UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO
CENTRO TECNOLÓGICO
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA AMBIENTAL

WANDERSON DE PAULA PINTO

O USO DA METODOLOGIA DE DADOS FALTANTES EM SÉRIES TEMPORAIS COM APLICAÇÃO A DADOS DE CONCENTRAÇÃO (PM_{10}) OBSERVADOS NA REGIÃO DA GRANDE VITÓRIA

VITÓRIA
2013

WANDERSON DE PAULA PINTO

O USO DA METODOLOGIA DE DADOS FALTANTES EM SÉRIES TEMPORAIS COM APLICAÇÃO A DADOS DE CONCENTRAÇÃO (PM_{10}) OBSERVADOS NA REGIÃO DA GRANDE VITÓRIA

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia Ambiental do Centro Tecnológico da Universidade Federal do Espírito Santo, como requisito parcial para obtenção do grau de Mestre em Engenharia Ambiental, na área de concentração Modelagem Matemática da Dispersão de Poluentes Atmosféricos em Ambientes Urbanos utilizando modelos estocásticos.
Orientador: Prof. Dr. Valdério Anselmo Reisen.

VITÓRIA
2013

Dados Internacionais de Catalogação-na-publicação (CIP)
(Biblioteca Setorial Tecnológica,
Universidade Federal do Espírito Santo, ES, Brasil)

Paula Pinto, Wanderson de, 1988-

P324u O uso da metodologia de dados faltantes em séries temporais com aplicação a dados de concentração (PM_{10}) observados na região da grande vitória / Wanderson de Paula Pinto. – 2013.

85 f. : il.

Orientador: Valdério Anselmo Reisen.

Dissertação (Mestrado em Engenharia Ambiental) – Universidade Federal do Espírito Santo, Centro Tecnológico.

1. Autocorrelação (Estatística). 2. Ausência de dados (Estatística). 3. Material particulado. 4. Análise de séries temporais. I. Reisen, Valdério Anselmo. II. Universidade Federal do Espírito Santo. Centro Tecnológico. III. Título.

CDU: 628

WANDERSON DE PAULA PINTO

O USO DA METODOLOGIA DE DADOS FALTANTES EM SÉRIES TEMPORAIS COM APLICAÇÃO A DADOS DE CONCENTRAÇÃO (PM_{10}) OBSERVADOS NA REGIÃO DA GRANDE VITÓRIA

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia Ambiental do Centro Tecnológico da Universidade Federal do Espírito Santo, como requisito parcial para obtenção do grau de Mestre em Engenharia Ambiental, na área de concentração Modelagem Matemática da Dispersão de Poluentes Atmosféricos em Ambientes Urbanos utilizando modelos estocásticos.

Aprovada em 27 de agosto de 2013.

COMISSÃO EXAMINADORA

Prof. Dr. Valdério Anselmo Reisen
Universidade Federal do Espírito Santo
Orientador

Prof.(a) Dra. Taciana Toledo de Almeida Albuquerque
Universidade Federal do Espírito Santo
Examinador Interno - PPGEA

Prof. Dr. Manuel Sena Junior
Universidade Federal de Pernanbuco
Examinador Externo - UFPE

Dedico este trabalho, a DEUS, por me dar forças para não desistir dos meus sonhos.

A Renata Helmer, o grande amor da minha vida, pela força, dedicação, compreenssão nos momentos difícies. Você foi o marco fundamental para esta vitória.

Agradecimentos

Agradeço primeiramente a DEUS, por me dar forças para conquistar mais esta vitória em minha vida.

Ao Prof. Valdério Anselmo Reisen, pela orientação sábia, cobrança e paciência.

À minha esposa Renata, pelo constante apoio, incentivo e principalmente por todos os momentos juntos.

Aos meus familiares, em especial a minha mãe Maria Lúcia e aos meus irmãos Diêniffer e Matheus por confiarem em mim.

Aos examinadores, Prof[a]. Taciana Toledo de Almeida Albuquerque e Prof. Manoel Sena Junior pelas sugestões e apontamentos para o enriquecimento do trabalho final.

À todos os professores e funcionários do PPGEA-UFES pelo apoio.

Aos colegas do NUMES, em especial aos amigos Adriano e Emerson.

Ao amigo Éder pelo apoio e incentivo.

À Fundação de Amparo à Pesquisas do Espírito Santo - FAPES, pela concessão de bolsa de estudo.

À todos aqueles que de alguma forma contribuíram para a realização deste trabalho.

Na maior parte das ciências um geração põe abaixo o que outra construiu, e que uma estabeleceu a outra desfaz. Somente na matemática é que cada geração constrói um novo andar sobre a antiga estrutura.

Hermann Hankel

Resumo

Dados da poluição atmosférica apresentam, em geral, observaçẽs faltantes. Esta pesquisa apresenta um estudo de metodologias para estimação da função de autocorrelação na presença de dados faltantes, baseados no trabalho de Yajima e Nishino (1999). Contempla também algumas técnicas para imputação de dados faltantes baseadas no uso do algoritmo EM, proposto por Dempster (1977), e nos modelos de séries temporais ARIMA de Box e Jenkins. Ensaios de simulações com quatro proporções de dados faltantes foram realizadas para comparar os erros quadráticos médios dos estimadores propostos. O estudo empírico evidenciou que o método de estimação sugerido apresenta bom desempenho em termos de medidas de erro quadrático médio. Como ilustração da metodologia proposta, duas séries temporais de concentrações de Material Particulado Inalável (PM_{10}) emitida na Região da Grande Vitória, E.S., Brasil, são analisadas.

Palavras-chave: PM_{10}, função de autocorrelação, dados faltantes.

Abstract

Data of air pollution have generally missing observations. This research presents a study of methods to estimate the autocorrelation function in the presence of missing data, based on the work of Yajima and Nishino (1999). There is also some techniques for imputation of missing data based on the use of the EM algorithm proposed by Dempster (1977), and the ARIMA time series models of Box and Jenkins. Testing simulations with frame proportions of missing data were performed to compare the mean square errors of the proposed estimators. The empirical study showed that the proposed estimation method has good performance in terms of mean squared error measures. As an illustration of the proposed methodology, two time series of concentrations of Inhalable Particulate Matter (PM_{10}) issued in the Region of Vitória, ES, Brazil, are analyzed.

Keywords: PM_{10}, autocorrelation function, missing data.

Lista de Figuras

3.1 Esquema do trato respiratório humano. 24

5.1 fac e facp para uma série temporal gerada por um processo $AR(1)$ com $\phi = 0.9$, $a_t \sim N(0,1)$ e tamanho da amostra $n = 300$. 46

5.2 fac e facp para uma série temporal gerada por um processo $MA(1)$, com $\theta = 0.5$, $a_t \sim N(0,1)$ e tamanho da amostra $n = 300$. 47

5.3 fac e facp para uma série temporal gerada por um processo $ARMA(1,1)$ com $\phi = 0.9$, $\theta = 0.5$, $a_t \sim N(0,1)$ e tamanho da amostra $n = 300$. 47

6.1 Localizaąão espacial das estações de monitoramento da qualidade do ar da RGV. 52

6.2 Parâmetros meteorológicos e poluentes monitorados em cada estação RAMQAR. 52

7.1 Série de concentração de PM_{10}, log da série, histograma da série e do logaritmo da mesma. 72

7.2 Estimaticas da ACF da série 1, com 5% de dados faltantes. 75

7.3 Estimaticas da ACF da série 1, com 40% de dados faltantes. 75

7.4 Série de concentração de PM_{10} com dados faltantes e histograma da mesma. . . 76

7.5 Estimaticas da ACF da série 2. 77

Lista de Tabelas

3.1 Principais poluentes regulamentados pela Resolução CONAMA n° 3 de 28/06/1990 e os seus efeitos sobre a saúde humana e o meio ambiente) 27

3.2 Padrões nacionais de qualidade do ar (Resolução CONAMA n° 3 de 28/06/1990) 30

3.3 Diretrizes da OMS 31

7.1 ACF teórica de um processo $AR(1)$. 60

7.2 Estimativas da ACF para um processo $AR(1)$, com $N = 100$ e $\phi = 0.3$, imputado, com quanto proporções de dados falantes. 60

7.3 Estimativas da ACF para um processo $AR(1)$, com $N = 100$ e $\phi = 0.5$, imputado, com quanto proporções de dados falantes. 61

7.4 Estimativas da ACF para um processo $AR(1)$, com $N = 100$ e $\phi = 0.7$, imputado, com quanto proporções de dados falantes. 61

7.5 Estimativas da ACF para um processo $AR(1)$, com $N = 100$ e $\phi = 0.95$, imputado, com quanto proporções de dados falantes. 62

7.6 Estimativas da função de autocorrelação obtidas para um processo $AR(1)$ com $\phi = 0.3$, através dos estimadores propostos, com tamanho da amostra $N = 100$. 64

7.7 Estimativas da função de autocorrelação obtidas para um processo $AR(1)$ com $\phi = 0.5$, através dos estimadores propostos, com tamanho da amostra $N = 100$. 64

7.8 Estimativas da função de autocorrelação obtidas para um processo $AR(1)$ com $\phi = 0.7$, através dos estimadores propostos, com tamanho da amostra $N = 100$. 65

7.9 Estimativas da função de autocorrelação obtidas para um processo $AR(1)$ com $\phi = 0.95$, através dos estimadores propostos, com tamanho da amostra $N = 100$. 65

7.10 ACF teórica de um processo $AR(1)$, com $\phi = 0.99$. 67

7.11 Estimativas da ACF para um processo $AR(1)$, com $\phi = 0.99$, através dos estimadores propostos e algoritmo EM, com tamanho de amostra $N = 100$. 67

7.12 Estimativas da ACF para um processo $AR(1)$, com $\phi = 0.3$, através dos estimadores propostos e algoritmo EM, com tamanho de amostra $N = 1000$. 68

7.13 Estimativas da ACF para um processo $AR(1)$, com $\phi = 0.99$, através dos estimadores propostos e algoritmo EM, com tamanho de amostra $N = 1000$. 68

7.14 Estimativas da ACF para um processo (Qui-Quadrado g.l=2) $AR(1)$, com $N = 100$ e $\phi = 0.7$, imputado, com quanto proporções de dados falantes. 69

7.15 Estimativas da função de autocorrelação obtidas para um processo $AR(1)$ (Qui-Quadrado g.l=2) com $\phi = 0.7$, através dos estimadores propostos, com tamanho da amostra $N = 100$. 69

7.16 Estimativas da ACF para um processo (t de Student g.l=3) $AR(1)$, com $N = 100$ e $\phi = 0.7$, imputado, com quanto proporções de dados falantes 70

7.17 Estimativas da função de autocorrelação obtidas para um processo $AR(1)$ (t de Student g.l=3) com $\phi = 0.7$, através dos estimadores propostos, com tamanho da amostra $N = 100$. 70

7.18 ACF teórica de um processo $ARMA(1,1)$, com $\phi = 0.7$ e $\theta = 0.3$. 70

7.19 Estimativas da função de autocorrelação obtidas para um processo $ARMA(1,1)$ com $\phi = 0.7$ e $\theta = 0.3$, através dos estimadores propostos, com tamanho da amostra $N = 100$. 71

7.20 Estatísticas descritivas da série de concentração das médias diárias PM_{10}. . . . 72

7.21 Performance dos métodos de imputação com quanto proporções de dados falantes, *log* da série de PM_{10}. 73

7.22 ACF da série 1 de concentrações PM_{10} com 5% de dados faltantes. 74

7.23 ACF da série 1 de concentrações PM_{10} com 40% de dados faltantes. 74

7.24 ACF da série 2 de concentrações PM_{10} com dados faltantes. 77

Sumário

Resumo **vi**

Abstract **vii**

Lista de Figuras **viii**

Lista de Tabelas **ix**

1 Introdução **15**

2 Objetivo **20**

2.1 Objetivo Geral 20

2.2 Objetivos Específicos 20

3 Revisão Bibliográfica **21**

3.1 Poluente atmosférico 21

3.2 Material Particulado (PM_{10}) 22

3.2.1 Efeitos sobre a saúde 23

3.3 Padrões de Qualidade do Ar 29

3.4 Modelagem e previsão da qualidade do ar 31

3.5 Revisão dos estudos que utilizaram ferramentas estatísticas para tratar dados faltantes em séries temporais de concentrações de poluentes atmosféricos . . . 33

4 Conceitos Básicos em Séries Temporais 36

4.1 Processos estocásticos . . . 36

4.2 Funções de autocovariância e autocorrelação . . . 37

4.3 Função de autocorrelação parcial . . . 38

4.4 Ruído branco . . . 39

4.5 Modelos estacionários . . . 40

4.5.1 autorregressivo (AR) . . . 40

4.5.2 Modelo de médias móveis (MA) . . . 40

4.5.3 Modelo autorregressivo e de médias móveis . . . 41

4.6 Modelos não estacionários . . . 43

4.6.1 Modelo Autorregressivo Integrado de Média Móvel [ARIMA (p,d,q)] . 43

4.7 Metodologia de modelagem . . . 44

5 Estimação da Função de Autocorrelação 46

5.1 Estimadores da função de autocorrelação na presença de dados faltantes . . . 48

6 Materiais e Métodos 51

6.1 Região de estudo . . . 51

6.2 Dados . . . 52

6.3 *Software* R . . . 53

6.4 Métodos de imputação . . . 53

6.4.1 Imputação por constantes 54

6.4.2 Imputação via algoritmo EM 55

6.5 Indicadores de performance 57

6.6 Recursos computacionais 58

6.7 Estudo das Simulações 58

7 Resultados **59**

7.1 Resultados 59

7.1.1 Simulações 59

7.1.2 Aplicações 72

8 Conclusões e Trabalhos Futuros **78**

8.1 Conclusões 78

8.2 Trabalhos Futuros 79

9 Referências Bibliográficas **80**

Capítulo 1

Introdução

A preocupação com efeitos da poluição do ar veio com o crescimento industrial iniciado no período da Revolução Industrial, devido a alguns episódios de alta concentração de poluentes ocorridos no início do século passado. Episódios de poluição excessiva que causaram aumento do número de mortes em algumas cidades da Europa e Estados Unidos. O primeiro episódio ocorreu em 1930, no vale de Meuse, Bélgica, anos mais tarde, um episódio semelhante ocorreu durante os últimos cinco dias do mês de outubro de 1948 na cidade de Donora, Pensilvânia. Porém, um dos mais graves episódios acerca dos efeitos deletérios dos poluentes do ar que se tem notícia ocorreu em Londres durante o inverno de 1952. Um evento de inversão térmica impediu a dispersão de poluentes, gerados então pelas indústrias e pelos aquecedores domiciliares que utilizavam carvão como combustível, e uma nuvem, composta por material particulado e enxofre, em concentrações muito acima do normal, permaneceu sobre a cidade por três dias, deixando cerca de 4 mil pessoas mortas por causa da poluição do ar. Esses episódios acarretaram em um aumento do número de óbitos em relação à média de óbitos em períodos semelhantes (BRAGA *et al.* 2005).

A poluição atmosférica é um fator inerente na vida humana. Nossos ancestrais já conviviam com a poluição natural oriunda das erupções vulcânicas e decomposição da matéria orgânica. Com a globalização, intensificou-se as atividades antropogênicas que contribuiu ainda mais para a deteriorazação da qualidade do ar. Com o crescimento populacional, desenvolvimento econômico e o crescimento da frota motorizada, as fontes de poluição multiplicaram-se agravando ainda mais o problema, mesmo em áreas não industrializadas (LIRA, 2009).

A poluição do ar é um tema extensivamente estudado nas últimas décadas e atualmente caracteriza-se como um fator importante na busca da prevenção do meio ambiente (LIRA, 2009). Segundo Gomes (2009) o monitoramento da qualidade do ar é realizado para determinar o nível de concentração dos poluentes presentes na atmosfera. Seus resultados permitem fazer um acompanhamento sistemático da qualidade do ar na região monitorada e também, os resultados, constituem elementos básicos para elaboração de diagnósticos da qualidade do ar, que podem subsidiar ações para o controle de emissões.

Lira (2009) defende que a avaliação da qualidade do ar não envolve somente o monitoramento da qualidade do ar, ela engloba a identificação das principais fontes que causam a poluição medida, os estudos de tendência, as estimativas de poluição em áreas não monitoradas, e a previsão de impacto na qualidade do ar de fontes ainda não instaladas. O conhecimento prévio dos níveis dos poluentes na atmosfera de uma região pode ser últil para fornecer dados para ativar ações de emergência durante períodos de estagnação atmosférica, quando os níveis de poluentes na atmosfera possam representar risco à saúde pública.

O monitoramento da qualidade do ar é realizado para determinar o nível de concentração dos poluentes presentes na atmosfera. Seus resultados permitem fazer um acompanhamento sistemático da qualidade do ar na região monitorada e também, constituem elementos básicos para elaboração de diagnósticos, que podem subsidiar ações para o controle de emissões.

A avaliação da qualidade do ar não envolve somente seu monitoramento, engloba também a identificação das principais fontes que causam a poluição medida, os estudos de tendência, as estimativas de poluição em áreas não monitoradas, e a previsão de impacto de fontes ainda não instaladas. O conhecimento prévio dos níveis dos poluentes na atmosfera de um região pode ser últil para fornecer dados para ativar ações de emergência durante períodos de estagnação atmosférica, quando os níveis de poluentes na atmosfera possam representar risco à saúde pública.

Fatores como densidade populacional, atividades industriais, aumento da frota veicular e condições meteorológicas contribuem para o aumento da poluição atmosférica. Um poluente muito comum nas regiões urbanas é o Material Particulado Inalável (PM_{10})(PEREZ e REYES, 2002). Define-se por material particulado as partículas de material sólido e líquido capazes de permanecer em suspensão na atmosfera devido a suas pequenas dimensões, como exemplos têm-se a poeira, a fuligem e as partículas de óleo (BRAGA *et al.* 2005). O termo PM_{10}, ou partículas inaláveis, se refere à fração do particulado que contém partículas com diâmetro aerodinâmico

menor que 10 μm (HOLGATE,*et al.* 1999).

Segundo Holgate *et al.* (1999) um nível elevado dos poluentes pode ocasionar desde irritação dos olhos, nariz e da garganta, bronquite e pneumonia até doenças respiratórias crônicas, câncer de pulmão, problemas cardíacos, etc. Diversos trabalhos científicos associaram a exposição a concentrações elevadas de PM_{10} na atmosfera a problemas de saúde (OSTRO *et al.*, 1996; ALMEIDA, 2006; NASCIMENTO *et al.*, 2006 e BARBOSA, 2009).

Para o gerenciamento da qualidade do ar é necessário conhecer as concentrações de poluentes e gerar previsões satisfatórias delas. A utilização de modelos de previsão é uma ferramenta importante para conhecer o comportamento e características de determinados poluentes, podendo, desta forma, prever possíveis picos de concentração. Para isto, pode-se fazer uso de duas classes de modelos, os experimentais e os matemáticos. Nesta última, têm-se os modelos determinísticos e os modelos estocásticos. O presente estudo se concentrará na classe de modelagem estocástica.

Com a difusão de metodologias para análises de séries temporais na literatura científica, diversos trabalhos foram publicados com análise de séries temporais de dados ambientais. Na área da poluição atmosférica, pode-se citar autores que usaram desta metodologia para estudar e analisar séries de concentrações de poluentes atmosféricos como Shively (1990), Robeson e Steyn (1990) e Goyal *et al.* (2006).

Vários métodos de previsão de séries temporais estão disponíveis na literatura, como o de médias móveis (MA), regressão linear com o tempo, suavização exponencial de Holt-Winters e os Modelos ARIMA (Modelo Autoregressivo Integrado de Média Móvel). Os modelos ARIMA proporcionam previsões probabilísticas, apresentado certa facilidade de implantação. Esses modelos apresentam diversas vantagens em relação aos outros modelos, como alisamento exponencial, em particular em sua capacidade de previsão e a grande quantidade de informações sobre mudanças relacionadas ao tempo (MISHA e DESAI, 2005).

Um problema frequente em séries temporais provenientes de monitoramentos da qualidade do ar é a presença de dados faltantes *(missing data)*. Estes dados ocorrem, geralmente, pois os equipamentos de medição das concentrações de contaminantes na atmosfera podem apresentar defeitos que impossibilitem seu funcionamento por algum tempo, ocasionando perda de dados. A análise de dados, incluindo apenas as observações disponíveis sem um tratamento estatístico

para os dados faltantes, pode produzir estimativa falsa da medida de efeito e subestimar sua precisão (JUNGER, 2008). Este problema tem sido extensivamente estudado. Várias metodologias foram desenvolvidas para tal, pode-se citar os trabalhos de Toda e Mackenzie, 1999; Yajima e Nishino, 1999; Dempster *et al.*, 1976; Robinson, 1981.

Entre as metodologias para o tratamento de dados faltantes em séries temporais estão os métodos de imputação usados em trabalhos que envolvem séries de concentração PM, dentre estes, Junninen *et al.* (2004), que avaliaram dois contextos (univariada e multivariada) e Plaia e Bandi (2006) propuseram uma metodologia de imputação de dados faltantes denominada de *Site-Depen-Dente* (SDEM). Os procedimentos de imputação de dados faltantes consistem em preencher os valores em falta e analisar o conjunto de dados resultantes usando métodos convencionais. Alguns procedimentos de imputação são simples e implementados na maioria dos aplicativos estatísticos. A principal desvantagem dos métodos de imputação é que, em sua maioria, a imprecisão devida à imputação não é contemplada na análise, portanto, a variância dos estimadores é subestimada (JUNGER, 2008).

