na literatura, o caso particular da distribuição Sichel, a distribuição PIG, para o ajuste de dados de contagem superdispersos com uma longa cauda positiva. A PIG é considerada como uma alternativa nos casos que a distribuição Binomial Negativa não bem se ajusta, demonstrado no estudo de simulação (tabelas 6.1 a 6.3). No entanto, as razões do bom desempenho da PIG pelo que se conhece são desconhecidas (SICHEL, 1982).

Nos últimos anos o emprego da classe GAMLSS em diversas áreas de pesquisa, como em medicina teve um aumento significativo. E de acordo com (BOHL et al., 2013) a eficiência do estimador não foi avaliada, ao analisar 14 trabalhos da (*pubmed*) e 34 trabalhos da (*web of Science*). Diante de tal fato (BOHL et al., 2013) realizou estudos com dados gerados de uma distribuição Gama e de uma distribuição GIG, além de utilizar dados reais contínuos com alta assimetria positiva para comparar o viés, a precisão e a cobertura dos estimadores da distribuição Gama dos GLM com os estimadores das distribuições Gama e GIG dos GAMLSS. No caso de dados seguindo uma distribuição Gama os desempenhos dos estimadores se mostraram similares. No entanto, no caso de dados seguindo uma distribuição GIG em certas cir-

cunståncias as estimativas da distribuição GIG se mostraram imprecisas ou errôneas e além disso, a distribuição GIG não bem se ajustou a dados gerados de uma distribuição GIG. Destaca-se que a principal função da distribuição GIG não é a modelagem de dados, mas a simulação de demais distribuição a partir de sua função de distribuição (ATKINSON, 1982). No geral, (BOHL et al., 2013) identificou que os estimadores da distribuição Gama dos GLM se mantiveram robustos em todos os momentos o que não ocorreu com os estimadores distribuição GIG do GAMLSS. Remetendo ao presente trabalho as conclusões apresentadas referentes ao modelo NB do GAMLSS são similares as obtidas com o ajuste de um modelo NB do GLM, com a vantagem de que os GAMLSS forneceu também a estimativa do parâmetro de dispersão.

No estudo realizado por (BOHL et al., 2013) são analisados apenas os estimadores das distribuições Gama e GIG dos GAMLSS, no entanto o autor destaca que diante da flexibilidade dessa classe de modelos pode existir outros modelos que bem se ajustam a dados contínuos com alta assimetria positiva e que tenham desempenhos melhores que os estimadores da distribuição Gama do GLM. Uma opção seria a comparação entre as distribuições Inversa Gaussiana e Gama, destacando que a distribuição Inversa Gaussiana é a distribuição marginal da PIG.

Contudo, se o interesse de um pesquisador for prever o comportamento da variável quantidade mensal de dias desfavoráveis à dispersão de poluentes na atmosfera para os próximos doze meses o modelo PVAR(p) é um proeminente candidato. Por outro lado, se o interesse de um pesquisador for identificar covariáveis que influenciam no evento de interesse e interpretação do modelo, o ajuste da distribuição Binomial Negativa é mais recomenda.

Vale salientar que o uso da classe de modelos VAR e da classe de modelos GAMLSS na modelagem das séries provenientes do monitoramento da qualidade do ar são promissoras, por serem poderosas metodologias estatísticas para a análise de dados multivariados/univariados com estrutura de regressão.

7.2 Sugestões para pesquisas futuras

Este trabalho apresentou algumas possibilidades para a modelagem de séries temporais de contagem, sendo estas pertencentes as classes VAR e GAMLSS. Como discutido, existem outras inúmeras possibilidades que são objeto de estudo de pesquisadores. Algumas opções de trabalhos futuros, incluem, mas não são limitadas a:

- Aplicar as técnicas da classe VAR para séries temporais não estacionárias, por exemplo o conjunto *Dados 2*, podendo ser possível, assim, comparar o desempenho do comporta-

mento previsto para a série temporal de contagem futura mediante o emprego dos modelos VAR e VEC;

- Incorporar um termo correspondente a uma unidade de tempo no ajuste do GAMLSS para representar que estamos modelando uma série temporal;
- Utilizar a classe de Modelos Univariados Periódicos Autorregressivos (PAR), que ajusta um modelo para cada mês, podendo, assim, ser analisado cada mês individualmente;
- Comparar o viés e a eficiência do modelo Poisson Inversa Gaussiana do GAMLSS com do modelo Binomial Negativa do GLM no ajuste de séries temporais de contagens assimétricas positivas.

APÊNDICE A – Alguns conceitos básicos de regressão e inferência

A.1 Critérios de seleção de modelo

A seleção de modelos é uma importante etapa, pois quando selecionado o modelo mais adequado ao conjunto de dados, mais informações podem ser extraídas dos dados. Alguns dos métodos usualmente utilizados para essa etapa são apresentados a seguir.

A.1.1 Desvio

Para avaliar a qualidade do ajuste de um modelo pode-se utilizar uma medida de discrepância, como a função *Deviance*, definida por

$$D(\mathbf{y}; \beta) = 2[\ell(\hat{\beta}_n) - \ell(\hat{\beta}_p)], \quad (\text{A.1})$$

em que $\ell(\hat{\beta}_n)$ e $\ell(\hat{\beta}_p)$ são as funções de verossimilhança maximizada para o modelo completo (ou saturado) e para o modelo sob pesquisa, em que p e n são os números de parâmetros desses modelos, respectivamente.

A vantagem do método *deviance* é que pode ser utilizado na comparação de dois modelos no teste da razão de verossimilhança e é aditivo para conjuntos aninhados de modelos (McCULLAGH; NELDER, 1989; MCEL DUFF, 2012).

