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

WANDERSON DE PAULA PINTO

ANÁLISE ESPECTRAL DE SÉRIES TEMPORAIS DE CONCENTRAÇÕES DE POLUENTES ATMOSFÉRICOS COM DADOS FALTANTES

VITÓRIA
2019

WANDERSON DE PAULA PINTO

ANÁLISE ESPECTRAL DE SÉRIES TEMPORAIS DE CONCENTRAÇÕES DE POLUENTES ATMOSFÉRICOS COM DADOS FALTANTES

Tese apresentada ao Programa de Pós-graduação em Engenharia Ambiental, do Centro Tecnológico, da Universidade Federal do Espírito Santo, como requisito parcial para obtenção do título de Doutor em Engenharia Ambiental, na área de concentração Poluição do Ar.
Orientador: Prof. Dr. Valdério Anselmo Reisen.

VITÓRIA
2019

Dedico este trabalho à minha esposa Renata.

Aos meus filhos João Pedro e João Vitor.

Aos meus Pais Maria Lúcia e José Maria (in memoriam).

Aos meus Avós Rosa, Sebastiana, Otacílio e Walfredo.

AGRADECIMENTOS

“O desejo profundo da humanidade pelo conhecimento é justificativa suficiente para nossa busca contínua.“

Stephen Hawking

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E/OU SIGLAS

ACAH	Análise de cluster aglomerativo hierárquico
ACOVF	Autocovariance function
ACP	Análise de componentes principais
ACPS	Análise de componentes principais supervisionada
AC	Análise de clusters
ACF	Autocorrelation fuction
AF	Análise fatorial
AO	Additive outliers
AQAMN	Air Quality Automatic Monitoring Network
AR	Autoregressive
ARCH	Modelos autorregressivos de heterocedasticidade condicional
ARFIMA	Modelo univariado autorregressivo fracionário integrado de médias móveis
ARMA	Autorregressivo de médias móveis
ARIMA	Autorregressivo integrado de médias móveis
ASAS	Alta Pressão Subtropical do Atlântico Sul
Aw	Clima tropical quente
B	Backshift operator
BEKK	Baba, Engle, Kraft e Kroner
BQM	Balanço químico de massa
C	Celsius
CAR	Carapina
CAM	Camburi
CCF	Cross-correlation function
CEASA	Centrais de Abastecimento do Espírito Santo
CETESB	Companhia Ambiental do Estado de São Paulo
CH_4	Metano
CO	Monóxido de carbono
CONAMA	Conselho Nacional do Meio Ambiente
COVNM	Compostos orgânicos voláteis não-metano
CP	Componente principal
CW	Centro/Ocidental
DGP	Data-generating processes
DV	Direção do vento
ECOSOFT	Ecosoft consultoria e softwares ambientais
ES	Espírito Santo
FA	Factorial analysis
FAR	False alarm rate
FBR	Função de base radial

LISTA DE ABREVIATURAS E/OU SIGLAS

FMP	Fatoração da matriz positiva
GARCH	Modelos generalizados autorregressivos de heterocedasticidade condicional
GLP	Gás liquefeito de petróleo
GVR	Greater Vitória Region
HC	Hidrocarbonetos
H_2O_2	Peridóxido de hidrogênio
H_2O	Óxido de hidrogênio
I	Identity matrix
IBGE	Instituto Brasileiro de Geografia e Estatística
IEMA	Instituto Estadual do Meio Ambiente e Recursos Hídricos
IPA	Índice de poluição do ar
LAR	Laranjeiras
LO	Additive levels outliers
LOSS	Loss rate
LM	Lagrange multiplier
MA	Moving average
MG	Minas Gerais
MI	Metas Intermediárias
MP_{10}	Material particulado com diâmetro igual ou inferior a 10 micrômetros
$MP_{2.5}$	Material particulado com diâmetro inferior a 2,5 micrômetros
NMHCs	Hidrocarbonetos não-metano
NO	Óxido nítrico
NO_2	Dióxido de nitrogênio
NO_X	Óxido de nitrogênio
P	Pressão atmosférica
PF	Padrões Finais
PP	Preciptação
PIB	Produto Interno Bruto
PM_{10}	Particulate matter particles with a diameter of 10 micrometers or less
$PM_{2.5}$	Particulate matter smaller than 2.5 micrometers in diameter
PSM	Pressão de superfície média
OMS	Organização Mundial de Saúde
O_3	Ozônio
PC	Principal component
PCA	Principal components analysis
PIB	Produto Interno Bruto
PP	Preciptação pluviométrica

LISTA DE ABREVIATURAS E/OU SIGLAS

POD	Probability of detection
PVC	Principal volatility components
R	Radiação solar
RAMQAr	Automática de Monitoramento da Qualidade do Ar da Grande Vitória
RCOV	Robust covariance
RES	Residuals
RGV	Região da Grande Vitória
RPCA	Robust principal component analysis
RPVC	Robust Principal Volatility Components Analysis
RMSE	Root Mean Squared Error
RTSE	Modelo de regressão com erros de séries temporais
S	Sul
SARFIMA	Seasonal autoregressive fractionally integrated moving average
SO_2	Dióxido de enxofre
T	Temperatura ambiente
TSA	Temperatura da superfície do ar
TSP	Temperatura da superfície da pele
TPS	Total suspended particles
TW	Tsuen Wan
UR	Umidade relativa
VAR	Modelo vetorial autorregressivo
VARMA	Modelo vetorial autorregressivo de médias móveis
VARFIMA	Modelo vetorial autorregressivo fracionário integrado de médias móveis
VOC	Volatile organic compound
VO	Additive volatility outliers
VV	Velocidade do vento
W	Oeste
WD	Wind direction
WHO	World Health Organization
WS	Wind speed

RESUMO

A poluição atmosférica tem afetado de forma significativa os seres vivos, mesmo quando seus valores estão abaixo do permitido pelas entidades regulamentadoras. Neste sentido, as questões relativas à qualidade do ar têm se tornado cada vez mais importantes, uma vez que vários problemas de saúde decorrem da poluição atmosférica. Dessa forma, diversos estudos aplicado técnicas de análise de séries temporais têm sido realizados, com o intuito de contribuir como ferramentas na tomada de decisões dos agentes públicos e privados no que diz respeito à prevenção de concentrações elevadas, ao controle da poluição atmosférica e à formulação de legislações para esse fim. Uma das metodologias estatísticas adotadas é a análise espectral, sendo a mesma utilizada para identificar propriedades do conjunto de dados, como por exemplo a sazonalidade. No entanto, observa-se que, entre os estudos que têm adotado esta técnica, uma característica comum é negligenciar a presença de dados faltantes *(missing data)*, que pode levar à subestimar a precisão dos resultados. Nota-se que nas séries temporais relacionadas à poluição atmosférica um problema frequente é a presença de dados faltantes, geralmente devido a falhas dos equipamentos de monitoramento. Assim, este documento concentra-se no estudo de metodologias usadas para estimar a função de autocorrelação e a densidade espectral de séries temporais univariadas na presença ou sem dados faltantes. Os estimadores sugeridos são baseadas na metodologia de Amplitude Modulada, proposta por Parzen (1963), e no periodograma de Lomb-Scargle (LOMB, 1976; SCARGLE, 1982). Além disso, foi proposto estimadores das funções de autocovariância e autocorrelação de séries temporais, considerando a conexão entre o domínio do tempo e da frequência por meio da relação entre a função de autocovariância e a densidade espectral. Desta forma, no primeiro artigo desta tese foram apresentados três métodos para estimação da função de autocorrelação de séries temporais univariadas estacionárias na presença de dados faltantes. As propriedades teóricas dos estimadores foram avaliadas e seus desempenhos para amostras finitas investigados através de um estudo de simulação numérica. Por fim, foi proposto a aplicação destas metodologias para avaliar uma série temporal de concentrações de MP_{10} da Região da Grande Vitória (RGV), Espírito Santo, Brasil, com dados faltantes. No segundo artigo é apresentado um método de estimação para as funções de autocorrelação e autocovariância de séries temporais considerando a conexão entre o domínio do tempo e da frequência. As propriedades assintóticas do método são avaliadas através de estudo de simulação de Monte Carlo para diferentes tamanhos amostrais e porcentagens de dados faltantes. Já no terceiro artigo, que é a principal contribuição desta tese, foram propostos dois métodos para estimar a função de densidade espectral de séries temporais estacionárias na presença de dados faltantes. Foi estudado o efeito da porcentagem de dados faltantes nos estimadores empregados. Os métodos foram analisados através de simulações e uma aplicação a dados reais de MP_{10} monitorados na RGV também foi considerada.

Palavras-chave: Séries Temporais. Análise Espectral. Dados Faltantes. Poluição do Ar. Periodograma de Lomb-Scargle.

ABSTRACT

Air pollution has significantly affected living beings, even when their values ??are below what is allowed by regulators. In this regard, air quality issues have become increasingly important as a number of health problems arise from air pollution. In this way, several studies applied time series analysis techniques have been carried out, aiming to contribute as tools in the decision making of the public and private agents with respect to the prevention of high concentrations, the control of air pollution and the formulation legislation for this purpose. One of the statistical methodologies adopted is the spectral analysis, which is used to identify properties of the dataset, such as seasonality. However, it is noted that among studies that have adopted this technique, a common feature is to neglect the presence of missing data, which may lead to underestimation of the accuracy of the results. Note that in the time series related to atmospheric pollution a frequent problem is the presence of missing data, usually due to the failure of the monitoring equipment. Thus, this paper concentrates on the study of methodologies used to estimate the autocorrelation function and the spectral density of univariate time series in the presence or absence of missing data. The suggested estimators are based on the Amplitude Modulated methodology, proposed by Parzen (1963), and in the Lomb-Scargle (LOMB, 1976; SCARGLE, 1982) periodogram. In addition, we proposed estimators of autocovarianance and autocorrelation functions of time series, considering the connection between the time domain and frequency by means of the relation between the autocovariance function and the spectral density. Thus, in the first article of this thesis were presented three methods to estimate the autocorrelation function of univariate stationary time series in the presence of missing data. The theoretical properties of the estimators were evaluated and their performances for finite samples investigated through a numerical simulation study. Finally, it was proposed the application of these methodologies to evaluate a time series of concentrations of PM_{10} of the Region of Greater Vitória (RGV), Espírito Santo, Brazil, with missing data. The second article presents an estimation method for the autocorrelation and autocovariance functions of time series considering the connection between time domain and frequency. The asymptotic properties of the method are evaluated through a Monte Carlo simulation study for different sample sizes and percentages of missing data. In the third article, which is the main contribution of this thesis, two methods were proposed to estimate the spectral density function of stationary time series in the presence of missing data. The effect of the percentage of missing data on the employed estimators was studied. The methods were analyzed through simulations and an application to actual PM_{10} data monitored at the RGV was also considered.

Keywords: Time series. Spectral Analysis. Missing Data. Air Pollution. Lomb-Scargle Periodogram.

SUMÁRIO

1 INTRODUÇÃO

A preocupação com os efeitos da poluição do ar veio com o crescimento industrial, iniciado no período da Revolução Industrial, que teve início na Inglaterra, em meados do século XVIII, devido a alguns episódios de alta concentração de poluentes, episódios esses, que causaram aumento do número de mortes em algumas cidades da Europa e dos Estados Unidos da América (EUA). O primeiro desses episódios ocorreu na Bélgica, em 1930, no vale do rio Meuse e resultou em, aproximadamente, seis mil pessoas com problemas respiratórios e 60 mortes. Anos mais tarde, em 1948, ocorreu um episódio semelhante ao da Bélgica na cidade de Donora, (EUA), que resultou em 20 óbitos e na metade da população local hospitalizada (JACOBS; BURGESS; ABBOTT, 2018). Porém, o terceiro, e mais conhecido entre os grandes desastres da poluição atmosférica, ocorreu na capital britânica, Londres, no ano de 1952. O episódio foi marcado por uma intensa inversão térmica, com duração de quatro dias, que acarretou a morte de quatro mil pessoas por doenças cardíacas e respiratórias, principalmente bronquite e pneumonia. O aumento do índice de mortalidade afetou pessoas de todas as idades, entretanto, foi maior entre as crianças, com destaque para a mortalidade de recém-nascidos, que quase duplicou, e de crianças de 1 a 12 meses, a qual mais do que dobrou e, entre os adultos com mais de 45 anos (LOGAN et al., 1953).

Desde então, o crescimento populacional, principalmente em áreas urbanas, e a intensificação das atividades de industrialização, aliado ao crescimento da fronta veicular em países em desenvolvimento, têm provocado situações de impacto ambiental que englobam desde problemas sociais, econômicos e de saneamento básico a poluição atmosférica (BRASSEUR et al., 1999). O ar poluído é uma mistura de partículas - material particulado[1] (MP) - e gases que são emitidos para a atmosfera principalmente por indústrias, veículos automotivos, termoelétricas, queima de biomassa e de combustíveis fósseis (ARBEX et al., 2012). Os poluentes podem ser classificados em primários e secundários. Os poluentes primários são emitidos diretamente para a atmosfera, e os secundários são resultantes de reações químicas entre os poluentes primários (BAIRD, 2002; ARBEX et al., 2012). Segundo Vingarzan (2004) e Oltmans et al. (2006) a industrialização, a urbanização e a queima de combustíveis fósseis, são os principais fatores que causam a degradação da qualidade do ar em diversas partes do mundo.

Os principais impactos sobre a saúde pública causados pela poluição atmosférica incluem aumento da taxa de mortalidade, aumento da taxa de doenças crônicas, alteração de importantes funções fisiológicas, aumento no custo das despesas com saúde, diminuição da produtividade e da qualidade de vida de uma forma geral (MIHELCIC; ZIMMERMAN; AUER, 2014; DERISIO, 2016). Estudos realizados, conforme descrito na sub-seção 3.2, têm mostrado que a poluição (principalmente por partículas inaláveis com diâmetro menor que 2,5 micrômetros

[1] Conjunto de poluentes constituídos de poeiras, fumaças e todo tipo de material sólido e líquido que se mantém suspenso na atmosfera por causa de seu pequeno tamanho.

($MP_{2,5}$), partículas inaláveis com diâmetro menor que 10 micrômetros (MP_{10}), monóxido de carbono (CO), dióxido de enxofre (SO_2), óxidos de nitrogênio (NO_X) e ozônio (O_3)), é um fator causador de doenças do aparelho respiratório, doenças cardiovasculares, doenças neurológicas, dentre outras. Em relação à poluição do ar, vale frisar que, os efeitos causados ao meio ambiente e à saúde da população por causa da emissão de poluentes atmosféricos, podem não ser apenas locais, pois dependem de fatores da região, como o relevo do entorno da fonte de emissão, as condições meteorológicas e a natureza dos poluentes. Isso significa que, esses poluentes podem viajar milhares de quilômetros pela atmosfera, distância suficiente para atravessar estados, países e até fronteiras de continentes, atingindo, assim, comunidades distantes do ponto de emissão (BERGIN et al., 2005; LEITE et al., 2011).

Em um estudo, publicado no ano de 2018 na revista *The Lancet*, Landrigan et al. (2018) afirmaram que durante décadas, a poluição e seus efeitos nocivos sobre a saúde das pessoas, o meio ambiente e o planeta foram negligenciados tanto pelos governos como pela agenda internacional de desenvolvimento. Segundo os autores, a poluição é a maior causa ambiental de doença e morte no mundo de hoje, responsável por cerca de nove milhões de mortes prematuras em 2015, o que representa, aproximadamente, 16% de todas as mortes. Neste contexto, autores como Matos (2012) Souza et al. (2014), Freitas et al. (2016), Nascimento et al. (2017), Tufik et al. (2017), Frauches et al. (2017), entre outros, demonstraram a relação entre poluentes atmosféricos na Região da Grande Vitória[2] (RGV) e problemas de saúde, o que justifica a escolha dos dados de poluição atmosférica desta região para o desenvolvimento desta pesquisa.

As mudanças ocorridas na RGV, como a implantação e ampliação de grandes projetos industriais, aumento dos empreendimentos imobiliários, aumento do consumo de energia e o aumento da frota veicular, tendem a elevar os índices de poluição atmosférica, mesmo com as interposições legais de controle de emissões (MONTE, 2016). Vale mencionar que desde o início da década de 1970, a indústria era apontada como a principal fonte de poluição na RGV. Porém, de acordo com o IEMA (INSTITUTO ESTADUAL DE MEIO AMBIENTE E RECURSOS HÍDRICOS DO ESTADO DO ESPÍRITO SANTO, 2014), o crescimento da frota veicular e os empreendimentos imobiliários, têm alterado o perfil da região nos últimos anos. O inventário de emissões atmosféricas da RGV (ECOSOFT CONSULTORIA E SOFTWARES AMBIENTAIS, 2011) estimou que os veículos e a indústria minero-siderúrgica são os principais responsáveis pelas emissões de CO e hidrocarbonetos e, a indústria minero-siderúrgica é responsável por mais de 70% da emissão de SO_2 para a atmosfera e por mais de 45% dos óxidos de nitrogênio (NO_x). Além disso, o inventário indicou como principais responsáveis pelos níveis de MP_{10} a ressuspensão do solo (69,3%), as indústrias (19,6%) e os veículos (escapamento e evaporativa) (3,9%). É importante destacar que, no ano de 2010, a população da RGV era de cerca de 1.565.393 habitantes que representam, aproximadamente, 45% da população total do estado do Espírito Santo, sendo que, 98,6% dessa população reside em área urbana (INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA, 2014). Como em áreas urbanas e

[2] A RGV é composta por cinco municípios, a saber: Cariacica, Serra, Viana, Vila Velha e Vitória.

industriais os índices de poluição são mais elevados, pode-se inferir que, 45% da polução do estado foi fortemente afetada pelas emissões de poluição na atmosfera (MONTE, 2016).

Nos estudos recentes relacionados com a poluição do ar, diversos autores, ver sub-seção 3.3, tem aplicado as técnicas de análise de séries temporais para avaliação de dados de poluição. Neste contexto, especial atenção tem sido dada aos modelos de séries temporais, principalmente, pela necessidade de modelos matemáticos capazes de estimar e prever os níveis de poluentes na atmosfera em determinada localidade.Vários métodos de modelagem e previsão de séries temporais estão disponíveis na literatura, como os autorregressivos (AR), os de médias móveis (MA), regressão linear com o tempo, suavização exponencial de Holt-Winters, os modelos autorregressivos integrados e de médias móveis (ARIMA) e os modelos autorregressivos integrados e de médias móveis sazonal multiplicativo (SARIMA), entre outros.

A análise de séries temporais é um tema bem discutido no meio acadêmico. Há, basicamente, dois enfoques usados na análise de séries temporais (MORETTIN; TOLOI, 2006). Em ambos, o objetivo é construir modelos para as séries, com os seguintes propósitos: investigar o mecanismo gerador da série temporal, fazer previsões de valores futuros da série, descrever apenas o comportamento da série, procurar periodicidades relevantes da série, entre outros. Em diversas áreas da ciência, destaca-se o primeiro enfoque em que, a análise é feita no domínio tempo, principalmente, no método proposto pelos estatísticos George Box e Gwilym Jenkins na década de 70. No segundo enfoque, pouco utilizado na análise de séries temporais de poluentes atmosféricos, a análise é conduzida no domínio da frequência[3].

Um problema frequente em séries temporais é a presença de dados faltantes[4] *(missing data)*, sobretudo na pesquisa ambiental (XIA et al., 1999), geralmente devido a falhas na aquisição de dados (PLAIA; BONDÌ, 2006). São vários os motivos que levam a este cenário, dentre os quais se podem citar: entrada de dados manual, medições feitas de forma incorreta, equipamentos com falhas operacionais e alto custo de coleta de dados (BUUREN; MULLIGEN; BRAND, 1994; LAKSHMINARAYAN; HARP; SAMAD, 1999; FARHANGFAR; KURGAN; PEDRYCZ, 2004; FARHANGFAR; KURGAN; PEDRYCZ, 2007; WU; WUN; CHOU, 2004; COLANTONIO et al., 2010).

É importante frisar que, 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). Na literatura existem diversos métodos para o tratamento dos dados faltantes em séries temporais (HARTLEY; HOCKING, 1971; BEALE; LITTLE, 1975; RUBIN, 1976; DEMPSTER; RUBIN, 1977; LITTLE, 1992; SCHAFER, 1997; JUNGER, 2008; LITTLE; RUBIN, 2014; JUNGER; LEON, 2015). Alguns procedimentos são simples e acabam produzindo estimativas viesadas e outros, mais sofisticados, dependem de

[3] A frequência indica a velocidade com que um fenômeno cíclico se repete. Assim, uma onda de alta frequência passa do pico ao vale ao pico num curto intervalo de tempo, ou curto período. Vale ressaltar, que o período é o inverso da frequência.