A utilização do algoritmo EM *(expectation-maximisation)* (DEMPSTER, 1977), é uma alternativa para preencher dados faltantes em séries temporais que são espacialmente referenciadas. O algoritmo EM consiste em um processo interativo envolvendo dois passos: previsão e estimação, que permitem calcular estimativas de máxima verossimilhança a partir de dados incompletos supondo que o padrão dos valores que estão faltando seja faltando ao acaso ou *missing at random*. Na imputação via algoritmo EM, incialmente as estimativas iniciais do vetor de médias e da estrutura de covariância são obtidos utilizando o conjunto incompleto de dados. Em seguida, atribui-se o mesmo modelo ARIMA (p,d,q) para cada uma das séries temporais provenientes das diferentes estações de monitoramento. Daí o procedimento é composto por: a) substituí-se os valores faltantes por alguma estimativa; b) estima-se os parâmetros do modelo ARIMA (p,d,q); c) atualiza-se os valores faltantes por estimativas obtidas a partir do modelo ARIMA ajustado; d) reestima-se os parâmetros do modelo ARIMA com os dados completos segundo o passo anterior (JUNGER, 2008).

Pode-se ainda, citar outros autores que exploraram essas metodologias para preencher observações faltantes em séries temporais de concentração de poluentes (PM_{10}) como Iglesias, Jorquera e Palma (2005), Norazian *et al.* (2008). Entretanto, há uma dificuldade em avaliar qual metodologia adotar, pois segundo Schafer (1997), alguns procedimentos são simples e acabam produ-

zindo estimativas viesadas, outros, mais sofisticados dependem de fortes pressupostos sobre o mecanismo gerador do padrão dos dados faltantes e complicadas implementações computacionais. Diante do problema exposto, neste trabalho, se faz uma investigação mais cuidadosa para avaliar diferentes metodologias para estimação de dados faltantes em séries temporais, estima-se a função de autocorrelação de séries considerando a presença de dados faltantes, com base nos estimadores propostos por Yajima e Nishino (1999). Por fim, como aplicação da metodologia testada, empiricamente, foi feito a análise de duas séries de concentração de material particulado inalável (PM_{10}).

Este trabalho é apresentado da seguinte forma: O capítulo 2 descreve os objetivos da pesquisa. No capítulo 3 tem-se a revisão bibliográfica. No capítulo 4 são apresentados alguns conceitos básicos usados no estudo de séries temporais. No capítulo 5 apresentamos os estimadores propostos por Yajima e Nishino (1999) para estimar a função de autocorrelação na presença de dados falatantes. No capítulo 6 é feito a descrição dos materiais e métodos utilizados no trabalho. Resultados de simulação e aplicações encontram-se no capítulos 7. Finalmente, no capítulo 8 contém as principais conclusões deste trabalho e sugestões para trabalhos futuros.

Capítulo 2

Objetivo

2.1 Objetivo Geral

O objetivo geral deste trabalho é avaliar diferentes metodologias para o tratamento de dados faltantes em séries temporais de concentração de material particulado inalável (PM_{10}) observados na Região da Grande Vitória.

2.2 Objetivos Específicos

- Utilizar metodologias para preencher dados faltantes em séries temporais das concentrações de material particulado inalável.
- Estudar diferentes metodologias, com quatro proporções, para tratamento de dados faltantes de concentração de PM_{10} monitorados na atmosfera urbana.
- Realizar um estudo empírico para avaliar as metodologias propostas para dados faltantes em séries temporais.
- Estimar a Função de Autocorrelação de séries temporais considerando a presença de dados faltantes (YAJIMA E NISHINO, 1999).

Capítulo 3

Revisão Bibliográfica

No presente capítulo são abortados alguns conceitos e estudos já realizados na área relacionada com o tema desta dissertação. Este capítulo está organizado em 5 seções e apresenta na primeira um breve resumo sobre a definição de poluente atmosférico. A segunda contém a definição de material particulado PM_{10} e os efeitos à saúde. Na terceira seção está descrito os Padrões de Qualidade do Ar. Na quarta é descrito estudos envolvendo modelagem e previsão da qualidade do ar e a quinta, descreve uma breve revisão dos principais estudos que utilizaram ferramentas estatísticas para preencher dados faltantes em séries temporais de concentrações de poluentes atmosféricos.

3.1 Poluente atmosférico

De acordo com a Resolução CONAMA nº 03, de junho de 1990, em seu artigo 1º, poluente atmosférico é qualquer forma de matéria ou energia com intensidade e quantidade, concentração ou características em desacordo com os níveis estabelecidos e que torne ou possam tornar o ar impróprio, nocivo ou ofensivo à saúde, inconveniente ao bem estar público, danoso aos materiais, flora ou fauna ou prejudicial à segurança, ao uso e gozo da propriedade e às atividades normais da comunidade (BRASIL, 1990).

As emissões de poluentes atmosféricos podem classificar-se em antropogênicas e naturais (LORA, 2002). As antropogênicas são aquelas provocadas pela ação do homem (indústria, transporte, geração de energia e outras). Já as naturais são causadas por processos naturais, tais como

emissões vulcânicas, processos microbiológicos, etc. Quanto a sua origem os poluentes são classificados em primários e secundários. Os poluentes primários são aqueles lançados diretamente na atmosfera, como resultado de processos industriais, gases de exaustão de motores de combustão interna, dentre outros. Como exemplo, pode-se citar os óxidos de enxofre SO_x, os óxidos de nitrogênio NO_x e particulados. Já os secundários são aqueles formados a partir de reações químicas, que ocorrem na atmosfera entre os poluentes primários.

3.2 Material Particulado (PM_{10})

No caso de poluição atmosférica, entende-se por material particulado as partículas de material sólido e líquido capazes de permanecer em suspensão na atmosfera devido a suas pequenas dimensões, como exemplos têm-se a poeira, fuligem e partículas de óleo (BRAGA *et al.* 2005; SEINFELD e PANDIS, 1998). O material particulado presente na atmosfera pode ser proveniente tanto de fontes naturais como de fontes antropogênicas, e sua forma e composição química pode ser bastante diversificada (BAIRD, 2002). As principais fontes de poluição por material particulado são indústrias, incineradores, veículos, atividades de construção civil, poeira, etc. O tamanho das partículas varia, $PM_{2.5}$ e PM_{10} para diâmetro aerodinâmico menor que 2.5 μm e 10 μm, respectivamente, e ainda, as partículas são classificadas como grossas ou finas, dependendo de seu diâmetro ser maior ou menor que 2.5 μm (BAIRD, 2002).

Partículas finas: são as partículas inaláveis com diâmetro aerodinâmico inferior a 2.5 μm. São as principais responsáveis pelos efeitos à saúde, uma vez que podem atingir o sistema respiratório inferior e as partículas grossas: são as partículas com diâmetro aerodinâmico entre 2.5 e 10 μm proveniente de fontes como, por exemplo, processos metalúrgicos ou estradas (HOLGATE*et al.*, 1999).

A principal preocupação com a presença de material particulado na atmosfera é que este poluente causa sérios danos à saúde dos seres humanos e aos animais. Outro problema do material particulado na atmosfera é que este poluente reduz a visibilidade (TRINDADE, 2009).

Estudos realizados confirmam em apontar que efeitos causados pela formação e o aumento da concentração de material particulado na atmosfera, estão relacionados ao aumento do risco de ocorrência de doenças respiratórias crônicas (GODISH, 1997)

O termo PM_{10}, ou partículas inaláveis, se refere à fração do particulado que contém partículas com diâmetro aerodinâmico menor que 10 μm (HOLGATE*et al.*, 1999).

3.2.1 Efeitos sobre a saúde

A poluição atmosférica tem afetado de forma significativa a vida dos seres vivos, mesmo quando seus valores estão abaixo do permitido pelos órgãos regulamentadores. As crianças e os idosos estão entre os grupos que têm se mostrado mais vulneráveis aos efeitos da poluição do ar (MARTINS, *et al.*, 2002).

Os efeitos dos poluentes atmosféricos variam em função do tempo de exposição e de suas concentrações. De forma geral, os efeitos podem ser classificados como agudos e crônicos (LIRA, 2009).

- **Agudos:** São de caráter temporários, estão relacionados a exposição a altas concentrações e os seus efeitos são imediatos.
- **Crônicos:** Os efeitos são de caráter permanente, está relacionado a exposição a baixas concentrações de poluentes e são a longo prazo.

O material particulado pode causar danos à saúde de humanos e de outros animais,mas estes por sua vez dependem da capacidade das partículas de penetrarem no sistema respiratório. As partículas de PM_{10} depositam-se principalmente no trato respiratório superior, enquanto que as partículas ultrafinas são capazes de atingir os alvéolos pulmonares (BORBOSA; TRINDADE, 2009).

Em geral, os mecanismos de defesa são adequados para remover as partículas inaladas maiores que 10 μm (GODISH, 1997). As partículas menores de 10 μm de diâmetro (PM_{10}) são chamadas de inaláveis, pois ficam retidas nas vias aéreas traqueobrônquica, conforme representado na figura 3.1 (TRINDADE, 2009).

Diversos estudos epidemiológicos têm demonstrado associações significativas entre as concentrações do poluente PM_{10} e saúde. As crianças e os idosos são os dois grupos etários mais suscetíveis aos efeitos da poluição atmosférica. Estudos mostram uma associação positiva entre mortalidade e morbidade (internações) por problemas respiratórios em crianças. Já entre os

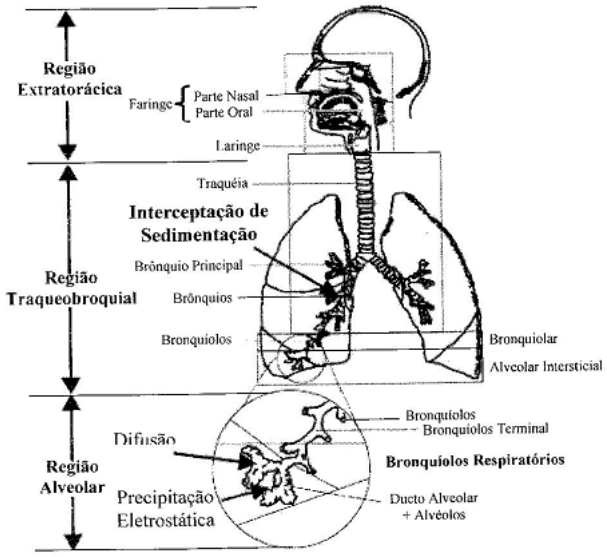

Figura 3.1: Esquema do trato respiratório humano.

Fonte: adaptado de Ruzer e Harley (apund TRINDADE, 2009, p. 26)

idosos, associa-se a poluição atmosférica ao aumento de mortalidade, por doenças respiratórias e cardiovasculares. Entre estes estudos podemos citar os seguintes.

Martins *et al.*, (2002) estudaram os efeitos causados pela poluição atmosférica na morbidade por gripe e por pneumonia em idosos no período de 1996 e 1998. Os dados diários de atendimentos por pneumonia e gripe em idosos foram obtidos em um pronto socorro médico de um hospital escola de referência no Município de São Paulo. Os níveis diários de CO, O_3, SO_2, NO_2 e PM_{10} foram obtidos na Companhia de Tecnologia de Saneamento Ambiental (CETESB), e os dados diários de temperatura mínima e umidade relativa do ar foram obtidos no Instituto Astronômico e Geofísico da USP. Utilizou-se o modelo aditivo generalizado de regressão de Poisson para verificar a relação entre a pneumonia, gripe e poluição atmosférica, tendo com variável dependente o número diário de atendimentos por pneumonia e gripe e como variável independente as concentrações médias diárias dos poluentes atmosféricos. Os resultados encontrados relataram que os poluentes O_3 e SO_2 estão diretamente associados à pneumonia e à gripe. Pôde-se observar que um aumento interquartil para os poluentes $O_3(38,80\mu g/m^3)$ e $SO_2(15,05\mu g/m^3)$ ocasionaram um acréscimo de 8,07% e 14,51%, respectivamente no número de atendimento por pneumonia e gripe em idosos, resultando que a poluição atmosférica promove efeitos adversos para a saúde de idosos.

Castro *et al.*, (2009) realizaram um estudo na cidade do Rio de Janeiro com uma amostra aleatória de 118 escolares, com idade entre seis e 15 anos, da rede pública, residentes até 2

Km do local do estudo. As informações sobre a qualidade do ar foram obtidas por meio de uma unidade móvel de monitoramento dos poluentes da Secretaria Municipal de Meio Ambiente do Rio de Janeiro, no local de estudo. Os autores ultilizaram dados dos poluentes PM_{10}, O_3, SO_2, NO_2 e CO como indicadores diários da poluição atmosférica das crianças sob a mesma condição de exposição. As condições metereológicas foram obtidas por meio de medidores localizados no Aeroporto do Galeão. Foram utilizadas as temperaturas mínima, média e máxima e a umidade relativa do ar. Os resultados encontrados mostram que mesmo dentro dos níveis acetáveis, a poluição atmosférica, principalmente por PM_{10} e NO_2, esteve associada à diminuição da função respiratória no Rio de Janeiro.

Nascimento *et al.*, (2006) usaram dados diários do número de internações por pneumonia na cidade de São José dos Campos - SP, dados de concentrações médias diárias dos poluentes SO_2, O_3 e PM_{10}, além de dados de dois parâmetros meteorológicos: temperatura e umidade relativa do ar. Os autores utilizaram modelos aditivos generalizados de regressão de Poisson para estimar a associação entre as internações por pneumonia e a poluição atmosférica. Os três poluentes apresentaram efeitos defasados nas internações por pneumonia, iniciada três a quatro dias após a exposição e decaindo rapidamente. Na estimativa de efeito acumulado de oito dias, observou-se, ao longo desse período, que para aumentos de 24,7 μ/m^3 na concentração de PM_{10}, houve um acréscimo de 9,8 % nas internações.

Braga *et al.*, (2007) avaliaram os efeitos agudos do PM_{10} sobre os atendimentos em pronto-socorro por doenças cardiovasculares e respiratórias no Município de Itabira - MG. Os resultados evidenciam que aumentos de 10 μ/m^3 de PM_{10} foram associados aos aumentos nos atendimentos por doenças respiratórias em torno de 4%, no dia corrente e no dia seguinte, para crianças menores de 13 anos, e de 12%, nos três dias subsequentes para os adolescentes entre 13 e 19 anos de idade. Já por doenças cardiovasculares, o efeito agudo, principalmente para os indivíduos com idade entre 45 e 64 anos.

Na mesma linha de investigação Barbosa (2009) realizou um estudo na RGV para analisar os efeitos da poluição do ar em crianças de 0 a 6 anos residente no Município de Serra. O autor analisou a associação entre as concentrações dos poluentes atmosféricos, material particulado (PM_{10}), dióxidos de nitrogênio (NO_2), ozônio (O_3) e o número de atendimentos hospitalares por doenças aéreas respiratórias em crianças, no período de janeiro de 2001 a dezembro de 2004, utilizando o Modelo Aditivo Generalizado e a técnica de Bootstrap. Foi considerado como des-

fecho (variável de interesse) os atendimentos por doenças aéreas respiratórias em crianças de 0 a 6 anos de idade, esses dados foram coletados no banco de dados do Hospital Infantil Nossa Senhora da Glória (HINSG), localizado no município de Vitória, foram considerados apenas os atendimentos em crianças residentes no município de Serra, as variáveis explicativas foram às concentrações médias diárias dos poluentes NO_2, O_3 e PM_{10} e as variáveis meteorológicas consideradas como fatores de confusão como médias diárias da temperatura máxima, média e mínima e umidade relativa do ar obtidos junto ao Instituto Estadual de Meio Ambiente e Recursos Hídricos (IEMA). Os resultados mostram que os níveis de concentrações dos poluentes em ambientes urbanos afeta diretamente os atendimentos hospitalares em crianças menores de 6 anos por doenças respiratórias.

Os principais poluentes atmosféricos, bem como o resumo de seus efeitos sobre saúde humana e o meio ambiente, são descritos na tabela 3.1.

Tabela 3.1: Principais poluentes regulamentados pela Resolução CONAMA n° 3 de 28/06/1990 e os seus efeitos sobre a saúde humana e o meio ambiente)

Poluente	**Características**	**Principais Fontes**	**Efeitos adversos à saúde**	**Efeitos gerais ao meio ambiente**
Monóxido de Carbono (CO)	Gás incolor, inodoro e insípido	Combustão incompleta de combustíveis fósseis (veículos automotores principalmente) e outros materiais que contenham carbono na sua composição	Combina-se rapidamente com a hemoglobina ocupando o lugar do oxigênio, podendo levar a morte por asfixia. A exposição crônica pode causar prejuízos ao sistema nervoso central, cardiovascular pulmonar e outros. Também pode afetar fetos causando peso reduzido no nascimento e desenvolvimento pós-natal retardado.	
Material Particulado (PTS e PM_{10})	São poeiras, fumaças e todo tipo de material sólido e líquido que, devido ao seu pequeno tamanho, se mantém suspenso na atmosfera.	Variam desde processos industriais, passando por veículos automotores e poeira de rua ressuspensa, até fontes naturais como pólen, aerossol marinho e solo.	As partículas inaláveis (PM_{10}) são as que causam maiores prejuízos à saúde, uma vez que não são retidas pelas defesas do organismo. Essas podem causar irritação nos olhos e na garganta, reduzindo a resistência às infecções e ainda provocando doenças crônicas. Além disso, atingem as partes mais profundas dos pulmões, transportando para o interior do sistema respiratório substâncias tóxicas e cancerígenas.	Alteração da visibilidade; alteração no balanço de nutrientes de lagos, rios e do solo; danificação da vegetação e alteração na diversidade do ecossistema. Além disso, pode causar danos estéticos (manchas e danificações de rochas e outros materiais).
Dióxido de Enxofre (SO_2)	Gás incolor com forte odor semelhante ao produzido na queima de palitos de fósforo).	Processos que utilizam queima de óleo combustível, refinaria de petróleo, veículos a diesel, polpa e papel.	A inalação, mesmo em concentrações muito baixas, provoca espasmos passageiros dos músculos lisos dos bronquíolos pulmonares. Em concentrações progressivamente maiores, causam o aumento da secreção mucosa nas vias respiratórias superiores, inflamações graves da mucosa e redução do movimento ciliar do trato respiratório. Pode, ainda, aumentar a inidência de rinite, faringite e bronquite.	Em certas condições, o SO_2 pode transformar-se em trióxido de enxofre (SO_3) e, com a umidade atmosférica, transforma-se em ácido sulfúrico, sendo assim um dos componentes da chuva ácida.

Óxidos de Nitrogênio (NO_x)	Gases.	Combustões em veículos automotores, industrias, usinas térmicas que utilizam óleo ou gás e incineradores.	O NO_2 é altamente tóxico ao homem, pois aumenta sua susceptibilidade aos problemas respiratórios em geral. Além disso, é irritante às mucosas e pode, nos pulmões, ser transformado em nitrosaminas (alguma das quais são carcinogênicas).	Pode levar a formação da chuva ácida e consequentemente danos à vegetação e agricultura. Além disso, contribui para formação do ozônio na troposfera; para o aquecimento global; formação de compostos quimiotóxicos e alteração da visibilidade.
Ozônio (O_3)	Gás incolor e inodoro nas concentrações ambientais, sendo o principal componente do *smog* fotoquímico.	Formação, na troposfera, a partir da reação dos hidrocarbonetos e óxidos de nitrogênio na presença de luz solar.	Provoca danos na estrutura pulmonar, reduzindo sua capacidade e diminuindo a resistência às infecções. Causa ainda, o agravamento de doenças respiratórias, aumentando a incidência de tosse, asma, irritações no trato respiratório superior e nos olhos.	É agressivo às plantas, agindo como inibidor da fotossíntese e produzindo lesões características nas folhas.

Fonte: Adaptado de IEMA, 2007

3.3 Padrões de Qualidade do Ar

Um padrão de qualidade do ar (PQAR) define legalmente o limite máximo para a concentração de um componente atmosférico que garanta a proteção da saúde e do bem-estar da população, bem como da fauna, flora, materiais e do meio ambiente em geral; sendo que a máxima concentração de um poluente é determinada em função de um período médio de tempo (IEMA, 2007).

Os padrões de qualidade do ar são baseados em estudos científicos dos efeitos produzidos por poluentes específicos e são fixados em níveis que possam propiciar uma margem de segurança adequada.

Os padrões nacionais de qualidade do ar foram estabelecidos pelo Instituto Brasileiro de Meio Ambiente (IBAMA) e aprovados pelo Conselho Nacional do Meio Ambiente (CONAMA), por meio da Resolução CONAMA 03/90. Essa Resolução define padrões de qualidade do ar como aquelas concentrações de poluentes atmosféricos que, ultrapassados, poderão afetar a saúde, a segurança e o bem-estar da população, assim como ocasionar danos à flora e à fauna, aos materiais e ao meio ambiente em geral (LIRA, 2009).

A Resolução CONAMA 03/90 estabelece dois tipos de padrões de qualidade do ar: os primários e os secundários.

- **Padrões primários de qualidade do ar**: As concentrações de poluentes que, ultrapassadas, poderão afetar a saúde da população. Podem ser entendidos como níveis máximos toleráveis de concentração de poluentes atmosféricos, constituindo-se em metas de curto e médio prazo.
- **Padrões secundários de qualidade do ar**: As concentrações de poluentes atmosféricos abaixo das quais se prevê o mínimo efeito adverso sobre o bem estar da população, assim como o mínimo dano à fauna e à flora, aos metais e ao meio ambiente em geral. Podem ser entendidos como níveis desejados de concentração de poluentes, constituindo-se em meta de longo prazo.

Segundo Hinrichs, Kleinbach e Reis (2011) o estabelecimento de padrões de qualidade do ar é uma tarefa complexa. Existe uma grande variação na susceptibilidade de diferentes pessoas a diferentes poluentes. Também existem efeitos sinérgicos a serem considerados, já que a poluição atmoférica atua somando-se aos efeitos de outras substâncias.