A.1.2 Critérios de informação

Os critérios de informação Akaike (AIC), Bayesiano (BIC) e Akaike corrigido (AICc) introduzidos por Akaike (1974), Schwarz (1978), Hurvich e Tsai (1989) respectivamente, quantificam a redução de variância dos resíduos do modelo com relação ao aumento do número de

parâmetros do modelo e ao mesmo tempo penalizam a inclusão de cada termo no modelo. Eles estão intimamente relacionados entre si e entre outros critérios, como o critério HQ introduzido por Quinn (1980) e o critério Erro Preditivo Final (FPE) (LIRA et al., 2012).

Os critérios AIC, BIC e AICc são amplamente utilizados principalmente na comparação de modelos de séries temporais e de regressão linear que utilizam o método de estimação de máxima verossimilhança. As funções dos critérios AIC, BIC, AICc, HQ e FPE são definidas por

$$AIC = -2\log L(\hat{\beta}) + 2p^*, \quad (A.2)$$

$$BIC = -2\log L(\hat{\beta}) + p^* \log(n), \quad (A.3)$$

$$AICc = AIC + \frac{2p^*(p^* + 1)}{n - p^* - 1}, \quad (A.4)$$

$$HQ(p) = \log(\det(\sum_{\varepsilon}(p))) + \frac{2(\log(\log(n)))}{n} pd^2 \quad (A.5)$$

e

$$FPE(p) = \left(\frac{n + p^*}{n - p^*}\right)^d \det(\sum_{\varepsilon}(p)), \quad (A.6)$$

em que $L(\cdot)$ é a função de verossimilhança, $\sum_{\varepsilon}(p) = n^{-1} \sum_{i=1}^n \hat{\boldsymbol{\varepsilon}}_t \hat{\boldsymbol{\varepsilon}}_t'$, p^* é o número total de parâmetros no modelo, p é a ordem ótima para o número de defasagens, n é o número de observações da variável de interesse e d é o conjunto de variáveis endógenas.

A.2 Análise de Multicolinearidade

Quando se trabalha com modelos de regressão multivariados é relevante verificar se as variáveis explicativas são correlacionadas. No caso de não haver qualquer relação entre elas,

conclui-se que são ortogonais.

Em estudos com dados reais é frequente a presença de não ortogonalidade. Se as variáveis forem muito correlacionadas, as inferências do modelo de regressão multivariado podem ser errôneas ou pouco confiáveis, isto é, podemos obter estimativas instáveis com altos erros padrões.

Quando uma ou mais variáveis explicativas são combinações lineares de outras, há presença de multicolinearidade e não existe um único estimador para os parâmetros do modelo. Então a matriz $X^T X$ é singular, em que X é a matriz de planejamento e X^T é a transposta da matriz X . A matriz X é definida por

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

A.2.1 Fator de Inflação da Variância - VIF

A presença de multicolinearidade pode ser identificada analisando-se o VIF, que mede a correlação de uma variável com todas as outras do modelo. A tolerância para a estimativa dos parâmetros é dada pelo inverso do VIF. O VIF é definido por

$$VIF_j = \frac{1}{1 - R_j^2}, \quad (\text{A.7})$$

em que R_j^2 é o R^2 da regressão da j -ésima variável explicativa sobre as demais variáveis explicativas do modelo.

Não há critério formal de decisão na análise do VIF. Usualmente considera-se indicativo de problemas de multicolinearidade se o valor do VIF for maior que dez.

A.3 Assimetria e Curtoses

É esperado que as observações de uma variável sejam distribuídas normalmente, no entanto na prática a assimetria e/ou quantidades de valores extremos não equiparáveis às esperadas da distribuição Normal são frequentes.

Sejam $\mathbf{y}_t = (y_{t1}, \dots, y_{tn})$, $\forall n \in \mathbb{N}$ e $n < \infty$ uma amostra aleatória com média e variância estimadas (ou estimativa do momento centrado de segundo ordem), denotada por $E(\mathbf{Y}_t) = \hat{\mu}_t$ e $Var(\mathbf{Y}_t) = \hat{\sigma}_t^2$, respectivamente. é definido o coeficiente de assimetria e seus critérios

$$\hat{A}(\mathbf{Y}_t) = \frac{\sum_{t=1}^n (Y_t - \hat{\mu}_t)^3}{(n-1)\hat{\sigma}_t^3}. \quad (\text{A.8})$$

A interpretação do $\hat{A}(\mathbf{Y}_t)$ é dada

- Se $\hat{A}(\mathbf{Y}_t) > 0$, indica uma distribuição com uma longa cauda à direita;
- Se $\hat{A}(\mathbf{Y}_t) < 0$, indica uma distribuição com uma longa cauda à esquerda;
- Se $\hat{A}(\mathbf{Y}_t) \simeq 0$, indica distribuição simétrica.

Analogamente, é definido o coeficiente de curtoses e seus critérios

$$\hat{K}(\mathbf{Y}_t) = \frac{\sum_{t=1}^n (Y_t - \hat{\mu}_t)^4}{(n-1)\hat{\sigma}_t^4}. \quad (\text{A.9})$$

A interpretação do $\hat{K}(\mathbf{Y}_t)$ é dada por

- Se $K(\mathbf{Y}_t) > 3$, os dados têm caudas pesadas ou longas se comparadas às caudas da distribuição Normal, então os dados são leptocúrticos;
- Se $K(\mathbf{Y}_t) < 3$, os dados têm caudas leves ou curtas se comparadas às caudas da distribuição Normal, então os dados são platicúrticos;
- Se $K(\mathbf{Y}_t) \simeq 3$, os dados têm caudas comparáveis às caudas da distribuição Normal, então os dados são mesocúrticos.