[4] Vale dizer que, segundo McKnight et al. (2007), de um modo geral, o termo dados faltantes significa que está faltando algum tipo de informação sobre o fenômeno em que estamos interessados. Normalmente, são observações que deveriam ter sido feitas, mas não foram, por algum motivo.

fortes pressupostos sobre o mecanismo gerador do padrão de dados faltantes e complicadas implementações computacionais. Rubin (1976) classifica dados incompletos de acordo com o mecanismo gerador do padrão de valores faltantes em dados "faltantes completamente ao acaso" ou MCAR (*missing completely at random*), dados "faltantes ao acaso" ou MAR (*missing at random*) e dados "faltantes não ao acaso" ou MNAR (*missing non at random*).

Uma definição mais formal dos mecanismos de dados faltantes pode ser apresentada considerando um conjunto de dados coletado Y, com m linhas, as quais representam os objetos, n colunas, as quais representam os atributos, e com $y_i = (y_{i1}, \cdots, y_{in})$, onde y_{ij} é o valor do atributo j para o objeto i. Dividido o conjunto de dados Y, em duas partes referentes aos dados observados e aos dados faltantes, tal que $Y = (Y_{obs}, Y_{fal})$. É preciso também definir uma matriz R, identificador de dados faltantes, com as mesmas dimensões de Y, em que $r_{ij} = 1$, se y_{ij} é observado, e $r_{ij} = 1$, caso contrário. O mecanismo de dados faltantes é caracterizado pela distribuição condicional de R dado Y, $P(R|Y)$, a qual pode ser do tipo MCAR, MAR ou MNAR (JUNGER, 2008; VERONEZE, 2011; LITTLE; RUBIN, 2014).

A hipótese de MCAR implica que os dados faltantes não dependem dos valores de Y, faltantes ou observados, tem-se $P(R|Y) = P(R)$, isso significa que a causa que levou aos dados faltantes é um evento aleatório. Nesse caso, os valores faltantes para uma variável são uma simples amostra aleatória dos dados dessa variável, ou seja, a distribuição dos valores faltantes é de mesma natureza dos valores observados (GILL et al., 2007; VERONEZE, 2011). Na hipótese de MAR, os dados faltantes não dependem dos valores de Y_{fal} e apenas dos valores de Y_{obs}, ou seja, a $P(R|Y) = P(R|Y_{obs})$. Neste caso, os valores faltantes de uma variável são como uma amostra aleatória simples dos dados para essa variável dentro de subgrupos definidos por valores observados, e a distribuição dos valores faltantes é a mesma que a distribuição dos valores observados dentro de cada subgrupo (ZHANG, 2003; VERONEZE, 2011). Já quando a distribuição de R depende dos dados faltantes contidos na matriz Y (Y_{fal}), podendo também depender dos dados observados, tem-se $P(R|Y) \neq P(R|Y_{obs})$, neste caso temos a hipótese de MNAR (ZHANG, 2003). Esta situação pode acontecer quando a causa dos dados faltantes numa variável é o próprio valor dela. Os mecanismos MCAR e MAR também são chamados de ignoráveis, enquanto o mecanismo MNAR é também chamado de não-ignorável. É importante ressaltar que, segundo McKnight et al. (2007) o termo ignorável não é utilizado para indicar que os pesquisadores podem ser indiferentes aos dados faltantes.

Vale dizer que, os mecanismos ignoráveis são considerados mais fáceis de lidar, pois seus efeitos nos modelos estatísticos estão disponíveis para o analista de dados. Ao contrário, quando o mecanismo é não-ignorável, não existe nenhuma informação dentro do conjunto de dados que permita modelar e compreender a maneira com que os dados faltantes aconteceram (MCKNIGHT et al., 2007; VERONEZE, 2011). Assim, os pressupostos de MAR para o mecanismo gerador dos dados faltantes podem ser mais realistas em estudos com séries temporais de dados de poluentes, já que esse é o cenário mais provável que se encontra no monitoramento de poluentes atmosféricos.

Na literatura existem vários métodos disponíveis para tratar o problema de dados faltantes,

dentre os quais estão os de imputação única, os de máxima verossimilhança (ML - do inglês *Maximum Likelihood*) e os de imputação múltipla. Nos métodos de imputação única, os dados faltantes são substituídos por valores factíveis. O princípio básico dos métodos ML é escolher como estimativa dos parâmetros aqueles valores que, se verdadeiros, maximizariam a probabilidade de observar o que, de fato, foi observado (ALLISON, 1987; VERONEZE, 2011). Os métodos de máxima verossimilhança objetivam a estimação dos parâmetros de modelos vinculados à distribuição dos dados. Já a imputação múltipla consiste em três passos, a saber: *(i)* São obtidos m bancos de dados completos por meio de técnicas adequadas de imputação; *(ii)* Separadamente, os m bancos são analisados por um método estatístico tradicional, como se realmente fossem conjuntos completos de dados; *(iii)* Os m resultados encontrados no passo *ii* são combinados de um jeito simples e apropriado para obter a chamada inferência da imputação repetida (RUBIN, 1976; NUNES; KLÜCK; FACHEL, 2009). Os resultados dessas análises são agregados, gerando estimativas únicas para os parâmetros de interesse. Um parâmetro de interesse é uma informação que se deseja conhecer sobre uma população, como por exemplo, a média populacional.

Entres os trabalhos que desenvolveram metodologias alternativas para a análise de séries temporais com dados faltantes, destacam-se Toda e McKenzie (1999), Junninen et al. (2004), Iglesias, Jorquera e Palma (2006), Plaia e Bondì (2006) e Norazian et al. (2008), utilizam a técnica de imputação para preencher os dados faltantes, já Metaxoglou e Smith (2007), Drake, Knapik e Leśkow (2014) e Junger e Leon (2015) sugerem a utilização de algoritmos do tipo Expectation-Maximization (EM) para tratar esse problema. Porém, esta abordagem tem a desvantagem de assumir uma distribuição específica (em geral, a normal) para os dados.

Alternativamente, uma abordagem promissora, proposta por Parzen (1963), é constituída da utilização de processos de amplitude modulada, onde a análise é realizada por meio de uma série temporal alternativa em que as observações presentes são substituídas por um e as faltantes por zeros. Neste contexto, destacam os trabalhos de Dunsmuir e Robinson (1981b) e Yajima e Nishino (1999) que estudaram o comportamento assintótico de diferentes estimadores da função de autocorrelação de processos de memória curta estacionários com dados faltantes. Baseados na função de densidade espectral assintótica de processos de amplitude modulada, Dunsmuir e Robinson (1981c) propuseram um estimador de Whittle para coeficientes do modelo ARMA e Dunsmuir e Robinson (1981a) estudaram a distribuição assintótica dessa metodologia. Outras referências que tratam da análise de séries temporais com dados faltantes são Bloomfield (1970), Bondon (2005), Ghazal e Elhassanein (2006), com destaque para os trabalhos de Bondon e Bahamonde (2012) que estudam a estimação de modelos autoregressivos condicionalmente heterocedásticos e Efromovich (2014) que proporam uma metodologia não paramétrica de estimação da densidade espectral.

Vale ressaltar que, uma série temporal é composta, em geral, por três componentes não observáveis, a saber: tendência, sazonalidade e aleatoriedade. A sazonalidade é uma componente difícil de ser modelada, pois é necessário compatibilizar a questão física do problema em estudo com a questão estatística. Define-se um fenômeno sazonal como aquele que ocorre re-

gularmente em períodos fixos de tempo (LATORRE; CARDOSO, 2001). Na análise de séries temporais, os modelos ARMA podem ser empregados quando a série em estudo está livre de tendência e de sazonalidade, os modelos ARIMA são utilizados quando há tendência e, para incorporar a componente de sazonalidade, utilizam-se os modelos SARIMA (LATORRE; CARDOSO, 2001; MORETTIN; TOLOI, 2006). Conforme Reisen et al. (2014), os modelos que descrevem de forma adequada o comportamento físico do dados são essenciais para a previsão precisa em qualquer área de aplicação. A sazonalidade é um fenômeno característico de alguns poluentes atmosféricos.

O fato das séries temporais apresentarem muitos ciclos de frequências e amplitudes[5] diferentes, ou seja, séries com propriedades de não estacionariedade e sazonalidade, é uma oportunidade para analisa-las no domínio da frequência. O uso da decomposição de séries temporais via análise espectral, surgiu como alternativa para indentificação das componentes das frequências desses ciclos. O espectro mostra a decomposição da variância de uma amostra de dados através de diferentes frequências. Assim, o espectro descreve as propriedades cíclicas de uma determinada série temporal. Supõe-se que as flutuações do processo subjacente são produzidas por um grande número de ciclos elementares de diferentes frequências, e que a contribuição de cada ciclo é constante em toda a amostra. O espectro então dá a contribuição relativa feita por estes ciclos elementares para a variância do processo global.

Na literatura, a análise espectral de séries temporais ambientais, baseia-se na avaliação de dados de médias horárias, diárias ou mensais como única ferramenta para investigar as periodicidades dos mesmos, conforme apresentado na sub-seção 3.3.3. Porém, o caso destes dados apresentarem margens significativas de dados faltantes ou ter intervalos de amostragem irregulares, ou até mesmo ambos, é talvez o principal motivo para a escassez de trabalhos com aplicação desta técnica a dados de poluição atmosférica com dados faltantes. A transformada rápida de Fourier (FFT) é um meio muito eficiente para calcular a transformada de Fourier discreta, mas uma limitação do algoritmo FFT e dos periodogramas comuns é que requer séries temporais igualmente espaçadas (PRIESTLEY, 1981; DILMAGHANI et al., 2007). Outra limitação do algoritmo FFT é que ele não tolera valores faltantes. Na análise de séries temporais amostradas em intervalos de tempo irregular ou séries com dados faltantes, geralmente é utilizado alguma metodologia para fazer a imputação para preencher os dados faltantes antes de fazer análise espectral via FFT. Isso, no entanto, não é inteiramente satisfatório porque modifica as propriedades estatísticas dos dados através da introdução de dados artificiais (LEROY, 2012).

Para superar essas limitações e, visando propor uma técnica que ofereça maior eficiência nas aplicações empíricas, nesta Tese sugerem-se duas metodologias para estimar a densidade espectral. A primeira foi proposta por (PARZEN, 1963), em que o problema da análise espectral de séries temporais estacionárias com dados faltantes é tratado com base no conceito de amplitude modulada. Se $\{X_t\}_{t\in\mathbb{Z}}$ é um processo estritamente estacionário com observações faltantes, $\{a_t\}$ uma sequência independente de $\{X_t\}$ e $Y_t = a_t X_t$ o processo observado, a estimação

[5] A amplitude refere-se à magnitude da distância vertical entre pico e vale numa função cíclica.

da função de densidade espectral $f_X(\lambda)$ do processo de interesse, $\{X_t\}$, é calculada utilizando observações do processo modulado $\{Y_t\}$.

A segunda metodologia considerada neste trabalho foi introduzida pela primeira vez em astrofísica e ela permite estimar o periodograma da séries temporal não igualmente espaçada, sem utilizar uma técnica de imputação para substituir os valores perdidos: O periodograma Lomb-Scargle. Ao estudar variáveis na Astronomia, Lomb (1976) propôs uma forma de encontrar periodicidades em dados não igualmente espaçados. Na tentativa de encontrar uma alternativa para imputar pseudo-dados em modelos sinusoidais, o autor propôs usar mínimos quadrados para curvas senoidais. Scargle (1982) prolongou o trabalho de Lomb definindo o periodograma Lomb-Scargle e derivando a distribuição nula para ele. Press e Rybicki (1989) propuseram uma formulação matemática prática para o periodograma Lomb-Scargle. Vale ressaltar, que são poucos os trabalhos que aplicaram o periodograma Lomb-Scargle em dados de poluição do ar (DILMAGHANI et al., 2007; HOCKE; KÄMPFER, 2009; DUTTON et al., 2010; BOWDALO; EVANS; SOFEN, 2016), e nos poucos encontrados, nota-se que não foi feito a avaliação da eficiência deste método para diferentes configurações de dados faltantes e propriedade estatísticas das séries temporais sob estudo.

Com base nas questões discutidas acima, os métodos no domínio do tempo e da frequência para computar a função de autocorrelação e densidade espectral de séries temporais com dados faltantes tornam-se o núcleo principal desta tese. Como primeira contribuição, foi avaliado, empiricamente, através de estudo de simulação de Monte Carlo, quatro métodos de estimação da função de autocorrelação de séries temporais univariadas estacionárias com dados faltantes. Na segunda contribuição, propõe-se um método de estimação das funções de matriz de autocorrelação e autocovariância de séries temporais univariadas estacionárias na presença de observações faltantes, a partir do domínio da frequência. A terceira contribuição foi propor dois métodos para estimar o periodograma de séries temporais na presença de dados faltantes. Como contribuições adicionais, as metodologias são aplicadas em dados de MP_{10} coletados nas estações da Rede Automática de Monitoramento da Qualidade do Ar da Grande Vitória, com o objetivo de analisar as séries de concentrações na presença de dados faltantes.

Essas contribuições são apresentadas na seção 5 em três artigos. Uma vez que todos os resultados teóricos são novos e devido a algumas complexidades analíticas, as propriedades assintóticas dos métodos de estimação não estão totalmente estabelecidas aqui e são deixadas para trabalhos futuros. No entanto, as investigações de tamanho de amostra finito, demonstram claramente que, os métodos funcionam muito bem e suportam seu uso em problemas reais.

No primeiro artigo, denominado "*Estimating the autocorrelation function in the presence of missing data: an application to PM_{10} concentrations*", foram apresentados três métodos para estimação da função de autocorrelação de séries temporais univariadas estacionárias na presença de dados faltantes. O primeiro método foi proposto por (PARZEN, 1963), os outros dois foram por (TAKEUCHI, 1995). As propriedades teóricas dos estimadores foram avaliadas e seus desempenhos para amostras finitas investigados através de um estudo de simulação numérica. Por fim, foi proposto a aplicação destas metodologias para avaliar séries temporais

de concentrações de MP_{10} com dados faltantes.

No segundo artigo, intitulado "*Spectral approaches for time series with missing data: an application to air pollution data*" foram propostos dois métodos para estimar a função de densidade espectral de séries temporais estacionárias. Foi estudado o efeito da porcentagem de dados faltantes nos estimadores empregados. Os métodos foram analisados através de simulações e uma aplicação a dados reais de MP_{10} monitorados na RGV também foi considerada.

O terceiro artigo, denominado "*The application of the spectral decomposition theorem to estimate the ACF function of stationary time series with missing data*", é apresentado um método de estimação para as funções de autocorrelação e autocovariância de séries temporais na presença dados faltantes no domínio da frequência. As propriedades assintóticas do método são avaliadas através de estudo de simulação de Monte Carlo para diferentes tamanhos amostrais e porcentagens de dados faltantes.

Por fim, destaca-se que, quatro trabalhos adicionais foram desenvolvidos e serviram de suporte para os três principais artigos desta tese (descritos no parágrafo anterior). Esses trabalhos estão apresentados no Capítulo 8, "Apêndice: estudos adicionais". O primeiro, denominado "Previsão da concentração de material particulado inalável, na Região da Grande Vitória, ES, Brasil, utilizando o modelo SARIMAX", foi publicado na Revista Engenharia Sanitária e Ambiental. O objetivo foi modelar e prever a concentração média diária de material particulado inálavel (MP_{10}), na RGV, Espírito Santo, Brasil, utilizando o modelo SARIMAX, para o período 01/01/2012 a 30/04/2015. Os dados deste estudo foram do tipo séries temporais de concentrações de MP_{10} e de variáveis meteorológicas (velocidade do vento, umidade relativa, precipitação pluvial e temperatura), obtidas junto ao Instituto Estadual de Meio Ambiente e Recursos Hídricos (IEMA), sendo escolhida a estação da Enseada do Suá para fazer o estudo de predição e previsão. Os resultados evidenciaram que em comparação com os modelos ARMA, o desempenho estatístico do modelo SARIMAX foi superior, no que diz respeito à predição de eventos de qualidade do ar regular. Dentre as variáveis meteorológicas avaliadas, a velocidade do vento e a precipitação pluvial foram significativas e melhoraram o ajuste do modelo. Em termos de previsão da qualidade do ar, os modelos de séries temporais mostraram resultados satisfatórios.

O segundo artigo adicional, denominado "Picos de concentração de poluição atmosférica na Região da Grande Vitória, ES, Brasil: uma aplicação da regressão logística", foi publicado na Revista Sociedade & Natureza. Esse trabalho objetivou avaliar os impactos das variáveis meteorológicas temperatura, umidade relativa, velocidade do vento e precipitação na probabilidade de ocorrência de picos/episódios de concentração de poluentes na RGV, Espírito Santo, Brasil, por meio do modelo Logit, para o período de 01/01/2012 a 31/12/2014. Os dados deste estudo foram do tipo séries temporais relativos às concentrações de PTS, MP_{10}, SO2, CO, NO2, O3 e às variáveis meteorológicas (velocidade do vento, umidade relativa, precipitação pluvial e temperatura). Os resultados mostraram que os índices de concentrações na RGV estão associados às mudanças das variáveis meteorológicas. Entre as variáveis meteorológicas avaliadas, a velocidade do vento e a precipitação foram mais significativas na redução da probabilidade de

ocorrência de classificação do ar como "não boa". O modelo Logit mostrou ser uma ferramenta estatística satisfatória para avaliar a qualidade do ar "não boa" da região de estudo.

Um terceiro artigo complementar, denominado *"Identification of periodic components in time series with missing data: an application to air pollution data"*, foi enviando para avaliação no Journal of Applied Statistics. Neste artigo, utilizou-se o periodograma Lomb-Scargle para estimar o espectro de séries temporais com dados faltantes, sem utilizar uma técnica de imputação para substituir os valores perdidos. Ensaios de Monte Carlo foram realizados para comparar a raiz do erro quadrático médio (REQM) do estimador proposto com os do método de estimação tradicional. O estudo empírico evidenciou que o método de estimação sugerido apresenta bom desempenho, em termos da REQM, para diferentes porcentagens de dados faltantes. É demonstrado, por simulações que, no cenário de séries temporais com dados faltantes, a metodologia padrão leva a resultados enganadores enquanto o método proposto não é afetado. A metodologia é aplicada para identificação de periodicidade de uma série de MP_{10} que possui características sazonais e concentrações ocasionais de grandes picos de poluentes.

O quarto artigo adicional, denominado "Análise estatística das concentrações de poluentes atmosféricos na Região da Grande Vitória, ES, Brasil, no período de 2008 a 2017", foi submetido para avaliação na Revista Engenharia Sanitária e Ambiental. Este estudo teve como objetivo avaliar, estatisticamente, os dados de séries temporais de MP_{10} e PTS na RGV, entre 2008 e 2017, verificando se as séries de cada poluente monitoradas em diferentes estações são geradas por um mesmo processo estocástico. Os resultados obtidos apresentam-se como um indicativo da necessidade de reformulação do projeto inicial da RAMQAr que, se somados a um estudo de dispersão de contaminantes, podem garantir a ampliação da área de cobertura da rede, com destaque para a reespacialização das estações já existentes, visando melhorar sua representatividade de dados e instalação de novas estações em locais ainda desprovidos de monitoramento.

Este trabalho está estruturado da seguinte forma: além desta introdução, o Capítulo 2 apresenta os objetivos da pesquisa. O Capítulo 3 destina-se à revisão de literatura. No Capítulo 4 são apresentados os materiais e métodos. O Capítulo 5 refere-se aos principais resultados e discussões, descritos em forma de dois artigos. As conclusões gerais estão descritas no Capítulo 6. O Capítulo 7 destina-se as referências. Por fim, no Capítulo 8, quatro estudos adicionais são apresentados, sendo que os mesmos serviram de suporte para os resultados principais da pesquisa.

2 OBJETIVOS

2.1 OBJETIVO GERAL

Este trabalho teve como objetivo geral propor métodos para a estimação da função de autocorrelação e da densidade espectral de séries temporais univariadas estacionárias na presença de dados faltantes, com aplicação para a poluição do ar, na Região da Grande Vitória, Espírito Santo.