Os poluentes e seus padrões de qualidade do ar, assim como o tempo de amostragem e os métodos de medição, fixados pela Resolução CONAMA nº 03 de 28/06/90 são apresentados na tabela 3.2. Os parâmetros regulamentados são: partículas totais em suspensão (PTS), partículas inaláveis (PM_{10}), dióxido de enxofre (SO_2), monóxido de carbono (CO), ozônio (O_3) e dióxido de nitrogênio (NO_x). As concentrações-padrão são expressas em ppm ou microgramas por metro cúbico.

Tabela 3.2: Padrões nacionais de qualidade do ar (Resolução CONAMA n° 3 de 28/06/1990)

Poluente	Tempo de Amostragem	Padrão Primário μgm^3	Padrão Secundário μgm^3	Método de Amostragem
PTS	24 horas[1]	240	150	separação/inercial
	MGA[2]	80	60	filtração
PM_{10}	24 horas[1]	150	150	separação/inercial
	MAA[3]	50	50	filtração
SO_2	24 horas	365	100	Infravermelho
	MAA[3]	80	40	não dispersivo
NO_2	1 hora	320	190	Quimiluminescência
	MAA[3]	100	100	
CO	1 hora	40.000/35 ppm	40.000/35 ppm	Pararosanilina
	8 horas	10.000 (9ppm)	10.000 (9ppm)	
O_3	1 hora[1]	160	160	Quimiluminescência

Fonte: CONAMA (1990)

(1) Não deve ser excedido mais que uma vez ao ano

(2) Média geométrica anual

(3) Média aritmética anual

A Organização Mundial de Saúde (OMS) estabelece diretrizes para a qualidade do ar com o objetivo de minimizar os riscos oferecidos à saúde pela emissão de poluentes (WHO, 2005). As diretrizes da Organização Mundial de Saúde (OMS), em relação ao poluente PM_{10} são apresentados na tabela 3.3.

Estas diretrizes não são padrões legais de qualidade do ar, elas têm o objetivo de prover uma base de informações de proteção à saúde pública e servem de orientação para o estabelecimento de padrões de qualidade do ar (LIRA, 2009).

Tabela 3.3: Diretrizes da OMS

Poluente	Tempo de amostragem	Concentração μ/m^3
$PM_{2.5}$	24 horas	25
	MAA[1]	10
PM_{10}	média de 24 horas	50
	MAA[1]	20
O_3	média de 8 horas	100
NO_2	média de 1 hora	200
	MAA[1]	40
SO_2	média de 10 minutos	500
	média de 24 horas	20

Fonte: Air Quality Guidelines, WHO, 2005

(1) Média aritmética anual

Tais valores de referência podem e devem considerar não apenas os aspectos de saúde e meio ambiente, mas também a viabilidade técnica, considerações econômicas e principalmente os fatores políticos e sociais.

3.4 Modelagem e previsão da qualidade do ar

A previsão da qualidade do ar pode ser utilizada como ferramenta de alerta sobre a concentração de poluentes na atmosfera e permitir a tomada de decisão quanto à adaptação de comportamento da população e grupos de risco, como crianças, idosos e pessoas com doenças respiratórias. Pode também servir para as autoridades competentes como informação para a preparação de planos para a redução de emissões e gerenciamento da qualidade do ar (GOMES, 2009).

Além disso na aplicação da metodologia de séries temporais, um dos principais objetivos é fazer a comparação entre os modelos quanto ao seu desempenho no processo de ajustamento aos dados em estudo e uma das formas de verificar essa qualidade é o estudo de previsão.

A seguir são citados alguns trabalhos que utilizaram modelos estatísticos para modelar e fazer previsão da qualidade do ar.

Agirre-Basurho, Ibarra-Berastegi e Madariaga (2006) utilizaram três modelos, um de Regressão Linear Múltipla e dois modelos de rede *perceptron* ou *Multilayer perceptron*, para modelar e prever a qualidade do ar da cidade de Bilbao, Espanha. Os modelos usaram como dados de entrada o fluxo de veículos e as variáveis meteorológicas, temperatura, umidade relativa, pressão, radiação, gradiente de temperatura, direção do vento e velocidade do vento, no período de 1993 à 1994 e, como saída prevista pelos modelos, tem-se a concentração de O_3 e NO2 com horizonte de previsão de 8 horas à frente. Os resultados mostram que os modelos de rede *perceptron* obtiveram resultados melhores para a previsão das concentrações de O_3 e NO2 quando comparados ao modelos de Regressão Linear Múltipla. Quanto ao desenpenho dos modelos de rede *perceptron* o que mais se destacou foi o modelo que considerou a sazonalidade da série das concentrações de O_3 e NO2.

Goyal, Chan e Jaiswal (2006) realizaram um estudo com três modelos estatísticos aplicados à média diária de concentração de PM_{10} medidos nas cidades de Delhi e Hong Kong. O trabalho objetivou desenvolver um modelo estatístico de previsão, das concentrações de PM_{10} e promover um estudo comparativo através do desempenho dos modelos. O modelo 1 é de regressão linear múltipla. O modelo 2 era um modelo de séries temporais ARIMA e o modelo 3 a combinação entre os modelos 1 e 2. Os dados de concentrações utilizados na modelagem foi de PM_{10}, e os parâmetros meteorológicos foram, velocidade do vento, temperatura, radiação solar e umidade relativa do ar, medidos no período de junho de 2000 a junho de 2001, para as duas cidades. Na comparação entre os modelos, as medidas de erro mostraram que o modelo 3 foi o que obteve o melhor desempenho. O estudo de previsão ocorreu apenas para a cidade de Delhi, e compreendeu o período de Junho de 2001 à junho de 2002. O modelo utilizado foi o 3, e os resultados da previsão mostrou-se satisfatório, já que o erro absoluto médio foi de 25% entre os valores previstos e os observados.

Gomes (2009) realizou um estudo de previsão de índices de qualidade do ar da Região da Grande Vitória - ES, Brasil, utilizando o modelo auto-regressivo de valores inteiros $INAR(p)$. O período de análise compreendeu 01/01/07 a 19/03/07 e obteve previsões do dia 20/03/07 a 25/03/07. Os poluentes investigados foram monóxido de carbono (CO), dióxido de nitrogênio (NO_x), dióxido de enxofre (SO_2) e ozônio (O_3). Para a escolha do modelo mais adequado o autor utilizou o critério de seleção automática para modelos $INAR(p)$ o $AICC_{INAR}$ que seleciona a melhor ordem p para cada modelo. Os resultados mostraram que todas as previsões para os índices de qualidade do ar foram classificadas como BOA corforme a Resolução CONAMA

03/90. Porém, baseados nas diretrizes da OMS (2005), a previsão do poluente SO_2 no dia 20/03/07, estação do Centro de Vila Velha, excedeu o valor de $20\mu g/m^3$ para média de 24 horas.

3.5 Revisão dos estudos que utilizaram ferramentas estatísticas para tratar dados faltantes em séries temporais de concentrações de poluentes atmosféricos

Diversos estudos têm utilizado ferramentas estatísticas para preencher dados faltantes em pesquisas relacionadas com séries temporais de contaminantes atmosféricos. Importantes estudos envolvendo essa questão estão listados abaixo.

Junninen *et al.* (2004) avaliaram dois contextos de imputação de dados aplicáveis a dados de concentração de poluentes. No estudo realizado, os dados foram avaliados no contexto de análise univariada (linear, *spline* e interpolação vizinho mais próximo) e multivariada (regressão baseados em imputação) (REGEM), vizinho mais próximo (NN), auto-organização de mapa (SOM) e *multi-layer perceptron* (MLP). Além disso, um procedimento de imputação múltipla foi considerado, para fazer a comparação entre os regimes de imputação única e múltipla. O objetivo dos autores era avaliar e comparar métodos de análise univariada e multivariada para imputação de dados faltantes em um conjunto de dados de qualidade do ar. Os conjuntos de dados utilizados na modelagem consistia em concentrações de NO_x, NO_2, O_3, PM_{10}, e CO, juntamente com quatro parâmetros meteorológicos: velocidade do vento, direção do vento, temperatura e umidade relativa. Os resultados mostraram que os métodos univariados são dependentes do comprimento do intervalo de tempo, ou seja, a quantidade de dados faltantes, e que seu desepenho também depende da variável em estudo. Os resultados obtidos com os métodos multivariados mostrou que tanto SOM e MLP apresentam um desempenho um pouco melhor que o método NN. A vantagem do SOM sobre outros métodos é que ele dependente menos da localização real dos dados faltantes, enquanto que as vantagens dos métodos NN são particularmente importantes em aplicações práticas, ou seja, é computacionalmente menos exigente e não gera novos valores os dados. Os resultados mostraram que, em geral, o significativo aumento em performances podem ser alcançados pela hibridização dos métodos multivariados,

e que a forma como esta combinação deve ser feita é dependente da variável a ser inspecionada.

Iglesias, Jorquera e Palma (2005) propuseram uma metodologia estatística para manipular regressões com longa dependência nos erros e dados faltantes. A estratégia de estimação foi desenvolvida através de uma abordagem Bayesiana e clássica. O estudo foi ilustrado com aplicação em um conjunto de dados de concentrações de poluentes atmosféricos da cidade de Santiago no Chile, para o período de 01 de janeiro de 1989 a 31 de dezembro de 1996, com um número muito alto de observações faltantes, 531 dias sem observação. Para correção dos dados faltantes os autores utilizaram o filtro de Kalman. Para explicar a variação nas concentrações de material particulado inalável (PM_{10}), os autores utilizaram como variáveis explanatórias da regressão, a velocidade do vento, a precipitação, as concentrações de CO_2 e SO_2. Essas variáveis mostraram-se estarem correlacionadas com a concentração do PM_{10}. Os resultados apresentam que a aplicação da metodologia aos dados reais, com a abordagem clássica, mostrou que a inferência pode ser distorcida se a longa dependência nos erros não for considerada.

Plaia e Bondi (2006) propuseram uma metodologia de imputação de dados faltantes no qual nomearam de método efeito *Site-Depen-Dente* (SDEM). O objetivo dos autores foi propor um método de imputação (SDEM) e comparar seu desempenho com outros métodos de imputação única e múltipla conhecidos na literatura. No estudo, foi considerado um conjunto de dados das concentrações de PM_{10} medidos em oito estações de monitoramento distribuidas na região metropolitana de Palermo, na Sicília, em 2003. Os resultados encontrados concordam, através dos indicadores de desempenho, em avaliar o método proposto (SDEM) como o melhor método entre os comparados no trabalho, independente da duração da lacuna de dados faltantes.

Junger (2008) discutiu questões teóricas e metodológicas para imputação de dados faltantes em séries temporais multivariadas. O objetivo do autor foi avaliar o efeito da poluição do ar em populações suscetíveis no município do Rio de Janeiro usando diferentes eventos de saúde, desenvolver metodologias de imputação e análise de séries temporais com a implementação de interfaces computacionais no ambiente R para análise de dados em epidemiologia ambiental. E ainda propor uma metodologia de imputação de dados faltantes em séries temporais de concentrações de poluentes e implementar a metodologia de imputação em uma biblioteca para o aplicativo de análise estatística R. O autor propôs alguns procedimentos de imputação de dados faltantes em séries temporais de concentrações de poluentes atmosféricos. No trabalho foi utilizado uma amostra de dados de concentrações de material particulado inalável (PM_{10}), ba-

seados no algoritmo EM *(expectation-maximisation)*. A trajetória temporal das séries foi modelada com o uso de *splines*, modelos de regressão ou modelos ARIMA com múltiplos regimes de covariâncias. Foi feito um estudo de simulações com diversas configurações de dados faltantes para avaliar a validade dos métodos utilizados. Os métodos são avaliados também quanto a sua performance por meio de indicadores de acurácia e concordância. Os resultados da avaliação das simulações permitem afirmar que: *(i)* a análise de dados considerando apenas as unidades de observações completas subestimaram o efeito do poluente no evento de saúde mesmo com pequena quantidade de dados faltantes, *(ii)* as imputações pela média e pela mediana apresentaram este efeito superestimado, grande dispersão das estimativas e baixa concordância dos valores imputados com os originais, *(iii)* os procedimentos multivariados apresentaram melhor desempenho e acurácia que os univariados, *(iv)* os métodos multivariados com ajuste do componente temporal apresentaram maior acurácia e precisão, *(v)* também apresentaram menores erros de previsão e maior concordância entre os valores imputados e os originais.

Os trabalhos revisados apresentaram metodologias para o tratamento de dados faltantes presentes em séries da concentração de poluentes atmosféricos e evidenciam a importância de conhecer e aplicar os métodos de imputação para estudar e modelar o comportamento da média da série de concentração de PM_{10}.

Capítulo 4

Conceitos Básicos em Séries Temporais

Nesta seção serão apresentados alguns conceitos básicos utilizados na análise de séries temporais e processos estacionários, que serão úteis na realização deste trabalho, baseados em modelos de Box e Jenkins (1970). Para mais detalhes de análise de séries temporais consultar em Brockwell e Davis (2006), Priestley (1983), Wei (2006) e Box *et al.*, (2008).

4.1 Processos estocásticos

Definição 1: Seja T um conjunto arbitrário. Um processo estocástico é uma família $X = X(t), t \in T$, tal que, para cada $t \in T$,$X(t)$ é uma variável aleatória.

Nestas condições, um processo estocástico é uma família de variáveis aleatórias $X(\Omega,t)$ indexadas no tempo, onde Ω é o espaço amostral e t o tempo. Para um dado t fixo, $X(\Omega,t)$ é uma variável aleatória. Para um dado espaço amostral Ω, $X(\Omega,t)$ é denominado realização ou trajetória do processo, ou ainda, uma série temporal. Assim, uma série temporal, pode ser definida como um registro de observações X_t de algum fenômeno medido sequencialmente em instantes de tempos t diferentes (MORETTIN e TOLOI, 2006).

Para um processo estocástico $\{X_t, t = 0, \pm 1, \pm 2, \pm, \cdots\}$ a função média do processo é definida por:

$$E(X_t) = \mu_t \tag{4.1}$$

onde $E(X_t)$ pode ser a média das concentrações de poluentes num modelo estocástico de análise ou previsão de contaminantes atmosféricos.

A variância do processo é dada por:

$$var(X_t) = \sigma_t^2 = E[(X_t - \mu_t)^2] \tag{4.2}$$

A relação entre duas variáveis aleatórias X_t e $X_{(t+k)}$ é dada pela covariância, expressa por:

$$\gamma(t,t+k) = cov[X_t, X_{(t+k)}] = E[(X_t - \mu_t)(X_{t+k} - \mu_t)] \tag{4.3}$$

e a correlação entre X_t e $X_{(t+k)}$ é dada pelo coeficiente de correlação:

$$\rho(t,t+k) = \frac{\gamma(t,t+k)}{\sqrt{\sigma_t^2 \sigma_{(t+k)}^2}} = \frac{E[(X_t - \mu_t)(X_{(t+k)} - \mu_t)]}{\sqrt{[(X_t - \mu_t)^2)(X_{(t+k)} - \mu_t)^2}} \tag{4.4}$$

Definição 2: Uma série temporal X_t diz-se fracamente estacionária ou estacionária de 2ª ordem se e somente se:

(i) $E(X_t) = \mu_t = \mu$, constante, independente de t.

(ii) $var(X_t) = \sigma^2$,constante, independente de t.

(iii) $\gamma(X_t, X_{(t+k)}) = cov(X_t, X_{(t+k)})$ depende apenas de k (não depende de t).

As autocorrelações $\rho(h)$ são obtidas normalizando as autocovariâncias através da sua divisão pelo produto dos respectivos desvios padrão, $\rho(h) = \frac{\gamma(h)}{\gamma(0)}$. Um exemplo de processo estacionário é o processo de ruído branco, que é uma sequência de variáveis aleatórias não-correlacionadas com média constante e variância constante ao longo do tempo.

4.2 Funções de autocovariância e autocorrelação

Para o processo estacionário X_t, em que $E(X_t) = \mu_t$, $var(X_t) = \sigma_t^2$, são constantes, e a covariância $cov(X_t, X_{(t+k)})$, depende apenas de k, a função de autocovariância é dado por:

$$\gamma_k = cov(X_t, X_{(t+k)}) = E[(X_t - \mu_t)(X_{(t+k)} - \mu_t)] \tag{4.5}$$

Onde γ_k define a função de autocovariância do processo, e k é o coeficiente de defasagem.

O coeficiente de autocorrelação de defasagem ρ_k é definido por:

$$\rho_k = \frac{E[(X_t - \mu_t)(X_{(t+k)} - \mu_t)]}{\sqrt{var(X_t)}\sqrt{var(X_{(t+k)}}} = \frac{cov(X_t, X_{(t+k)})}{\sqrt{var(X_t)}\sqrt{var(X_{(t+k)}}} \tag{4.6}$$

Observe que, se o processo for estacionário, teremos que $var(X_t) = var(X_{(t+k)})$. Logo, o coeficiente de autocorrelação para processos estacionários pode se expresso como:

$$\rho_k = \frac{cov(X_t, X_{(t+k)})}{var(X_t)} = \frac{\gamma_k}{\gamma_0} \tag{4.7}$$

A função de autocorrelação representa a covariância e a correlação entre X_t e $X_{(t+k)}$ separadas apenas por uma distância de tempo k.

As propriedades das funções de autocovariância e autocorrelação são:

(I) $\gamma_0 = var(X_t), \rho_0 = 1$.

(II) $\quad |\gamma_k| \leq \gamma_0, \quad |\rho_k| \leq 1$, para todo k inteiro.

(III) $\gamma_k = \gamma_{(-k)} e \rho_k = \rho_{(-k)}$, para todo k inteiro.

A propriedade (III), é consequência dos pares X_t, $X_{(t+k)}$ e X_t, $X_{(t-k)}$ terem o mesmo operador de defasagem, o que mostra que a função de autocorrelação é simétrica em relação a $k = 0$.

4.3 Função de autocorrelação parcial

Seja X_t um processo estacionário. Para investigar a correlação entre X_t e $X_{(t+k)}$, após a remoção dos efeitos que sobre elas produzem as variáveis intermediárias $X_{(t+1)}$, $X_{(t+2)}, \cdots, X_{(t+k-1)}$, define-se a função de autocorrelação parcial expressa por:

$$\phi_{kk} = \frac{corr[X_t, X_{(t+k)}]}{(X_{(t+1)}, \cdots, X_{(t+k-1)})} \tag{4.8}$$

De forma geral a função de autocorrelação parcial para $k = 1, 2, \cdots$, porde ser definida como:

$$\phi_{kk} = \frac{\begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{(k-2)} & \rho_1 \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{(K-3)} & \rho_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho_{(k-1)} & \rho_{(k-2)} & \rho_{(k-3)} & \cdots & \rho_1 & \rho_k \end{bmatrix}}{\begin{bmatrix} 1 & \rho_1 & \cdots & \rho_{k-2} & \rho_{k-1} \\ \rho_1 & 1 & \cdots & \rho_{k-3} & \rho_{k-2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho_{k-1} & \rho_{k-1} & \cdots & \rho_1 & 1 \end{bmatrix}} \tag{4.9}$$

A prova dos resultados acima podem ser encontrados em Wei (2006).

4.4 Ruído branco

Um processo a_t é chamado de ruído branco quando é formado por uma sequência de variáveis aleatórias não correlacionadas com a mesma distribuição, com média constante $E(a_t) = \mu_a$, geralmente considerada zero, isto é, $\mu_a = 0$, variância constante $var(a_t) = \sigma_a^2$ e covariância $\gamma_k = cov(a_t, a_{(t+k)}) = 0$, para todo inteiro $k \neq 0$.

Definição 3: O processo de ruído branco a_t é estacionário com

$$\text{função de autocovariância } \gamma_k = \begin{cases} \sigma_a^2 & \text{para } k = 0 \\ 0 & \text{para } k \neq 0 \end{cases}$$

$$\text{função de autocorrelação } \rho_k = \begin{cases} 1 & \text{para } k = 0 \\ 0 & \text{para } k \neq 0 \end{cases}$$

$$\text{e função de autocorrelação parcial } \phi_{kk} = \begin{cases} 1 & \text{para } k = 0 \\ 0 & \text{para } k \neq 0 \end{cases}$$

4.5 Modelos estacionários

Os modelos estacionários nos estudos de séries temporais caracterizam-se quando os dados flutuam em torno de uma média constante, independente da variância e do tempo, as flutuações continuam constantes ao longo do tempo. Nesta seção são apresentados alguns modelos estacionários. São eles, os modelos auto-regressivos (AR) e média móvel (MA) que são casos especiais do modelo autorregressivo e de média móvel (ARMA).

4.5.1 autorregressivo (AR)

O modelo autorregressivo de ordem p, denotado por $AR(p)$, pode ser expresso na seguinte forma:

$$X_t = \phi_1 X_{(t-1)} + \phi_2 X_{(t-2)} + \cdots + \phi_p X_{(t-p)} + a_t \tag{4.10}$$

Se definirmos o operador autorregressivo estacionário de ordem p

$$\phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p \tag{4.11}$$

então pode-se escrever

$$\phi_p(B) X_t = a_t. \tag{4.12}$$

O modelo $AR(p)$ é sempre invertível e estacionário quando as raízes do polinômio autorregressivo de ordem p, $\phi_p(B)$, estão fora do círculo unitário.

4.5.2 Modelo de médias móveis (MA)

O Modelo de médias móveis de ordem q, denotado por $MA(q)$, é expresso na seguinte forma:

$$X_t = \mu_t + a_t - \theta_1 a_{(t-1)} - \cdots - \theta_q - a_{(t-q)} \tag{4.13}$$

e sendo $\widetilde{(X_t)} = X_t - \mu$, teremos

$$\widetilde{(X_t)} = (1 - \theta_1 B - \cdots - \theta_q B^q) a_t = \theta(B) a_t, \tag{4.14}$$

onde

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q \tag{4.15}$$

é o operador de médias móveis de ordem q.