O desvio padrão da estimativa amostral k é $\sqrt{\frac{24}{n}}$ para um ruído branco Gaussiano (GROENEVELD; MEEDEN, 1984; RIGBY; STASINOPOULOS, 2006).

A.4 Medidas de correlação entre duas variáveis

Em aplicações que envolvem duas ou mais variáveis é comum o interesse em conhecer se existe uma relação entre elas. Inicialmente, é prudente a análise de um diagrama de dispersão das variáveis antes da análise de qualquer medida de correlação.

Nesta seção são apresentadas três medidas de dependência entre duas variáveis: o coeficiente de correlação linear de Pearson, o coeficiente de correlação de postos de Spearman e o coeficiente de correlação de postos de Kendall.

A.4.1 Coeficiente de correlação linear de Pearson

O coeficiente de correlação linear de Pearson é um método paramétrico frequentemente utilizado para medir a correlação entre duas variáveis. Ele é um método paramétrico no sentido que supõe que a distribuição teórica da variável é conhecida. Esse método admite que (i) as variáveis são aleatórias, (ii) a relação entre elas é linear e (iii) a distribuição conjunta delas é uma distribuição Normal Bivariada (BUNCHAFT; KELLNER, 1999).

Sejam X_1 e X_2 duas variáveis as quais as suposições (i) a (iii) são válidas. A formulação do coeficiente de correlação linear de Pearson é dada por

$$\rho(X_1, X_2) = \frac{\text{cov}(X_1, X_2)}{\sqrt{\sigma_{X_1}^2 \sigma_{X_2}^2}}, \quad (\text{A.10})$$

em que $-1 < \rho(X_1, X_2) < 1$, $\text{cov}(X_1, X_2) = E[(X_1 - \mu_{X_1})(X_2 - \mu_{X_2})]$, $\sigma_X^2 = \sum_{i=1}^n \frac{(x_i - \mu_X)^2}{n}$ e $\mu_X = \sum_{i=1}^n \frac{x_i}{n}$.

A interpretação do $\rho(X_1, X_2)$ é dada por

- (i) Se $\rho(X_1, X_2) = 1$ existe relação linear perfeita positiva entre X_1 e X_2 ;
- (ii) Se $\rho(X_1, X_2) = -1$ existe relação linear perfeita negativa entre X_1 e X_2 ;
- (iii) Se $-1 < \rho(X_1, X_2) < 1$ o coeficiente de correlação pode ser avaliado qualitativamente;
- (iv) Se $\rho(X_1, X_2) = 0$ não existe uma relação linear entre X_1 e X_2 .

Ressaltando que a hipótese de normalidade bivariada é imprescindível para amostras consideradas pequenas. Porém à medida que aumenta o tamanho da amostra essa hipótese tem a sua importância minimizada, de acordo com o Teorema do Limite Central (TLC) para distribuição multivariadas (JOHNSON; WICHERN, 1988).

A.4.2 Coeficientes de correlação de postos de Spearman e de Kendall

Os coeficientes de correlação de postos de Spearman e de Kendall são métodos não paramétricos, no sentido de que não há suposições formuladas sobre a forma da distribuição das variáveis. São denominados por ρ_S e τ , respectivamente (SPRENT, 2007).

Sejam $(X_1, Y_1), \dots, (X_n, Y_n)$ pares independentes de uma amostra aleatória de uma população bivariada e $R_i = \text{posto}(X_i)$ e $S_i = \text{posto}(Y_i)$, para $i = 1, \dots, n$. A formulação do ρ_S e do τ são dadas, respectivamente por

$$\rho_S = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}, \quad (\text{A.11})$$

em que $-1 < \rho_S < 1$, $d_i = (R_i - S_i)$ é a diferença entre os postos e n é o número de pares ordenados, para $i = 1, \dots, n$. A formulação ρ_S é baseada na formulação do ρ .

$$\tau = \frac{n_c - n_d}{\frac{n(n-1)}{2}}, \quad (\text{A.12})$$

em que $-1 < \tau < 1$, $\frac{n(n-1)}{2}$ é o número total de pares de postos, n_c é o número de pares concordantes e n_d é o número de pares discordantes.

A interpretação do τ é dada por

- (i) Se $\tau = 1$ todos os pares de postos são concordantes, então, $n_c = \frac{n(n-1)}{2}$ e $n_d = 0$;
- (ii) Se $\tau = -1$ todos os pares de postos são discordantes, então, $n_c = 0$ e $n_d = \frac{n(n-1)}{2}$;
- (iii) Se $\tau = 0$, espera-se uma mistura de concordantes e discordantes.

O ρ_S e o τ são medidas de correlação utilizadas para observações ordinais, de forma que seja possível atribuir postos a cada uma das observações.

As medidas de correlação ordinal, ρ_S e τ , não podem ser interpretadas da mesma forma que o coeficiente de correlação linear de Pearson ρ . Pelo fato de ρ_S e τ não serem coeficientes que representam necessariamente a tendência linear. Nesse caso são considerados como índices de monotonicidade (LIRA, 2004).

As relações entre as observações das variáveis podem ser interpretadas por

- (i) Se $x_i < x_j$ sempre que $y_i < y_j$ ou $x_i > x_j$ sempre que $y_i > y_j$ é interpretado como uma relação direta e perfeita, com coeficiente igual a 1;
- (ii) Se $x_i < x_j$ sempre que $y_i > y_j$ ou $x_i > x_j$ sempre que $y_i < y_j$ é interpretado como uma relação indireta e perfeita, com coeficiente igual a -1;
- (iii) Se não ocorrer nem (i) e nem (ii), o coeficiente estará entre -1 e 1;
- (iv) Se o coeficiente for nulo, então X e Y são independentes.