2.2 OBJETIVOS ESPECÍFICOS

De forma específica, pretendeu-se:

a) Propor uma metodologia para estimação da densidade espectral de séries temporais com dados faltantes e avaliar empiricamente, através da simulações de Monte Carlo, as propriedades assintóticas dos estimadores;

b) Propor métodos para estimar as funções de autocovariância e autocorrelação de séries temporais univariadas na presença de dados faltantes a partir da conexão entre domínios de tempo e frequência;

c) Avaliar por meio de simulação numérica os métodos de estimação da função de autocorrelação, de séries temporais univariadas estacionárias na presença de dados faltantes, baseados na abordagem de Amplitude Modulada proposta por (PARZEN, 1963);

d) Aplicar a técnica de análise espectral para analisar e interpretar o comportamento dos dados de PM_{10}, na presença de observações faltantes, medidos nas nove estações da Rede Automática de Monitoramento da Qualidade do Ar da Região da Grande Vitória, ES, Brasil.

3 REVISÃO DE LITERATURA

3.1 POLUIÇÃO ATMOSFÉRICA

A Resolução nº 491, de novembro de 2018, do Conselho Nacional do Meio Ambiente (CONAMA, 2018), caracteriza como poluente atmosférico qualquer forma de matéria em quantidade, concentração, tempo ou outras características, que tornem ou possam tornar o ar: i) impróprio ou nocivo à saúde; ii) inconveniente ao bem-estar público; iii) danoso aos materiais, à fauna e à flora; ou, iv) prejudicial à segurança, ao uso e ao gozo da propriedade e às atividades normais da comunidade.

No que tange às emissões de poluentes, essas podem ser classificadas em antropogênicas e naturais. Quanto às antropogênicas, as mesmas são decorrentes das ações do homem (indústria, transporte, geração de energia e outras). Já as naturais originam-se de processos naturais, como emissões vulcânicas e processos microbiológicos. Além disso, os poluentes podem ser classificados, segundo a sua origem, em primários e secundários. Os primários são lançados, diretamente na atmosfera, pelas fontes de emissão, como por exemplo: SO_2, CO, NO_X e hidrocarbonetos (HC). Já os secundários formam-se na atmosfera por meio da reação química entre poluentes primários ou desses com constituintes naturais da atmosfera. Pode-se citar como exemplo: O_3 e peridóxido de hidrogênio (H_2O_2).

Conforme Dallarosa (2005), os poluentes atmosféricos podem ser divididos de forma genérica em três grupos de substâncias: sólidos, líquidos e gasosos. Entretanto, em função da grande interação que ocorre entre essas fases, pode-se restringí-los a dois grupos: particulados e gases. De acordo com Godish (1991), particulados e gases são a principal forma de ocorrência de poluentes na atmosfera.

A seguir encontra-se uma breve descrição dos principais poluentes atmosféricos.

- Dióxido de enxofre (SO_2): gás tóxico e incolor, pode ser emitido por fontes naturais ou por fontes antropogênicas e pode reagir com outros compostos na atmosfera, formando material particulado de diâmetro reduzido;

- Dióxido de nitrogênio (NO_2): gás poluente altamente oxidante, sendo que sua presença na atmosfera é fator preponderante na formação do ozônio troposférico;

- Hidrocarbonetos (HC): compostos formados de carbono e hidrogênio e que podem se apresentar na forma de gases, partículas finas ou gotas;

- Material particulado (MP): mistura complexa de sólidos com diâmetro reduzido. Seus componentes apresentam características físicas e químicas variadas. Geralmente, o material particulado é classificado de acordo com o diâmetro das partículas, em função da relação existente entre diâmetro e possibilidade de penetração no trato respiratório;

- Monóxido de carbono (CO): gás inodoro e incolor, formado no processo de queima de combustíveis;

- Ozônio (O_3): poluente secundário, formado a partir de outros poluentes atmosféricos, e altamente oxidante na troposfera (camada inferior da atmosfera).

A Resolução CONAMA nº 491/2018 CONAMA (2018) define padrão de qualidade do ar como:

- um dos instrumentos de gestão da qualidade do ar, determinado como valor de concentração de um poluente específico na atmosfera, associado a um intervalo de tempo de exposição, para que o meio ambiente e a saúde da população sejam preservados em relação aos riscos de danos causados pela poluição atmosférica;

- padrões de qualidade do ar intermediários - PI: padrões estabelecidos como valores temporários a serem cumpridos em etapas;

- padrão de qualidade do ar final - PF: valores guia definidos pela Organização Mundial da Saúde - OMS em 2005.

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

Os poluentes e seus padrões de qualidade do ar fixados pela Resolução CONAMA nº 491/2018 são apresentados na Tabela 1. Os parâmetros regulamentados são: Material Particulado (MP_{10}): partículas de material sólido ou líquido suspensas no ar, na forma de poeira, neblina, aerossol, fuligem, entre outros, com diâmetro aerodinâmico equivalente de corte de 10 micrômetros; Material Particulado ($MP_{2,5}$): partículas de material sólido ou líquido suspensas no ar, na forma de poeira, neblina, aerossol, fuligem, entre outros, com diâmetro aerodinâmico equivalente de corte de 2,5 micrômetros; Partículas Totais em Suspensão (PTS): partículas de material sólido ou líquido suspensas no ar, na forma de poeira, neblina, aerossol, fuligem, entre outros, com diâmetro aerodinâmico equivalente de corte de 50 micrômetros; Dióxido de enxofre (SO_2); Dióxido de nitrogênio (NO_2); Ozônio (O_3); Fumaça; e, Monóxido de carbono (CO). As concentrações-padrão são expressas em ppm ou $\mu g/m^3$.

O Governo do Estado do Espírito Santo, por meio do Decreto nº 3463-R, de 16 de dezembro de 2013 (GOVERNO DO ESTADO DO ESPÍRITO SANTO, 2013), estabeleceu os padrões estaduais de qualidade do ar, em que, além dos padrões já descritos e estabelecidos na Resolução CONAMA n° 491/2018 (excessão feita à fumaça), estão incluídas as partículas sedimentadas (poeira sedimentada). Além disso, o decreto inseriu o conceito de Metas Intermediárias (MI), que são estabelecidas como valores temporários a serem cumpridos em etapas, visando à melhoria gradativa da qualidade do ar, e Padrões Finais (PF), que representam os alvos de longo

prazo. No mais, foram criadas três MI que levam ao gradual atendimento dos padrões finais, estabelecidos com base nas diretrizes da OMS para os poluentes de interesse investigados por essa organização, uma estratégia semelhante à adotada pelo estado de São Paulo, em abril de 2013 (para mais detalhes ver Instituto Estadual de Meio Ambiente e Recursos Hídricos do Estado do Espírito Santo (2014)) (MONTE, 2016).

Vale frisar que, em relação aos padrões de qualidade do ar, existe uma diferença entre os valores estabelecidos pela Resolução CONAMA nº 491, em vigor (PI-1), e as diretrizes da OMS, uma vez que as diretrizes da OMS (revisadas em 2005) levam em conta os diversos estudos científicos realizados a partir de 1990. Na Tabela 1 são apresentados os padrões nacionais e estaduais de qualidade do ar e as diretrizes da OMS. É importante ressaltar que, as diretrizes da OMS 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). 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.

3.2 POLUIÇÃO ATMOSFÉRICA E 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 temporário, estão relacionados à exposição a altas concentrações e os seus efeitos são imediatos.

- **Crônicos:** os efeitos são de caráter permanente, estão relacionados à exposição a baixas concentrações de poluentes e são de longo prazo.

Segundo Holgate et al. (1999), um nível elevado dos poluentes pode ocasionar desde irritação dos olhos, nariz e garganta, bronquite e pneumonia, até doenças respiratórias crônicas, câncer de pulmão, problemas cardíacos, entre outros. Conforme Cançado et al. (2006), isso ocorre,

> pois a poluição do ar causa uma resposta inflamatória no aparelho respiratório induzida pela ação de substâncias oxidantes, as quais acarretam aumento da produção, da acidez, da viscosidade e da consistência do muco produzido pelas vias aéreas, levando, consequentemente, à diminuição da resposta e/ou eficácia do sistema mucociliar. O aumento da poluição do ar tem sido associado ao aumento da viscosidade sanguínea, de marcadores inflamatórios (proteína C reativa, fibrinogênio) e da progressão de arteriosclerose, a alterações da coagulação, à redução da variabilidade da frequência

Tabela 1: Padrões nacionais e estaduais de qualidade do ar e diretrizes da OMS

			$MP_{2.5}$ [µg/m^3]	MP_{10} [µg/m^3]	PTS [µg/m^3]	PS [g/m^2. 30 dias]	SO_2 [µg/m^3]	NO_2 [µg/m^3]	O_3 [µg/m^3]	CO [ppm]	FUMAÇA [µg/m^3]	Pb[a] [µg/m^3]
Padrões de Qualidade do Ar (CONAMA n° 491/2018)	Curta exposição	PI-1	60 (24h)	120 (24h)	-	-	125 (24h)	260 (1h)[b]	140 (8h)[c]	-	120 (24h)	-
		PI-2	50 (24h)	100 (24h)	-	-	50 (24h)	240 (1h)[b]	130 (8h)[c]	-	100 (24h)	-
		PI-3	37 (24h)	75 (24h)	-	-	30 (24h)	220 (1h)[b]	120 (8h)[c]	-	75 (24h)	-
		PF	25 (24h)	50 (24h)	240 (24h)	-	20 (24h)	200 (1h)[b]	100 (8h)[c]	09 (8h)[c]	50 (24h)	-
	Longa exposição	PI-1	20 (ano)[d]	40 (ano)[d]	-	-	40 (ano)[d]	60 (ano)[d]	-	-	40 (ano)[d]	-
		PI-2	17 (ano)[d]	35 (ano)[d]	-	-	30 (ano)[d]	50 (ano)[d]	-	-	35 (ano)[d]	-
		PI-3	15 (ano)[d]	30 (ano)[d]	-	-	20 (ano)[b]	45 (ano)[d]	-	-	30 (ano)[d]	-
		PF	10 (ano)[d]	20 (ano)[d]	80 (ano)[e]	-	-	40 (ano)[d]	-	-	20 (ano)[d]	0,5 (ano)[d]
Metas e padrão estadual (Decreto n° 3463-R 2013)	Curta exposição	MI1-ES	-	120 (24h)	180 (24h)	14	60 (24h)	240 (1h)	140 (8h)	-	-	-
		MI2-ES	50 (24h)	80 (24h)	170 (24h)	-	40 (24h)	220 (1h)	120 (8h)	-	-	-
		MI3-ES	37 (24h)	60 (24h)	160 (24h)	-	30 (24h)	210 (1h)	110 (8h)	-	-	-
		PF-ES	25 (24h)	50 (24h)	150 (24h)	-	20 (24h)	200 (1h)	100 (8h)	10.000 (8h) 30.000 (1h)	-	-
	Longa exposição	MI1-ES	-	45 (ano)[d]	65 (ano)[e]	-	40 (ano)[d]	50 (ano)[d]	-	-	-	-
		MI2-ES	20 (ano)[d]	33 (ano)[d]	63 (ano)[e]	-	30 (ano)[d]	45 (ano)[d]	-	-	-	-
		MI3-ES	15 (ano)[d]	25 (ano)[d]	62 (ano)[e]	-	20 (ano)[d]	42 (ano)[d]	-	-	-	-
		PF-Es	10 (ano)[d]	20 (ano)[d]	60 (ano)[e]	-	-	40 (ano)[d]	-	-	-	
Diretriz OMS	Curta exposição		25 (24h)	50 (24h)	-	-	20 (24h) 500 (10min)	200 (1h)	100 (8h)	10.000 (8h) 30.000 (1h)	-	-
	Longa exposição		10 (ano)[d]	20 (ano)[d]	-	-	-	40 (ano)[d]	-	-	-	-

Nota: 1) O tempo de média considerado para o cálculo da concentração do poluente está indicado entre parênteses; e, 2) [a] Medido nas partículas totais em suspensão; [b] Média horária; [c] Máxima média móvel obtida no dia; [d] Média aritmética anual; [e] Média geométrica anual.

Fonte: Iema (2014), OMS (2005), CONAMA (2018).

> cardíaca (indicador de risco para arritmia e morte súbita), à vasoconstricção e ao aumento da pressão arterial, todos fatores de risco para doenças cardiovasculares.

Diversos estudos epidemiológicos têm demonstrado associações significativas entre a exposição às concentrações elevadas de poluentes atmosféricos e a problemas de saúde (OSTRO et al., 1995; OSTRO et al., 1996a; OSTRO et al., 1996b; OSTRO et al., 1999; OSTRO; HURLEY; LIPSETT, 1999; MARTINS et al., 2002; GOUVEIA et al., 2003; NASCIMENTO et al., 2006; BRAGA et al., 2007; SOUZA et al., 2014; ESTRELLA et al., 2018; LIU et al., 2018; TIBUAKUU et al., 2018). Autores como Brunekreef e Holgate (2002), Maynard (2004), World Health Organization (2005), Liu et al. (2018), Tibuakuu et al. (2018), entre outros, demonstraram a relação entre os poluentes legislados (MP_{10}, CO, SO_2, NO_X e O_3) e os problemas de saúde. No ano de 2012, por exemplo, a morte de 4,3 milhões pessoas foi atribuída à poluição atmosférica (WORLD HEALTH ORGANIZATION, 2014).

Pope III et al. (2002) avaliaram a relação entre a exposição a longo prazo à poluição do ar por partículas finas e todas as causas de morte, câncer de pulmão e mortalidade cardiopulmonar. Os resultados evidenciaram que a poluição relacionada a partículas finas e óxido de enxofre foi associada a todas as causas, câncer de pulmão e mortalidade cardiopulmonar. Cada aumento de 10 $\mu g/m^3$ na poluição do ar por partículas finas foi associado a um aumento de, aproximadamente, 4%, 6% e 8% do risco de mortalidade por todas as causas, cardiopulmonar e câncer de pulmão, respectivamente. Medidas da fração de partículas grossas e partículas totais em suspensão não foram consistentemente associadas com a mortalidade. Os autores concluíram que a exposição a longo prazo à poluição do ar por partículas finas relacionada à combustão é um importante fator de risco ambiental para a mortalidade por causa cardiopulmonar e por câncer de pulmão. Estudos semelhantes recentes, também evidenciam a associação entre os poluentes atmosféricos, a mortalidade e o agravamento de doenças respiratórias (HOEK et al., 2013; XIONG et al., 2015; TO et al., 2016) e cardiovasculares (CHANG; CHEN; YANG, 2015; BRAVO et al., 2016; ZÚÑIGA et al., 2016).

Martins et al. (2002) estudaram os efeitos causados pela poluição atmosférica na morbidade por gripe e por pneumonia em idosos no período de 1996 a 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 MP_{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 (Universidade de São Paulo). Utilizou-se o modelo aditivo generalizado de regressão de Poisson para verificar a relação entre a pneumonia, gripe e poluição atmosférica, tendo como 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 atendimentos por pneumonia e gripe em idosos, in-

dicando que a poluição atmosférica promove efeitos adversos para a saúde de idosos.

Braga et al. (2007) avaliaram os efeitos agudos do MP_{10} sobre os atendimentos em pronto-socorro por doenças cardiovasculares e respiratórias no Município de Itabira, MG. Os resultados evidenciam que elevações de 10 $\mu g/m^3$ de MP_{10} foram associadas 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, houve um efeito agudo, principalmente para os indivíduos com idade entre 45 e 64 anos.

Castro et al. (2009) realizaram um estudo na cidade do Rio de Janeiro em uma amostra aleatória de 118 escolares, com idade entre 6 e 15 anos, da rede pública, residentes até dois quilômetros 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 utilizaram dados dos poluentes MP_{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 meteorológicas 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 aceitáveis, a poluição atmosférica, principalmente por MP_{10} e NO_2, esteve associada à diminuição da função respiratória no Rio de Janeiro.

Negrete et al. (2010) avaliaram os efeitos da exposição à poluição do ar sobre as internações hospitalares por Insuficiência Cardíaca Congestiva (ICC) de adultos e idosos na cidade de Santo André, São Paulo, para o período de 2000 a 2007. Conforme os resultados obtidos pelos autores, para um aumento de 24,6 $\mu g/m^3$ de MP_{10}, foi observado um aumento de 3,0% nas internações por ICC no mesmo dia da exposição.

Na Região da Grande Vitória (RGV), Souza et al. (2014) realizaram um estudo, cujo objetivo foi investigar a associação entre concentrações dos poluentes atmosféricos e atendimentos diários por causas respiratórias em crianças. Foram analisadas as contagens diárias de admissões hospitalares de crianças menores de seis anos e as concentrações diárias dos poluentes atmosféricos MP_{10}, SO_2, NO_2, O_3 e CO, de janeiro de 2005 a dezembro de 2010. Os autores combinaram duas técnicas para a análise estatística: modelo de regressão de Poisson em modelos aditivos generalizados e análise de componentes principais. Os resultados mostraram que, o aumento de 10,49 $\mu g/m^3$ (intervalo interquartílico) nos níveis do poluente MP_{10}, levou a um aumento de 3,0% do valor do risco relativo estimado por meio do modelo aditivo generalizado, enquanto no modelo aditivo generalizado usual a estimativa foi de 2,0%. Ainda, segundo os resultados, existe uma relação significativa entre os níveis de concentração dos poluentes e o número de atendimentos hospitalares em crianças menores de seis anos, mesmo em um ambiente com níveis abaixo dos padrões recomendados pelo CONAMA e pela OMS.

Freitas et al. (2016) analisaram o impacto da poluição atmosférica na morbidade respiratória e cardiovascular de crianças e adultos em Vitória, ES. Foi realizado um estudo utilizando modelos de séries temporais via regressão de Poisson a partir de dados de hospitalizações e poluentes

em Vitória, de 2001 a 2006. Foram testadas como variáveis independentes o MP_{10}, o SO_2 e o O_3, em defasagem simples e acumulada até cinco dias. Introduziram-se temperatura, umidade e variáveis indicadoras dos dias da semana e feriados da cidade como variáveis de controle nos modelos. Os autores observaram que, para cada incremento de 10 $\mu g/m^3$ dos poluentes MP_{10}, SO_2 e O_3, houve aumentos no risco relativo percentual (RR%) para as hospitalizações por doenças respiratórias totais de 9,67; 6,98 e 1,93, respectivamente.

Nascimento et al. (2017) em um estudo realizado na RGV, objetivaram analisar a associação entre a concentração de material particulado fino na atmosfera e atendimento hospitalar por doenças respiratórias agudas em crianças. Os autores avaliaram dados de contagem diária de atendimentos ambulatoriais e hospitalizações por doenças respiratórias (CID-10) em crianças de zero a 12 anos em três hospitais da RGV. Para a coleta de material particulado fino, foram utilizados amostradores portáteis de partículas instalados em seis locais na região estudada. O Modelo Aditivo Generalizado com distribuição de Poisson, ajustado para efeitos das covariáveis preditoras, foi utilizado para avaliar a relação entre os desfechos respiratórios e a concentração de material particulado fino. De acordo com os resultados, foi identificada associação positiva entre atendimentos ambulatoriais e hospitalizações de crianças com até 12 anos devido à doenças respiratórias agudas e a concentração de material particulado fino na atmosfera.

Calderón-Garcidueñas et al. (2016) realizaram um estudo de revisão para examinar o impacto da poluição do ar no desenvolvimento do cérebro de crianças e as consequências clínicas, cognitivas, estruturais, cerebrais e metabólicas. Consequências potenciais a longo prazo para os cérebros dos adultos e os efeitos sobre a esclerose múltipla (EM) também foram discutidos. Segundo os autores, a neuroinflamação difusa, o dano à unidade neurovascular e a produção de autoanticorpos para as proteínas neurais e de junção estreita, são achados preocupantes em crianças cronicamente expostas a concentrações acima dos padrões atuais de O_3 e $MP_{2,5}$ e, podem constituir fatores de risco significativos para o desenvolvimento da doença de Alzheimer mais tarde na vida.

Bowe et al. (2017) avaliaram a relação entre a poluição atmosférica e saúde renal. Os autores usaram modelos de sobrevivência para avaliar a associação entre concentrações de MP_{10}, NO_2 e CO e risco de incidência da taxa de filtração glomerular estimada (eTFG) inferior a 60 mL/min por 1,73 m^2 , doença renal crônica incidente, declínio de eGFR de 30% ou mais e doença renal terminal. No estudo, os autores trataram a exposição como variável no tempo quando ela era atualizada anualmente e quando os participantes da coorte se mudavam. Segundo os resultados, a exposição a concentrações mais altas de MP_{10} , NO_2 e CO está associada ao aumento do risco de doença renal crônica incidente, declínio de eGFR e doença renal terminal.