4.5.3 Modelo autorregressivo e de médias móveis

Seja X_t $(t = 1, 2, 3, ...)$ um processo que satisfaz a equação em diferenças dada por

$$X_t - \phi_1 X_{(t-1)} - \cdots - \phi_p X_{(t-p)} = a_t - \phi_1 a_{(t-1)} - \cdots - \phi_q a_{(t-q)}, \tag{4.16}$$

ou,

$$\Phi_p(B) X_t = \Theta_q(B) a_t, \tag{4.17}$$

onde a_t é ruído branco, i. e., $a_t \sim RB(0, \sigma_a^2)$, $\Phi_p(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$ e $\Theta_q(B) = 1 - \theta_1 B - \cdots - \theta_q B^q$. O processo X_t defnido em (4.17) é chamado de processo autorregressivo e de médias móveis de ordem (p, q), denotado por $ARMA(p, q)$.

O processo (4.17) é estacionário e invertível se as raízes de $\Phi_p(B)$ e $\Theta_q(B)$ não são comuns e encontram-se fora do círculo unitário. Para maiores detalhes consultar em Wei (2006), Brockwell e Davis (2002) e Box *et al.* (2008).

Função de autocorrelação de um processo ARMA(p,q)

Multiplicando ambos os membros de (4.16) por X_{t-k} e aplicando o valor esperado, vemos que a função de autocovariância satisfaz a equação de diferença

$$\gamma_k = \phi_1\gamma_{k-1} + \cdots + \phi_p\gamma_{k-p} + \gamma_{Xa}(k) - \theta_1\gamma_{Xa}(k-1) - \cdots - \theta_q\gamma_{Xa}(k-q) \tag{4.18}$$

onde $\gamma_{Xa}(k)$ é a função de covariância cruzada entre X_t e a_t definida por $\gamma_{Xa}(k) = E[X_{t-k}a_t]$. Uma vez que X_{t-k} depende apenas de choques que ocorreram até o instante $t-k$ através da representação de média móvel infinita $X_{t-k} = \psi(B)a_{t-k} = \sum_{j=0}^{\infty}\psi_j a_{t-k-j}$, segue-se que

$$\gamma_{Xa}(k) = \begin{cases} 0 & k > 0 \\ \psi_{-k}\sigma_a^2 & k \leq 0 \end{cases} \tag{4.19}$$

de modo que (4.18)pode ser expresso como

$$\gamma_k = \phi_1\gamma_{k-1} + \cdots + \phi_p\gamma_{k-p} - \sigma_a^2(\theta_k\psi_0 + \theta_{k+1}\psi_1 + \cdots + \theta_q\psi_{q-k}). \tag{4.20}$$

Vemos que (4.20) implica

$$\gamma_k = \phi_1\gamma_{k-1} + \phi_2\gamma_{k-2} + \cdots + \phi_p\gamma_{k-p} \quad k \geq q+1 \tag{4.21}$$

e, por conseguinte

$$\rho_k = \phi_1\rho_{k-1} + \phi_2\rho_{k-2} + \cdots + \phi_p\rho_{k-p} \quad k \geq q+1 \tag{4.22}$$

Assim, para o processo ARMA(p, q), as autocorrelações de *lags* $1, 2, \cdots, q$ serão afetadas diretamente pelos parâmetros de médias móveis, mas para $k > q$ as mesmas comportam-se como nos modelos autorregressivos.

Pode-se verificar que se $q < p$ a ACF consiste numa mistura de exponenciais e/ou de senóides amortecidas. Entretanto, se $q \geq p$, os primeiros $q-p+1$ valores $\rho_0, \rho_1, \cdots, \rho_{q-p}$ não seguirão este padrão (BOX, JENKINS e REINSEL, 2008).

4.6 Modelos não estacionários

Os modelos estudados na seção 2.3.2 são apropriados para descrever séries que se desenvolvem no tempo em torno de uma média constante, isto é, séries estacionárias. Muitas séries associadas com fenômenos ambientais, não são estacionárias. Nesta seção, será apresentado uma revisão dos modelos não-estacionários. Muitos modelos não-estacionários podem ser tornar em estacionários, como é o caso do modelo autorregressivo integrado de média móvel, $ARIMA(p,d,q)$, de onde se obtém como resultado o modelo $ARMA(p,q)$ (PAULA, 2002).

Os padrões não-estacionários de uma série temporal resultam em autocorrelações positivas que dominam um diagrama de autocorrelação. É importante remover a não-estacionariedade antes de ser proceder à construção de um modelo de série temporal. Os processos que estabilizam a média e a variância são os métodos de diferenciações e os processos de transformação para a variância (MORETTIN e TOLOI, 2006).

4.6.1 Modelo Autorregressivo Integrado de Média Móvel [ARIMA (p,d,q)]

Seja d um inteiro não negativo. X_t é um processo autoregressivo integrado e de médias móveis ARIMA(p, d, q) se X_t satisfaz a equação em diferenças da forma

Uma série temporal X_t é dita um processo $ARIMA(p,d,q)$ se,

$$\phi_p(B)(1-B)^d X_t = \theta_0 + \theta_q(B)a_t, \quad a_t \sim RB(0,\sigma_a^2) \tag{4.23}$$

onde $\phi_p(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$ e $\phi_q(B) = 1 - \phi_1 B - \cdots - \phi_q B^q$ são respectivamente polinômios autorregressivo estacionário e médias móveis invertível. Portanto, X_t é processo não-estacionário que depois de diferenciado ($\nabla^d X_t (d \geq 1)$ e $\nabla = 1 - B$) se transforma em processo estacionário e invertível $ARMA(p,q)$.

Quando $d = 0$ o processo X_t é estacionário e θ_0 está relacionado com a média $\mu = EX_t$,

$$\theta_0 = \mu(1 - \phi_1 - \cdots - \phi_p). \tag{4.24}$$

Se $\mu = 0$, pode omitir-se θ_0; quando $\mu \neq 0$, θ_0 é parâmetro que deve ser estimado.

Para $d = 0$, o modelo $ARIMA(p,d,q)$ pode ser escrito na forma,

$$\Phi(B)U_t = \Theta_q(B)a_t \tag{4.25}$$

onde $U_t = (1-B)^d X_t$ é o processo estacionário $ARMA(p,q)$

Quando $d \geq 1$, o processo X_t não é um processo estacionário. Os modelos *ARIMA* (autorregressivos integrados e de médias móveis) são capazes de descrever os processos de geração de uma variedade de séries.

4.7 Metodologia de modelagem

No estudo de séries temporais o objetivo principal é encontrar um modelo apropriado que descreva o fenômeno gerador de cada série estudada. Nesta seção, aborda-se as etapas da metodologia de Box e Jenkins (1970) para o processo de escolha do melhor modelo que se ajusta ao conjunto de dado (PM_{10}). Neste trabalho, a série temporal X_t, representará as concentrações de PM_{10}.

A metodologia Box-Jenkins aplicada neste trabalho esta dividida nas seguintes etapas:

1. Identificação:

- Comportamento geral da série;
- Transformação dos dados para estabilizar a variância;
- Diferenciação dos dados para obter a série estacionária;
- Seleção de um modelo a partir da observação da FAC e FACP.

2. Estimação:

- Estimar os parâmetros do modelo;

- Selecionar o melhor modelo através do critério de Akaike (AIC);
- Avaliação do diagnóstico: na aplicação da metodologia Box-Jenkins, após a identificação do modelo é fundamental fazer a avaliação da qualidade estatística do modelo e análise residual gráfica, verificando se as suposições não forma violadas. Caso contrário volta-se para a identificação do modelo. Esse procedimento será repetido até que um modelo adeguado seja aceito.

3. Previsão.

A previsão de passos à frente é o cálculo do valor esperado de uma futura observação condicionada a valores passados e ao valor presente da variável, ou seja,

$$\widehat{X}_t(h) = E(X_{(t+h)} \mid X_t, X_{(t-1)}, \cdots). \tag{4.26}$$

Onde $\widehat{X}_t(h)$ é o valor estimado da variável X_t no horizonte de h períodos de tempos futuros com base em t observações passadas. O valor de X_{t+h} é calculado com o modelo que melhor se ajusta aos dados, $ARMA(p,q)$ ou $ARIMA(p,d,q)$.

Na aplicação da metodologia de séries temporais realizada neste trabalho, um dos principais objetivos foi fazer a comparação entre os modelos quanto ao seu desempenho no processo de ajustamento aos dados em estudo e uma das formas de verificar essa qualidade é o estudo de previsão.

Capítulo 5

Estimação da Função de Autocorrelação

A estimativa da função de autocorrelção de uma série temporal é um dos aspectos mais importantes na identificação e construção do modelo (YAJIMA e NISHINO, 1999). Os processos $AR(p)$, $MA(p)$ e $ARMA(p,q)$, descritos no capítulo 4, apresentam a função de autocorrelação (fac) e função de autocorrelação parcial (facp) com características especiais. Assim,

- Um processo $AR(p)$ tem fac decaindo de acordo com exponenciais e/ou senóides amortecidas, infinita em extensão. A facp tem um decaimento brusco para zero a partir de um certo *lag* k.

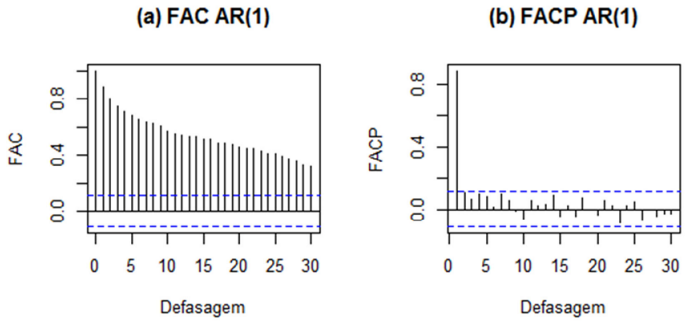

Figura 5.1: fac e facp para uma série temporal gerada por um processo $AR(1)$ com $\phi = 0.9$, $a_t \sim N(0,1)$ e tamanho da amostra $n = 300$.

- Um processo $MA(q)$ tem fac finita, no sentido que ela apresenta um corte após o *lag* q. A facp se comporta de maneira similar à fac de um processo $AR(p)$: é denominada por exponenciais ou senóides amortecidas.

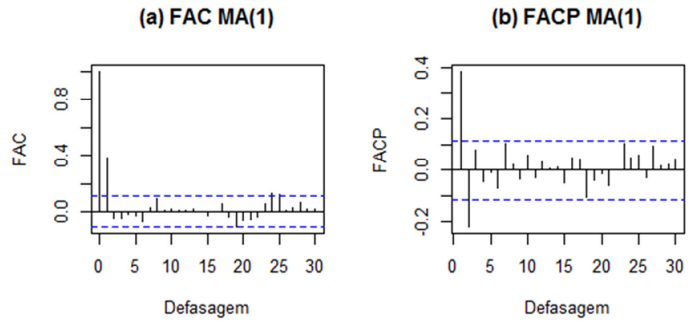

Figura 5.2: fac e facp para uma série temporal gerada por um processo $MA(1)$, com $\theta = 0.5$, $a_t \sim N(0,1)$ e tamanho da amostra $n = 300$.

- Um processo $ARMA(p,q)$ tem fac infinita em extensão, a qual decai de acordo com exponenciais ou senóides amortecidas após o *lag* $q-p$. A facp se comporta como a facp de um pocesso MA puro.

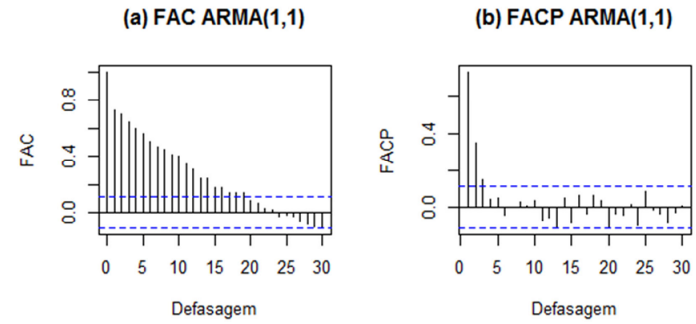

Figura 5.3: fac e facp para uma série temporal gerada por um processo $ARMA(1,1)$ com $\phi = 0.9$, $\theta = 0.5$, $a_t \sim N(0,1)$ e tamanho da amostra $n = 300$.

Estas obsevações são úteis no procedimento de identificação do modelo que será ajustado aos dados observados (WEI,2006).

A estimativa da função de autocorrelção de uma série temporal é um dos aspectos mais importantes na identificação e construção do modelo (YAJIMA e NISHINO, 1999). Porém, o processo de estimação da função autocorrelação através de (7), apresenta problema quando existem dados faltantes. Uma solução consiste na exculsão das unidades de observação com valores não aferidos em uma ou mais covariáveis; também conhecido como análise de dados completos. Porém, de acordo com os trabalhos de Greenland e Finkle (1995) e Greenland e Rothman (1998) dependendo do número de covariáveis no modelo e da quantidade de dados faltantes, pode haver perda significativa da precisão dos estimadores. Para Box *at al.*, (1994), no caso de análise de séries temporais, esta abordagem é mais problemática, pois a exclusão de

dados faltantes pode alterar consideravelmente as estruturas de dependêcia temporal, tendência e sazonalidade.

Além deste método há, na literatura, vários procedimentos para estimar parâmetros de modelos com dados faltantes. Os principais procedimentos podem ser classificados em dois grupos: os baseados em modelos e os baseados em imputação (LITTLE E RUBIN, 1989; JUGER, 2008). Neste trabalho focamos nos métodos de imputação e na estimação da função de autocorrelação na presença de dados faltantes, através dos estimadores estudos em Yajima e Nishino (1999).

Consideramos três estimadores para a função de autocorrelação na presença de dados faltantes, o primeiro originalmente foi proposto por Parsen (1963), o segundo é um tipo de estimador de mínimos quadrados (YAJIMA e NISHINO, 1999). Finalmente, o terceiro estimador é um coeficiente de correlação amostral baseado em todos os pares observados. Os dois últimos foram propostos por Takeuchi (1995). Os três estimadosres foram estudados em Yajima e Nishino (1999), segundo eles estes estimadores têm as mesmas propriedades assintóticas sobre uma amostra completa, mas eles podem comportar-se de forma diferente um do outro na presença de dados faltantes.

5.1 Estimadores da função de autocorrelação na presença de dados faltantes

Seja $\{X_t\}$ um processo definido em (2.3.1). Parzen (1963) formulou um modelo de séries temporais com dados faltantes, no qual os dados observados $\{Y(n)\}(n = 1,2,\cdots,N)$ são expressos por $Y(n) = a(n)X(n)$ em que $\{a(n), n = 1,2,\cdots\}$ representa o estado da observação,

$$\begin{cases} a(n) = 1 & \text{se} \quad X(n) \quad é \quad observado, \\ a(n) = 0 & \text{se} \quad X(n) \quad não \quad é \quad observado. \end{cases}$$

Seja

$$\overline{a} = \frac{1}{N}\sum_{n=1}^{N} a(n), \tag{5.1}$$

$$C_a(l) = \frac{1}{N}\sum_{n=1}^{N-l} a(n)a(n+l), \tag{5.2}$$

$$C_Y(l) = \frac{1}{N}\sum_{n=1}^{N-l} Y(n)Y(n+l). \tag{5.3}$$

Em termos dessas quantidades, Parzen (1963) define a estimativa de $\gamma_X(l)$ como

$$\widehat{\gamma}_X(l) = \frac{C_Y(l)}{C_a(l)}. \tag{5.4}$$

Se $X(n)$ tem observações faltantes, a média μ pode ser estimada por $\widehat{\mu} = \frac{\sum_{n=1}^{n} Y(n)}{\sum_{n=1}^{n} a(n)}$, nesse caso, em $\widehat{\gamma}_X(l)$, $Y(n)$ deve ser subtituido por $Y(n) - \widehat{\mu}$. Assumi-se que $X(n)$ e $a(n)$ são independentes. Em seguida, defini-se

$$\gamma_X(l) = Cov(X(n), X(n+l)), \tag{5.5}$$

$$\rho_X(l) = Cor(X(n), X(n+l)) = \frac{\gamma_X(l)}{\gamma_X(0)}, \tag{5.6}$$

$$\hat{\gamma}_a(l) = \gamma_a(l) + \mu_a^2, \tag{5.7}$$

onde $\gamma_X(l)$ e $\rho_X(l)$ são definidos da mesma forma de um processo estacinário.

Yajima e Nishino (1999) estudaram os seguintes estimadores da função de autocorrelação $\{\rho_x(l)\}$ para séries com dados faltantes. O primeiro foi proposto por Parzen (1963) e, mais tarde, suas propriedades assintóticas foram investigados sob várias hipóteses em $\varepsilon(n)$ por Dunsmuir e Robinson (1981). Denotamos-lo por $\hat{\rho}_{PDR}(l)$,

$$\hat{\rho}_{PDR}(l) = \frac{C_Y(l)/C_a(l)}{C_Y(0)/C_a(0)} = \frac{\sum_{n=1}^{N-l} Y(n)Y(n+l) / \sum_{n=1}^{N-l} a(n)a(n+l)}{\sum_{n=1}^{N} Y(n)^2 / \sum_{n=1}^{N} a(n)^2}. \tag{5.8}$$

O numerador e o denominador são estimadores de $\gamma_x(l)$ e $\gamma_x(0)$, respectivamente, $\hat{\rho}_{PDR}(l)$ é interpretado como uma espécie de o estimador Yule-Walker.

O segundo foi proposto por Takeuchi (1995) e adaptado independentemente por Shin e Sarkar (1995) como o valor inicial de um processo de Newton-Raphson para obter o método da máxima probabilidade de um modelo AR (1). Este estimador é definido por

$$\hat{\rho}_{SST}(l) = \frac{\sum_{n=1}^{N-l} Y(n)Y(n+l)}{\sum_{n=1}^{N-l} a(n+l)Y(n)^2}. \tag{5.9}$$

Denotando $\hat{\rho}_{SST}(l) = \sum_{n=1}^{N-l} a(n)a(n+l)X(n)X(n+l)/\sum_{n=1}^{N-l} a(n)a(n+l)X(n)^2$, vemos que este é um estimador de mínimos quadrados para o modelo, que consiste de todos os pares observados $(X(n+l),X(n))$, onde $X(n+l)$ é uma varável dependente e $X(n)$ uma variável independente (YAJIMA e NISHINO, 1999).

Finalmente, o terceiro estimador foi também proposto por Takeuchi (1995) e definido por

$$\hat{\rho}_T(l) = \frac{\sum_{n=1}^{N-l} Y(n)Y(n+l)}{\sqrt{\sum_{n=1}^{N-l} a(n+l)Y(n)^2}\sqrt{\sum_{n=1}^{N-l} a(n)Y(n+l)^2}}. \tag{5.10}$$

O estimador $\hat{\rho}_T(l)$ é um coeficiente de correlação amostral baseado em todos os pares observados $(X(n+l),X(n))$.

Capítulo 6

Materiais e Métodos

6.1 Região de estudo

Este trabalho utilizou, para fazer aplicações da metodologia estudada, dados coletados na Região da Grande Vitória (RGV), constituída pelos municípios de Vitória, Vila Velha, Cariacica, Serra e Viana, estão localizadas na região sudeste do Estado do Espírito Santo. Segundo o Instituto Brasileiro de Geografia e Estatística (IBGE, 2010) a região da Grande Vitória é constituída de 1.475.332 habitantes, abrange uma área de 1.461 Km^2, sendo um dos principais pólos de desenvolvimento urbano e industrial do estado. A região sofre com diversos tipos de problemas ambientais, dentre os quais está a deterioração da qualidade do ar, devido às emissões atmosféricas por indústrias e pela frota veicular.

A RGV possui uma Rede Automática de Monitoramento da Qualidade do Ar (RAMQAR) inaugurada em julho de 2000, de propriedade do Instituto Estadual de Meio Ambiente e Recursos Hídricos (IEMA). A referida rede é distribuída em oito estações localizadas nos municípios que compõem a RGV, da seguinte forma: o município Serra com duas estações localizadas nas regiões de Laranjeiras e Carapina; o município Vitória com três estações localizadas nas regiões de Jardim Camburi, Enseada do Suá e Centro de Vitória. O município de Vila Velha apresenta duas estações localizadas nas regiões do Ibes e Centro de Vila Velha e o município de Cariacica com uma estação em Cariacica. A localização espacial das estações de monitoramento da RAMQAR está ilustrada na Figura 6.1.

A RAMQAR monitora os seguintes poluentes: Partículas Totais em Suspensão (PTS); Partículas

Inaláveis (PM_{10}); Ozônio (O_3); Óxido de Nitrogênio (NO_x); Monóxido de Carbono (CO) e Hidrocarbonetos (HC). Realiza-se também o monitoramento dos seguintes parâmetros meteorológicos: Direção dos ventos (DV); Velocidade dos ventos (VV); Precipitação pluviométrica (PP); Umidade relativa do ar (UR); Temperatura (T); Pressão atmosférica (P) e Radiação solar (I). Os poluentes e parâmetros meteorológicos monitorados em cada estação RAMQAR encontram-se na Figura 6.2.

Figura 6.1: Localizaąão espacial das estações de monitoramento da qualidade do ar da RGV.

Fonte: Google Earth.

Estações	PTS	PM_{10}	SO_2	CO	NO_x	HC	O_3	Meteorologia
Estação Laranjeiras	■	■	■	■	■		■	
Estação Carapina	■	■						DV, VV, UR PP, P, T, I
Estação Jardim Camburi	■	■	■		■			
Estação Enseado do Suá	■	■	■	■	■	■	■	DV, VV
Estação Vitória Centro	■	■	■	■	■	■		
Estação Ibes	■	■	■	■	■	■	■	DV, VV
Estação Vila Velha		■	■					
Estação Cariacica	■	■	■	■	■		■	DV, VV, T

Figura 6.2: Parâmetros meteorológicos e poluentes monitorados em cada estação RAMQAR.