A.4.3 Teste de Filliben

O teste de Filliben (1975) é utilizado para identificar a normalidade dos dados. Esse teste consiste em calcular a correlação entre as estatísticas de ordem e as medianas ordenadas da distribuição normal padrão. As hipóteses testadas são:

H_0 : os dados seguem uma distribuição normal.

H_1 : os dados não seguem uma distribuição normal.

A estatística de teste é dada por,

$$\text{corr}(X, M) = \frac{\sum(X_i - X_m)(M_i - M_m)}{\sqrt{\sum(X_i - X_m)^2 \sum(M_i - M_m)^2}}, \quad (\text{A.13})$$

em que X_i e X_m é a i-ésima e m-ésima observação de uma amostra ordenada e M_i e M_m é a i-ésima e m-ésima mediana ordenada de uma população seguindo uma distribuição normal padrão.

Quanto mais próximo de um for o valor da estatística de teste, mais garantia se tem da normalidade dos dados, sendo que se o valor tabelado R for maior do que a estatística de teste $\text{corr}(X, M)$, há evidências para rejeitarmos a hipótese de normalidade dos dados para um nível de significância α .

APÊNDICE B – Ajustes sem covariáveis - Dados de jan/1999 a dez/2010

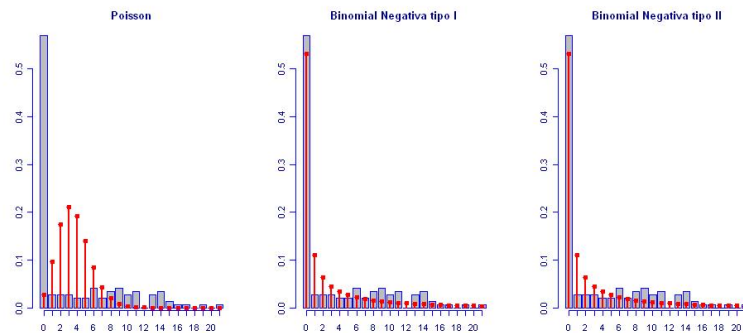


Figura B.1: Ajuste PO, NBI e NBII para os dados da variável dias desfavoráveis.

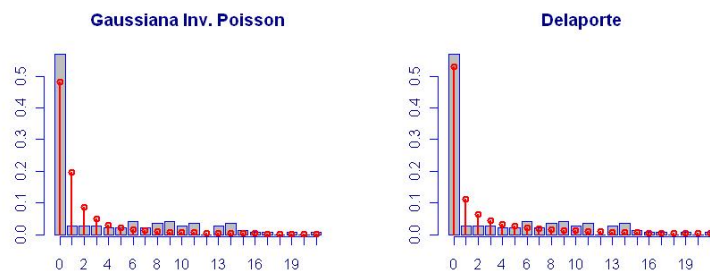


Figura B.2: Ajuste PIG e DEL para os dados da variável dias desfavoráveis.

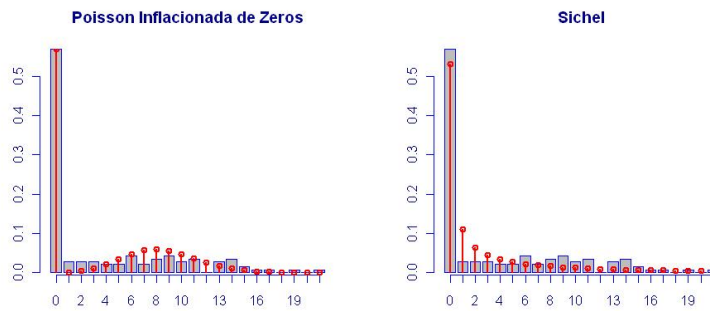


Figura B.3: Ajuste ZIP e SI para os dados da variável dias desfavoráveis.

Tabela B.1: Valores dos critérios de informação - dados de jan/1999 a dez/2010.

	DEL	NBI	NBII	PO	ZIP	PIG	SICHEL
AIC	611.7	609.7	609.7	1296.5	615.6	660.2	611.7
BIC	620.6	615.6	615.6	1299.5	621.5	666.2	620.6

Tabela B.2: Média e variância dos dias desfavoráveis e o intervalo de confiança para o coeficiente estimado.

$\hat{\mu}$	$\hat{\sigma}^2$	Intervalo de confiança
3.6	214.6	$1.28 \pm 1.96(0.18)=(0.9,1.6)$

Tabela B.3: Modelo NB estimado.

Parâmetro	Estimativa	Exp(estimativa)	Erro	p-valor
$\hat{\mu}$	1.28	3.6	0.18	< 0.001
$\hat{\sigma}$	2.79	16.28	0.25	< 0.001

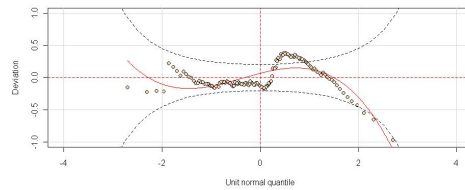


Figura B.4: Desvios *versus* quantis normalizados (gráfico de envelope) - modelo sem regressão NB.