Wang et al. (2019) investigaram as diferenças urbano-rurais e sexuais nos riscos de aumento de dez cânceres mais comuns na China, relacionados à alta concentração de $MP_{2,5}$ no lado sudeste da linhagem de Hu. De acordo com os resultados do trabalho, os riscos aumentados de câncer de pulmão, ovário, próstata e leucemia estão intimamente associados ao aumento da exposição ao $MP_{2,5}$, além disso, o $MP_{2,5}$ afeta, significativamente, os riscos de câncer de pulmão e leucemia na área rural. Segundo os autores, os resultados demonstram que, os maiores

riscos para o câncer de pulmão e leucemia com o aumento da exposição ao $MP_{2,5}$, são mais significativos para o sexo feminino.

No mais, estudos recentes comprovaram ainda os efeitos da exposição à poluição atmosférica por material particulado na performance cognitiva humana (ZHANG; CHEN; ZHANG, 2018) e, também, a associação entre o aumento do risco de incidência de diabetes melitus e a exposição ao ar poluído por material particulado (BOWE et al., 2018). Raz et al. (2014) exploraram a associação entre a exposição materna à poluição do ar por $MP_{2,5}$ e as chances do transtorno do espectro do autismo em seu filho. Os autores concluíram que maior exposição materna ao $MP_{2,5}$ durante a gravidez, particularmente no terceiro trimestre, foi associada a maiores chances de uma criança ter transtorno do espectro do autismo. Segundo Béjot et al. (2018), a poluição do ar deve ser considerada como um novo fator de risco cerebrovascular e neurodegenerativo modificável. Esse enorme problema de saúde pública mundial requer políticas de saúde ambiental capazes de reduzir a poluição do ar e, portanto, o ônus do AVC 9Acidente Vascular Cerebral) e da demência.

Os principais poluentes atmosféricos, bem como o resumo de seus efeitos sobre saúde humana e o meio ambiente, são descritos na Tabela 2

Tabela 2: Principais poluentes regulamentados pela Resolução CONAMA n° 491 de 19/11/2018 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 MP_{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 (MP_{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 incidência 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, indústrias, usinas térmicas que utilizam óleo ou gás e incineradores.	Dentre os NO_X, 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, a 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 ESTADO DA ARTE SOBRE O USO DE TÉCNICAS ESTATÍSTICAS NA POLUIÇÃO ATMOSFÉRICA

O aumento dos níveis de poluição atmosférica ocorrido nos últimos anos tem feito com que, cada vez mais, os pesquisadores voltem suas atenções para essa problemática, que afeta a população com um todo. 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 e, o presente estudo se concentrou na classe de modelagem estocástica. Uma vez que o objetivo desta pesquisa foi avaliar a técnica de análise espectral de séries temporais com dados faltantes e sazonalidade, para aplicação em dados de poluentes atmosféricos, esta seção visa descrever alguns estudos que aplicaram técnicas estatísticas para a análise de séries temporais de concentrações de poluentes atmosféricos, metodologias para o tratamento de dados faltantes e análise no domínio da frequência (análise espectral) da poluição atmosférica[6].

3.3.1 Estado da arte sobre o uso de técnicas estatísticas para análise de séries temporais de poluição atmosférica

Jorquera et al. (1998) realizaram previsões para o nível máximo de concentração de O_3 diário, na cidade de Santigo, Chile, utilizando modelos de séries temporais (ARMAX), de redes neurais e o modelo fuzzy. Rao e Zurbenko (1994) propuseram um método estatístico para filtrar ou moderar a influência das flutuações meteorológicas nas concentrações de O_3. O uso desta técnica para avaliar tendências na qualidade do ar ambiente do O_3 foi demonstrado com dados de O_3 de um local de monitoramento em Nova Jersey. Já Liu e Johnson (2002) fizeram previsões para picos diários de concentração de O_3, em Milwaukee, Estados Unidos, no período de 1987 a 1993, por meio do modelo de regressão com erros de séries temporais (RTSE), através da inclusão de variáveis exógenas, o que os autores denominaram de principal componente (PC) com gatilho. Liu e Johnson (2003) estudaram os picos diários de concentração de O_3, para Milwaukee, EUA (Estados Unidos da América), no período de 1999 a 2002, utilizando o PC com gatilho na abordagem de Box-Jenkins com RTSE.

Chaloulakou et al. (2003) mediram, no centro de Atenas, Grécia, durante o período de 1 de junho de 1999 a 31 de maio de 2000, as concentrações de $MP_{2,5}$ e MP_{10}. O conjunto de dados foi utilizado para estabelecer níveis base de concentração de $MP_{2,5}$ e MP_{10}, que poderiam ser usados no futuro para avaliar a eficácia das estratégias implementadas de controle de emissões, comparar os níveis de concentração de $MP_{2,5}$ e MP_{10} observados com os padrões de partículas

[6]Cabe mencionar que a técnica de análise espectral não foi desenvolvida recentemente, porém, salvo engano, foram encontrados poucos trabalhos com aplicação dessa técnica em estudos sobre poluição atmosférica.

ambientais da União Europeia e dos EUA, dentre outros. Por fim, os autores utilizaram os dados coletados para investigar a relação entre material particulado e parâmetros meteorológicos, como velocidade e direção do vento local e, desenvolver um modelo de regressão para prever a média diária de MP_{10} em Atenas. Os resultados do estudo sublinharam a importância das fontes de emissão locais, principalmente do tráfego, que são responsáveis pelos elevados níveis de concentração de $MP_{2,5}$ e MP_{10} observados durante este período de amostragem.

Agirre-Basurko, Ibarra-Berastegi e Madariaga (2006) utilizaram três modelos, um de regressão linear múltipla e dois modelos de redes neurais, 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 a 1994. Como saída prevista para os modelos, adotou-se as concentrações de O_3 e NO_2, com horizonte de previsão de oito horas à frente. Os resultados mostraram que os modelos de redes neurais obtiveram resultados melhores para a previsão das concentrações de O_3 e NO_2 quando comparados ao modelo de regressão linear múltipla. Quanto ao desempenho dos modelos de redes neurais, o que mais se destacou foi o modelo que considerou a sazonalidade das séries de concentrações de O_3 e NO_2.

Goyal, Chan e Jaiswal (2006) realizaram um estudo com três modelos estatísticos aplicados à média diária de concentração de MP_{10}, medido nas cidades de Delhi e Hong Kong. O trabalho objetivou desenvolver um modelo estatístico de previsão das concentrações de MP10 e promover um estudo comparativo através do desempenho dos modelos, a saber: i) modelo de regressão linear múltipla (modelo 1); ii) modelo de séries temporais ARIMA (modelo 2); e iii) combinação entre os modelos 1 e 2 (modelo 3). Além do MP_{10}, alguns parâmetros meteorológicos foram adotados, como a velocidade do vento, a temperatura, a radiação solar e a umidade relativa do ar, medidos no período de junho de 2000 a junho de 2001. 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 a junho de 2002. O modelo 3 foi utilizado, e os resultados da previsão foram satisfatórios.

Sousa et al. (2006) avaliaram o desempenho de três métodos estatísticos, a saber: modelo de séries temporais, regressão linear múltipla e modelos de redes neurais artificiais de feedforward, para prever as concentrações médias diárias de O_3 na cidade do Porto, em Portugal. Os estudos foram realizados para o ano de 2002 e seus respectivos quatro trimestres, separadamente. Os resultados evidenciaram que os melhores índices de desempenho foram obtidos com modelos de redes neurais artificiais feedforward na etapa de validação. Os autores concluíram que, os modelos de redes neurais artificial feedforward foram mais eficientes para prever as concentrações de O_3 no local de estudo.

Arain et al. (2007) desenvolveram uma metodologia para incluir os efeitos do fluxo de vento em modelos de regressão de uso da terra (LUR) para prever as concentrações de NO_2 para estudos de exposição em saúde. Segundo os autores, o NO_2 é amplamente utilizado em estudos

de saúde como um indicador da poluição do ar gerada pelo tráfego em áreas urbanas. Os resultados apontaram que a inclusão da direção do vento observada, interpolada de alta resolução de uma rede de 38 estações meteorológicas, em um modelo LUR, melhorou as estimativas de concentração de NO_2 em áreas densamente povoadas, de alto tráfego e industriais/comerciais na área urbana de Toronto-Hamilton (THUA) de Ontário, Canadá. O estudo também demonstrou que os campos de vento podem ser integrados na estrutura de regressão do uso da terra. Tal integração teve uma influência perceptível na previsão geral do modelo e, talvez mais importante, na avaliação dos efeitos na saúde sobre a distribuição espacial relativa da poluição do tráfego em todo o THUA. Os autores sugeriram que a metodologia desenvolvida no estudo pode ser aplicada em outras grandes áreas urbanas em todo o mundo.

Vardoulakis e Kassomenos (2008) aplicaram a técnica de análise de correlação e regressão para avaliar dados de contrações de MP_{10} monitorados entre 2001 e 2003 em Atenas (Grécia) e Birmingham (Reino Unido). Os dados de MP_{10} de Atenas e Birmingham foram analisados quanto às relações com outros poluentes (NO_X , CO, O_3 e SO_2) e parâmetros meteorológicos (velocidade do vento, temperatura, umidade relativa , precipitação, radiação solar e pressão atmosférica). Correlações positivas significativas entre MP_{10} e NO_X , CO e radiação solar foram observadas nos locais de monitoramento selecionados durante as estações frias. Por outro lado, correlações negativas entre MP_{10} e O_3, velocidade do vento e precipitação foram observadas durante as mesmas estações. No entanto, segundo os autores, essas correlações tornaram-se mais fracas durante as estações quentes, devido à formação de aerossóis secundários e maior ressuspensão de poeira do solo. Além disso, os autores empregaram a técnica de análise de componentes principais e a de regressão para quantificar a contribuição das fontes de não-combustão para os níveis de MP_{10} observados. Os resultados evidenciaram que esta contribuição variou entre 45% e 70% em Birmingham e, entre 41% e 74% em Atenas. Por fim, foi verificado que o transporte de longo alcance de partículas da Europa continental teve um efeito marcante sobre os níveis de MP_{10} em Birmingham, enquanto o clima local teve uma influência mais forte sobre os níveis de MP_{10} em Atenas.

Pires et al. (2008) aplicaram a técnica de análise de componentes principais para identificar locais similares de poluição atmosférica e as fontes emissoras na região metropolitana de Porto, Portugal. Pires et al. (2009) realizaram um estudo em Portugal com o objetivo de mostrar como a análise de componentes principais pode ser usada para identificar medições redundantes em redes de monitoramento da qualidade do ar. Segundo os autores, o bom desempenho obtido pelos modelos mostrou que os locais de monitoramento selecionados pelo procedimento apresentado no estudo foram suficientes para inferir as concentrações de poluentes atmosféricos na região definida pelos locais iniciais de monitoramento. Além disso, os analisadores de poluentes do ar correspondentes às medições redundantes podem ser instalados em regiões não monitoradas, permitindo a ampliação da rede de monitoramento da qualidade do ar.

Gomes (2009) realizou um estudo de previsão de índices de qualidade do ar da RGV, ES, Brasil, utilizando o modelo autorregressivo de valores inteiros INAR(p). O período de análise foi de 01/01/07 a 19/03/07, sendo as previsões datadas de 20/03/07 a 25/03/07. Os poluentes inves-

tigados foram CO, NO_X, SO_2 e O_3. Para a escolha do modelo mais adequado o autor utilizou o critério de seleção automática para modelos INAR(p) e o AICCINAR, 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, conforme a Resolução CONAMA 03/1990 (CONAMA, 1990). Porém, baseados nas diretrizes da OMS (WORLD HEALTH ORGANIZATION, 2005), a previsão do poluente SO_2, para o dia 20/03/07, estação do Centro de Vila Velha, excedeu o valor de 20 $\mu g/m^3$ para média de 24 horas, ou seja, mesmo estando dentro o limite do padrão nacional,na época, essa concentração é prejudicial para a saúde humana.

Liu (2009) realizaram simulações para as concentrações de MP_{10}, na cidade de Ta-Liao, China. O autor adotou o modelo de regressão com erros de séries temporais (RTSE), incluindo uma variável explicativa resultante da análise de componentes principais (ACP) para completar a simulação de MP_{10} (denominada de "CP trigger"). Segundo o autor, as variáveis O_3, temperatura do ponto de orvalho, NO_X, direção do vento e ACP foram significantes nos modelos RTSE na maior parte do tempo. Os resultados demonstraram que as predições são melhores quando da presença da ACP. Estudos semelhantes foram realizados por Liu (2007) e Liu et al. (2013), no contexto de regressão múltipla e dos modelos Box-Jenkins (BOX; JENKINS; REINSEL, 2008) de séries temporais.

Lyra, Oda-Souza e Viola (2011) ajustaram dois modelos de regressão linear múltipla à concentração média de 24 h de MP_{10}. Os dados de concentração de MP_{10} e os dados dos elementos meteorológicos (temperatura e umidade do ar, precipitação pluvial, velocidade do vento e pressão atmosférica) foram monitorados em São Cristóvão, na cidade do Rio de Janeiro, entre 01/05/02 e 31/08/03. Os resultados apontaram que, dentre os elementos meteorológicos avaliados, a umidade relativa do ar e a precipitação pluvial explicam a maior parte da variabilidade do MP_{10} na cidade do Rio de Janeiro. Quando considerada a concentração do material particulado do dia anterior nas avaliações, esta variável é a mais significativa para a variação do material particulado do dia. Os autores concluíram que, em termos de previsão da qualidade do ar, os modelos mostram resultados satisfatórios, e podem ser utilizados operacionalmente.

Leite et al. (2011) realizaram um estudo cujo objetivo foi analisar a qualidade do ar atmosférico de Uberlândia, em Minas Gerais, por meio de modelos de regressão logística simples. Foram utilizados os dados de MP_{10} do período de 2003 a 2008. As variáveis preditoras referentes ao clima (umidade relativa, velocidade do vento, precipitação diária e temperatura média) e ao fluxo de veículos foram utilizadas de forma contínua no modelo logístico. Já as variáveis relacionadas ao dia da semana e a estação do ano foram codificadas de forma binária. Os autores concluíram que existe relação significativa de atributos climáticos e variáveis temporais com a qualidade do ar de Uberlândia. Além disso, modelos logísticos simples podem ser usados para calcular probabilidades de se obter qualidade do ar considerada boa.

Paschalidou et al. (2011) empregaram dois tipos de modelos de redes neurais artificiais (NN) utilizando as técnicas multicamadas perceptron (MLP) e radial basis function (RBF), bem como um modelo baseado na análise de regressão de componentes principais (PCRA), para prever as concentrações horárias de MP_{10} em quatro áreas urbanas (Larnaca, Limassol, Nicosia e

Paphos)do Chipre. Segundo os autores, o desenvolvimento do modelo foi baseado em uma variedade de parâmetros meteorológicos e poluentes correspondentes ao período de dois anos, entre julho de 2006 e junho de 2008, e a avaliação do modelo foi obtida através do uso de uma série de instrumentos e metodologias de avaliação bem estabelecidos. Os resultados da avaliação revelaram que os modelos MLP NN exibem o melhor desempenho de previsão com valores do coeficiente de correção variando entre 0,65 e 0,76, enquanto que os modelos NN RBF e PCRA revelam um desempenho bastante fraco, com valores do coeficiente de correlação entre 0,37-0,43 e 0,33-0,38, respectivamente. Os autores concluíram que, os modelos avaliados no estudo podem fornecer às autoridades locais previsões confiáveis sobre a qualidade do ar.

Gripa et al. (2012) compararam dois modelos, um de séries temporais e um de regressão linear múltipla, para modelagem e previsão das concentrações médias de MP_{10}, monitoradas na RGV, ES, com a incorporação de fatores meteorológicos. Ambos os modelos evidenciaram resultados semelhantes. No entanto, o modelo de regressão apresentou medidas de previsão das concentrações médias de MP_{10} um pouco melhores do que a do modelo de séries temporais.

Jian et al. (2012) utilizaram o modelo ARIMA para investigar o efeito de fatores meteorológicos nas concentrações de partículas submicrométricas (UFP) e partículas com diâmetro menor ou igual a 1,0 micrometro ($MP_{1,0}$), sob condições de tráfego intenso em Hangzhou, China, uma cidade com um rápido aumento de frota de veículos rodoviários. Os resultados do modelo ARIMA indicaram que a pressão barométrica e a velocidade do vento foram anti-correlacionadas enquanto que a temperatura e a umidade relativa correlacionaram-se positivamente com as concentrações do número de UFP e de massa do $MP_{1,0}$. De acordo com os resultados, os fatores meteorológicos pressão barométrica, velocidade do vento, temperatura e umidade relativa foram preditores significativos para a concentração do número de UFP e de massa do $MP_{1,0}$. Segundo os autores, os modelos desenvolvidos no trabalho podem ser úteis em futuros estudos de larga escala na China.

Arhami, Kamali e Rajabi (2013) apontam que progressos recentes no desenvolvimento de metamodelos de redes neurais artificiais abriram o caminho para o uso confiável desses modelos na previsão de concentrações de poluentes atmosféricos na atmosfera urbana. Os autores desenvolveram uma estudo na cidade de Teerã, Irã, em que empregaram as redes neurais artificiais para construir predições dos poluentes NO_X, NO_2, NO, O_3, CO e MP_{10}, com base nos dados coletados em uma estação de monitoramento na área central, densamente povoada, da cidade. Os autores utilizaram como variáveis explicativas a velocidade do vento, a temperatura do ar, a umidade relativa e a direção do vento. A natureza complexa das condições da fonte de poluentes foi refletida através do uso de hora do dia e mês do ano, como variáveis de entrada, e o desenvolvimento de diferentes modelos para cada dia da semana. Os resultados mostraram que modelos de redes neurais artificiais podem ser usados como metamodelos confiáveis para a previsão de poluentes atmosféricos horários em ambientes urbanos.

Kanaroglou et al. (2013) desenvolveram um modelo de regressão do uso da terra para as

concentrações de poluição do ar com SO_2. Foram utilizados no estudo dados de monitoramento móvel coletados em Hamilton, Ontário, Canadá, entre 2005 e 2010. As concentrações observadas de SO_2 foram regredidas em relação a um conjunto abrangente de variáveis de uso e transporte da terra. Os autores usaram um modelo de regressão de mínimos quadrados ordinários e um modelo autorregressivo simultâneo de erro espacial. De acordo com os resultados, o uso do modelo autorregressivo espacial foi eficaz para remover a autocorrelação espacial que ocorreu nos resíduos do modelo de regressão de mínimos quadrados ordinários. Estudos semelhantes foram conduzidos por Jerrett et al. (2005), Hoek et al. (2008), Atari et al. (2008), Wheeler et al. (2008), Adamkiewicz et al. (2010), Allen et al. (2011) e Gulliver et al. (2011). Esta técnica é muito utilizada pois, segundo os autores, o objetivo é desenvolver um modelo capaz de prever as concentrações de poluentes em locais não monitorados. Segundo Kanaroglou et al. (2013), a modelagem de regressão do uso do solo é uma ferramenta poderosa para atribuir a distribuição espacial das concentrações de poluentes do ar. Uma vez que o modelo é desenvolvido, ele pode ser aplicado para estimar as concentrações de poluição do ar em locais dentro da área de estudo.

Reisen et al. (2014) modelaram a média diária de concentração de material particulado inalável, na cidade de Cariacica, Espírito Santo, Brasil, utilizando um processo integrado fracionado sazonal, com volatilidade. Segundo os autores, as estimativas fracionárias evidenciaram que a série é estacionária na média e apresentou o fenômeno de memória longa a longo prazo e, também, nos períodos sazonais. Uma propriedade de variância não-constante também foi encontrada nos dados. Neste contexto, foi considerado o modelo SARFIMA (modelo autoregressivo sazonal fracionário integrado de média móvel) com inovações do tipo GARCH para a modelagem de previsão de MP_{10}. Os autores concluíram que o modelo ajustado capturou bem a dinâmica da série. Os intervalos de previsão fora da amostra foram melhorados considerando-se os erros heteroscedásticos e foram capazes de capturar os períodos de maior volatilidade.

Monte, Albuquerque e Reisen (2015) realizaram um estudo para estimar e prever a concentração horária de O_3 na RGV, Espírito Santo, Brasil, utilizando um modelo ARMAX-GARCH, no período 01/01/2011 a 31/12/2011. O estudo utilizou dados cedidos pelo Instituto Estadual de Meio Ambiente e Recursos Hídricos (IEMA). De acordo com os resultados apresentados, as previsões horárias para o dia 31/12/2011 foram muito próximas dos valores observados. Observou-se também que as estimativas seguiram a trajetória diária da concentração de O_3. Os autores concluíram que o modelo ARMAX-GARCH é mais eficaz na predição de episódios de poluição de O_3, em comparação aos modelos ARMA e ARMAX.