Fonte: Adaptado de Relatório da Qualidade do Ar da Região da Grande Vitória, 2010.

6.2 Dados

Uma das dificuldades na avaliação de procedimentos de imputação de dados faltantes é que geralmente não há como comparar os valores imputados com os valores reais. A geração de dados simulados com variáveis correlacionadas e com dependência temporal não é trivial e o melhor

modelo pode não ser capaz de capturar toda a dinâmica inerente ao processo estocástico subjacente (JUNGER, 2008). Por isto, neste trabalho, pretende-se usar, além dos dados simulados dados reais de concentração média de material particulado inalável.

Os valores de concentração média de material particulado (PM_{10}) serão obtidos junto ao Instituto Estadual de Meio Ambiente e Recursos Hídricos (IEMA). Todos os dados de concentração serão fornecidos para tempos de média de 24 horas em $\mu g/m^3$.

Os métodos propostos neste trabalho assumem que os dados têm distribuição normal. Entretanto, dados ambientais geralmente não seguem esta distribuição. Assim, todas as imputações serão realizadas usando o logaritmo natural dos dados originais para melhor aproximação da distribuição normal e estabilidade da variância para os métodos baseados em regressão (Box e Cox, 1964).

6.3 *Software* R

O ambiente R é disponibilizado sobre os termos da GNU: *General Public License*, primeira comunidade de compartilhamento de *softwares livres*. A página principal é http://www.r-project.org, localizada em Viena, Áustria. R é em grande parte um veículo para o desenvolvimento de novos métodos interativos de análise de dados.

O *software* R possui um grande número de procedimentos estatísticos convencionais, entre eles estão os modelos lineares, modelos de regressão não linear, análise de séries temporais, testes estatísticos paramétricos e não paramétricos, análise multivariada, etc. Tem uma grande quantidade de funções para o desenvolvimento de ambiente gráfico e criação de diversos tipos de apresentação de dados (REISEN e SILVA, 2011).

6.4 Métodos de imputação

Segundo Schafer (1997) imputação é um termo genérico para o preenchimento de dados faltantes com valores plausíveis. Nesta seção, os procedimentos de imputação única e o algoritmo EM *(expectation-maximisation)* utilizados no trabalho serão abordados.

A seguinte seção descreve a metodologia teórica para imputação de dados faltantes em séries temporais. Os métodos de imputação são baseados no algoritmo EM (Dempster *et al.,* 1977), e nos trabalhos de Junninen *et at.,* (2004) e Plaia e Bondi (2006).

6.4.1 Imputação por constantes

Os métodos de imputação por constantes são os mais comuns dentre os métodos de imputação única (IU), o princípio deste método é imputar um valor para cada dado faltante da base de dados e, então, analisa-la como se não houvesse dados faltantes (MCKNIGHT *et al.,* 2007).

Os métodos de imputação por constantes são os mais comuns dentre os métodos de IU. De forma geral, esses métodos substituem todos os valores faltantes de uma variável por um único valor, uma constante. A seguir, serão apresentados os dois métodos de IU utilizados neste trabalho, média e mediana.

Devido à sua facilidade de implementação a imputação da média se torna um método muito comum e bastante utilizado (MYRTVEIT *et al.,* 2001). Nesta técnica, a média dos valores de um atributo que contém dados faltantes é usada para preencher os seus os espaços com dados faltantes (FARHANGFAR *et al.,* 2004). Mas, na maioria dos casos este método não é eficaz para o tratamento, pois os valores extremos ficam sub-representados, o que implica na perda de variabilidade, ou seja, a variância das variáveis com dados faltantes é subestimada (MCKNIGHT *et al.,* 2007).

Segundo Mcknight (2007) o efeito prejudicial deste método é reduzido somente em bases de dados com uma pequena porcentagem de dados faltantes. A média é a melhor medida de tendência central para variáveis normalmente distribuídas, assim quando não trata-se de uma distribuição normal, os resultados deste método não apresentam bons resultados (VERONEZE, 2011). Outro método de IU é a mediana, a imputação da mediana é bem parecida com a imputação da média e apresenta as mesmas vantagens e desvantagens. Porém, segundo Veroneze (2011) este método é uma alternativa melhor para variáveis que não são normalmente distribuídas, pois a mediana representa melhor a tendência central de uma distribuição que possui grandes desvios da distribuição normal.

6.4.2 Imputação via algoritmo EM

O nome *Expectatin Mmaximization* (EM) vem de seus dois principais passos: Esperança e Maximização. A formalização deste método foi proposta por Dempster *et al.,* (1977). Segundo Little e Rubin (2002) o algoritmo EM é um método geral para obter estimativas de máxima verossimilhança em bases de dados incompletos. Como estas estimativas podem ser difíceis de obter para bases de dados complexas, é necessário um procedimento para reduzir esta dificuldade, que é o objetivo do algoritmo EM (MACKNIGHT *et al.,* 2007).

Neste trabalho será apresentado uma resumida descrição do algoritmo EM, maiores detalhes podem ser encontrados em Dempster *et al.,* (1977) e Junger (2008).

O algoritmo EM pode ser utilizado quando queremos estimar um conjunto de parâmetros, que tem uma determinada distribuição de probabilidades, a qual é obtida utilizando apenas uma parte dos dados. Suponha que $A = A_1, A_2, \cdots, A_m$ represente as m unidades amostrais para os quais os valores são conhecidos, e $B = B_1, B_2, \cdots, B_n$ os valores não conhecidas (Exemplo adaptado de ASSUNÇÃO, 2012).Rapidamente, pode-se explicar o algoritmo da seguinte maneira:

1. Escolher valores iniciais para os parâmetros do modelo considerando (por exemplo, médias e matriz de covariância para o modelo normal multivariado) (VERONEZE, 2011).
2. Faça até a convergência:
   (a) **Esperança**: calcular o valor esperado do logaritmo da função de verossimilhança em relação a distribuição condicional de A dado B sob a atual estimativa dos prarâmetros, ou seja, imputar valores para os dados faltantes baseando-se nos valores dos parâmetros.
   (b) **Maximização**: encontra o valor do parâmetro que maximiza a verossimilhança baseada em todos os dados, ou seja, estimar novos valores dos parâmetros.

O método converge quando a diferença entre os valores estimados dos parâmetros em duas interações consecutivas é menor que um pré-estabelecido (VERONEZE, 2011).

Neste trabalho utilizou o algoritmo EM proposto por Junger (2008), na qual vez uso da plataforma *msdti* do R, que foi implementada pelo autor. A seguir será feito a descrição do funcionamento do algoritmo EM. Tal descrição foi baseado nos trabalhos de Junger (2008) e Dempster et al. (1977).

Seja $X_t, (t = 1, \cdots, n)$, a t-ésima realização do vetor aleatório X, com distribuição normal multivariada, com m componentes não observados. O vetor X_t pode ser arranjado de forma que os m componentes faltantes sejam colocados nas primeiras posições, ou seja, $X_t = x_{t1}, \cdots, x_{tm}$,- $x_{t(m+1)}, \cdots, x_{tp})^T$, e representado como $X_t = (X_{t1}, X_{t2})^T$. Considere B janelas com diferentes regimes de covariâncias ao longo do tempo. A estimativa do vetor média no instante t e janela $b, b = (1, \cdots, B)$, pode ser particionado seguindo a mesma configuração dos componentes de X_t, isto é,

$$\widetilde{\mu}_t = \begin{bmatrix} \widetilde{\mu}_{t1} \\ \widetilde{\mu}_{t2} \end{bmatrix}, \quad \text{e} \quad \widetilde{\Sigma}_b = \begin{bmatrix} \widetilde{\Sigma}_{b11} & \widetilde{\Sigma}_{b12} \\ \widetilde{\Sigma}_{b21} & \widetilde{\Sigma}_{b22} \end{bmatrix}. \tag{6.1}$$

O algoritmo de imputação consiste em *(i)* substituir os valores faltantes por valores estimados, *(ii)* estimar os parâmetros μ e Σ do modelo normal subjacente (as estimativas de μ são usadas apenas para estimar Σ) e o nível de cada série temporal univariada μ_t (usado para imputar os dados faltantes), *(iii)* reestimar os valores faltantes considerando os parâmetros atualizados e o nível de cada série temporal. Este processo é repetido até que os valores estimados cessem de variar (Junger, 2008).

As estimativas iniciais $\widetilde{\mu}_0$ e $\widetilde{\Sigma}_0$ são respectivamente o vetor média e a matriz de covariâncias amostrais considerando apenas os dados observados. Na iteração $(k+1)$ do passo E do algoritmo EM, os valores faltantes são imputados como a média condicional aos valores observados e os parâmetros estimados na interação anterior dada por

$$\widetilde{X}_{t1}^{(k+1)} = E[X_{t1}|X_{t2}, \widetilde{\mu}_t^{(k)}, \widetilde{\Sigma}_b^{(k)}] = \widetilde{\mu}_{t1}^{(k)} + \widetilde{\Sigma}_{b12}^{(k)} \widetilde{\Sigma}_{b22}^{(k)-1} (X_{t2} - \widetilde{\mu}_{t2}^{(k)}) \tag{6.2}$$

e as contribuições para as covariâncias são dadas por

$$\widetilde{X_{t1}X_{t2}}^{(k+1)} = E[X_{t1}X_{t1}^T|X_{t2}, \widetilde{\mu}_{t1}^k, \widetilde{\Sigma}_b^{(k)}] = \widetilde{\Sigma}_{b11}^{(k)} - \widetilde{\Sigma}_{b12}^{(k)} \widetilde{\Sigma}_{b22}^{(k)-1} \widetilde{\Sigma}_{b21}^{(k)} + X_{t1}X_{t1}^T \tag{6.3}$$

e

$$\widetilde{X_{t1}X_{t2}}^{(k+1)} = E[X_{t1}X_{t1}^T|X_{t2}, \widetilde{\mu}_{t1}^k, \widetilde{\Sigma}_b^{(k)}] = \widetilde{X}_{t1}\widetilde{X}_{t2}^T. \tag{6.4}$$

Segundo Junger (2008) no passo M, são computadas as estimativas de máxima verossimilhança revisadas de μ e Σ, considerando implícito o índice da interação $(k+1)$, $\widetilde{\mu}_b = \Sigma_{t=1}^{nb} \widetilde{X}_{bt}/n_b$ e

$\widetilde{\Sigma}_b = \Sigma_{t=1}^{nb} \widetilde{X_{bt} X_{bt}^T} / n_b - \widetilde{\mu}_b \widetilde{\mu}_b^T$. A estimativa de $\widetilde{\mu}_b$ é usada apenas para o cálculo de $\widetilde{\Sigma}_b$.

A contribuição do componente temporal de cada série univariada é estimado de modo que se faz necessários modelos adicionais para estimação de μ_t. Segundo Junger (2008) neste método de imputação, estão implementadas três opções de estimação do nível das séries temporais: modelo auto-regressivo integrado de médias móveis (ARIMA), spline cúbica não paramétrica e modelo aditivo generalizado (MAG). No presente trabalho foi utilizado para fazer a estimação de μ_t o modelo ARIMA. A estimativa do nível para a variável X_j no instante t é a previsão um passo a frente no modelo ARIMA dado por $\widetilde{\mu}_t = E[X_{jt}|X_{j(t-1)}, X_{j(t-2)}, \cdots]$. A estimativa do nível é calculada usando as informações passadas de X_j (BOX *et al.*, 1994).

6.5 Indicadores de performance

Para avaliar os métodos em termos de qualidade da imputação em uma única replicação de um padrão escolhido ao acaso foram utilizados os indicadores de performance propostos por Junnien *et al.,* (2004). Nas equações seguintes, N denota o número de valores faltantes no conjunto de dados modelados, X_i são os valores reais, $\widetilde{X}_i$ são os valores imputados, $i = 1, \cdots, m$, $\overline{X}$ é a média dos valores reais e $\overline{\widetilde{X}}$ é a média dos vaores imputados.

1. Coeficiente de correlação de Pearson (r) entre os valores observados e imputados:

$$\widehat{r} = \left[\frac{1}{N} \frac{\sum_{i=1}^{N} \left[(X_i - \overline{X})(\widetilde{X}_i - \overline{\widetilde{X}}) \right]}{\hat{\sigma}_X \hat{\sigma}_{\widetilde{X}}} \right], \tag{6.5}$$

2. Raiz do erro quadratico médio:

$$\widehat{RMSE} = \left(\frac{1}{N} \sum_{i=1}^{N} \left[\widetilde{X}_i - X_i \right]^2 \right)^{\frac{1}{2}}, \tag{6.6}$$

3. Erro absoluto médio:

$$\widehat{MAE} = \frac{1}{N} \sum_{i=1}^{N} \left| X_i - \widetilde{X}_i \right| \tag{6.7}$$

O coeficiente de correlação de Pearson é o indicador mais comum para avaliar o desempenho de métodos de imputação. A raiz do erro quadrático médio será utilizado para estimar a média

geral do erro de cada imputação. O erro médio absoluto será usado como uma medida mais sensível do erro do modelo, pois é menos influenciada por grandes diferenças entre os valores originais e os imputados (JUNGER, 2008).

6.6 Recursos computacionais

As rotinas para a simulação de porcentagens de dados faltantes, imputações, estimação da função de autocorrelação e respectivas análises foram implementadas usando o *software* R (seção 6.3). O algoritmo EM normal multivariado utilizado neste trabalho está implementado na biblioteca R *mtsdi (multivariante time-series data imputation)* desenvolvida por Junger (2008). A biblioteca *mtsdi* é uma coleção de rotinas para a imputação de dados faltantes em séries temporais multivariadas, com isso não se fez necessário implementar tal metodologia.

6.7 Estudo das Simulações

Com o objetivo de avaliar a validade dos métodos de imputação e dos estimadores da função de autocorrelação na presença de dados faltentes propostos neste trabalho será feito um estudo de simulação baseado na geração de padrões representativos de diversos cenários de dados faltantes.

O trabalho de Greenlad e Rothman (1998) indica que para uma pequena proporção de dados faltantes e um grande número de observações a análise de dados completos produz bons resultados. Desta forma neste trabalho serão investigadas proporções iguais a 5%, 10%, 20%, 30% e 40% de dados faltantes e o cenário de 5% será incluído como referência. A proporção de 40%, por outro lado, serve para avaliar os métodos de imputação e estimação de ACF sob condições extremas de informação perdida.

Cada cenário de dados faltantes foi replicado 100 vezes e imputado usando procedimentos univariados e multivariados comuns na literatura e aqueles propostos neste trabalho com diferentes ajustes do componente temporal. Entre os univariados usados a mediana e a média. Os multivariados incluem o algoritmo EM para a distribuição normal.

Capítulo 7

Resultados

7.1 Resultados

Com o objetivo de avaliar a validade dos métodos de imputação e dos estimadores da função de autocorrelação na presença de dados faltantes propostos neste trabalho, esta seção dedica-se à apresentação de resultados de simulação e aplicação da metodologia a um conjunto de dados reais.

7.1.1 Simulações

Para avaliar o desempenho da metodologia proposta para imputar os dados faltantes e estimar a função de autocorrelação na presença deles, esta seção dedica-se à apresentação de resultados de simulação. Nos estudos de simulação foram comparadas as estimativas da função de autocorrelção obtidas através dos estimadores $\hat{\rho}_{PDR}(l)$, $\hat{\rho}_{SST}(l)$ e $\hat{\rho}_{T}(l)$, definidos em (5.8), (5.9) e (5.10), respectivamente, com o valor teórico.

Realizou-se 100 replicações com tamanho amostral $N = 100$ e $N = 1000$. Os dados gerados são provenientes de um processo $AR(1)$ (seção 4.5), $a_t \sim N(0,1)$, com $\phi = 0.3$, 0.5, 0.7, 0.95 e 0.99. As séries geradas foram investigadas com proporções iguais à $5\%, 15\%, 30\%$ e 40% de dados faltantes gerados. Os resultados de simulação foram obtidos através da linguagem de programação R versão 2.15.1, calculando para cada experimento a média das estimativas da função de autocorrelação, e o erro quadrático médio. Para avaliar os métodos de imputação,

cada série com dados faltantes foi imputada usando procedimentos univariados (média e mediana) e multivariados (algoritmo EM), com a série completa obteve-se as estimativas da função de autocorrelação.

O trabalho de Greenlad e Rothman (1998) indica que, para uma pequena proporção de dados faltantes e um grande número de observações, a análise de dados completos produz bons resultados. Desta forma, a porcentagem de 5% dados faltantes foi incluída como referência. A porcentagem de 40%, por outro lado, serve para avaliar os métodos de imputação e estimação de ACF sob condições extremas de informação perdida. A tabela 7.1 contenpla os valores teóricos para o processo $AR(1)$ com diferentes valores para ϕ, já as tabelas 7.2, 7.3, 7.4 e 7.5 contém os resultados de estimação da função de autocorrelação obtidas para a amostra imputada.

Tabela 7.1: ACF teórica de um processo $AR(1)$.

ϕ		$\rho(1)$	$\rho(2)$	$\rho(3)$	$\rho(4)$	$\rho(5)$
0.3	$\rho(l)$	0.3	0.09	0.027	0.008	0.002
0.5	$\rho(l)$	0.5	0.25	0.125	0.062	0.031
0.7	$\rho(l)$	0.7	0.49	0.343	0.240	0.168
0.95	$\rho(l)$	0.95	0.902	0.857	0.814	0.773

Tabela 7.2: Estimativas da ACF para um processo $AR(1)$, com $N = 100$ e $\phi = 0.3$, imputado, com quanto proporções de dados falantes.

P[1]	Método de imputação		$\hat{\rho}(1)$	$\hat{\rho}(2)$	$\hat{\rho}(3)$
5%	Média	$\hat{\rho}(l)$	0.27242	0.06580	0.01047
		$\widehat{EQM}$	0.00950	0.01154	0.01195
	Mediana	$\hat{\rho}(l)$	0.27266	0.06598	0.01046
		$\widehat{EQM}$	0.00948	0.01152	0.01199
	Algoritmo EM	$\hat{\rho}(l)$	0.29310	0.08293	0.02423
		$\widehat{EQM}$	0.00950	0.01170	0.01236
15%	Média	$\hat{\rho}(l)$	0.25359	0.09407	0.02713
		$\widehat{EQM}$	0.01152	0.00783	0.01056
	Mediana	$\hat{\rho}(l)$	0.25342	0.09340	0.02596
		$\widehat{EQM}$	0.01152	0.00783	0.01079
	Algoritmo EM	$\hat{\rho}(l)$	0.31238	0.15848	0.08347
		$\widehat{EQM}$	0.01024	0.01449	0.01511
30%	Média	$\hat{\rho}(l)$	0.18063	0.05223	0.01640
		$\widehat{EQM}$	0.02087	0.00973	0.00951
	Mediana	$\hat{\rho}(l)$	0.18188	0.05257	0.01672
		$\widehat{EQM}$	0.02055	0.00959	0.00933
	Algoritmo EM	$\hat{\rho}(l)$	0.31480	0.18501	0.13048
		$\widehat{EQM}$	0.00986	0.02203	0.02508
40%	Média	$\hat{\rho}(l)$	0.15322	0.00949	0.00399
		$\widehat{EQM}$	0.03030	0.01492	0.01120
	Mediana	$\hat{\rho}(l)$	0.15251	0.00849	0.00252
		$\widehat{EQM}$	0.03045	0.01510	0.01104
	Algoritmo EM	$\hat{\rho}(l)$	0.34839	0.20737	0.16558
		$\widehat{EQM}$	0.01689	0.03106	0.03565

(1) P = porcentagem de dados faltantes

Tabela 7.3: Estimativas da ACF para um processo $AR(1)$, com $N = 100$ e $\phi = 0.5$, imputado, com quanto proporções de dados falantes.

P[1]	Método de imputação		$\hat{\rho}(1)$	$\hat{\rho}(2)$	$\hat{\rho}(3)$
5%	Média	$\hat{\rho}(l)$	0.44414	0.20872	0.09967
		EQM	0.01169	0.01365	0.01526
	Mediana	$\hat{\rho}(l)$	0.44375	0.20861	0.09954
		EQM	0.01173	0.01368	0.01532
	Algoritmo EM	$\hat{\rho}(l)$	0.46844	0.23617	0.12239
		EQM	0.00957	0.01273	0.01526
15%	Média	$\hat{\rho}(l)$	0.40230	0.17702	0.08176
		EQM	0.01712	0.01810	0.01306
	Mediana	$\hat{\rho}(l)$	0.40176	0.17724	0.08171
		EQM	0.01739	0.01793	0.01282
	Algoritmo EM	$\hat{\rho}(l)$	0.47530	0.25637	0.14499
		EQM	0.00855	0.01444	0.01447
30%	Média	$\hat{\rho}(l)$	0.33723	0.15236	0.05328
		EQM	0.03585	0.01880	0.01479
	Mediana	$\hat{\rho}(l)$	0.33610	0.15362	0.05385
		EQM	0.03661	0.01859	0.01477
	Algoritmo EM	$\hat{\rho}(l)$	0.48683	0.30652	0.19683
		EQM	0.00956	0.01534	0.01783
40%	Média	$\hat{\rho}(l)$	0.26410	0.12050	0.06944
		EQM	0.06375	0.02589	0.01306
	Mediana	$\hat{\rho}(l)$	0.26268	0.12056	0.06874
		EQM	0.06461	0.02565	0.01290
	Algoritmo EM	$\hat{\rho}(l)$	0.49661	0.35780	0.28185
		EQM	0.01023	0.02642	0.04031

(1) P = porcentagem de dados faltantes

Tabela 7.4: Estimativas da ACF para um processo $AR(1)$, com $N = 100$ e $\phi = 0.7$, imputado, com quanto proporções de dados falantes.