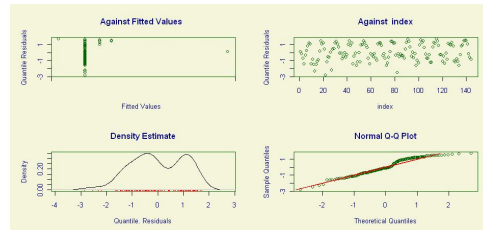


Figura B.5: Resíduos *versus* valores ajustados, resíduos *versus* índices, estimativa do núcleo da densidade e quantis amostrais *versus* quantis teóricos (gráfico QQ) - modelo sem regressão NB.

Tabela B.4: Resumo dos quantis dos resíduos para o ajuste do modelo sem regressão NB.

$\hat{\mu}$	$\hat{\sigma}^2$	$\hat{A}(Y)$	$\hat{K}(Y)$	Teste de Filliben
-0.0003	1.08	-0.2	2.17	1

```
R > library(gamlss)
R > library(gamlss.dist)
R > par(mfrow = c(1, 3))
R > mPO <- histDist(Y, "PO", main = "Poisson")
R > mNBI <- histDist(Y, "NBI", main = "Binomial Negativa tipo I")
R > mNBII <- histDist(Y, "NBII", main = "Binomial Negativa tipo II")
R > par(mfrow = c(1, 2))
R > mPIG <- histDist(Y, "PIG", main = "Gaussiana Inversa Poisson")
R > mDEL <- histDist(Y, "DEL", main = "Delaporte")
R > par(mfrow = c(1, 2))
R > mZIP <- histDist(Y, "ZIP", main = "Poisson Inflacionada de Zeros")
R > mSichel <- histDist(Y, "SICHEL", main = "Sichel")
R > summary(mNBII)
```


APÊNDICE C – Ajustes sem covariáveis - Dados de maio a setembro, 2001 a 2010

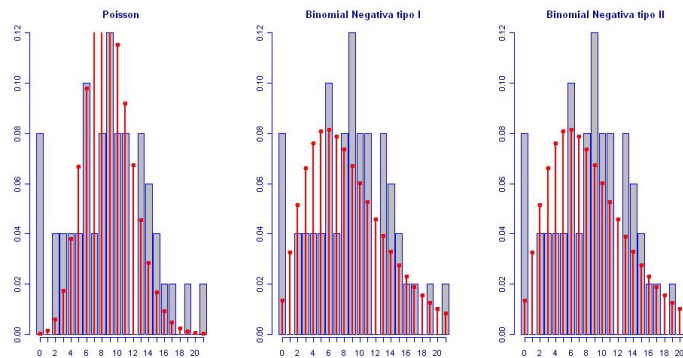


Figura C.1: Ajuste PO, NBI e NBII para os dados da variável dias desfavoráveis.

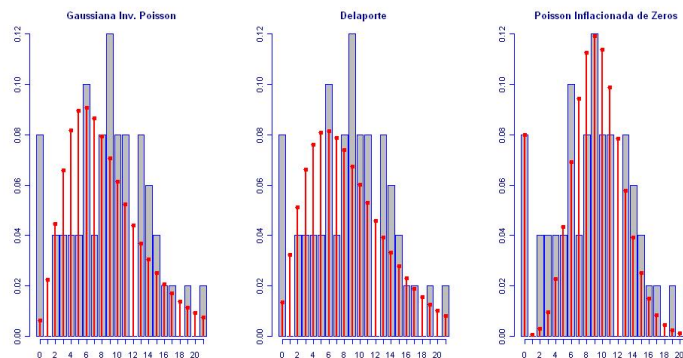


Figura C.2: Ajuste PIG, DEL e SI para os dados da variável dias desfavoráveis.

Tabela C.1: Valores dos critérios de informação - dados de maio a setembro, 2001 a 2010.

	DEL	NBI	NBII	PO	ZIP	PIG
AIC	312.6	310.6	310.6	354.5	311.2	316.0
BIC	318.4	314.4	314.4	356.4	315.0	319.9

Tabela C.2: Média e variância dos dias desfavoráveis e do intervalo de confiança para o intercepto.

$\hat{\mu}$	$\hat{\sigma}^2$	Intervalo de confiança para μ
8.8	32.1	$2.17 \pm 1.96(0.09)=(2.0,2.3)$

Tabela C.3: Modelo NB estimado.

Parâmetro	Estimativa	Exp(estimativa)	Erro	p-valor
$\hat{\mu}$	2.17	8.8	0.09	< 0.001
$\hat{\sigma}$	-1.20	0.3	0.31	< 0.001

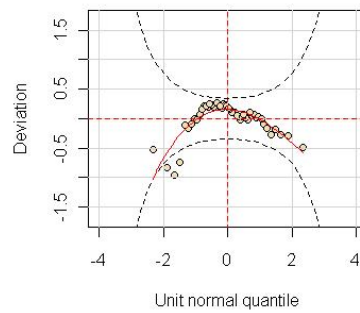


Figura C.3: Desvios *versus* quantis normalizados (gráfico de envelope) - modelo sem regressão NB.

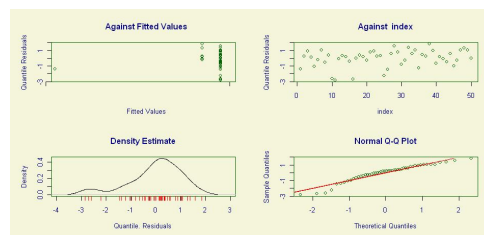


Figura C.4: Resíduos *versus* valores ajustados, resíduos *versus* índices, estimativa do núcleo da densidade e quantis amostrais *versus* quantis teóricos (gráfico QQ) - modelo sem regressão NB.