Monte, Albuquerque e Reisen (2016) objetivaram verificar os impactos das variáveis meteorológicas temperatura, umidade relativa, velocidade do vento e precipitação sobre a qualidade do ar, na RGV, Espírito Santo, Brasil, considerando o poluente MP_{10}, por meio do modelo Logit. O período de estudo foi de janeiro de 2005 a dezembro de 2010, onde a qualidade do ar foi classificada como "não boa" e "boa". Também foram estimados os efeitos dos dias da semana e das estações do ano sobre a probabilidade de ocorrência de qualidade do ar "não boa". Os autores concluíram que fatores meteorológicos como a precipitação pluviométrica e a veloci-

dade do vento contribuíram significantemente para a redução da probabilidade de ocorrência de qualidade do ar "não boa". Observaram também que, nos finais de semana, quando a produção industrial diminui, reduz-se os serviços logísticos e o fluxo de carros é menor, a chance de ocorrer qualidade do ar "não boa" é fortemente reduzida, quando comparado aos dias de semana. Além disso, notou-se que nas estações do outono e do inverno, a probabilidade de se verificar qualidade do ar "não boa" caiu de maneira relevante, sendo que na primavera e no verão notou-se uma elevação desta probabilidade.

Monte, Albuquerque e Reisen (2017) utilizaram e a metodologia vetorial autorregressiva (VAR) e o teste de causalidade de Granger para verificar as inter-relações entre as concentrações de O_3 e NO_2 na RGV, Espírito Santo, Brasil. Segundo os autores, os resultados revelaram que as concentrações de O_3 e de NO_2 da Região (estação) de Laranjeiras foram as menos afetadas por concentrações de outras estações. Devido à localização, as concentrações de O_3 e NO_2 da Enseada do Suá tiveram significativa influência de outras regiões, especialmente de Jardim Camburi, Ibes e Vitória - Centro. A concentração de O_3 na região do Ibes foi fortemente influenciada pelas concentrações de O_3 e de NO_2 da Enseada do Suá. Além disso, as concentrações em Cariacica sofreram impactos relevantes das concentrações da Enseada do Suá, provavelmente, devido à direção do vento Norte/Nordeste, predominante na RGV.

Sarnaglia, Monroy e Vitória (2018) consideraram a modelagem e previsão de concentrações máximas diárias de O_3 por hora em Laranjeiras, Serra, Brasil, através de modelos de regressão dinâmica. A fim de levar em conta a assimetria natural e o rigor excessivo dos dados, os autores utilizaram um modelo de regressão linear com erros autorregressivos e inovações após um membro da família da mistura de escala de distribuições de desvio-normal. Poluentes e variáveis meteorológicas foram considerados como preditores, juntamente com alguns fatores determinísticos, isto é, dias da semana e estações do ano. Segundos os resultados do trabalho, o modelo estimado foi capaz de explicar satisfatoriamente a estrutura de correlação da série temporal do O_3.

Reisen et al. (2018) avaliaram uma metodologia para a estimativa de modelos sazonais de séries temporais de memória longa na presença de "outliers". O trabalho objetivou propor um estimador semiparamétrico robusto para os parâmetros do modelo SARFIMA, através do uso de um periodograma robusto. Os autores demostraram, através de simulações, que a metodologia robusta se comporta como a clássica para estimar os parâmetros de memória longa, se não houver outliers. Por outro lado, no cenário contaminado (presença de outliers), a metodologia padrão leva a resultados enganosos, enquanto o método proposto não é afetado. A metodologia foi aplicada para modelar e prever concentrações de SO_2 na RGV.

Reisen et al. (2019) consideraram a modelagem de fatores para séries temporais de alta dimensão contaminadas por outliers aditivos. O estimador do número de fatores foi obtido por uma análise própria de uma matriz de covariância robusta definida não-negativa. Os autores realizaram estudo de Monte Carlo para analisar o desempenho do estimador robusto do número de fatores sob os cenários de séries temporais multivariadas, com e sem outliers aditivos. Por fim, foi realizado uma aplicação na qual a análise fatorial robusta foi empregada para reduzir

a dimensionalidade dos dados de MP_{10} e, portanto, identificar o comportamento poluidor do poluente em questão.

3.3.2 Estado da arte sobre o uso de metodologias para tratar dados faltantes em séries temporais de poluição atmosférica

Um dos primeiros estudos sobre dados faltantes em séries de concentrações de poluentes atmosféricos foi realizado por Davison e Hemphill (1987). Os autores realizaram um estudo sobre a análise estatística de dados de O_3 quando faltavam medições. No trabalho, dois métodos são propostos para lidar com observações faltantes que, segundo os autores, frequentemente apresentam problemas na análise estatística dos dados ambientais do O_3. O primeiro baseia-se na superação dos dados em relação aos limites e fornece uma classe flexível e geral de modelos para análise estatística de dados de poluição do ar. O segundo usa os valores medidos de variáveis relacionadas para imputar os dados faltantes. Na época, os autores aplicaram os métodos propostos para realizar avaliação de dados monitorados no Texas, EUA.

Junninen et al. (2004) avaliaram dois contextos de imputação para aplicação em dados de concentrações de poluentes. No estudo realizado, os dados foram avaliados no contexto de análise univariada (linear, *spline* e interpolação pelo vizinho mais próximo) e multivariada (regressão baseados em imputação, vizinho mais próximo, auto-organização de mapa e *multi-layer perceptron*). 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 foi 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 consistiram em séries de concentrações de NO_X, NO_2, O_3, MP_{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 desempenho também depende da variável em estudo. Os resultados obtidos com os métodos multivariados evidenciaram que tanto a auto-organização de mapa quanto o *multi-layer perceptron* apresentam um desempenho um pouco melhor que o método vizinho mais próximo. A vantagem do método auto-organização de mapa sobre os demais, é que ele depende menos da localização real dos dados faltantes, enquanto que as vantagens dos métodos vizinho mais próximo 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, um significativo aumento em performances pode ser alcançado pela hibridização dos métodos multivariados e que, a forma como esta combinação deve ser feita depende da variável sob estudo.

Iglesias, Jorquera e Palma (2006) propuseram uma metodologia estatística para manipular regressões com longa dependência nos erros e dados faltantes. A estratégia de estimação foi desenvolvida através das abordagens 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. A fim de explicar a variação nas concentrações de MP_{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 correlacionadas com a concentração de MP_{10}. De acordo com os resultados apresentados, 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 Bondì (2006) propuseram uma metodologia de imputação de dados faltantes a 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 MP_{10} medidos em oito estações de monitoramento distribuídas 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.

Plaia e Bondì (2010) realizaram um estudo cujo o objetivo foi propor um novo método de imputação de regressão (único) que, considerando a estrutura e as características particulares do conjunto de dados, cria um conjunto de dados "completo" que pode ser analisado por qualquer pesquisador em diferentes ocasiões e usando diferentes técnicas. Os autores utilizaram um conjunto de dados de MP_{10} registrados em Palermo em 2003 para simular situações de dados faltantes, a fim de avaliar o desempenho do método de imputação por meio de indicadores de desempenho. Brown, Harris e Cox (2013) discutiram as consequências da falta de dados durante as atividades de monitoramento da qualidade do ar e o cálculo da concentração média anual de massa de poluentes ambientais. Além disso, os autores realizaram uma descrição matemática de um método para a determinação da concentração média anual usando a média simples.

Gómez-Carracedo et al. (2014) utilizaram um conjuntos de dados de qualidade do ar (NO, NO_2, NO_X, CO, O_3, MP_{10}, $MP_{2,5}$ e $MP_{1,0}$) de uma estação de imersão automática situada numa zona costeira suburbana próxima da cidade da Corunha, Espanha, com taxas de dados faltantes variando de 4% a 24%, para verificar se grandes diferenças ocorrem quando distintos métodos de imputação (Média Incondicional, Mediana modificada, Baseado em componentes principais, Expectation maximization (EM) e Imputação múltipla) atualmente utilizados, são aplicados para o preenchimento dos dados faltantes. Todos os métodos foram executados de forma semelhante, embora a imputação múltipla tenha gerado valores imputados mais dispersos. De acordo com os resultados apresentados, as principais diferenças ocorreram quando uma variável com valores ausentes correlacionou-se mal com as outras características e quando uma variável apresentou cargas relevantes em vários fatores não rotacionados, o que, algumas vezes alterou a ordem dos fatores rotacionados. Os melhores resultados foram obtidos com o algoritmo EM.

Zainuri, Jemain e Muda (2015) apresentaram vários métodos de imputação para dados de

qualidade do ar, especificamente na Malásia. Os autores objetivaram selecionar o melhor método de imputação e comparar se havia alguma diferença nos métodos utilizados entre as estações de monitoramento na Malásia. Foram simulados, de forma aleatória, dados faltantes para vários casos com 5, 10, 15, 20, 25 e 30% de informações perdidas. Os métodos utilizados no trabalho para imputação foram a substituição média e mediana, algoritmo EM, decomposição em valores singulares (SVD), método K-vizinhos mais próximos (KNN) e método sequencial K-vizinho mais próximo (SKNN). O desempenho das imputações foi comparado usando os indicadores de desempenho coeficiente de correlação (R), índice de concordância (d) e erro absoluto médio (MAE). Com base nos resultados obtidos, os autores concluíram que EM, KNN e SKNN são os três melhores métodos.

Junger e Leon (2015) discutiram questões teóricas e metodológicas para imputação de dados faltantes em séries temporais multivariadas de poluentes atmosféricos. Os autores apresentaram um método baseado em imputação que usa o algoritmo EM sob a suposição de distribuição normal. Diferentes abordagens foram consideradas para a filtragem da componente temporal. Foi realizado um estudo de simulação para avaliar a validade e o desempenho do método proposto em comparação com alguns métodos frequentemente utilizados. As simulações mostraram que, quando a quantidade de dados faltantes era de apenas 5%, a análise completa dos dados produziu resultados satisfatórios. independentemente do mecanismo gerador dos dados ausentes, enquanto a validade começou a degenerar quando a proporção de valores faltantes excedia 10%. Segundo os resultados, o método de imputação proposto pelos autores apresentou boa acurácia e precisão em diferentes contextos com relação aos padrões de observações ausentes. A maioria das imputações obteve resultados válidos, mesmo sob ausência não aleatória. Os métodos propostos no trabalho foram implementados como um pacote denominado *multivariate time series data imputation* (*mtsdi*) no software estatístico R.

Zakaria e Noor (2018) realizaram uma estudo sobre métodos de imputação para preenchimento de dados faltantes em séries de concentrações de poluentes atmosféricos na Malásia. No estudo, os dados de monitoramento horário de CO, O_3, MP_{10}, SO_2, NO_X, NO_2, temperatura ambiente e umidade foram utilizadas para avaliar quatro métodos de imputação (*Mean Top Bottom*, Regressão Linear, Imputação Múltipla e Vizinho Mais Próximo). As observações dos poluentes do ar foram simuladas com quatro percentagens de dados perdidos, isto é, 5%, 10%, 15% e 20%. No trabalho, foram utilizadas medidas de desempenho para descrever a adequação do ajuste dos métodos de imputação, a saber: o Erro Absoluto Médio, a Raiz do Erro Quadrático Médio, o Coeficiente de Determinação e o Índice de Concordância. A partir dos resultados das medidas de desempenho, os autores concluíram que, o método *Mean Top Bottom* de imputação foi o mais adequado para o preenchimento dos valores faltantes nos dados de poluentes atmosféricos. Por outro lado, a Regressão Linear foi o método que apresentou os piores resultados para imputação dos dados.

Miller et al. (2018) avaliaram quatro métodos de imputação de dados faltantes para abordar os dados de concentração de BTEX (benzeno, tolueno, etilbenzeno e xilenos), medidos em Windsor e Sarnia, Ontário, Canadá, no outono de 2005. Os métodos de imputação avaliados

foram: imputação do valor médio, ponderação inversa da distância, proporções interespécies e regressão. As concentrações e relações entre espécies foram geralmente similares entre as duas cidades. Usando essas cidades industrializadas como estudos de caso, os autores demonstraram que as técnicas de proporções interespécies ou regressão dos dados para os quais há informação completa, junto com uma concentração medida (i.e. benzeno) para prever resultados de concentrações perdidas (i.e. TEX), que existe uma boa concordância entre os valores previstos e medidos. Os autores apontam que, na ausência de quaisquer concentrações conhecidas, o método ponderação inversa da distância pode fornecer uma concordância razoável entre as concentrações observadas e estimadas para as espécies BTEX, e foi superior à imputação de valor médio que não foi capaz de preservar a tendência espacial.

3.3.3 Estado da arte sobre o uso da análise espectral de séries temporais de poluição atmosférica

Rao, Samson e Peddada (1976) realizaram uma comparação entre as características espectrais da concentração de SO_2 e a velocidade do vento no domínio da frequência (análise espectral) em Long Island, Nova Iorque. Os resultados evidenciaram que, durante o inverno, os dois espectros têm um pico dominante correspondente à escala de tempo sinótico, indicando que as variações climáticas sinóticas são responsáveis pelas oscilações de longo período do poluente. Schlink, Herbarth e Tetzlaff (1997) desenvolveram um método de previsão de séries temporais para permitir o aviso antecipado do smog no inverno. Os autores construíram, com base na análise espectral, um modelo de componente para a série temporal de concentrações de SO_2 usando essencialmente um algoritmo recursivo de Kalman. Salcedo et al. (1999) utilizaram as técnicas da análise espectral e de séries temporais para a identificação de variações a longo prazo na média (tendência) e de componentes cíclicos ou periódicos de níveis ambientais de poluição atmosférica da cidade de Porto em Portugal. Já Kelsall, Zeger e Samet (1999), motivados por um estudo da associação entre as contagens da mortalidade diária e a poluição do ar, apresentaram uma abordagem de estimação, no domínio da frequência, para modelos log-lineares que respondem tanto pela superdispersão quanto pela autocorrelação. Os autores aplicaram os métodos para estimar a associação entre as contagens de mortalidade e a concentração de partículas transportadas pelo ar na Filadélfia, EUA, para os anos de 1974-1988.

Hies et al. (2000) apresentaram um método, utilizando análise espectral, para analisar séries temporais de carbono elementar para diferentes fontes de poluição atmosférica em Berlim. O estudo compreendeu o período de 01 de abril de 1994 a 31 de março de 1995. As séries temporais de médias diárias de carbono elementar em vários locais urbanos foram avaliadas com os respectivos espectros. Periodicidades típicas, e bem conhecidas, causadas por influências antropogênicas e meteorológicas foram identificadas usando espectros de coerência e fase. Os autores mostraram que, o aquecimento doméstico pela combustão do carvão aparece com uma periodicidade de 365 dias, o tráfego contribui com picos de 3,5; 4,6 e 7 dias no espectro e o carbono elementar de longo alcance elevado pode ser identificado como picos característicos com

periodicidades na gama de 13 a 42 dias. Como as amplitudes relativas das várias influências variam dependendo da localização do ponto de medição na área urbana, segundo os autores, o uso de espectros estimados ajuda a encontrar a influência do tráfego, aquecimento doméstico do carvão e transporte de longo alcance na concentração de carbono elementar.

Sebald et al. (2000) utilizaram a análise espectral para investigar os processos de formação e decomposição do ozônio troposférico. O conjunto de dados considerado foi o das concentrações médias horárias de O_3, NO_X e parâmetros meteorológicos, como temperatura, velocidade do vento, umidade relativa do ar, máximo diurno da temperatura do ponto de orvalho e pressão barométrica. Os parâmetros foram medidos continuamente de 1993 a 1995 nas redes monitoramento da qualidade do ar e de parâmetros meteorológicos do Deutscher Wetterdienst (Serviço Meteorológico Alemão). Para detectar a razão para concentrações extremamente altas de O_3 no verão, os autores dividiram os dados em duas componentes, a saber: componente sazonal de baixa frequência e de alta frequência. A segunda componente foi avaliada utilizando o espectro de densidade de potência correspondente. Mostrou-se que as variações meteorológicas de grande escala e de escala sinótica afetam as concentrações de O_3 em todos os locais de monitoramento. Os autores concluíram que, no conjunto de dados de 1993-1995 esta influência contribuiu para picos com durações de ciclo de 16-35 dias no espectro de densidade de potência estimado a partir dos dados de O_3.

Marr e Harley (2002) aplicaram a análise espectral para descrever a história e a distribuição espacial das diferenças de dia da semana nas concentrações de O_3 ambiente, NO_X e COV, através da análise de medições realizadas por duas décadas, de locais localizados em toda a Califórnia. Segundo os autores, a análise espectral das séries temporais de concentração mostrou que padrões semanais nas concentrações de O_3, tipicamente com valores mais altos nos finais de semana, se tornaram mais difundidos na Califórnia entre 1980 e 1999. Em contraste, um forte padrão semanal das concentrações de NO_X estão presentes durante todo o período, e padrões semanais em concentrações de COV, embora não tão evidentes, também estiveram presentes durante todo o período de 20 anos.

Fuentes (2002) apresentaram novas abordagens e ferramentas espectrais para estimar a estrutura espacial de um processo não estacionário. Mais especificamente, os autores propuseram uma abordagem para a análise espectral de processos espaciais não-estacionários que é baseada no conceito de espectros espaciais, isto é, funções espectrais que são dependentes do espaço. Esta noção de espectro espacial generaliza a definição de espectros para processos estacionários, e sob certas condições, o espectro espacial em cada local pode ser estimado a partir de uma única realização do processo espacial. Segundo os autores, a motivação para realização do trabalho foi a modelagem e previsão de concentrações de O_3 ao longo de diferentes fronteiras geopolíticas, nos EUA, para avaliação da conformidade com os padrões de qualidade do ar ambiente.

Cvitaš et al. (2004) analisaram, fazendo uso da aplicação da transformada de Fourier, os dados obtidos durante muitos anos de monitorização contínua de O_3 em 12 estações selecionadas da rede EUROTRAC-TOR (Eureka environmental project: European experiment on Transport

and transformation of environmentally relevant trace Constituents in the troposphere over Europe, subproject Tropospheric Ozone Research). A média das transformações ao longo dos anos, as funções de autocorrelação e transformadas de Fourier foram calculadas para cada local. Os resultados confirmaram a existência de uma variação comum nas frações de volume de O_3 com períodos quase que variando entre 7 e 44 dias. Segundo os autores, estas frequências estão provavelmente relacionadas com influências meteorológicas da escala sinótica cíclica-quase cíclica.

Kandlikar (2007) utilizou métodos espectrais para analisar mudanças na qualidade do ar em um único local de monitoramento em Déli, Índia, de 2000 a 2006. Cálculos de densidade espectral de potência de dados de concentração diária de MP_{10}, CO, NO_X e SO_X revelam a presença de tendências e oscilações periódicas para todos os poluentes. Os autores aplicaram a Análise do Espectro Singular (SSA) para decompor os dados diários em tendências não lineares, estatisticamente significativas, ciclos sazonais e outras oscilações. Períodos de reduções acentuadas foram observados para as concentrações de SO_X e CO em 2001 e 2002, respectivamente. Foi observado que a queda acentuada, tanto na tendência quanto na amplitude do ciclo sazonal de CO, coincide com a mudança para o Gás Natural Comprimido como combustível na frota de transporte público de Nova Déli. Mudanças observadas nas concentrações de SO_X e MP_{10} foram provavelmente causadas por fontes não relacionadas ao tráfego de veículos.

Choi et al. (2008) investigaram a dependência regional antropogênica na variação semanal de poluentes do ar e sua relação com as condições meteorológicas sobre a China para os verões de 2001-2005. A análise espectral foi aplicada às observações diárias locais de concentrações de MP_{10} e precipitação de 31 estações terrestres, estimativas de reanálise de variáveis atmosféricas regionais e captação de nuvens por satélite. Os resultados confirmam a presença da interação entre MP_{10} e as condições meteorológicas na camada limite, e sugerem uma possível ligação da formação de nuvens à concentração MP_{10} em uma escala semanal. Um estudo semelhante foi realizado por Kai et al. (2008) que usaram a análise espectral e mais dois métodos para investigar propriedades de escala de tempo em séries temporais de índices de poluição do ar (SO_2, NO_2 e MP_{10}) em Xangai, na China.