P[1]	Método de imputação		$\hat{\rho}(1)$	$\hat{\rho}(2)$	$\hat{\rho}(3)$
5%	Média	$\hat{\rho}(l)$	0.62592	0.40205	0.24622
		EQM	0.01233	0.02173	0.02941
	Mediana	$\hat{\rho}(l)$	0.62575	0.40182	0.24609
		EQM	0.01243	0.02187	0.24609
	Algoritmo EM	$\hat{\rho}(l)$	0.65454	0.43361	0.27852
		EQM	0.00882	0.01723	0.02432
15%	Média	$\hat{\rho}(l)$	0.55361	0.35128	0.21513
		EQM	0.02745	0.02994	0.03391
	Mediana	$\hat{\rho}(l)$	0.55298	0.35119	0.21477
		EQM	0.02773	0.02995	0.03423
	Algoritmo EM	$\hat{\rho}(l)$	0.64390	0.44689	0.30271
		EQM	0.00968	0.01439	0.02140
30%	Média	$\hat{\rho}(l)$	0.45628	0.30498	0.19987
		EQM	0.06594	0.04487	0.03223
	Mediana	$\hat{\rho}(l)$	0.45320	0.30269	0.19720
		EQM	0.06768	0.04687	0.03307
	Algoritmo EM	$\hat{\rho}(l)$	0.66123	0.51616	0.39294
		EQM	0.00826	0.01356	0.02118
40%	Média	$\hat{\rho}(l)$	0.36919	0.24589	0.14570
		EQM	0.11795	0.07165	0.05142
	Mediana	$\hat{\rho}(l)$	0.36341	0.24186	0.14342
		EQM	0.12253	0.07358	0.05192
	Algoritmo EM	$\hat{\rho}(l)$	0.65387	0.51968	0.40569
		EQM	0.01088	0.01605	0.02455

(1) P = porcentagem de dados faltantes

Tabela 7.5: Estimativas da ACF para um processo $AR(1)$, com $N = 100$ e $\phi = 0.95$, imputado, com quanto proporções de dados falantes.

P[1]	Método de imputação		$\hat{\rho}(1)$	$\hat{\rho}(2)$	$\hat{\rho}(3)$
5%	Média	$\hat{\rho}(l)$	0.83938	0.74119	0.65476
		EQM	0.01546	0.03506	0.05716
	Mediana	$\hat{\rho}(l)$	0.83779	0.73956	0.65310
		EQM	0.01604	0.03573	0.05786
	Algoritmo EM	$\hat{\rho}(l)$	0.88424	0.78941	0.70081
		EQM	0.00727	0.02125	0.04021
15%	Média	$\hat{\rho}(l)$	0.74625	0.65394	0.57211
		EQM	0.04578	0.07029	0.09603
	Mediana	$\hat{\rho}(l)$	0.74077	0.65066	0.56948
		EQM	0.04860	0.07221	0.09776
	Algoritmo EM	$\hat{\rho}(l)$	0.88087	0.79100	0.70743
		EQM	0.00779	0.02087	0.03771
30%	Média	$\hat{\rho}(l)$	0.60654	0.52856	0.46731
		EQM	0.12257	0.14773	0.16240
	Mediana	$\hat{\rho}(l)$	0.58028	0.50395	0.44687
		EQM	0.14233	0.16773	0.17963
	Algoritmo EM	$\hat{\rho}(l)$	0.87456	0.80046	0.72350
		EQM	0.00871	0.01719	0.03003
40%	Média	$\hat{\rho}(l)$	0.51134	0.43109	0.38380
		EQM	0.20285	0.23431	0.23874
	Mediana	$\hat{\rho}(l)$	0.47460	0.39747	0.35404
		EQM	0.24057	0.27068	0.27020
	Algoritmo EM	$\hat{\rho}(l)$	0.88560	0.82121	0.75147
		EQM	0.00733	0.01351	0.02319

(1) P = porcentagem de dados faltantes

Devido a sua facilidade de implementação, a imputação pela média se torna um método muito comum e bastante utilizado (MYRTVEIT *et al.,* 2001). Nesta técnica, a média dos valores de um conjunto de dados que contém dados faltantes é usada para preencher seus os espaços com dados faltantes (FARHANGFAR *et al.,* 2004). Mas, os resultados de simulação evidenciam que este método não é eficaz para o tratamento, pois os valores extremos ficam sub-representados, o que implica na perda de variabilidade, ou seja, a variância das variáveis com dados faltantes é subestimada.

Pecebe-se que o efeito prejudicial deste método é reduzido somente em amostras com uma pequena porcentagem de dados faltantes (5%). A média é a melhor medida de tendência central para variáveis normalmente distribuídas, assim quando não se trata de uma distribuição normal, os resultados deste método não apresentam bons resultados. Outro método testado foi a mediana, a imputação da mediana é bem parecida com a imputação da média e apresenta as mesmas vantagens e desvantagens. Porém, este método é uma alternativa melhor para variáveis que não são normalmente distribuídas, pois a mediana representa melhor a tendência central de uma distribuição que possui grandes desvios da distribuição normal.

O terceiro método de imputação utilizado no trabalho foi o algoritmo EM *(Expectatin Maximi-*

zation), o nome vem de seus dois principais passos: Esperança e Maximização. A formalização deste método foi proposta por Dempster *et al.,* (1977). Segundo Little e Rubin (2002) o algoritmo EM é um método geral para obter estimativas de máxima verossimilhança em bases de dados incompletos. Como estas estimativas podem ser difíceis de obter para bases de dados complexas, é necessário um procedimento para reduzir esta dificuldade, que é o objetivo do algoritmo EM (MACKNIGHT *et al.,* 2007). O algoritmo EM foi utilizado para estimar os dados faltantes, que tem uma determinada distribuição de probabilidades, a qual foi obtida utilizando a parte não faltante.

As estimativas obtidas para a ACF das séries imputadas pelo algoritmo EM apresentaram resultados muito próximos do valor de referência, o teórico. O procedimento apresentou uma ligeira tendência a superestimar os valores da ACF à medida que a porcentagem de dados faltantes aumentava (Tabelas 7.2, 7.3, 7.4 e 7.5). De forma geral, a imputação pelo algoritmo EM apresentou boas estimativas com todas as porcentagens de dados faltentes avaliadas.

Os resultados de simulações (Tabelas 7.2, 7.3 e 7.5) evidenciam que com 5% de dados faltantes, todos os procedimentos testados para fazer as imputações pruduziram boas estimativas para ACF. Neste caso, a quantidade de dados faltantes é muito pequena para prejudicar a eficiência da estimação da função de autocorrelação. Mesmo com pequenas quantidades de dados faltantes, a imputação pela média e mediana deve ser evitada, devido à perda de variabilidade dos dados. A validade da análise, com estes dois métodos, começa a degenerar com o aumento da porcentagem de dados faltantes.

As tabelas 7.6, 7.7, 7.8 e 7.9 contêm os resultados de estimação da função de autocorrelação obtidas através dos estimadores $\hat{\rho}_{PDR}(l)$, $\hat{\rho}_{SST}(l)$ e $\hat{\rho}_{T}(l)$. Os resultados numéricos evidenciam a insensibilidade dos estimadores em relação à porcentagem de dados faltantes. Na medida que a quantidade de dados faltantes aumenta os três estimadores mantêm-se praticamente inalterados.

Os resultados apresentados nas tabelas 7.6, 7.7, 7.8 e 7.9, indicam que, ambos os estimadores forneceram estimativas para a função de autocorrelação bem próximas, independendo da quantidade de dados faltantes testada (5%, 15%, 30% e 40%). Empiricamente os estimadores propostos têm EQM, obtidos através dos valores estimados e os teóricos, significamente pequenos. Características estas que indicam que tal metodologia pode ser aplicada em amostras com grandes e pequenas porcentagens de dados faltantes.

Tabela 7.6: Estimativas da função de autocorrelação obtidas para um processo $AR(1)$ com $\phi = 0.3$, através dos estimadores propostos, com tamanho da amostra $N = 100$.

P[1]	Estimador		$\hat{\rho}(1)$	$\hat{\rho}(2)$	$\hat{\rho}(3)$
5%	$\hat{\rho}_{PDR}(l)$	Média	0.29665	0.06237	0.02182
		$\widehat{EQM}$	0.00855	0.01329	0.01426
	$\hat{\rho}_{SST}(l)$	Média	0.29768	0.06248	0.02082
		$\widehat{EQM}$	0.00861	0.01337	0.01457
	$\hat{\rho}_{T}(l)$	Média	0.29767	0.06175	0.02171
		$\widehat{EQM}$	0.00833	0.01316	0.01441
15%	$\hat{\rho}_{PDR}(l)$	Média	0.29363	0.07212	0.0479
		$\widehat{EQM}$	0.01673	0.01799	0.01732
	$\hat{\rho}_{SST}(l)$	Média	0.29209	0.07168	0.04925
		$\widehat{EQM}$	0.01615	0.01851	0.01753
	$\hat{\rho}_{T}(l)$	Média	0.29305	0.07111	0.04751
		$\widehat{EQM}$	0.01558	0.01797	0.01734
30%	$\hat{\rho}_{PDR}(l)$	Média	0.29524	0.06503	0.00754
		$\widehat{EQM}$	0.01894	0.02112	0.02107
	$\hat{\rho}_{SST}(l)$	Média	0.29450	0.06656	0.00777
		$\widehat{EQM}$	0.01900	0.02185	0.02028
	$\hat{\rho}_{T}(l)$	Média	0.29536	0.06683	0.00743
		$\widehat{EQM}$	0.01869	0.02065	0.02095
40%	$\hat{\rho}_{PDR}(l)$	Média	0.28963	0.07883	0.03725
		$\widehat{EQM}$	0.02720	0.02502	0.02722
	$\hat{\rho}_{SST}(l)$	Média	0.27969	0.07941	0.03655
		$\widehat{EQM}$	0.02697	0.02387	0.02490
	$\hat{\rho}_{T}(l)$	Média	0.27825	0.07800	0.03916
		$\widehat{EQM}$	0.02433	0.02462	0.02866

(1) P = porcentagem de dados faltantes

Tabela 7.7: Estimativas da função de autocorrelação obtidas para um processo $AR(1)$ com $\phi = 0.5$, através dos estimadores propostos, com tamanho da amostra $N = 100$.

P[1]	Estimador		$\hat{\rho}(1)$	$\hat{\rho}(2)$	$\hat{\rho}(3)$
5%	$\hat{\rho}_{PDR}(l)$	Média	0.48794	0.23618	0.11573
		EQM	0.00888	0.01574	0.01746
	$\hat{\rho}_{SST}(l)$	Média	0.48828	0.23513	0.11466
		EQM	0.00812	0.01565	0.01728
	$\hat{\rho}_{T}(l)$	Média	0.48722	0.23656	0.11652
		EQM	0.00825	0.01575	0.01750
15%	$\hat{\rho}_{PDR}(l)$	Média	0.48919	0.24548	0.13386
		EQM	0.01081	0.01714	0.02088
	$\hat{\rho}_{SST}(l)$	Média	0.48872	0.24551	0.13274
		EQM	0.00904	0.01687	0.02088
	$\hat{\rho}_{T}(l)$	Média	0.48969	0.24409	0.13295
		EQM	0.00951	0.01645	0.02061
30%	$\hat{\rho}_{PDR}(l)$	Média	0.48520	0.24287	0.12032
		EQM	0.01983	0.02303	0.02617
	$\hat{\rho}_{SST}(l)$	Média	0.48126	0.24541	0.12203
		EQM	0.01688	0.02249	0.02613
	$\hat{\rho}_{T}(l)$	Média	0.48069	0.24344	0.12004
		EQM	0.01525	0.02261	0.02502
40%	$\hat{\rho}_{PDR}(l)$	Média	0.47660	0.21089	0.10759
		EQM	0.03345	0.03562	0.03950
	$\hat{\rho}_{SST}(l)$	Média	0.47279	0.21504	0.10327
		EQM	0.02798	0.03621	0.03746
	$\hat{\rho}_{T}(l)$	Média	0.47550	0.21124	0.10395
		EQM	0.02617	0.03438	0.03921

(1) P = porcentagem de dados faltantes

Tabela 7.8: Estimativas da função de autocorrelação obtidas para um processo $AR(1)$ com $\phi = 0.7$, através dos estimadores propostos, com tamanho da amostra $N = 100$.

P[1]	Estimador		$\hat{\rho}(1)$	$\hat{\rho}(2)$	$\hat{\rho}(3)$
5%	$\hat{\rho}_{PDR}(l)$	Média	0.67926	0.47265	0.32244
		$\widehat{EQM}$	0.00665	0.01620	0.02304
	$\hat{\rho}_{SST}(l)$	Média	0.68166	0.47399	0.32253
		$\widehat{EQM}$	0.00578	0.01537	0.02283
	$\hat{\rho}_{T}(l)$	Média	0.67945	0.47040	0.32068
		$\widehat{EQM}$	0.00592	0.01553	0.02279
15%	$\hat{\rho}_{PDR}(l)$	Média	0.67237	0.47092	0.32806
		$\widehat{EQM}$	0.01135	0.01839	0.02447
	$\hat{\rho}_{SST}(l)$	Média	0.67593	0.47413	0.32918
		$\widehat{EQM}$	0.00924	0.01777	0.02381
	$\hat{\rho}_{T}(l)$	Média	0.67521	0.46717	0.32477
		$\widehat{EQM}$	0.00884	0.01708	0.02294
30%	$\hat{\rho}_{PDR}(l)$	Média	0.68142	0.45688	0.31622
		$\widehat{EQM}$	0.01677	0.02397	0.02924
	$\hat{\rho}_{SST}(l)$	Média	0.69265	0.47001	0.32045
		$\widehat{EQM}$	0.01219	0.02531	0.02916
	$\hat{\rho}_{T}(l)$	Média	0.68448	0.45895	0.31921
		$\widehat{EQM}$	0.00906	0.02050	0.02791
40%	$\hat{\rho}_{PDR}(l)$	Média	0.67318	0.44828	0.30541
		$\widehat{EQM}$	0.02411	0.03466	0.03529
	$\hat{\rho}_{SST}(l)$	Média	0.68077	0.46526	0.31082
		$\widehat{EQM}$	0.02013	0.04530	0.03554
	$\hat{\rho}_{T}(l)$	Média	0.67445	0.44168	0.30210
		$\widehat{EQM}$	0.01492	0.02898	0.03263

(1) P = porcentagem de dados faltantes

Tabela 7.9: Estimativas da função de autocorrelação obtidas para um processo $AR(1)$ com $\phi = 0.95$, através dos estimadores propostos, com tamanho da amostra $N = 100$.

P[1]	Estimador		$\hat{\rho}(1)$	$\hat{\rho}(2)$	$\hat{\rho}(3)$
5%	$\hat{\rho}_{PDR}(l)$	Média	0.93118	0.89921	0.86982
		EQM	0.00320	0.00481	0.00720
	$\hat{\rho}_{SST}(l)$	Média	0.93261	0.90061	0.87100
		EQM	0.00208	0.00367	0.00607
	$\hat{\rho}_{T}(l)$	Média	0.92803	0.89520	0.86438
		EQM	0.00210	0.00343	0.00555
15%	$\hat{\rho}_{PDR}(l)$	Média	0.91768	0.88327	0.85057
		EQM	0.00725	0.00846	0.01106
	$\hat{\rho}_{SST}(l)$	Média	0.92679	0.89358	0.86037
		EQM	0.00382	0.00657	0.00982
	$\hat{\rho}_{T}(l)$	Média	0.92114	0.88489	0.85110
		EQM	0.00316	0.00486	0.00720
30%	$\hat{\rho}_{PDR}(l)$	Média	0.91742	0.87671	0.84465
		EQM	0.00946	0.01087	0.01178
	$\hat{\rho}_{SST}(l)$	Média	0.92433	0.88883	0.85922
		EQM	0.00462	0.00949	0.01144
	$\hat{\rho}_{T}(l)$	Média	0.91917	0.88191	0.84723
		EQM	0.00337	0.00529	0.00739
40%	$\hat{\rho}_{PDR}(l)$	Média	0.94505	0.89923	0.86644
		EQM	0.01777	0.01588	0.01746
	$\hat{\rho}_{SST}(l)$	Média	0.92729	0.88220	0.85100
		EQM	0.00493	0.01104	0.01281
	$\hat{\rho}_{T}(l)$	Média	0.92884	0.89560	0.86184
		EQM	0.00241	0.00371	0.00598

(1) P = porcentagem de dados faltantes

O procedimento de imputação EM testado neste trabalho apresentou bons resultados em diversas situações de dados faltantes. Mesmo com bons resultados, deve-se considerar que dados imputados são apenas estimativas dos valores que seriam observados. Segundo Junger (2008) a imputação de amostras com padrões complexos com uma pocentagem grande de dados faltantes deve receber atenção especial. Desta forma, os resultados empíricos sugerem que a metodologia proposta neste trabalho mostra-se como uma alternativa para tal problema, podendo ser aplicada a conjuntos de dados com observações faltantes, não comprometendo a análise estatística.

As tabelas 7.10 e 7.11, contém os valores teóricos para um processo $AR(1)$ com $\phi = 0.99$ e os resultados da estimação da ACF através dos estimadores propostos e algoritmo EM, respectivamente. Tais resultados evidenciam que o comportamento da ACF, obtida através dos estimadores proportos, são melhor que os resultados obtidos com o algoritmo EM, quando o processo está próximo da não estacionariedade, $\phi = 0.99$.

Já nas tabelas 7.12 e 7.13 é apresentado a comparação do desempenho dos métodos sob amostras grandes $(N = 1000)$ e com ϕ próximo de zero e um ($\phi = 0.3$ e $\phi = 0.99$). Fica evidente que os resultados das estimativas da ACF obtidos com os estimadores propostos, independete do valor de ϕ, são melhores do que os obtidos com o algoritmo EM, para amostras com 40% de dados faltantes.

Para avaliar os métodos sob pressupostos de não assimetria dos dados, o processo AR(1) foi ajustado em dados simulados por uma distribuição Qui-Quadrado $(g.l = 2)$. Os resuldados (Tabelas 7.14 e 7.15) mostram que os estimadores se comportam bem para dados simulados provinientes de distribuições não asimétricas. Os resultados (Tabelas 7.14, 7.15, 7.16 e 7.17) sugerem que os estimadores $\hat{\rho}_{PDR}(l)$, $\hat{\rho}_{SST}(l)$ e $\hat{\rho}_{T}(l)$ mantém as propriedades, mesmo quando aplicados em dados simulados com distribuição de probabilidade diferente da distribuição normal. Já nas tabelas 7.18 e 7.19, é apresentado os valores teóricos da ACF para um processo $ARMA(1,1)$, com $\phi = 0.7$ e $\theta = 0.3$, resultados das estimativas para ACF, através dos estimadore propostos e algoritmo EM, para o processo, sendo que, o principal objetivo dos resultados contidos na tabela 7.19 é mostrar a importancia da especificação da ordem do modelo para o algoritmo EM, que não é o caso dos estimadores propostos.

Tabela 7.10: ACF teórica de um processo $AR(1)$, com $\phi = 0.99$.

	$\rho(1)$	$\rho(2)$	$\rho(3)$	$\rho(4)$	$\rho(5)$
$\rho(l)$	0.99	0.9801	0.9702	0.9605	0.9509

Tabela 7.11: Estimativas da ACF para um processo $AR(1)$, com $\phi = 0.99$, através dos estimadores propostos e algoritmo EM, com tamanho de amostra $N = 100$.

P[1]	Estimador		$\hat{\rho}(1)$	$\hat{\rho}(2)$	$\hat{\rho}(3)$	$\hat{\rho}(4)$	$\hat{\rho}(5)$
5%	Algoritmo EM	Média	0.92068	0.85021	0.78154	0.71763	0.65759
		$\widehat{EQM}$	0.00637	0.02207	0.04691	0.07702	0.11034
	$\hat{\rho}_{PDR}(l)$	Média	0.97843	0.96376	0.94999	0.93651	0.92371
		$\widehat{EQM}$	0.00159	0.00281	0.00440	0.00626	0.00831
	$\hat{\rho}_{SST}(l)$	Média	0.97234	0.95992	0.94849	0.93768	0.92736
		$\widehat{EQM}$	0.00143	0.00273	0.00439	0.00628	0.00842
	$\hat{\rho}_{T}(l)$	Média	0.96829	0.95333	0.93880	0.92480	0.91152
		$\widehat{EQM}$	0.00132	0.00254	0.00416	0.00607	0.00807
40%	Algoritmo EM	$\hat{\rho}(l)$	0.91434	0.86316	0.80883	0.75353	0.69547
		$\widehat{EQM}$	0.00842	0.01983	0.03616	0.05697	0.08505
	$\hat{\rho}_{PDR}(l)$	Média	0.98335	0.95502	0.93495	0.91446	0.89678
		$\widehat{EQM}$	0.01665	0.01462	0.01618	0.01950	0.02223
	$\hat{\rho}_{SST}(l)$	Média	0.96734	0.94570	0.93286	0.91331	0.89612
		$\widehat{EQM}$	0.00316	0.00936	0.01415	0.01853	0.02214
	$\hat{\rho}_{T}(l)$	Média	0.95945	0.93995	0.92172	0.90405	0.88655
		$\widehat{EQM}$	0.00263	0.00489	0.00753	0.01058	0.01419

(1) P = porcentagem de dados faltantes

Tabela 7.12: Estimativas da ACF para um processo $AR(1)$, com $\phi = 0.3$, através dos estimadores propostos e algoritmo EM, com tamanho de amostra $N = 1000$.