Tabela C.4: Resumo dos quantis dos resíduos para o modelo NB.

$\hat{\mu}$	$\hat{\sigma}^2$	$\hat{A}(Y)$	$\hat{K}(Y)$	Teste de Filliben
-0.015	1.15	-0.9	3.5	0.96

APÊNDICE D – Simulação

Dados 1

```
# numero de observacoes da amostra
N <- n

# parametro pre fixado do modelo AR(1)
rho <- 0.6

# ajuste de um AR(1)
log.lambda <- 1 + arima.sim(model=list(ar=rho),n)

# gera dados de uma NB
y <- rpois(n, lambda=exp(log.lambda))
```

Dados 2

```
library(COUNT)
nobs <- n # numero de observacoes da amostra
x1 <- runif(nobs) # gera dados da distribuicao Uniforme (0,1)
x2 <- runif(nobs) # gera dados da distribuicao Uniforme (0,1)
xb <- 0.5 +0.75*x1 -1.25*x2 # preditor linear
a <- 0.75 # parametro Alpha
ia <- 1/a # inversa do parametro Alpha
exb <- exp(xb) # inversa do ajuste do preditor linear
xg <- rgamma(nobs,a,a,ia) # gera de dados de uma distribuição Gama(GA)
xbg <- exb*xg # mistura das PO e GA
y <- rpois(nobs, xbg) # gera dados de uma NBII
```

Dados 3

```
N <- n # numero de observacoes da amostra
y <- rnbinom(n, sigma=5, media = 8.78) # gera dados de uma NB-C
```

Algoritmo para simulação usando a geração de dados 1

```
library(gamlss)
library(tseries)
simula <- function(n) {
  pos <- 1 # inicializacao da posicao
  min <- integer(7) # vetor de AIC
  vetfunc <- integer(7) # contador
  N <- m # quantidade de numeros aleatorios gerados
  rho <- 0.6 # parametro pre fixado do modelo AR(1)

  # ajuste de um AR(1)
  log.lambda <- 1 + arima.sim(model=list(ar=rho), n=N)

  for (i in 1:n) {
    # gera dados de uma NB
    y <- rpois(N, lambda=exp(log.lambda))

    # Ajuste dos modelos
    mPO <- histDist(y, "PO", main ="Poisson")
    mNBI <- histDist(y, "NBI", main ="Binomial Negativa tipo I")
    mNBII <- histDist(y, "NBII", main ="Binomial Negativa tipo II")
    mPIG <- histDist(y, "PIG", main ="Gaussiana Inv. Poisson")
    mDEL <- histDist(y, "DEL", main ="Delaporte")
    mZIP <- histDist(y, "ZIP", main ="Poisson Inflacionada de Zeros")
    mSichel <- histDist(y,"SICHEL", main ="Sichel")

    print(AIC(mPO, mNBI,mNBII, mPIG,mDEL,mZIP, mSichel)) /* imprimi os AIC */

    min[1] = AIC(mPO)
```

```
min[2] = AIC(mNBI)
min[3] = AIC(mNBII)
min[4] = AIC(mPIG)
min[5] = AIC(mDEL)
min[6] = AIC(mZIP)
min[7] = AIC(mSichel)

# percorre os sete valores obtidos de AIC
for(k in 1:7) {
  if(min[1] > min[k]) {
    min[1] = min[k] # identifica o menor AIC
    pos = k        # obtem a posicao do menor
  }
}

vetfunc[pos] = vetfunc[pos] + 1
print(vetfunc) # imprimir a quantidade de vezes que
               # cada modelo obteve menor AIC
print(min)    # imprimir os menores valores AIC
}
}

simula(1000)
```


Referências Bibliográficas

- AHRENS, C. D. *Meteorology today an introduction to weather, climate, and the environment*. 9 ed. ed. [S.l.]: Brooks/ Cole, USA, 2009.
- AKAIKE, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, v. 19, n. 6, p. 716 – 723, 1974.
- AKANTZILIOTOU, K.; RIGBY, R. A.; STASINOPOULOS, D. M. The r implementation of generalized additive models for location scale and shape. In: STASINOPOULOS, M.; TOULOUMI, G. (Ed.). *Statistical modelling in Society: Proceedings of the 17th International Workshop on statistical modelling*. Chania, Greece: [s.n.], 2002. p. 75–83.
- ATKINSON, A. C. The simulation of generalized inverse Gaussian and hyperbolic random variables. *SIAM J. SCI. STAT. COMPUT.*, v. 3(4), p. 502–515, 1982.
- BELLANDER, T. et al. Using geographic information systems to assess individual historical exposure to air pollution from traffic and house heating in stockholm. *Environ Health Perspect*, v. 109(6), p. 633–639, 2001.
- BOHL, A. A. et al. Are generalized additive models for location, scale, and shape an improvement on existing models for estimating skewed and heteroskedastic cost data? *Health Services and Outcomes Research Methodology*, v. 13, n. 1, p. 18–38, 2013.
- BRANDT, P. T.; APPLEBY, J. *Conflict phases and processes: Bayesian Markov-switching models of endogenous systems*. Chicago, Illinois: Annual Meeting of the Midwest Political Science Association, 2007.
- BUNCHAFT, G.; KELLNER, S. R. O. *Estatística sem mistérios*. 2 ed, v.2. ed. [S.l.]: Petrópolis: Vozes, 1999.
- BUUREN, S. V.; FREDRIKS, M. Worm plot: a simple diagnostic device for modelling growth reference curves. *Statistics in Medicine*, v. 20, n. 8, p. 1259–1277, 2001.
- CAMALIER, L.; COX, W.; DOLWICK, P. The effects of meteorology on ozone in urban areas and their use in assessing ozone trends. *Atmospheric Environment*, v. 41, p. 7127–7137, 2007.
- CAVALCANTI, M. A. F. H. Identificação de modelos var e causalidade de granger: uma nota de advertência. *Computer Methods and Programs in Biomedicine*, v. 97, p. 168–197, 2010.
- CETESB. *Qualidade do ar no estado de São Paulo*. [S.l.], mai 2001-2012. Relatórios de Qualidade do Ar no Estado de São Paulo.
- CETESB. *Estudo do comportamento do ozônio na região metropolitana de São Paulo*. [S.l.], 2002. Diretoria de recursos hídricos e engenharia ambiental - Departamento de qualidade ambiental.