Tchepel e Borrego (2010) utilizaram a análise espectral para compreender os processos físicos subjacentes e a influência das fontes de emissão na variabilidade das concentrações de poluentes atmosféricos na região metropolitana do Porto (Portugal). Foram utilizadas séries temporais anuais de poluentes relacionados ao tráfego (CO e MP_{10}). Utilizaram-se, como métodos, o periodograma e a análise espectral bivariada para identificar as contribuições das flutuações de curto prazo (períodos de 12 e 24 h) nas concentrações de CO e MP_{10}. Os autores concluíram que as análises realizadas revelaram a influência distinta das emissões de tráfego local e de longo alcance e que a metodologia mostrou ser uma ferramenta poderosa para a análise das causas da poluição do ar.

Tchepel et al. (2010) propuseram uma abordagem para a estimativa de concentrações de fundo usando dados medidos de qualidade do ar decompostos em componentes de linha de base e de curto prazo da área urbana de Lisboa (Portugal). Para isso, os autores utilizaram a

densidade espectral para dados de monitoramento da qualidade do ar com base na análise da série de Fourier. Depois, as flutuações de curto prazo, associadas à influência das emissões locais e as condições de dispersão, foram extraídas das medições originais usando um filtro de média móvel iterativo e levando em conta a contribuição de frequências mais altas determinadas a partir da análise espectral. No trabalho, os autores aplicaram esta metodologia para definir as concentrações de fundo de material particulado (MP_{10}) usadas como dados de entrada para um modelo CFD de escala local, e comparada com uma abordagem alternativa usando concentrações de fundo fornecidas por um sistema de modelagem de qualidade do ar de mesoescala. Os resultados apresentaram um melhor desempenho para o modelo em microescala quando inicializado por séries temporais decompostas, o que demonstrou a importância da metodologia proposta na redução da incerteza das previsões do modelo. Conforme apontou o estudo, a decomposição das medições de qualidade do ar e a remoção das flutuações de curto prazo discutidas no trabalho, é uma técnica valiosa para determinar concentrações de fundo representativas.

Kumar e Ridder (2010) estimaram um modelo de heterocedasticidade condicional autorregressivo generalizado (GARCH) associado com o método FFT-ARIMA (transformada rápida de Fourier-autorregressivo integrado de média móvel), para prever os episódios de concentração de O_3 em duas cidades europeias, Bruxelas e Londres. Segundo os autores, a FFT tem sido usada para modelar as periodicidades (a periodicidade anual é especialmente distinta) exibidas pela série de tempo. Os resultados revelaram que modelar a concentração de O_3 por meio do modelo GARCH, além de melhorar os intervalos de confiança das previsões de curto prazo, também proporcionou maior acurácia na probabilidade de previsão de episódios críticos de concentração de O_3.

He e Lu (2012) realizaram uma investigação sobre a variação periódica dos níveis de poluentes em uma interseção de tráfego típica de Hong Kong. Foram realizadas medições de CO, CO_2 e MP. Os dados medidos mostraram variações periódicas com os intervalos do sinal de trânsito. A abordagem da densidade espectral de potência (PSD) foi usada para inspecionar as tendências e as oscilações periódicas dos poluentes medidos. A análise do espectro singular foi aplicada para decompor os dados medidos em tendências e oscilações não lineares estatisticamente significativas no processo. A partir dos resultados, a maioria das tendências apresentaram inclinação a aumentar devido à hora do rush próxima, durante o experimento. Além disso, segundo os autores, todas as oscilações mudaram regularmente com um período de 136 segundos, o que é coincidente com o período do sinal de tráfego e a frequência calculada usando o PSD.

Iordache e Dunea (2013) realizaram um estudo na Romênia com o objetivo de fazer a aplicação da análise de espectro cruzado (CSA) para revelar as correlações entre séries de poluentes atmosféricos e séries meteorológicas em diferentes frequências. Os dados foram coletados nas cidades de Brasov, Ploiesti e Târgoviste no período de novembro de 2011 a março de 2012, com uma hora de amostragem. Todos os valores de amplitude cruzada computados foram interpretados como uma medida de covariância entre os respectivos componentes de frequência nas duas séries, selecionando cinco picos mais altos com os correspondentes períodos, que foram filtra-

dos e classificados de forma decrescente. Os resultados mostraram uma contribuição importante das flutuações de curto prazo (períodos de 12 e 24 horas) para a variância total dos poluentes analisados (NO_2, SO_2 e partículas suspensas), com dependência significativa do fator meteorológico (radiação incidente, temperatura do ar ou umidade relativa) e menos nas condições do local. Segundo os autores, esse tipo de análise de dados pode auxiliar na identificação de constantes ou padrões, caracterizando as interações entre variáveis, o que é uma ferramenta útil para parametrização e calibração de modelos de poluição do ar.

Fajardo et al. (2018) com objetivo de estudar séries temporais com memória longa, sugeriram a utilização de um periodograma alternativo, denominado M-periodograma, que é obtido relacionando o periodograma a um problema de regressão e usando um M-estimador para os coeficientes do modelo de regressão. Segundo os autores, além de ser um periodograma alternativo atraente para séries temporais de memória longa, ele também é resistente a valores discrepantes aditivos. O desempenho de robustez do estimador foi investigado através de simulação de Monte Carlo. Como uma aplicação prática, os autores investigaram o efeito de observações atípicas em dados de poluição do ar, MP_{10}. Os autores apontam que, além da importância de modelar e prever este poluente, a séries de MP_{10} apresentaram, em geral, características interessantes, como sazonalidade, assimetria, e também altos níveis de poluição, que podem ser considerados como observações atípicas no contexto do trabalho.

Vale dizer que, salvo engano, foram encontrados poucos trabalhos que aplicaram as técnicas de análise espectral para avaliar séries temporais de poluentes atmosféricos com dados faltantes. Entre os trabalhos encontrados, destacam-se os de (DILMAGHANI et al., 2007; HOCKE; KÄMPFER, 2009; DUTTON et al., 2010; MOSHENBERG; LERNER; FISHBAIN, 2015; BOWDALO; EVANS; SOFEN, 2016). Dilmaghani et al. (2007), Hocke e Kämpfer (2009), Dutton et al. (2010) e Bowdalo, Evans e Sofen (2016) utilizaram o periodograma de Lomb-Scargle para identificar periodicidades em dados de poluição atmosférica com dados faltantes. Já Moshenberg, Lerner e Fishbain (2015) apresentaram o teorema de amostragem discreta para a tarefa de atribuir dados faltantes em séries temporais longitudinais de qualidade do ar. Dentro do contexto do teorema de amostragem discreta, dois esquemas espectrais para preenchimento de valores faltantes são apresentados - um método baseado em Transformação de Cosseno Discreta (DCT) e outro na Decomposição Variável Única em Cluster (K-SVD). Os autores concluíram que, os métodos espectrais são uma ótima opção para a imputação de dados de qualidade do ar, que deve ser considerada, especialmente, quando os padrões de dados ausentes são desconhecidos.

As diversas aplicações que podem ser feitas com a junção das técnicas clássicas de análise de séries temporais no domínio do tempo e análise espectral, fazem com que esta área de estudo seja umas das mais dinâmicas na avaliação de dados de concentrações de poluentes atmosféricos. Dessa forma, justifica-se a continuidade dos estudos relativos ao tema. Além disso, na maioria dos estudos que utilizam a análise espectral para avaliar dados de poluição atmosférica, os dados faltantes são preenchidos utilizando técnicas de imputação única ou múltipla, para posterior análise. Segundo Junger (2008), a principal desvantagem dos proce-

dimentos baseados em imputação é que em sua maioria a imprecisão devida à imputação não é contemplada na análise e, portanto, a variância dos estimadores é subestimada. São poucos os trabalhos que propuseram estimadores para a função de densidade espectral na presença de dados faltantes, sendo esse o foco principal desta pesquisa. Do mais, por se tratar de um assunto relativamente recente dentro da poluição do ar, a exploração tende a trazer novos resultados.

4 MATERIAIS E MÉTODOS

4.1 REGIÃO DE ESTUDO E REDE DE MONITORAMENTO

A área de estudo compreenderá a RGV, Espírito Santo, Brasil, localizada na costa sul do oceano Atlântico [latitude 20°19 S (Sul), longitude 40°20 W (Oeste)]. A RGV é constituída pelos municípios de Vitória, Vila Velha, Cariacica, Serra e Viana. Por estar situada na região litorânea, a RGV apresenta clima tropical quente (Aw), de acordo com a classificação climática de Köppen (KÖPPEN, 1900), possuindo inverno ameno e seco, e verão chuvoso e quente. As temperaturas médias variam entre 24° C (Celsius) e 30° C.

No ano de 2010, a população do Espírito Santo era de 3.514.952 (INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA, 2014). Deste total 1.687.704 estava residindo na região metropolitana do estado, que é composta pelos municípios da Grande Vitória, mais Fundão e Guarapari. Tomando-se somente a RGV, a população chegou a 1.565.393, o que representa cerca de 45% da população capixaba. Segundo o Instituto Brasileiro de Geografia e Estatística (INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA, 2014) a RGV abrange uma área de 1.461 Km^2, sendo um dos principais polos de desenvolvimento urbano e industrial do estado. A região sofre com diversos tipos de problemas ambientais, entre os quais está à deterioração da qualidade do ar, devido às emissões atmosféricas por indústrias, pela frota veicular e ressuspensão do solo causada pelo vento e tráfego veicular.

Vale ressaltar que a RGV possui uma Rede Automática de Monitoramento da Qualidade do Ar (RAMQAR) inaugurada em julho de 2000, de propriedade do IEMA. Esta rede é distribuída em nove estações distribuídas por quatro municípios da RGV, a saber: o município Serra com três estações localizadas nas regiões de Laranjeiras, Carapina e Cidade Continental; 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 1.

Na Tabela 3, são apresentadas os códigos de identificação das estações junto ao IEMA, seus respectivos códigos de identificação neste estudo (ID), a localização das estações de acordo com os bairros em que estão alocadas, o ano de início da operação de cada estação e suas coordenadas planas (UTM).

A RAMQAR monitora os seguintes poluentes: Partículas Totais em Suspensão (PTS); Material Particulado Inalável (MP_{10}); Material Particulado Respirável ($MP_{2,5}$), Dióxido de Enxofre (SO_2); Ozônio (O_3); Óxidos de Nitrogênio (NO_X); Dióxido de Nitrogênio (NO_2); Monóxido de Carbono (CO) e Hidrocarbonetos totais (HCT), Hidrocarbonetos não Metano (HCnM), Metano (CH_4), Monóxido de Nitrogênio (NO). E, ainda, realiza-se o monitoramento dos seguintes parâmetros meteorológicos: Direção escalar do vento (DV); Velocidade escalar do vento (VV); Precipitação pluviométrica (PP); Umidade relativa do ar (UR); Precipitação Pluviométrica (PP);

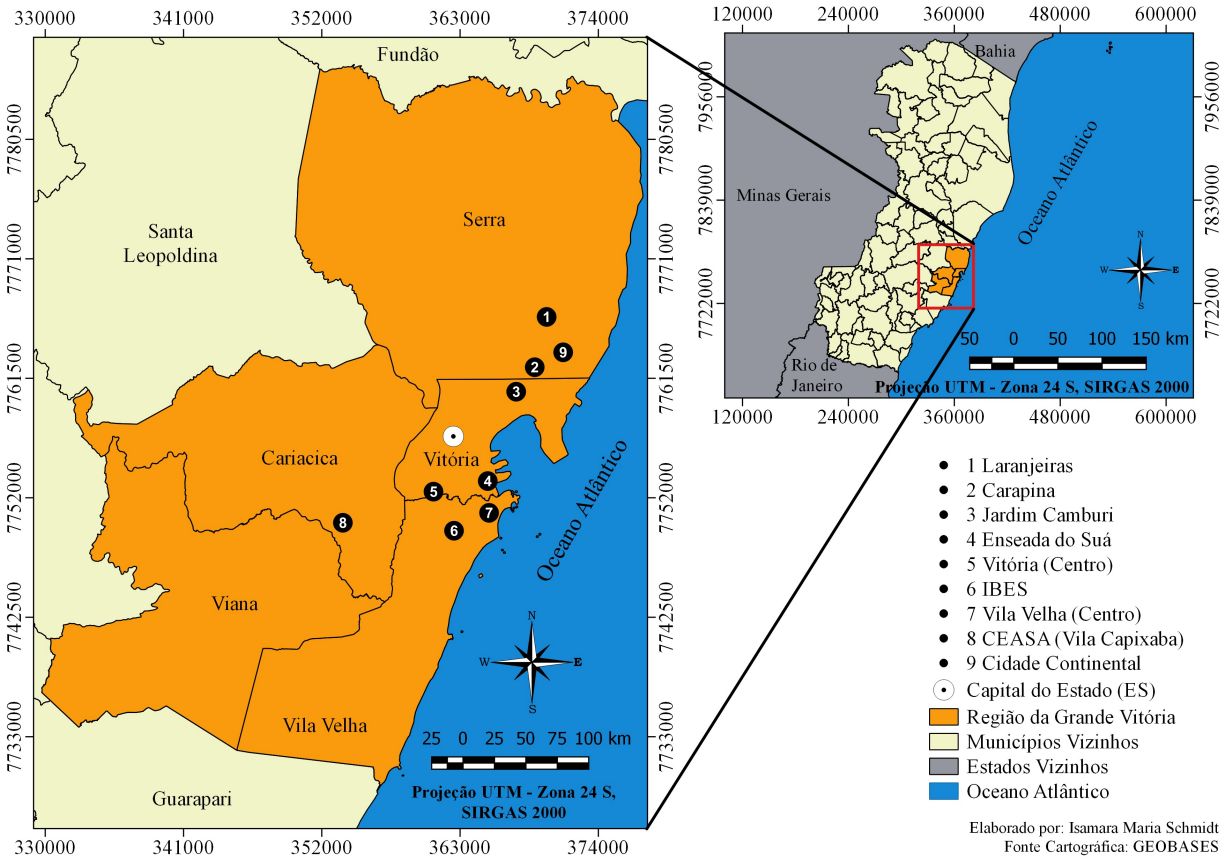


Figura 1: Estações de monitoramento da qualidade do ar na Grande Vitória.

Tabela 3: Localização das estações da RAMQAr

Estação	Localização	Bairro	Início da Operação	Coordenadas (m)	
				Leste	Norte
RAMQAr 1	Hospital Dório Silva	Laranjeiras	2000	369917	7766305
RAMQAr 2	ArcelorMittal Tubarão	Carapina	2000	368945	7762315
RAMQAr 3	Unidade de Saúde	Jardim Camburi	2000	367429	7760371
RAMQAr 4	Corpo de Bombeiros	Enseada do Suá	2000	365266	7753279
RAMQAr 5	Ministério da Fazenda	Centro (Vitória)	2005	360857	7752450
RAMQAr 6	4º Batalhão da Polícia Militar	IBES	2000	362532	7749346
RAMQAr 7	Ao lado do Colégio Marista	Centro (Vila Velha)	2000	365354	7750721
RAMQAr 8	CEASA	Vila Capixaba (CEASA)	2000	353697	7749998
RAMQAr 9	ArcelorMittal Tubarão	Cidade Continental	2011	371218	7763588

Fonte: Instituto Estadual de Meio Ambiente e Recursos Hídricos do Estado do Espírito Santo (2018)

Temperatura do ar (T); Pressão atmosférica (P); Desvio padrão da direção do vento (SIGT); e, Radiação solar (RS). Os poluentes e parâmetros meteorológicos monitorados em cada estação RAMQAR encontram-se na Tabela 4.

Tabela 4: Poluentes e parâmetros meteorológicos monitorados nas estações da RAMQAR

Estação	$MP_{2.5}$	PTS	MP_{10}	SO_2	CO	NO	NO_2	NO_X	HCT	O_3	CH_4	HCnM	Meteorologia
RAMQAR 1		X	X	X	X	X	X	X		X			
RAMQAR 2		X	X										DV,VV,UR,PP,P,T,RS,SIGT
RAMQAR 3		X	X	X		X	X	X					
RAMQAR 4	X	X	X	X	X	X	X	X	X	X	X	X	DV,SIGT,VV
RAMQAR 5		X	X	X	X	X	X	X	X		X	X	
RAMQAR 6	X	X	X	X	X	X	X	X	X	X	X	X	DV,SIGT,T,UR,VV
RAMQAR 7			X	X									
RAMQAR 8		X	X	X	X	X	X	X		X			DV,SIGT,UR,VV,T
RAMQAR 9		X	X	X	X	X		X		X			DV,VV

Fonte: Instituto Estadual de Meio Ambiente e Recursos Hídricos do Estado do Espírito Santo (2018)

4.2 DADOS

Para fins de aplicação da metodologia estudada neste trabalho e avaliação da qualidade do ar na RGV, foi adotado nos estudos principais: (i) Primeiro artigo: as concentrações de MP_{10} medidas na estação de monitoramento localizada em Enseada do Suá, Vitória e (ii) Terceiro artigo: as concentrações de MP_{10} medidas nas noves estações de monitoramento da RAMQAR da RGV. Para o primeiro artigo principal, as concentrações referem-se ao período de 01 de janeiro de 2008 a 31 de dezembro de 2017 e para o terceiro, ao período de 01 de janeiro de 2008 a 31 dezembro de 2018.

Já para os estudos adicionais foi adotado: i) no primeiro artigo: as concentrações de MP_{10} da estação de monitoramento da Enseada do Suá; as variáveis meteorológicas, temperatura, humidade relativa, precipitação e velocidade do vento, monitoradas nas estações de Carapina e Cariacica; ii) segundo artigo: as concentrações máximas de PTS, MP_{10}, SO_2, CO, NO_2 e O_3 medidas entre as estações que fazem o monitoramento do poluente; as variáveis meteorológicas, temperatura, humidade relativa, precipitação e velocidade do vento; iii) terceiro artigo: as médias diárias de concentrações de MP_{10} medidas nas estações de Carapina, Laranjeiras, Jardim Camburi, Enseada do Sul, Centro de Vitória, Ibes, Centro de Vila Velha e Cariacica; e, iv) Quarto artigo: as concentrações de MP_{10} das estações Laranjeiras (E1), Carapina (E2), Jardim Camburi, Enseada do Suá, Vitória (Centro), IBES e Cariacica e as concentrações PTS das estações Jardim Camburi, Enseada do Suá, Centro de Vitória, IBES e Cariacica. Para o primeiro artigo adicional as concentrações referem-se ao período de 01 de janeiro de 2012 a abril de 2015. Para o segundo, ao período de 01 de janeiro de 2012 a 31 de dezembro de 2014 e, para o terceiro, ao período de 01 janeiro de 2010 a 31 dezembro de 2016, e, para o quarto, ao período de 01 de janeiro de 2008 a 31 de dezembro de 2017, sendo a periodicidade diária e medida em $\mu g/m^3$.

o período de 01 de janeiro de 2008 a 31 de dezembro de 2017

4.3 RECURSOS COMPUTACIONAIS

As rotinas para a simulação, aplicação da metodologia proposta, desenvovimento de novos códigos e toda análise será utilizando o *software* R (R Development Core Team, 2016). 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. O R é em grande parte um veículo para o desenvolvimento de novos métodos interativos de análise de dados. 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; SILVA, 2011).

5 RESULTADOS E DISCUSSÕES

Nesta seção encontram-se os três principais artigos resultantes desta tese.

Estimating the autocorrelation function in the presence of missing data: an application to PM_{10} concentrations

Wanderson de Paula Pinto[1], Valdério Aselmo Reisen[1,2], Edson Zambon Monte[3], Manoel Raimundo de Sena Junior[4]

[1] *Graduate Program in Environmental Engineering (PPGEA) - Federal University of Espírito Santo, Brazil.*
[2] *Department of Statistics, Federal University of Espírito Santo, Brazil.*
[3] *Department of Economics, Econometric Reseach Group (GPE), Federal University of Espírito Santo, Brazil.*
[4] *Department of Statistics, Federal University of Pernambuco, Recife (PE), Brazil.*

Abstract

Atmospheric pollution data generally has missing information. This paper presents a study of different methodologies to estimate the autocorrelation function (ACF) of a time series with missing data. We consider the estimators proposed by Yajima & Nishino (1999) and an imputation technique. The mean squared errors of the different estimators are compared through Monte Carlo simulations for different proportions of missing data. As an application, the autocorrelation function of the Inhalable Particulate Matter concentrations (PM_{10}) was estimated. Additionally, a spectral analysis of the series under study was considered. The statistical methodology discussed in this paper can be very useful in practical situations when dealing with time series with missing data.