P[1]	Estimador		$\hat{\rho}(1)$	$\hat{\rho}(2)$	$\hat{\rho}(3)$	$\hat{\rho}(4)$	$\hat{\rho}(5)$
5%	Algoritmo EM	Média	0.28235	0.08559	0.03154	0.00816	0.00385
		$\widehat{EQM}$	0.00104	0.00095	0.00118	0.00127	0.00106
	$\hat{\rho}_{PDR}(l)$	Média	0.29709	0.08992	0.03307	0.00803	0.00371
		$\widehat{EQM}$	0.00082	0.00102	0.00134	0.00144	0.00117
	$\hat{\rho}_{SST}(l)$	Média	0.29634	0.08969	0.03297	0.00806	0.00375
		$\widehat{EQM}$	0.00084	0.00101	0.00133	0.00143	0.00117
	$\hat{\rho}_{T}(l)$	Média	0.29688	0.08993	0.03306	0.00801	0.00368
		$\widehat{EQM}$	0.00083	0.00103	0.00134	0.00142	0.00117
40%	Algoritmo EM	$\hat{\rho}(l)$	0.19986	0.07755	0.04745	0.03416	0.03123
		$\widehat{EQM}$	0.01103	0.00158	0.00173	0.00183	0.00200
	$\hat{\rho}_{PDR}(l)$	Média	0.29612	0.08533	0.03463	0.01270	0.00662
		$\widehat{EQM}$	0.00249	0.00330	0.00335	0.00289	0.00277
	$\hat{\rho}_{SST}(l)$	Média	0.29461	0.08421	0.03490	0.01271	0.00677
		$\widehat{EQM}$	0.00246	0.00321	0.00331	0.00280	0.00276
	$\hat{\rho}_{T}(l)$	Média	0.29596	0.08484	0.03476	0.01208	0.00643
		$\widehat{EQM}$	0.00231	0.00326	0.00333	0.00286	0.00273

(1) P = porcentagem de dados faltantes

Tabela 7.13: Estimativas da ACF para um processo $AR(1)$, com $\phi = 0.99$, através dos estimadores propostos e algoritmo EM, com tamanho de amostra $N = 1000$.

P[1]	Estimador		$\hat{\rho}(1)$	$\hat{\rho}(2)$	$\hat{\rho}(3)$	$\hat{\rho}(4)$	$\hat{\rho}(5)$
5%	Algoritmo EM	Média	0.97310	0.95824	0.94317	0.92854	0.91400
		$\widehat{EQM}$	0.00040	0.00077	0.00129	0.00192	0.00265
	$\hat{\rho}_{PDR}(l)$	Média	0.98808	0.98204	0.97592	0.96973	0.96359
		$\widehat{EQM}$	0.00014	0.00017	0.00026	0.00040	0.00057
	$\hat{\rho}_{SST}(l)$	Média	0.98772	0.98204	0.97619	0.97022	0.96431
		$\widehat{EQM}$	0.00008	0.00009	0.00016	0.00029	0.00045
	$\hat{\rho}_{T}(l)$	Média	0.98713	0.98084	0.97456	0.96830	0.96205
		$\widehat{EQM}$	0.00009	0.00011	0.00014	0.00025	0.00041
40%	Algoritmo EM	$\hat{\rho}(l)$	0.91273	0.90334	0.89621	0.88750	0.87896
		$\widehat{EQM}$	0.00728	0.00744	0.00729	0.00744	0.00754
	$\hat{\rho}_{PDR}(l)$	Média	0.98624	0.97954	0.97370	0.96714	0.96048
		$\widehat{EQM}$	0.00219	0.00186	0.00168	0.00164	0.00171
	$\hat{\rho}_{SST}(l)$	Média	0.98705	0.97926	0.97432	0.96882	0.96256
		$\widehat{EQM}$	0.00010	0.00061	0.00097	0.00131	0.00141
	$\hat{\rho}_{T}(l)$	Média	0.98671	0.98005	0.97339	0.96681	0.96023
		$\widehat{EQM}$	0.00010	0.00011	0.00016	0.00028	0.00043

(1) P = porcentagem de dados faltantes

Tabela 7.14: Estimativas da ACF para um processo (Qui-Quadrado g.l=2) $AR(1)$, com $N = 100$ e $\phi = 0.7$, imputado, com quanto proporções de dados falantes.

P[1]	Método de imputação		$\hat{\rho}(1)$	$\hat{\rho}(2)$	$\hat{\rho}(3)$	$\hat{\rho}(4)$	$\hat{\rho}(5)$
5%	Média	$\hat{\rho}(l)$	0.64058	0.42765	0.28784	0.18534	0.11017
		$\widehat{EQM}$	0.00839	0.01418	0.01800	0.02073	0.02003
	Mediana	$\hat{\rho}(l)$	0.63987	0.42736	0.28798	0.18536	0.11004
		$\widehat{EQM}$	0.00861	0.01434	0.01814	0.02088	0.02008
	Algoritmo EM	$\hat{\rho}(l)$	0.67604	0.46089	0.31723	0.21161	0.13123
		$\widehat{EQM}$	0.00481	0.01164	0.01652	0.01950	0.01885
40%	Média	$\hat{\rho}(l)$	0.38956	0.24736	0.16378	0.10197	0.06275
		$\widehat{EQM}$	0.10557	0.06987	0.04591	0.03061	0.02139
	Mediana	$\hat{\rho}(l)$	0.39185	0.24840	0.16480	0.10166	0.06305
		$\widehat{EQM}$	0.10424	0.06948	0.04554	0.03067	0.02127
	Algoritmo EM	$\hat{\rho}(l)$	0.66511	0.52925	0.41693	0.32515	0.24564
		$\widehat{EQM}$	0.00954	0.01618	0.02391	0.02714	0.02566

(1) P = porcentagem de dados faltantes

Tabela 7.15: Estimativas da função de autocorrelação obtidas para um processo $AR(1)$ (Qui-Quadrado g.l=2) com $\phi = 0.7$, através dos estimadores propostos, com tamanho da amostra $N = 100$.

P[1]	Estimador		$\hat{\rho}(1)$	$\hat{\rho}(2)$	$\hat{\rho}(3)$	$\hat{\rho}(4)$	$\hat{\rho}(5)$
5%	$\hat{\rho}_{PDR}(l)$	Média	0.67314	0.43805	0.28119	0.17769	0.09670
		$\widehat{EQM}$	0.00637	0.01493	0.02165	0.02557	0.02985
	$\hat{\rho}_{SST}(l)$	Média	0.66860	0.43514	0.27933	0.17616	0.09574
		$\widehat{EQM}$	0.00585	0.01442	0.02131	0.02503	0.02933
	$\hat{\rho}_{T}(l)$	Média	0.68073	0.44604	0.28786	0.18248	0.09879
		$\widehat{EQM}$	0.00543	0.01382	0.02183	0.02576	0.03049
40%	$\hat{\rho}_{PDR}(l)$	Média	0.68265	0.44585	0.27919	0.14597	0.09223
		$\widehat{EQM}$	0.03148	0.03686	0.04500	0.04675	0.04014
	$\hat{\rho}_{SST}(l)$	Média	0.67195	0.44401	0.28173	0.14911	0.09197
		$\widehat{EQM}$	0.01652	0.03515	0.04711	0.04740	0.03866
	$\hat{\rho}_{T}(l)$	Média	0.68364	0.44056	0.27897	0.14837	0.09386
		$\widehat{EQM}$	0.01609	0.02869	0.04221	0.04737	0.04136

(1) P = porcentagem de dados faltantes

Tabela 7.16: Estimativas da ACF para um processo (t de Student g.l=3) $AR(1)$, com $N = 100$ e $\phi = 0.7$, imputado, com quanto proporções de dados falantes

P[1]	Método de imputação		$\hat{\rho}(1)$	$\hat{\rho}(2)$	$\hat{\rho}(3)$	$\hat{\rho}(4)$	$\hat{\rho}(5)$
5%	Média	$\hat{\rho}(l)$	0.63110	0.41564	0.27191	0.16827	0.09627
		$\widehat{EQM}$	0.01123	0.02000	0.02174	0.02263	0.02409
	Mediana	$\hat{\rho}(l)$	0.63090	0.41538	0.27172	0.16827	0.09615
		$\widehat{EQM}$	0.01129	0.02009	0.02172	0.02267	0.02398
	Algoritmo EM	$\hat{\rho}(l)$	0.66285	0.44590	0.30134	0.19328	0.11735
		$\widehat{EQM}$	0.00753	0.01631	0.01969	0.02105	0.02244
40%	Média	$\hat{\rho}(l)$	0.37201	0.23998	0.14480	0.10065	0.06449
		$\widehat{EQM}$	0.11949	0.07390	0.05243	0.03190	0.02295
	Mediana	$\hat{\rho}(l)$	0.36754	0.23485	0.14217	0.09787	0.06264
		$\widehat{EQM}$	0.12321	0.07682	0.05372	0.03264	0.02332
	Algoritmo EM	$\hat{\rho}(l)$	0.65394	0.53145	0.42303	0.33220	0.25942
		$\widehat{EQM}$	**0.01108**	**0.01747**	**0.02667**	**0.02960**	**0.03001**

(1) P = porcentagem de dados faltantes

Tabela 7.17: Estimativas da função de autocorrelação obtidas para um processo $AR(1)$ (t de Student g.l=3) com $\phi = 0.7$, através dos estimadores propostos, com tamanho da amostra $N = 100$.

P[1]	Estimador		$\hat{\rho}(1)$	$\hat{\rho}(2)$	$\hat{\rho}(3)$	$\hat{\rho}(4)$	$\hat{\rho}(5)$
5%	$\hat{\rho}_{PDR}(l)$	Média	0.68564	0.46869	0.33205	0.23633	0.17039
		$\widehat{EQM}$	0.00786	0.01704	0.02352	0.02806	0.02729
	$\hat{\rho}_{SST}(l)$	Média	0.68283	0.46713	0.33192	0.23517	0.16987
		$\widehat{EQM}$	0.00658	0.01636	0.02341	0.02816	0.02723
	$\hat{\rho}_{T}(l)$	Média	0.68062	0.46628	0.33088	0.23371	0.17001
		$\widehat{EQM}$	0.00706	0.01634	0.02312	0.02726	0.02695
40%	$\hat{\rho}_{PDR}(l)$	Média	0.66414	0.47292	0.32435	0.24503	0.17006
		$\widehat{EQM}$	0.03702	0.02867	0.03487	0.03803	0.04207
	$\hat{\rho}_{SST}(l)$	Média	0.68845	0.52013	0.34617	0.26449	0.18996
		$\widehat{EQM}$	0.01566	0.07588	0.04839	0.05453	0.07369
	$\hat{\rho}_{T}(l)$	Média	0.67821	0.46236	0.33362	0.23885	0.17244
		$\widehat{EQM}$	0.01624	0.02319	0.03601	0.03583	0.04070

(1) P = porcentagem de dados faltantes

Tabela 7.18: ACF teórica de um processo $ARMA(1,1)$, com $\phi = 0.7$ e $\theta = 0.3$.

	$\rho(1)$	$\rho(2)$	$\rho(3)$	$\rho(4)$	$\rho(5)$
$\rho(l)$	0.720	0.504	0.352	0.246	0.172

Tabela 7.19: Estimativas da função de autocorrelação obtidas para um processo $ARMA(1,1)$ com $\phi = 0.7$ e $\theta = 0.3$, através dos estimadores propostos, com tamanho da amostra $N = 100$.

P[1]	Estimador		$\hat{\rho}(1)$	$\hat{\rho}(2)$	$\hat{\rho}(3)$	$\hat{\rho}(4)$	$\hat{\rho}(5)$
5%	$EM_{erro}(l)^2$	Média	0.73182	0.48037	0.31162	0.20558	0.13869
		$\widehat{EQM}$	0.00358	0.01019	0.01607	0.02188	0.02422
	$EM(l)^3$	Média	0.75830	0.50452	0.32533	0.19778	0.09984
		$\widehat{EQM}$	0.00436	0.01040	0.01815	0.02577	0.03162
	$\hat{\rho}_{PDR}(l)$	Média	0.78806	0.54149	0.36616	0.24383	0.15094
		$\widehat{EQM}$	0.00828	0.01130	0.01757	0.02448	0.02935
	$\hat{\rho}_{SST}(l)$	Média	0.79189	0.54339	0.36608	0.24407	0.15118
		$\widehat{EQM}$	0.00749	0.01067	0.01643	0.02382	0.02900
	$\hat{\rho}_T(l)$	Média	0.78769	0.53753	0.36386	0.24170	0.14972
		$\widehat{EQM}$	0.00698	0.01039	0.01664	0.02354	0.02852
40%	$EM_{erro}(l)^2$	Média	0.49936	0.34211	0.24981	0.17121	0.12664
		$\widehat{EQM}$	0.05993	0.04091	0.02981	0.03033	0.02497
	$EM(l)^3$	Média	0.73080	0.58512	0.46269	0.36415	0.27277
		$\widehat{EQM}$	0.00624	0.01888	0.02960	0.03373	0.03224
	$\hat{\rho}_{PDR}(l)$	Média	0.80270	0.56341	0.36966	0.28115	0.19438
		$\widehat{EQM}$	0.02253	0.02493	0.02870	0.03978	0.04063
	$\hat{\rho}_{SST}(l)$	Média	0.79985	0.57503	0.36834	0.28340	0.19465
		$\widehat{EQM}$	0.01430	0.03704	0.02868	0.04434	0.04542
	$\hat{\rho}_T(l)$	Média	0.79508	0.56124	0.37631	0.28508	0.19693
		$\widehat{EQM}$	0.00977	0.01808	0.02799	0.03854	0.04398

(1) P = porcentagem de dados faltantes

(2) EM_{erro} - com erro de especificação (EM-AR(1))

(3) EM - com especificação correta (EM-ARMA(1,1))

7.1.2 Aplicações

Como ilustração, nesta seção, é apresentado duas aplicações da metodologia descrita nas Seções 4 e 5. Os dados (séries temporais) são observações obtidas na estação de monitoramento do bairro de Jardim Camburi, Vistória-ES, Brasil. A primeria utiliza os dados de concentrações médias diárias de PM_{10}, medidas em $\mu g/m^3$. Foram consideradas 731 observações compreendidas entre 01 de janeiro de 2003 e 31 de dezembro de 2004. Com esta série fez a análise dos metodos de imputação, propostos na seção 3.2, e estimação da função de autocorrelação na presença de dados faltantes, seção 5, com as seguintes proporções de dados faltantes $5\%, 15\%, 30\%$ e 40%. A tabela 7.22 contém as medidas sumárias para a série de concentrações de PM_{10}.

Tabela 7.20: Estatísticas descritivas da série de concentração das médias diárias PM_{10}.

	Média	Desvio	Mínimo	Máximo	Mediana
Série de PM_{10}	27.593	7.816	7.083	67.830	27.333

A metodologia testada neste trabalho assume que o comportamento dos dados segue uma distribuição normal. Entretando, geralmente, dados de concentrações de PM_{10}, não apresentam variância não constante, e por esse motivo, diversos trabalhores apontam as transformações de Box-Cox (1964) aos dados como melhor opção para estabilizar a variância da série. A Figura 7.1 mostra o gráfico da série de concentração de PM_{10} e do logaritmo dessa série.

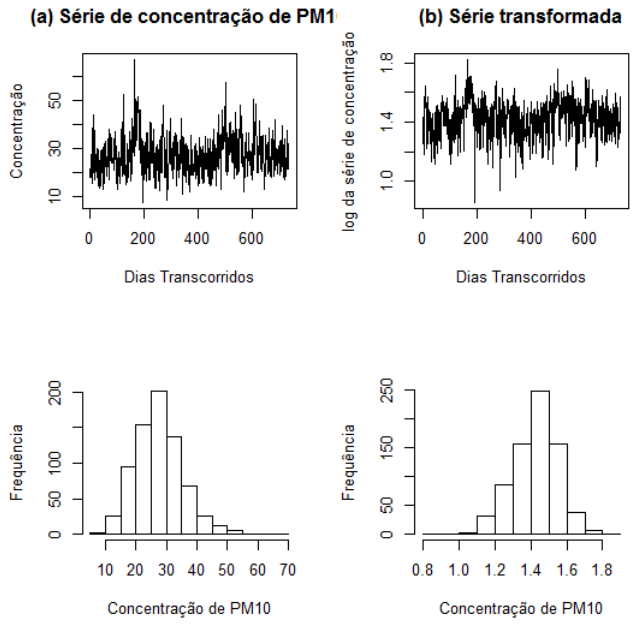

Figura 7.1: Série de concentração de PM_{10}, log da série, histograma da série e do logaritmo da mesma.

Com o objetivo de quantificar e exemplificar o efeito provocado, aos métodos de imputação, pela porcentagem de dados faltantes, foi avaliada a performance para quatros conficurações da série transformada. Na aplicação foi feita uma única replicação, foram utilizados os indicadores sugeridos no trabalho de Junnien *et al.,* (2004)(Seção 6.5).

O coeficiente de correlação de Pearson é o indicador mais comum para avaliar o desempenho de métodos de imputação. A raiz do erro quadrático médio foi utilizada para estimar o erro de cada imputação. O erro absoluto médio foi usado como uma medida mais sensível do erro da imputação, pois é menos influenciado por grandes diferenças entre os valores originais e os imputados. A tabela 7.23 apresenta os resultados para os indicadores de performance.

Tabela 7.21: Performance dos métodos de imputação com quanto proporções de dados falantes, *log* da série de PM_{10}.

P[1]	Método de imputação	$\widehat{r}^3$	$\widehat{RMSE}$	$\widehat{MAE}$
5%	Média	0.90752	0.05863	0.01757
	Mediana	0.97855	0.02616	0.00638
	Algoritmo EM	0.75694	0.08770	0.05158
15%	Média	0.50257	0.20527	0.10216
	Mediana	0.87362	0.06295	0.02638
	Algoritmo EM	0.67821	0.09754	0.06126
30%	Média	0.18902	0.52008	0.36561
	Mediana	0.08267	1.01664	0.72032
	Algoritmo EM	0.62641	0.10253	0.07158
40%	Média	0.08671	0.72482	0.57097
	Mediana	0.03686	1.13764	0.90049
	Algoritmo EM	0.53115	0.11195	0.07977

(1) P = porcentagem de dados faltantes

(3) $\widehat{r}$ = Coeficiente de correlação de Pearson

De acordo com os resultados da tabela 7.23, observa-se um gradiente de crescimento ou decrescimento nos indicadores em função da quantidade de dados faltantes. Através da Tabela 7.23 nota-se que ambos procedimentos forneceram bons resultados para a porcentagem de 5% de dados faltantes. Os métodos de imputação pela média e mediana consiste em substituir o valor faltante por uma constante, logo estes métodos apresentam uma Ineficiência quando aplicados aos dados com uma porcentagem maior que 5% de dados faltantes. O método de imputação

pelo algoritmo EM, apresentou bom desenpenho com valores baixos de $\widehat{RMSE}$ e $\widehat{MAE}$. Os coeficientes de correlação foram superiores a 0.53. As tabelas 7.24 e 7.25 contém as estimativas para a função de autocorrelação da série 1, com os valores de referência para as proporções de dados faltantes, obtidas através das metodologias estuda no presente trabalho.

Tabela 7.22: ACF da série 1 de concentrações PM_{10} com 5% de dados faltantes.

$\hat{\rho}(l)$	$\hat{\rho}(l)$	$\hat{\rho}(l)EM$	$\hat{\rho}_{PDR}(l)$	$\hat{\rho}_{SST}(l)$	$\hat{\rho}_{T}(l)$
$\hat{\rho}(1)$	0.470	0.471	0.477	0.477	0.479
$\hat{\rho}(2)$	0.236	0.243	0.234	0.234	0.236
$\hat{\rho}(3)$	0.152	0.170	0.145	0.144	0.145
$\hat{\rho}(4)$	0.141	0.166	0.127	0.128	0.126
$\hat{\rho}(5)$	0.107	0.122	0.104	0.103	0.107
$\hat{\rho}(6)$	0.232	0.236	0.230	0.229	0.229
$\hat{\rho}(7)$	0.338	0.335	0.340	0.337	0.338
$\hat{\rho}(8)$	0.180	0.172	0.189	0.187	0.188
$\hat{\rho}(9)$	0.083	0.095	0.088	0.087	0.087
$\hat{\rho}(10)$	0.062	0.088	0.072	0.071	0.072

Tabela 7.23: ACF da série 1 de concentrações PM_{10} com 40% de dados faltantes.

$\hat{\rho}(l)$	$\hat{\rho}(l)$	$\hat{\rho}(l)EM$	$\hat{\rho}_{PDR}(l)$	$\hat{\rho}_{SST}(l)$	$\hat{\rho}_{T}(l)$
$\hat{\rho}(1)$	0.470	0.382	0.472	0.510	0.488
$\hat{\rho}(2)$	0.236	0.265	0.305	0.337	0.267
$\hat{\rho}(3)$	0.152	0.185	0.172	0.182	0.166
$\hat{\rho}(4)$	0.141	0.152	0.162	0.172	0.153
$\hat{\rho}(5)$	0.107	0.134	0.084	0.087	0.083
$\hat{\rho}(6)$	0.232	0.239	0.223	0.230	0.230
$\hat{\rho}(7)$	0.338	0.312	0.352	0.370	0.351
$\hat{\rho}(8)$	0.180	0.196	0.134	0.146	0.132
$\hat{\rho}(9)$	0.083	0.133	0.095	0.100	0.086
$\hat{\rho}(10)$	0.062	0.117	0.020	0.021	0.020

As figuras 7.2(a), 7.2(b), 7.2(c), 7.2(d) e 7.2(e) mostram, respectivamente, as estimativas da ACF obtidas, (a) com a série de PM_{10} completa, as demais foram obtidas com a série faltando 5% dos dados, (b) uso do EM para fazer as imputações, (c) com o estimador $\hat{\rho}_{PDR}(l)$, (d) com

$\hat{\rho}_{SST}(l)$ e (e) através de $\hat{\rho}_{T}(l)$.