CHATFIELD, C. *The Analysis of Time Series: An Introduction*. 6th. ed. [S.l.]: Chapman e Hall CRC., 2004.

COLE, T. J.; GREEN, P. Smoothing reference centile curves: The lms method and penalized likelihood. *Statistics in Medicine*, v. 11, p. 1305–1319, 1992.

CORDEIRO, G. M.; DEMÉTRIO, C. G. B. *Modelos Lineares Generalizados e Extensões*. [S.l.], jul 2008. Departamento de Estatística e Informática, UFRPE e Departamento de Ciências Exatas, ESALQ, USP.

CORDEIRO, G. M.; RODRIGUES, J.; CASTRO, M. The exponential COM-Poisson distribution. *Statistical Papers*, v. 53(3), p. 653–664, 2012.

DAMGHANI, B. M. et al. The misleading value of measured correlation. *Wilmott*, v. 2012, n. 62, p. 64–73, November 2012.

DELAPORTE, P. Quelques problemes de statistique mathematique poses par l'assurance automobile et le bonus non sinistre. *Bulletin Trimestriel de l'Institut des Actuaire Francais*, v. 227, p. 87–102, 1959.

DIEBOLD, F. X. *Elements of Forecasting*. [S.l.]: South-Western College Publishing, 1998.

DUNN, P. K.; SMYTH, G. K. Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, v. 5, n. 3, p. 236–244, 1996.

FAHRMEIR, L.; TUTZ, G. Bayesian inference for generalized additive mixed models based on markov random field priors. *Appl. Statist.*, v. 50, p. 201–220, 2001.

FLORENCIO, L. A. *Engenharia de avaliações com base em modelos GAMLSS*. Tese (Doutorado) — Universidade Federal de Pernambuco, Fevereiro 2010.

GREENE, W. H. *Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binomial Regression Models*. [S.l.], 1994.

GROENEVELD, R.; MEEDEN, G. Measuring skewness and kurtosis. *The Statistician*, v. 33, p. 391–399, 1984.

HAMILTON, J. D. *Times Series Analysis*. [S.l.]: Princeton University Press, Princeton, 1994.

HARVEY, A. C. *Forecasting Structural Time Series Models and the Kalman Filter*. [S.l.]: Cambridge: Cambridge University Press, 1989.

HASTIE, T. J.; TIBSHIRANI, R. J. *Generalized Additive Models*. [S.l.]: Chapman & Hall/CRC, 1990.

HATEMI, J. A. Tests for cointegration with two unknown regime shifts with an application to financial market integration. *Empirical Economics*, v. 35, p. 497–505, 2008.

HILBE, J. M. *Negative Binomial Regression*. 2nd. ed. [S.l.]: Cambridge: United Kingdom at the University Press, 2011.

HILBE, J. M. *Using R to Create Synthetic Discrete Response Regression Models*. [S.l.], julho 2011. Arizona State University, e Jet Propulsion Laboratory, California Institute of Technology.

- HUDSON, I. L.; KIM, S.; KEATLEY, M. R. Climate effects and temperature thresholds for eucalypt flowering: a GAMLSS ZIP approach. In: CHAN, F.; MARINOVA, D.; ANDERSSON, R. (Ed.). *MODSIM2011, 19th International Congress on Modelling and Simulation*. [S.l.]: 19th International Congress on Modelling and Simulation, Perth, Australia, 2011. p. 2647–2653.
- HURVICH, C. M.; TSAI, C. Regression and time series model selection in small samples. *Biometrika*, v. 76, n. 2, p. 297–307, 1989.
- JERRETT, M. et al. A review and evaluation of intraurban air pollution exposure models. *Journal of Exposure Analysis and Environmental Epidemiology*, v. 15, p. 185–204, 2005.
- JOHANSEN, S.; JUSELIUS, K. Maximum likelihood estimation and inference on cointegration - with applications to the demand for money. *Oxford Bulletin of Economics and Statistics*, v. 52, n. 2, p. 169–210, 1990.
- JOHNSON, R. A.; WICHERN, D. W. *Applied multivariate statistical analysis*. 2nd. ed. [S.l.]: New Jersey: Prentice Hall International, 1988.
- JONES, R. H.; BRELSFORD, W. Time series with periodic structure. *Biometrika*, v. 54, p. 403–408, 1967.
- KEDEM, B.; FOKIANOS, K. *Regression Models for Time Series Analysis*. [S.l.]: Wiley series in probability and statistics, 2002.
- LAMBERT, D. Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, v. 34, p. 1–14, 1992.
- LEE, A. H. et al. Modeling young driver motor vehicle crashes: data with extra zeros. *Elsevier*, v. 34(4), p. 515–521, 2002.
- LIRA, S. A. *Análise de correlação: abordagem teórica e de construção dos coeficientes com aplicações*. Tese (Doutorado) — Universidade Federal do Paraná, 2004.
- LIRA, T. S. et al. Handbook of environment and waste management. In: _____. [S.l.]: World Scientific, 2012. cap. Air Quality Modeling and Prediction.
- LORD, D.; GEEDIPALLY, S. R. The negative binomial-Lindley distribution as a tool for analyzing crash data characterized by a large amount of zeros. *Accident Analysis & Prevention*, v. 43(5), p. 1738–1742, 2011.
- LORD, D.; MANNERING, F. The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Elsevier*, v. 44(5), p. 291–305, 2010.
- LORD, D.; WASHINGTON, S.; IVAN, J. Poisson, Poisson-gamma and zero inflated regression models of motor vehicle crashes: Balancing statistical fit and theory. *Accident Analysis & Prevention*, v. 37(1), p. 35–46, 2005.
- LORD, D.; WASHINGTON, S. P.; IVAN, J. N. Further notes on the application of zero-inflated models in highway safety. *Accident Analysis & Prevention*, v. 39, n. 1, p. 53–57, 2007.
- LÜTKEPOHL, H. New introduction to multiple time series analysis. *Econometric Theory*, v. 22(5), p. 961–967, 2006.