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

1. Introduction

The analysis of time series is a topic well discussed in the academic environment. There are two approaches relate to time series tools: time and frequency domain analysis. In both, the objective is to construct models with the following main purposes: to investigate the generating mechanism of the time series; to forecast future values; to describe the behaviour of the series; to detect the relevant periodicities; among others. Several methods to modeling and prediction the time series are available in the literature, such as moving average (MA), linear regression in time, exponential smoothing of Holt-Winters and ARIMA (Integrated Autoregressive Mobile Average Model) models.

A frequent problem in time series is the presence of missing data, specially in the environmental area, usually due to data acquisition failures (Xia et al., 1999), (Plaia & Bondì, 2006). There are several reasons for this scenario, among which we can cite: manual data entry, measurements made incorrectly, equipment with operational failures and high cost of data collection (Van Buuren et al., 1994, Lakshminarayan et al., 1999, Farhangfar et al., 2004, 2007, Wu et al., 2004, Colantonio et al., 2010).

Data analysis including only available data, without statistical treatment for missing data, may produce inaccurate estimates, see for example, the discussion in (Junger, 2008, Junger & Leon, 2015). In the literature there are several methods for the treatment of missing data in time series (Hartley & Hocking, 1971, Beale & Little, 1975, Rubin, 1976, Dempster & Rubin, 1977, Little, 1992, Schafer, 1997b, Junger, 2008, Little & Rubin, 2014, Junger & Leon, 2015). Some procedures are simple and end up producing biased estimates and others more sophisticated depend on strong assumptions about the mechanism of generating missing data patterns and complicated computational implementations. Rubin (1976) classified incomplete data according to their generating mechanisms. Data can be missing at random (MAR), missing completely at random (MCAR) or missing not at random (MNAR).

In the literature, there are several available methods to treat time series with missing data, such as, single imputation, maximum likelihood and multiple imputation. In the single imputation methods, the missing data are replaced by feasible values such as mean or median values. The maximum likelihood methods aim of estimating the parameters of the models linked to the data distribution. Multiple imputation method consists of three steps, namely: (i) Complete

databases are obtained through appropriate imputation techniques; (ii) Separately, m banks are analyzed by a traditional statistical method, as if they were indeed complete sets of data; (iii) The results found in step ii are combined in a simple and appropriate way to obtain the so-called repeated imputation inference (Rubin, 1976, Nunes et al., 2009). The results of these analyses are aggregated, generating unique estimates for the parameters of interest, which is the information that one wishes to know about the parameter of the population, such as the mean. According to Junger (2008), Junger & Leon (2015), the main disadvantage of the single imputation is by the fact that the imprecision due to imputation procedure is not considered in the analysis and, therefore, the variances of the estimators are underestimated. Unlike single imputation methods, multiple imputation methods do not tend to underestimate variance (Little & Rubin, 2014).

As examples of papers that have developed alternative methodologies for the analysis of time series with missing data one can mention Toda & Makenzie (1999), Junninen et al. (2004), Iglesias et al. (2005), Plaia & Bondì (2006) and Norazian et al. (2008). Metaxoglou & Smith (2007), Drake et al. (2014) and Junger & Leon (2015) suggest the use of Expectation-Maximization (EM) algorithms (see, Dempster & Rubin (1977), for details of the algorithm EM) to address this problem. However, this approach has the disadvantage of assuming a specific (usually normal) distribution for the data.

Parzen (1963) proposed a promising approach which makes the use of amplitude modulation technique, where the analyses are performed by means of an alternative time series in which the present observations are replaced by one and missing by zeros. Dunsmuir & Robinson (1981b) and Yajima & Nishino (1999) have studied the asymptotic behaviour of different estimators for the autocorrelation function of stationary short-memory processes with missing data. Based on the asymptotic spectral density function of amplitude modulation technique, Dunsmuir & Robinson (1981c) proposed a Whittle estimator for ARMA model coefficients and Dunsmuir & Robinson (1981a) studied the asymptotic distribution of this methodology. Other references that deal with the analysis of time series with missing data are Bloomfield (1970), Bondon (2005), Ghazal & Elhassanein (2006), with emphasis on the works of Bondon & Bahamonde (2012), who studied the estimation of conditionally heteroscedastic autoregressive models, and Efromovich (2014) who proposed a non-parametric estimation of spectral density.

In the context of Air pollution area, Junninen et al. (2004) and Plaia & Bondì (2006) proposed methodologies for imputation of missing values in air quality data sets. In the former work, univariate and multivariate methods were evaluated; in the second one, the Site-Dependent Effect method (SDEM) was proposed. Iglesias et al. (2005) and Norazian et al. (2008) have also explored methodologies for filling missing observations in air pollution data (with application specifically for PM_{10} pollutant).

Although there are many simple methodologies suggested in the literature to fill missing data in time series, it is unclear which one to adopt since they produce, in general, estimates with very large bias (Schafer (1997a)). However, the methodologies which seems to give more precisely estimates depend on complicated computer implementations and require strong assumptions about the generating pattern of the missing data.

Based on the above discussion, the main objective of this paper is to explore and evaluate the methodologies to estimate the autocorrelation function of incomplete series proposed by Parzen (1963) and Yajima & Nishino (1999) and to compare with a standard imputation method which is widely used in the Air Quality area. These methodologies are applied to the data of PM_{10} concentrations. To the best of our knowledge, these methods have not yet been explored by practitioners, specially those in the air pollution area. Therefore, this papers also contributes to the insertion of these methodologies in the applied area.

This paper is organized as follows. Section 2 presents the statistical methods. Section 3 presents simulation and an empirical studies and Section 4 deals with the analysis of PM_{10} concentrations. Some conclusions are drawn in Section 5.

2. Time Series Models

Consider the following process X_t ($t = 1, 2, \dots$) with expansion

$$X_t = \mu + \sum_{j=0}^{\infty} \beta_j \epsilon_{t-j}, \tag{1}$$

where μ is the mean of X_t, i.e, $E(X_t) = \mu$,

$$\sum_{j=0}^{\infty} \beta_j^2 < \infty, \tag{2}$$

and ϵ_t is a white noise process with $E(\epsilon_t) = 0$ and $\text{Var}(\epsilon_t) = \sigma_\epsilon^2$. X_t is defined as a second order stationary time series with variance $Var(X_t) = E(X_t - \mu)^2 = \sigma^2$ and autocovariance

$$\gamma(l) = Cov(X_t, X_{t+l}) = E(X_t - \mu)(X_{t+l} - \mu), \quad l = \pm 1, \cdots \tag{3}$$

which is a function only of the time difference l. The autocorrelation (ACF) between X_t and X_{t+l} is given by

$$\rho(l) = \frac{Cov(X_t, X_{t+l})}{\sqrt{Var(X_t)}\sqrt{Var(X_{t+l})}} = \frac{\gamma(l)}{\gamma(0)}, \tag{4}$$

where $Var(X_t) = Var(X_{t+l}) = \gamma(0)$.

A special case that satisfies Equation (1) is the $ARMA(p, q)$ (Autoregressive Moving Average Processes) defined as:

$$X_t = \mu + \sum_{j=1}^{p} \phi_j (X_{t-j} - \mu) + \epsilon_t + \sum_{j=1}^{q} \theta_j \epsilon_{t-j}, \tag{5}$$

where the polynomials $\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$ and $\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q$ with real coefficients have no common zeros, neither $\phi(z)$ nor $\theta(z)$ has zeros in the closed unit disk $\{z \in \mathbb{C} : |z| \leq 1\}$. For further details, see Brockwell & Davis (2002), Wei (2006) and Box et al. (2008), among others.

Given a sample $X_1, \ldots, X_n$ of X_t, the usual estimator of $\gamma(l)$ is the sample autocovariance function

$$\hat{\gamma}(l) = n^{-1} \sum_{t=1}^{n-l} (X_t - \overline{X})(X_{t+l} - \overline{X}), \tag{6}$$

where $\overline{X} = n^{-1} \sum_{t=1}^{n} X_t$ is the sample mean and the ACF is estimated by

$$\hat{\rho}(l) = \frac{\hat{\gamma}(l)}{\hat{\gamma}(0)} \tag{7}$$

where $\hat{\gamma}(0)$ is the estimate of the variance of the process.

The estimators discussed above are only possible to be computed when the time series is complete, that is, there is no missing data. One way to obtain ACF estimators when there are missing values, is to use one the method suggested by Parzen (1963), usually called the amplitude modulated estimation method. This is summarized as follows.

Let the process Y_t be defined as

$$Y_t = a_t X_t,$$

where X_t is assumed to be defined for all time, a_t is given by

$$a_t = \begin{cases} 1 & \text{if } X_t \text{ is observed,} \\ 0 & \text{if } X_t \text{ is missing.} \end{cases}$$

Y_t is X_t when it is observed, and zero when the value of X_t is missing.

In practice, missing values may occur regularly or randomly. Parzen (1963) consider the case of periodic sampling where the observed data consist of repeated groups of A consecutive observations followed by B missed observations. It is reference concentrate on non-parametric spectral analysis of time series with missing values, while a parametric approach is considered by Dunsmuir and Robinson (1981b). In the time domain, asymptotic properties of nonparametric estimators of autocovariances and autocorrelations of amplitude odulated time series are established by (Dunsmuir & Robinson, 1981b) and (Yajima & Nishino, 1999). These results can be used to build Yule-Walker type estimators for an AR process with missing observations (Bondon & Bahamonde, 2012).

Now, let

$$\bar{a} = n^{-1} \sum_{t=1}^{n} a_t,$$

$$C_a(l) = n^{-1} \sum_{t=1}^{n-l} a_t a_{t+l},$$

$$C_Y(l) = n^{-1} \sum_{t=1}^{n-l} Y_t Y_{t+l}.$$

Parzen (1963), in terms of these quantities, defines an estimate of $\gamma(l)$ as:

$$\hat{\gamma_X}(l) = \frac{C_Y(l)}{C_a(l)}. \tag{8}$$

If X_t has nonzero mean (μ), then it may be estimated by $\hat{\mu}_X = \frac{\sum_{t=1}^{n} Y_t}{\sum_{t=1}^{n} a_t}$. In this case, to calculate $\hat{\gamma}_X(l)$, Y_t should be replaced by $Y_t - \hat{\mu}_X$.

The following set of assumptions base the definitions of the time series ACF estimators with missing observations through the use of the apline modulation process proposed by Parzen (1963).

Assumption 1. *a_t is a real, deterministic sequence, asymptotically stationary, satisfying $\sum_t |a_{t+1} - a_t| < \infty$, and X_t is a strictly stationary process with $E\{|X_t|^k\} < \infty$, $k > 0$.*

Assumption 2. *a_t and X_t are independent, strictly stationary processes, with $E\{|a_t|^k\} < \infty$, $k > 0$ and $E\{|X_t|^k\} < \infty$, $k > 0$ (Yajima & Nishino, 1999, Toloi & Morettin, 1993).*

Assumption 3. *a_t and X_t is a sequence of independent and identically distributed random variables with a non-zero mean μ_a and strictly positive variance σ_a^2. X_t is as strictly stationary process with $E\{|X_t|^k\} < \infty$, $k > 0$, independent of a_t (Toloi & Morettin, 1993).*

Yajima & Nishino (1999) suggest three estimators for the autocorrelation function $\rho_x(l)$ for time series with missing data. The first one was proposed by Parzen (1963), the asymptotic properties of which were later investigated by Dunsmuir & Robinson (1981b). We denote it by $\hat{\rho}_{PDR}(l)$,

$$\hat{\rho}_{PDR}(l) = \frac{C_Y(l)/C_a(l)}{C_Y(0)/C_a(0)} = \frac{\sum_{t=1}^{n-l} Y_t Y_{t+l} / \sum_{t=1}^{n-l} a_t a_{t+l}}{\sum_{t=1}^{n} Y_t^2 / \sum_{t=1}^{n} a_t^2}. \tag{9}$$

In equation (9), the numerator and the denominator are estimators of $\gamma_x(l)$ and $\gamma_x(0)$, respectively, and $\hat{\rho}_{PDR}(l)$ is interpreted as a kind of Yule-Walker estimator.

The second one was proposed by Takeuchi (1995) and independently adjusted by Shin & Sarkar (1995) as the initial value of a Newton-Raphson procedure for obtaining the method of maximum probability of an $AR(1)$ model. The estimator is defined by

$$\hat{\rho}_{SST}(l) = \frac{\sum_{t=1}^{n-l} Y_t Y_{t+l}}{\sum_{t=1}^{n-l} a_{t+l} Y_t^2} = \frac{\sum_{t=1}^{n-l} a_t a_{t+l} X_t X_{t+l}}{\sum_{t=1}^{n-l} a_t a_{t+l} X_t^2}, \tag{10}$$

one can see it is a least-square estimator of the model, consisting of all observed pairs (X_{t+l}, X_t), where (X_{t+l}) is a dependent variable and X_t denotes an independent variable (Yajima & Nishino, 1999).

Finally, the third estimator was also proposed by Takeuchi (1995) and it is defined by

$$\hat{\rho}_T(l) = \frac{\sum_{t=1}^{n-l} Y_t Y_{t+l}}{\sqrt{\sum_{t=1}^{n-l} a_{t+l} Y_t^2} \sqrt{\sum_{t=1}^{n-l} a_t Y_{t+l}^2}}. \tag{11}$$

The estimator $\hat{\rho}_T(l)$ is the sample correlation coefficient based on all observed pairs (X_{t+l}, X_t). The study of these estimations, in the context of $ARMA(p, q)$ models, is one of the purposes of this paper, on the basis of different

empirical properties of missing data. These estimators have the same limiting distribution under complete sampling (Yajima & Nishino, 1999). From an analytical point of view, the properties of the estimators were discussed in Yajima & Nishino (1999) and in Dunsmuir & Robinson (1981b) for both long and short memory processes, respectively, that is, in the context of the Box and Jenkins models. Long memory can be modeled by the $ARFIMA(p, d, q)$ process where d is the fractional parameter(see, for example, Reisen (1994), Lévy-Leduc et al. (2011a) and Lévy-Leduc et al. (2011b) among others). Stationary models $ARMA(p, q)$ belong to that class of the ARFIMA models with $d = 0$.

It is noteworthy that all the approaches discussed above, for the analysis of stationary processes in the time domain, can be extended to the frequency domain. In section 4, an example of this type of application will be presented for frequency domain analysis of a time series of PM_{10} concentrations.

3. A simulation study

The empirical behaviour of the estimators was analysed using Monte Carlo simulations. The time series were generated from an autoregressive process of moving average (ARMA) process with $\epsilon_t \sim N(0, 1)$. We obtained 1000 replications (RE) of sample sizes $N = 100$ and $N = 1000$. The model parameters are displayed in the tables.

The series were investigated with ratios of 5%, 10%, 15%, 30%, and 40% of missing data. Greenland & Rothman (1998) indicates that, for a small proportion of missing data and for a large number of observations, the analysis of the complete data generally produces good results. In this way, the percentage of 5% of missing data was considered as reference of the smallest amount of tolerable missing information. The percentage of 40%, on the other hand, was used to evaluate the estimation methods of the autocorrelation function under extreme conditions of missing information.

As mentioned in the Section 1, data with incomplete observations are quite common in practice and many procedures for making missing data imputation have been developed and are available in the literature to solve or mitigate this problem. In this direction, for the purposes of comparison, the Expectation Maximization (EM) algorithm is also considered in this work to perform imputation. This method was proposed by Dempster & Rubin (1977). According to Little & Rubin (2014), the EM algorithm is a general method for obtaining maximum likelihood estimates for incomplete databases. Since these estimates may be difficult to obtain for complex databases, a procedure to reduce this difficulty is required. This is the goal of the EM algorithm (Mcknight et al., 2007). After imputing the data, the sample autocorrelation function defined in equation (6) was computed.

The EM algorithm can be used when one wants to estimate a set of parameters that follow a certain probability distribution, which is obtained using only part of the data. Suppose $A = A_1, A_2, \cdots, A_m$ represents the m sample units for which values are known, and $B = B_1, B_2, \cdots, B_n$ represents the units for which values are not known.Therefore, it is possible to explain the algorithm as follows:

1. Choose initial values for the model parameters (e.g., mean and covariance matrix for the multivariate normal model).
2. Do until convergence:
   (a) **Expectation**: calculate the expected log-likelihood function with respect to the conditional distribution of A given B, under the current estimates of parameters; that is, impute values for missing data based on the values of the parameters;
   (b) **Maximization**: find the values of the parameters that maximize the likelihood based for all data;that is, estimate new values for the parameters.

Convergence is attained when the difference between the estimated values of the parameters in two consecutive iterations is smaller than a pre-established difference. For a more detailed study of the algorithm EM as follows references can be consulted: Dempster & Rubin (1977) and Junger & Leon (2015).

It is noteworthy that the algorithm EM was chosen as the reference method for imputation due to the study performed by Junger & Leon (2015). The authors proposed some procedures for imputation of data in multivariate time series, e.g., daily concentrations of atmospheric pollutants, based on the MS algorithm. The temporal trajectory of the series is modeled with the use of splines, regression models or ARIMA (autoregressive integrated moving average) models with multiple covariance regimes. The authors performed a simulation with several missing data configurations to evaluate the validity of the proposed methods and those available as standard in most statistical

analysis applications, such as: imputation by the mean, the median, the nearest neighbor and the conditional mean. The methods were evaluated for their performance by means of accuracy and agreement indicators. In addition, the authors included a penalization criterion for the information lost in order to contemplate in the study model the uncertainty introduced by the imputation. According to results, the proposed imputation method exhibited good accuracy and precision in different configurations with respect to the patterns of missing observations.

All the codes were implemented using the software R (R Core Team, 2017). The normal multivariate EM algorithm is implemented in the R mtsdi (multivariate time-series data imputation) library, which was developed by Junger (2008) and Junger & Leon (2015). The mtsdi library is a collection of routines for imputing missing data in time series.

The tables display display the mean over 1000 replications of the sample bias and mean squared error (MSE). The ACF estimates were obtained with the classical estimator of ACF (7), for imputed series, and with the estimators $\hat{\rho}_{PDR}(l)$, $\hat{\rho}_{SST}(l)$ and $\hat{\rho}_{T}(l)$ (9), (10) and (11), respectively, for the series with missing observations. The values presented in bold correspond to the smallest men squared error.

From the results presented in Tables 1 and 2, it is inferred that the estimates obtained for the ACF of series imputed using the MS algorithm and estimated by the sub-study methods proposed by Yajima & Nishino (1999), presented results very close to the true values. However, the imputation procedure showed a slight tendency to overestimate the ACF values, for $\phi = 0.3$, and a slight tendency to overestimate the ACF values for $\phi = 0.7$, as the percentage of lost data increases. In addition, the results of the simulation study show that with 5% of data lost, all procedures yielded to good estimates for the ACF. In this case, the amount of data lost is too small to prevent the efficiency of the estimation of the autocorrelation function.

The results shown in Tables 1 and 2 were obtained for a large sample ($N = 1000$) for $\phi = 0.3$ and $\phi = 0.7$. It should be noted that the results of the ACF estimates obtained with the evaluated estimators, for $\phi = 0.7$, present a better performance than those obtained with the series imputed with the algorithm EM, regardless of the percentage of missing data. Watch the fact that the algorithm is sensitive to a set of non-stationary data, that is, it produces unsatisfactory results for ϕ close to one. The results indicate that both estimators ($\hat{\rho}_{PDR}(l)$, $\hat{\rho}_{SST}(l)$ and $\hat{\rho}_{T}(l)$ of the ACF in the presence of missing data provide estimates very close to the theoretical autocorrelation function, regardless of the percentage of missing data (5%, 10%, 15%, 30% and 40%). Empirically, the proposed estimators have significantly small MSEs (Tables 1 and 2), obtained from the estimated and theoretical values. These characteristics indicate that these methodologies can be applied to samples with large and small percentages of missing data. It is observed that the estimate of the ACF obtained with the series imputed by the algorithm EM tends to underestimate the value of the ACF for large series, fact very visible for $\phi = 0.7$.