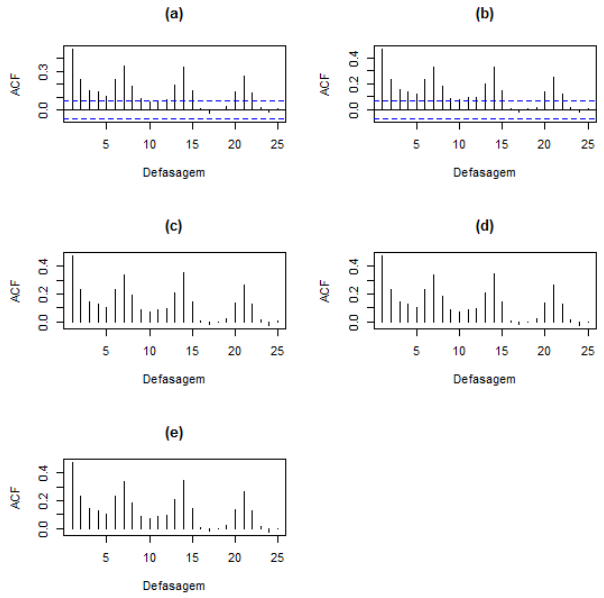

Figura 7.2: Estimaticas da ACF da série 1, com 5% de dados faltantes.

As figuras 7.3(a), 7.3(b), 7.3(c), 7.3(d) e 7.3(e) mostram, respectivamente, as estimativas da ACF obtidas, (a) com a série de PM_{10} completa, as demais foram obtidas com a série faltando 40% dos dados, (b) uso do EM para fazer as imputações, (c) com o estimador $\hat{\rho}_{PDR}(l)$, (d) com $\hat{\rho}_{SST}(l)$ e (e) através de $\hat{\rho}_{T}(l)$.

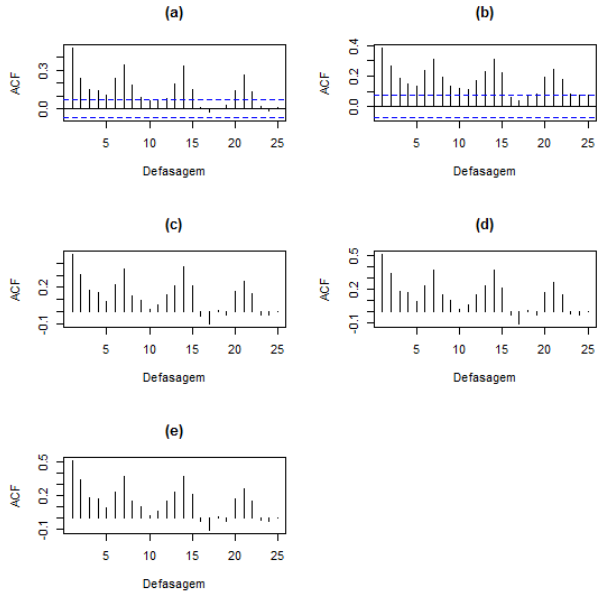

Figura 7.3: Estimaticas da ACF da série 1, com 40% de dados faltantes.

Por meio das tabelas 7.24 e 7.25 e das figuras 7.2 e 7.3, nota-se que os estimadores $\hat{\rho}_{PDR}(l)$, $\hat{\rho}_{SST}(l)$ e $\hat{\rho}_{T}(l)$ forneceram boas estimativas quando comparadas com as obtidas através da

amostra completa. Já as estimativas obtidas com a série imputada via EM, foram boas para a proporção de 5% de dados faltantes, quando tal foi aumentada para 40% o procedimento apresentou uma tendência a superestimar os valores da ACF, o que não aconteceu com os estimadores propostos.

A segunda aplicação utiliza a série de concentração médias diárias de PM_{10}, denominada série 2, medidas em μg/m^3. O período aqui considerado se estende de 01 de janeiro de 2005 e 31 de dezembro de 2006, são 730 observações, sendo que 57 são dados faltantes. A figura 7.4 mostra a série de PM_{10} com *missind data* e o histograma. A tabela 7.26 contém as estimativas para a função de autocorrelação da série 2.

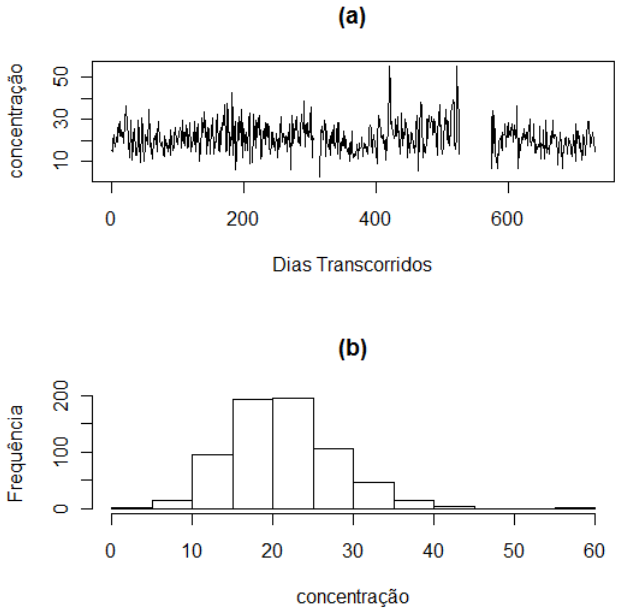

Figura 7.4: Série de concentração de PM_{10} com dados faltantes e histograma da mesma.

As figuras 7.5(a), 7.5(b), 7.5(c), 7.5(d) e 7.5(e) mostram, respectivamente, as estimativas da ACF obtidas, (a) com a exclusão dos dados faltantes, (b) com o uso do EM para fazer as imputações, (c) com o estimador $\hat{\rho}_{PDR}(l)$, (d) com $\hat{\rho}_{SST}(l)$ e (e) através de $\hat{\rho}_T(l)$.

Dos resultados anteriores (aplicação 2), pode-se concluir que o métodos testados constituem-se uma boa opção para estimar a função de autocorreção na presença de dados faltantes; adicionalmente, a metodologia sugerida na Seção 5.1 é adequada para estimação da ACF de séries com mémória curta e com a propriedade de sazonalidade, no contexto de modelos de BOX e JENKINS.

Tabela 7.24: ACF da série 2 de concentrações PM_{10} com dados faltantes.

$\hat{\rho}(l)$	$\hat{\rho}(l)EM$	$\hat{\rho}_{PDR}(l)$	$\hat{\rho}_{SST}(l)$	$\hat{\rho}_{T}(l)$
$\hat{\rho}(1)$	0.412	0.411	0.412	0.415
$\hat{\rho}(2)$	0.147	0.139	0.141	0.141
$\hat{\rho}(3)$	0.033	0.020	0.024	0.021
$\hat{\rho}(4)$	0.035	0.018	0.019	0.019
$\hat{\rho}(5)$	0.048	0.033	0.032	0.034
$\hat{\rho}(6)$	0.106	0.093	0.092	0.096
$\hat{\rho}(7)$	0.150	0.142	0.139	0.146
$\hat{\rho}(8)$	0.094	0.085	0.083	0.087
$\hat{\rho}(9)$	0.085	0.072	0.070	0.074
$\hat{\rho}(10)$	0.111	0.097	0.093	0.099

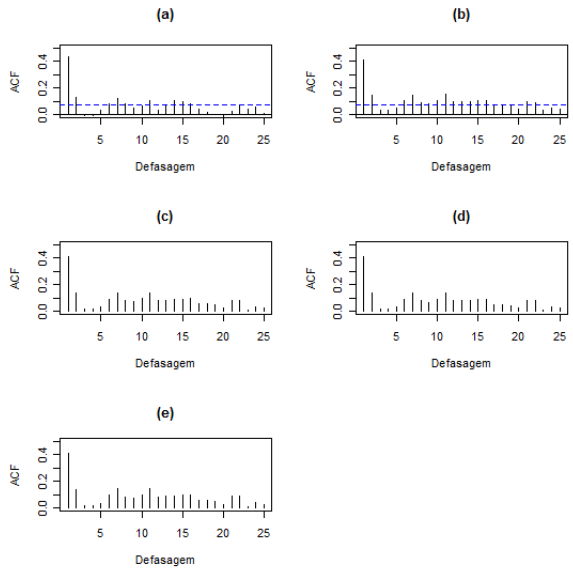

Figura 7.5: Estimaticas da ACF da série 2.

Capítulo 8

Conclusões e Trabalhos Futuros

8.1 Conclusões

Esta dissertação apresenta um estudo de metodologias para dados faltantes em séries temporais. As metodologias são baseadas nos métodos de imputação e nos estimadores da função de autocorrelação na presença de dados faltantes. Os resultados de simulação indicam que, empiricamente, em geral, os métodos de imputação pela média e mediana são ineficazes para amostras com uma porcentagem de dados faltantes superior a 5%. Além disso, a análise de permance dos métodos de imputação, baseando em uma única replicação, mostram que os métodos IU aprsentam baixa qualidade. Estes métodos mostraram baixa correlação dos valores reais em relação aos imputados, para amostras com porcentagem de dados faltantes superior a 30%. Já para o algoritmo EM, o estudo de simulação mostrou que tal metodologia apresenta boas estimativas para todos as quatro conficurações de dados faltantes, entretanto, fica evidente que para processos perto da não-estacionariedade, este método tende a subestimar os valores da ACF.

As estimativas da ACF obtiadas com os estimadores $\hat{\rho}_{PDR}(l)$, $\hat{\rho}_{SST}(l)$ e $\hat{\rho}_{T}(l)$ (YAJIMA e NISHINO, 1999) e com o algoritmo EM (DEMPSTER, 1977) tem EQM signifcativamente pequenos comparados com os valores teóricos para os processor estudados. Os resultados da aplicação 1, sugerem que o método EM tendem a superestimar os valores da ACF. A investigação empírica foi feita em diferentes cenários, destacando o efeito da porcentagem de dados faltantes sobre os estimadores. As séries de concentrações de PM_{10}, aplicação 1 e 2, apresentam sazonalidade. Através dos resultados das aplicações, percebe-se que a metodologias testadas

neste trabalho representa bem tal fenômeno, mesmo sob condições estremas de dados faltantes (40%). A metodologia estudada neste trabalho mostrou ser uma alternativa para o tratamento de dados faltantes em séries temporais de concentrações de poluentes atmosféricos, podendo ser aplicada a conjunto de dados contendo deferentes porcentagens de dados faltantes.

8.2 Trabalhos Futuros

- Implementar os estimadores propostos em uma plataforma do R;
- Implementar, em uma plataforma do R, o algoritmo EM para modelos ARIMA com d fracionário;
- Utilizar técnicas de modelagens que efetuem a estimação dos parâmetros dos modelos, para séries com memória longa, assumindo a presença de dados faltantes. Neste contexto, podemos citar Dunsmuir e Robinson (1981).

Capítulo 9

Referências Bibliográficas

AGIRRE-BASURKO, E.; IBARRA-BERASTEGI, G.; MADARIAGA, I., 2006. Regression and multilayer perceptron-based models to forecast hourly O_3 and NO_2 leves in the Bilbao area. *Environmental Modelling e Software*, 21, 430:446.

ALMEIDA, M. A. I., 2006. Modelo Aditivo Generalizado (MAG) no estudo da relação entre o número de atendimentos hospitalares por causas respiratórias e a qualidade do ar. Dissertação de Mestrado, Vitória: Programa de Pós-Graduação em Engenharia Ambiental: Universidade Federal do Espírito Santo.

ASSUNÇÃO, F., 2012. Estratégias para tratamento de variáveis com dados faltantes durante o desenvolvimento de modelos preditivos. Dissertação de Mestrado, São Paulo: Instituto de Matemática e Estatística: Universidade de São Paulo.

BAIRD, C. Química Ambiental. 2ª Edição. Porto Alegre, RS: Bookman, 2002.

BARBOSA, G.C., 2009. O modelo aditivo generalizado e a técnica de Bootstrap: um estudo entre o número de atendimento hospitalar por causas respiratórias e a qualidade do ar. 60 f. Dissertação (Mestrado) Programa de Pós-Graduação em Engenharia Ambiental do Centro Tecnológico, UFES, Vitória, 2009.

BOX G. COX DR., 1964. An analysis of transformations. *Journal of the Royal Statistical Society*. B.; 26(2): 211-252.

BOX, G. E. P. e JENKINS, G. M., 1970. Time Series Analysis, Forecasting and Control. Holden-Day.

BOX, G. E. P.; JENKINS, G. M.; REINSEL, G. C., 2008. Time Series Analysis, Forecasting and Control. Holden-Day.

BOX, G. E. P.; JENKINS, G. M.; REINSEL, G. C.. Time Series Analysis, Forecasting and Control. 3 ed. New Jersey, Prentice Hall, 1994.

BRAGA A. L. S., *et al.*, 2007. Associação entre poluição atmosférica e doenças respiratórias e cardiovasculares na cidade de Itabira, Minas Gerais, Brasil. Cadernos de Saúde Pública. vol. 23, suppl.4, pp. S570-S578.

BRAGA, B. *et al.*,. Introdução à engenharia ambiental: O desafio do desenvolvimento sustentável. 2. ed. São Paulo: Pearson Prentice Hall, 2005.

Brockwell, P. e Davis, R., 2002. Introduction to Time Series and Forecasting, second edn, Springer Verlag.

Brockwell, P. e Davis, R., 2006. Time Series: Theory and Methods, second edn, Springer Verlag.

CASTRO, H. A., *et al.*, 2009. Efeitos da poluição do ar na função respiratória de escolares, Rio de Janeiro, RJ. Revista de Saúde Pública 43, 26-34.

CONSELHO NACIONAL DE MEIO AMBIENTE - CONAMA. Resolução CONAMA Nº 03, de 28 de junho de 1990. Diário Oficial da União, 28 de junho. 1990.

DEMPSTER A, Laird N, Rubin D. Maximum Likelihood from Incomplete Data via the Algorithm EM. *Journal of the Royal Statistical Society*, B. 1977;39:1-38.

DUNSMUIR, W. e ROBINSON, P. M., 1981. Estimation of times series models in the presence of missing data. *Journal of the American Statistical Association*, 76, No. 375, pp. 560-568.

FARHANGFAR, A., KURGAN, L., PEDRYCZ, W., 2004. Experimental anlysis of methods for imputation of missing values in databases. *Proceedings of SPIE*, Orlando, vol. 5421, pp.172-182.

GODISH, T. Air quality. Boca Raton: CRC Press, LLC, 1997.

GOMES K. S., 2009. Modelagem INAR(p) para a previsão de índices de qualidade do ar. 71 f. Dissertação (Mestrado) Programa de Pós-Graduação em Engenharia Ambiental do Centro Tecnológico, UFES, Vitória.

GOYAL, P., CHAN, A. T. e JAISWAL, N., 2006. Statistical models for the prediction of respirable suspended particulate matter in urban cities. *Atmospheric Environment* 40, 2068-2077.

GREENLAND S, e FINKLE W. D., 1995. A critical look at methods for handling missing covariates in epidemiologic regression analyses. *American Journal of Epidemiology*, 1255-1264.

GREENLAND S, e ROTHMAN K. J., 1998. Modern epidemiology. 2 ed. Philadelphia, Lippincott-Raven.

HINRICHS, R.A., KLEINBACH, M., BELICO DOS REIS, L., Energia e Meio Ambiente. Cengage Learning, Tradução da 4a Edição norte-americana, 2011.

HOLGATE, S. T.; SAMET, J. M.; KOREN, H. S.; MAYNARD, R. L. Air Pollution and Health. San Diego, EUA. Academic Press, 1999.

IBGE, Instituto Brasileiro de Geografia e Estatística: Censo 2010. Disponível em: http://www.-ibge.gov.br/home/estatistica/populacao/censo2010/resultados_dou/ES2010.pdf. Acesso em: 01 dezembro 2011.

IGLESIAS, P., JORQUERA, H. e PALMA, W., 2005 . Data analysis using regression models with missing observations and long memory: an application study. *Computational statistics e Data Analysis* 50, 2028-2043.

INSTITUTO ESTADUAL DE MEIO AMBIENTE E RECURSOS HÍDRICOS. Relatório da qualidade do ar na região da grande vitória 2007. Cariacica: 2008.

JUNGER, W. L., 2008. Análise, imputação de dados e interfaces computacionais em estudos de séries temporais epidemiológicas. 178 f. Tese (Doutorado) - Programa de Pós-graduação em Saúde Coletiva, Universidade do Estado do Rio de Janeiro, Rio de Janeiro.

JUNNINEN, H., NISKAA, H., TUPPURAINENC, K., RUUSKANENA, J. e KOLEH-MAINEN, M., 2004. Methods for imputation of missing values in air quality data sets. *Atmospheric Environment* 38, 2895-2907.

LIRA, T. S., 2009. Modelagem e previsão da qualidade do ar na cidade de Uberlândia - MG. Tese de Doutorado, Uberlândia: Programa de Pós-Graduação em Engenharia Química: Universidade Federal de Uberlândia.

LITTLE, R.J.A. e RUBIN D.B., 1989. Statiscal analysis with missing data. New York, Wiley.

LORA, E. Eduardo Silva. Prevenção e controle da poluição nos setores energético, industrial e de transporte. 2.ed. Rio de Janeiro: Interciência, 2002.

MARTINS, L. C., *et al.*, 2002. Poluição atmosférica e atendimentos por pneumonia e gripe em São Paulo, Brasil. Revista de Saúde Pública, vol. 36, p. 88-94.

MCKNIGHT, P. C., MCKNIGHT, K. M., FIGUEREDO, A. J., Missing data: a gentle introduction. New York. The Guilford Press, 2007.

MISHARA, A. K. e DESAI V. R., 2005. Drought forecasting using stochastic model. *Stochastic Enviromental Rescarch and Risk Assessment*. vol.19. pp. 326-339.

MORETTIN, P. A. e TOLOI, C. M. C.. Análise de séries temporais. 2. ed. São Paulo: Egard Blucher, 2006.

MYRTVEIT, I., STENSRUD, E., OLSSOM, U. H., 2001. Analyzing data sets with missing data: an empirical evaluation of imputation methods and likelihood-based methods. *IEEE Transactions On Software Engineering*, vol. 27, no. 11, pp. 999-1013.

NASCIMENTO, L.F.C.; PEREIRA, L.A.A.; BRAGA, A.L.F.; MÓDOLOA, M.C.C. e CARVALHO, J.A.C., 2006. Efeitos da poluição atmosférica na saúde infantil em São José dos Campos, SP. Revista de Saúde Pública, vol. 40, p. 77-82.

NORAZIAN, M. Noor. *et al.*, 2008. Estimation of missing values in air pollution data using single imputation techniques. *ScienceAsia*. vol. 34: 341-345.

OSTRO, B., SANCHES, J. M., ARANDA, C. e ESKELAND, G. S., 1996. Air pollution and mortality: results from a study of Santiago, Chile. *Journal of Exposure Analysis and Environmental Epidemiology* 6, 97-114.

PARZEN, E., 1963. On spectral analysis with missing observations and amplitude modulation. *Sankhyã*, Ser. A. 25, 383-392.

PAULA, R. R. C., 2002. Modelagem estocástica das flutuações turbulentas de poluentes atmosféricos. 142 f. Dissertação (Mestrado) Programa de Pós-Graduação em Engenharia Ambiental do Centro Tecnológico, UFES, Vitória.

PEREZ, P. e REYES, J., 2002. Prediction of maximum of 24-h average of PM_{10} concentrations 30 h in advance in Santiago, Chile. *Atmospheric Environment* 36, 4555-4561.

PLAIA A, BONDÌ AL., 2006. Single imputation method of missing values in environmental pollution data sets. *Atmospheric Environment* 40: 7316-7330.

PRIESTLEY, M.B., 1983. Spectral Analysis in Time Series. Academic Press.

RAISEN, V. A. e SILVA, A. N. O uso da linguagem R para cálculos de estatística básica. Vitória, ES: EDUFES, 2011.

Relatório de Monitoramento da Qualidade do Ar na Região da Grande Vitória 2000 - 2009. UFES. Departamento de Engenharia Ambiental (2010).

ROBESON S.M. e STEYN, D.G., 1990. Evaluation and comparison of statistical forecast models for daily maximum ozone concentrations. *Atmospheric Environment*. Part B. 24(2), 303-312.

SCHAFER, J.L., Analysis of incomplete multivariante data. London, Chapman Hall, 1997.

SEINFELD, J. H. e PANDIS, S. N., 1998. Atmospheric chemistry and physics - from air pollution to climate change. New York: John Wiley & Sons, 1998.

SHIN, D. W. e SARKAR, S., 1995. Estimation for stationary AR(1) models with non-consecutively observed samples. *Sankhyã*, Ser. A, 57, 287-298.

SHIVELY, T. S., 1990. An analysis of the long-term trend in ozone data from two Houston, Texas monitoring sites Atmospheric Environment. Part B. *Urban Atmosphere* 24(2), 293-301.

TAKEUCHI, K., 1995. A comment on Recent development of economic data analysis at the 63rd anual meeting of Japan Satistical Society.

TODA, H. Y., MAKENZIE, C.R., 1999. LM tests for unit roots in the presence of missing observations: small sample evidence. *Mathematics and computers in simulation*, vol. 48: 457-468.

TRINDADE, C. C., 2009. Avaliação do uso de diferentes modelos receptores com dados de PM2,5: balanço químico de massa (BQM) e fatoração de matriz positiva (FMP). 143 f. Dissertação (Mestrado) Programa de Pós-Graduação em Engenharia Ambiental do Centro Tecnológico, UFES, Vitória.

VERONEZE, R., 2007. Tratamento de dados faltantes empregando biclusterização com imputação múltipla. Dissertação de Mestrado, Campinas: Programa de Pós-Graduação em Engenharia Elétrica e de Computação: Universidade Estatual de Campinas.

WHO. Air quality guidelines: global update 2005. Copenhagem 2006.

WEI. W. Time Series Analysis: univariate and multivariate methods. Pearson. Boston. 2006. 2 ed.

YAJIMA,Y. e NISHINO, H., 1999. Estimation of the autocorrelation function of a stationary time series with missing observations. *Sankhyã*. Ser. A 61, 189-207.