MARGARIDO, M. A. Teste de cointegração de johansen utilizando o SAS. *Agric. São Paulo*, v. 51(1), p. 87–101, 2004.

MCCULLAGH, P.; NELDER, J. *Generalized Linear Models*. 2nd. ed. [S.l.]: Chapman and Hall, London, 1989.

MCELDUFF, F. C. *Models for Discrete Epidemiological and Clinical Data*. Tese (Doutorado) — UCL Institute of Child Health, University College London., 2012.

MEDEIROS, A.; GOUVEIA, N. Relação entre baixo peso ao nascer e a poluição do ar no município de São Paulo. *Rev. Saúde Pública*, v. 39(6), p. 965–972, 2005.

MORETTIN, P. A.; TOLOI, C. M. C. *Análise de Séries Temporais*. 2nd. ed. [S.l.]: São Paulo: Egard Blucher, 2006.

MORGAN, B. J. T.; PALMER, K. J.; RIDOUT, M. S. Score test oddities. *The American Statistician*, v. 61, p. 285–288, 2007.

NELDER, J. A.; WEDDERBURN, W. M. Generalized linear models. *Journal of Royal Statistical Society - Series A*, v. 135, n. 3, p. 370, 1972.

PAGANO, M. On periodic and multiple autoregressions. *The Annals of Statistics*, v. 6, p. 1310–1317, 1978.

PFUFF, B. Var, svar and svec models: Implementation within r package vars. *Journal of Statistical Software*, v. 27(4), p. 32, 2008.

QUINN, B. G. Order determination for a multivariate autoregression. *Journal of Royal Statistical Society - Series B*, v. 42, n. 2, p. 182–185, 1980.

R Development Core Team. *R: A language and environment for Statistical Computing*. [S.l.], 2009. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3, URL <http://www.R-project.org>.

RIGBY, R. A.; STASINOPOULOS, D. M. A semi-parametric additive model for variance heterogeneity. *Statistical Computing*, v. 6, p. 57–65, 1996a.

RIGBY, R. A.; STASINOPOULOS, D. M. Generalized additive models for location, scale and shape. *Appl. Statist*, v. 54, n. 6, p. 507–554, 2005.

RIGBY, R. A.; STASINOPOULOS, D. M. *Statistical Modelling using GAMLSS in R*. 2006. 1-87 p.

RIGBY, R. A.; STASINOPOULOS, D. M. Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software*, v. 23, n. 7, p. 01–46, 2007.

SCHWARZ, G. E. Estimating the dimension of a model. *Annals of Statistics*, v. 6, n. 2, p. 461–464, 1978.

SICHEL, H. S. On a family of discrete distributions particularly suited to represent long-tailed frequency data. In: *In Proceedings of the Third Symposium on Mathematical Statistics*. [S.l.: s.n.], 1971. p. 51–97.

- SICHEL, H. S. On a distribution representing sentence-length in written prose. *J. R. Statist. Soc.*, v. 137, p. 25–34, 1974.
- SICHEL, H. S. On a distribution law for word frequency. *J. Amer. Statist. Ass.*, v. 70, p. 542–547, 1975.
- SICHEL, H. S. Repeat-buying and the generalized inverse Gaussian-Poisson distribution. *Appl. Statist.*, v. 31 (3), p. 193–204, 1982.
- SIMS, C. A. Macroeconomics and reality. *Econometrica*, v. 48, p. 1–48, 1980.
- SIMS, C. A.; ZHA, T. Bayesian methods for dynamic multivariate models. *International Economic Review*, v. 39(4), p. 949–968, 1998.
- SPRENT, P. *Applied Non Parametric Statistical Methods, Second Edition*. 4nd. ed. [S.I.]: Chapman & Hall/CRC Texts in Statistical Science, 2007.
- TROUTMAN, B. M. Some results in periodic autoregression. *Biometrika*, v. 66, p. 219–228, 1979.
- URSU, E.; DUCHESNE, P. On modelling and diagnostic checking of vector periodic autoregressive time series models. *Time Series Analysis*, v. 30, p. 198–207, 2008.
- WEST, M.; HARRISON, J.; MIGON, H. Dynamic generalized linear models and bayesian forecasting. *Journal of the American Statistical Association*, v. 80, p. 73–83, 1985.
- WILLMOTA, G. E.; SUNDTB, B. On evaluation of the delaporte distribution and related distributions. *Scandinavian Actuarial Journal*, v. 2, p. 101–113, 1989.