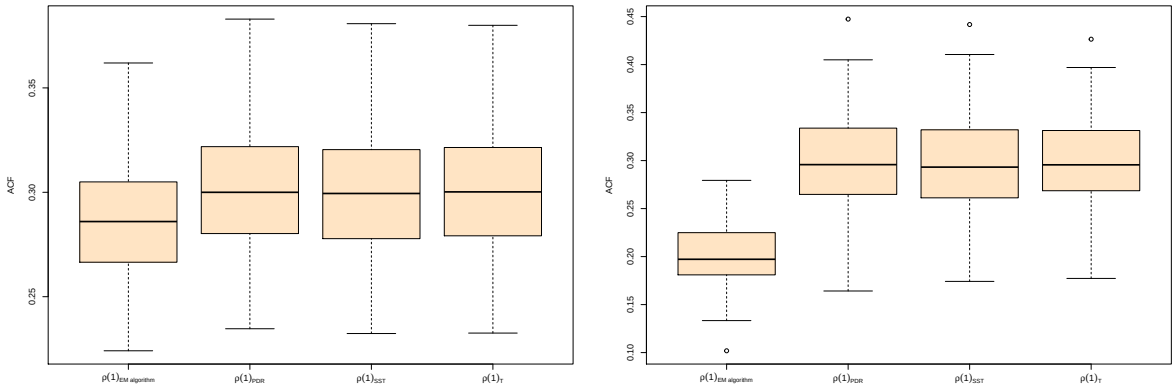
It is important to say that the imputation procedure by the algorithm EM tested in this study showed good results in several fault situations. However, even with good results, it should be considered that the imputed data are only estimates of the values that would be observed. According to Junger (2008), samples imputed with complex patterns and with a large percentage of missing data should receive special attention. Therefore, the empirical results suggest that the methodology proposed in this work is shown as an alternative to the problem and can be applied to data sets with missing observations, without compromising the statistical analysis.

The empirical study was extended to perform the process $ARMA(1,1)$, with $\phi = 0.3$ and $\phi = 0.7$, for $N = 1000$. In addition, to evaluate the methods under the assumption of non-asymmetric data, the $ARMA(1,0)$ process, with $\phi = 0.7$, was adjusted for simulated data by a chi-square distribution ($df = 2$) . The results in Tables 3 and 4 show that the estimators behave well for data generated from non-asymmetric distributions and data from an ARMA(1,1) process. In these two cases, the efficiency of the methods under the two conditions of missing data references (5% and 40%) was evaluated. The results also suggest that the estimators $\hat{\rho}_{PDR}(l)$, $\hat{\rho}_{SST}(l)$ and $\hat{\rho}_{T}(l)$ maintain its properties, even when applied to simulated data with probability distributions different from a normal distribution.

Table 1: Estimates of the ACF for an $ARMA(1,0)$ process with $\phi = 0.3$, using the proposed estimators and EM algorithm with sample size $N = 1000$ (theoretical ACF: $\rho(1) = 0.3$, $\rho(2) = 0.09$, $\rho(3) = 0.027$, $\rho(4) = 0.0081$ and $\rho(5) = 0.00243$).

P[1]	Estimator		$\hat{\rho}(1)$	$\hat{\rho}(2)$	$\hat{\rho}(3)$	$\hat{\rho}(4)$	$\hat{\rho}(5)$
5%	EM algorithm	Mean	0.29071	0.08727	**0.02764**	0.00990	0.00905
		MSE	0.00102	0.00122	0.00155	0.00149	0.00137
	$\hat{\rho}_{PDR}(l)$	Mean	0.30548	0.09102	0.02819	**0.00964**	0.00894
		MSE	0.00102	0.00131	0.00169	0.00163	0.00150
	$\hat{\rho}_{SST}(l)$	Mean	0.30533	0.09102	0.02824	0.00971	0.00896
		MSE	0.00103	0.00132	0.00170	0.00164	0.00151
	$\hat{\rho}_{T}(l)$	Mean	**0.30531**	**0.09101**	0.02812	0.00965	**0.00891**
		MSE	0.00102	0.00131	0.00169	0.00163	0.00150
10%	EM algorithm	Mean	0.26910	0.07989	0.02563	0.01067	0.00623
		MSE	0.00193	0.00134	0.00123	0.00128	0.00128
	$\hat{\rho}_{PDR}(l)$	Mean	0.29718	0.08597	0.02560	0.00862	0.00381
		MSE	0.00126	0.00153	0.00151	0.00158	0.00163
	$\hat{\rho}_{SST}(l)$	Mean	0.29723	**0.08608**	**0.02561**	0.00869	0.00392
		MSE	0.00123	0.00154	0.00150	0.00156	0.00163
	$\hat{\rho}_{T}(l)$	Mean	**0.29731**	0.08607	0.02556	**0.00853**	**0.00379**
		MSE	0.00121	0.00153	0.00151	0.00159	0.00164
15%	EM algorithm	Mean	0.25972	0.08132	0.02879	**0.00947**	0.00685
		MSE	0.00237	0.00128	0.00116	0.00109	0.00144
	$\hat{\rho}_{PDR}(l)$	Mean	0.30038	0.08961	0.02733	0.00428	0.00071
		MSE	0.00102	0.00163	0.00151	0.00146	0.00193
	$\hat{\rho}_{SST}(l)$	Mean	0.30088	**0.08967**	**0.02726**	0.00424	0.00007
		MSE	0.00101	0.00162	0.00149	0.00147	0.00193
	$\hat{\rho}_{T}(l)$	Mean	**0.30030**	0.08951	0.02733	0.00435	**0.00077**
		MSE	0.00098	0.00161	0.00151	0.00144	0.00192
30%	EM algorithm	Mean	0.21853	0.08037	0.03854	0.02720	0.02444
		MSE	0.00766	0.00113	0.00130	0.00146	0.00144
	$\hat{\rho}_{PDR}(l)$	Mean	0.29203	**0.09128**	0.03111	0.01410	**0.01083**
		MSE	0.00202	0.00186	0.00220	0.00207	0.00190
	$\hat{\rho}_{SST}(l)$	Mean	0.29152	0.09131	**0.03091**	0.01449	0.01095
		MSE	0.00186	0.00185	0.00219	0.00210	0.00188
	$\hat{\rho}_{T}(l)$	Mean	**0.29299**	0.09139	0.03109	**0.01387**	0.01084
		MSE	0.00185	0.00184	0.00218	0.00206	0.0018
40%	EM algorithm	Mean	0.19751	0.07973	0.03917	0.02528	0.02392
		MSE	0.01150	0.00112	0.00111	0.00137	0.00165
	$\hat{\rho}_{PDR}(l)$	Mean	0.29414	0.09168	0.02345	0.00043	-0.00351
		MSE	0.00281	0.00264	0.00257	0.00265	0.00280
	$\hat{\rho}_{SST}(l)$	Mean	0.29624	0.09257	0.02314	0.00002	-0.00368
		MSE	0.00256	0.00267	0.00257	0.00265	0.00281
	$\hat{\rho}_{T}(l)$	Mean	**0.29641**	**0.09128**	**0.02371**	**0.00066**	**-0.00348**
		MSE	0.00247	0.00259	0.00258	0.00263	0.00283

(1) P = percent of missing data.



(a) (b)

Figure 1: Box Plot estimates of the ACF for an $ARMA(1,0)$ process with $\phi = 0.3$, using the proposed estimators and EM algorithm with sample size $N = 1000$, with (a) 5% and (b) 40% missing data.

Table 2: Estimates of the ACF for an $ARMA(1,0)$ process with $\phi = 0.7$, using the proposed estimators and EM algorithm with sample size $N = 1000$ (theoretical ACF: $\rho(1) = 0.7$, $\rho(2) = 0.49$, $\rho(3) = 0.343$, $\rho(4) = 0.2401$ and $\rho(5) = 0.16807$).

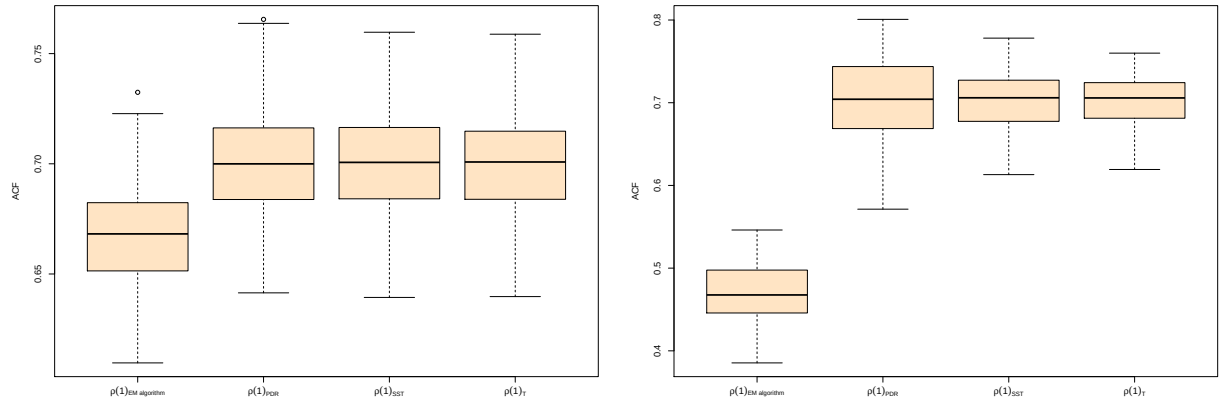
P[1]	Estimator		$\hat{\rho}(1)$	$\hat{\rho}(2)$	$\hat{\rho}(3)$	$\hat{\rho}(4)$	$\hat{\rho}(5)$
5%	EM algorithm	Mean	0.66646	0.46321	0.32368	0.22695	0.15966
		MSE	0.00167	0.00202	0.00233	0.00243	0.00215
	$\hat{\rho}_{PDR}(l)$	Mean	0.69964	0.48664	0.34033	0.23884	0.16782
		MSE	0.00062	0.00147	0.00222	0.00262	0.00244
	$\hat{\rho}_{SST}(l)$	Mean	0.69989	**0.48695**	**0.34070**	**0.23912**	**0.16807**
		MSE	0.00058	0.00145	0.00221	0.00261	0.00246
	$\hat{\rho}_{T}(l)$	Mean	**0.69991**	0.48664	0.34008	0.23872	0.16791
		MSE	0.00054	0.00140	0.00220	0.00262	0.00244
10%	EM algorithm	Mean	0.63507	0.44339	0.31021	0.21660	0.15325
		MSE	0.00485	0.00345	0.00300	0.00265	0.00249
	$\hat{\rho}_{PDR}(l)$	Mean	**0.69837**	**0.48492**	0.33553	0.23198	0.16118
		MSE	0.00074	0.00154	0.00225	0.00250	0.00259
	$\hat{\rho}_{SST}(l)$	Mean	0.69773	0.48473	0.33554	0.23190	0.16107
		MSE	0.00065	0.00152	0.00226	0.00249	0.00260
	$\hat{\rho}_{T}(l)$	Mean	0.69766	0.48486	**0.33585**	**0.23175**	**0.16126**
		MSE	0.00058	0.00146	0.00224	0.00249	0.00258
15%	EM algorithm	Mean	0.60324	0.42617	0.30445	0.21939	0.15735
		MSE	0.00991	0.00521	0.00312	0.00252	0.00254
	$\hat{\rho}_{PDR}(l)$	Mean	0.69558	0.48460	0.34079	0.23972	0.16647
		MSE	0.00082	0.00167	0.00221	0.00267	0.00302
	$\hat{\rho}_{SST}(l)$	Mean	**0.69702**	**0.48593**	**0.34204**	**0.24048**	**0.16714**
		MSE	0.00067	0.00166	0.00229	0.00274	0.00310
	$\hat{\rho}_{T}(l)$	Mean	0.69689	0.48500	0.34028	0.23909	0.16630
		MSE	0.00006	0.00158	0.00224	0.00268	0.00307
30%	EM algorithm	Mean	0.52290	0.37979	0.27965	0.20767	0.15826
		MSE	0.03246	0.01337	0.00574	0.00261	0.00196
	$\hat{\rho}_{PDR}(l)$	Mean	0.70251	0.49431	0.34790	**0.24410**	0.17096
		MSE	0.00193	0.00209	0.00304	0.00231	0.00274
	$\hat{\rho}_{SST}(l)$	Mean	0.70267	**0.49338**	0.34812	0.2442	**0.17054**
		MSE	0.00123	0.00201	0.00297	0.00221	0.00277
	$\hat{\rho}_{T}(l)$	Mean	**0.70075**	0.49438	**0.34764**	0.24405	0.17081
		MSE	0.00092	0.00180	0.00291	0.00231	0.00275
40%	EM algorithm	Mean	0.47083	0.34796	0.26369	0.20339	0.16225
		MSE	0.05347	0.02137	0.00788	0.00347	0.00259
	$\hat{\rho}_{PDR}(l)$	Mean	0.70359	0.49102	0.34641	0.24229	0.17439
		MSE	0.00232	0.00260	0.00309	0.00414	0.00519
	$\hat{\rho}_{SST}(l)$	Mean	**0.70056**	**0.49062**	0.34615	0.24209	**0.17314**
		MSE	0.00129	0.00314	0.00324	0.00420	0.00493
	$\hat{\rho}_{T}(l)$	Mean	0.70245	0.49183	**0.34581**	**0.24157**	0.17393
		MSE	0.00095	0.00174	0.00285	0.00395	0.00497

(1) P = percent of missing data.

Table 3: Estimates of the ACF for an $ARMA(1,1)$ process with $\phi = 0.7$ and $\theta = 0.3$, using the proposed estimators and EM algorithm with sample size $N = 1000$ (theoretical ACF: $\rho(1) = 0.72$, $\rho(2) = 0.504$, $\rho(3) = 0.352$, $\rho(4) = 0.246$ and $\rho(5) = 0.1728$).

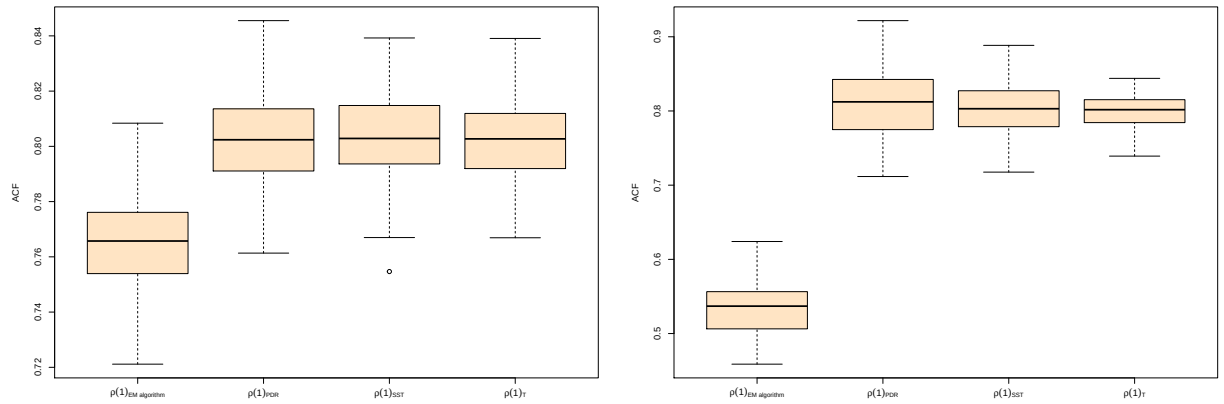
P[1]	Estimator		$\hat{\rho}(1)$	$\hat{\rho}(2)$	$\hat{\rho}(3)$	$\hat{\rho}(4)$	$\hat{\rho}(5)$
5%	EM algorithm	Mean	**0.76035**	**0.52873**	**0.36877**	**0.25591**	**0.17890**
		MSE	0.00190	0.00156	0.00192	0.00255	0.00307
	$\hat{\rho}_{PDR}(l)$	Mean	0.79752	0.55511	0.38722	0.26866	0.18754
		MSE	0.00634	0.00365	0.00300	0.00311	0.00343
	$\hat{\rho}_{SST}(l)$	Mean	0.79790	0.55527	0.38750	0.26891	0.18773
		MSE	0.00632	0.00362	0.00301	0.00313	0.00346
	$\hat{\rho}_{T}(l)$	Mean	0.79798	0.55475	0.38709	0.26872	0.18752
		MSE	0.00631	0.00354	0.00294	0.00311	0.00341
40%	EM algorithm	Mean	0.52847	0.39399	0.29736	**0.22652**	**0.18101**
		MSE	0.03789	0.01352	0.00467	0.00268	0.00241
	$\hat{\rho}_{PDR}(l)$	Mean	0.80251	0.56643	0.40011	0.27716	0.19866
		MSE	0.00940	0.00647	0.00531	0.00540	0.00520
	$\hat{\rho}_{SST}(l)$	Mean	0.80180	0.56974	**0.39968**	0.27763	0.19963
		MSE	0.00782	0.00742	0.00570	0.00564	0.00553
	$\hat{\rho}_{T}(l)$	Mean	**0.80106**	**0.56549**	0.40036	0.27644	0.19898
		MSE	0.00715	0.00584	0.00514	0.00516	0.00525

(1) P = percent of missing data.



(a) (b)

Figure 2: Box Plot estimates of the ACF for an $ARMA(1,0)$ process with $\phi = 0.7$, using the proposed estimators and EM algorithm with sample size $N = 1000$, with (a) 5% and (b) 40% missing data.



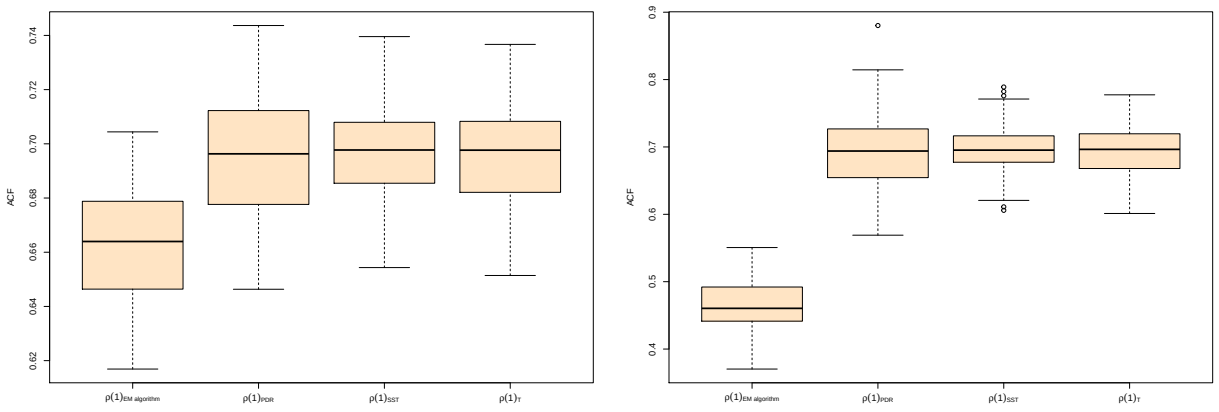
(a) (b)

Figure 3: Box Plot estimates of the ACF for an $ARMA(1,1)$ process with $\phi = 0.7$ and $\theta = 0.3$, using the proposed estimators and EM algorithm with sample size $N = 1000$, with (a) 5% and (b) 40% missing data.

Table 4: Estimates of the autocorrelation function obtained for an $ARMA(1,0)$ process (Chi-Square g.f. = 2) with $\phi = 0.7$, using the proposed estimators with sample size $N = 1000$ (theoretical ACF: $\rho(1) = 0.7$, $\rho(2) = 0.49$, $\rho(3) = 0.343$, $\rho(4) = 0.2401$ and $\rho(5) = 0.16807$).

P[1]	Estimator		$\hat{\rho}(1)$	$\hat{\rho}(2)$	$\hat{\rho}(3)$	$\hat{\rho}(4)$	$\hat{\rho}(5)$
5%	EM algorithm	Mean	0.66589	0.46710	0.32489	0.22835	0.15899
		MSE	0.00164	0.00167	0.00185	0.00220	0.00247
	$\hat{\rho}_{PDR}(l)$	Mean	0.69759	0.48795	0.33769	0.23550	0.16239
		MSE	0.00054	0.00128	0.00170	0.00228	0.00267
	$\hat{\rho}_{SST}(l)$	Mean	**0.69787**	**0.48815**	**0.33802**	**0.23578**	**0.16267**
		MSE	0.00051	0.00135	0.00180	0.00239	0.00276
	$\hat{\rho}_{T}(l)$	Mean	0.69728	0.48661	0.33712	0.23560	0.16237
		MSE	0.00048	0.00122	0.00167	0.00229	0.00267
40%	EM algorithm	Mean	0.46795	0.34364	0.25914	0.19951	0.15804
		MSE	0.05554	0.02315	0.00892	0.00418	**0.00302**
	$\hat{\rho}_{PDR}(l)$	Mean	0.69934	0.48323	0.33622	0.23264	0.16276
		MSE	0.00319	0.00284	0.00304	0.00406	0.00416
	$\hat{\rho}_{SST}(l)$	Mean	0.70376	**0.48675**	**0.33916**	**0.23467**	**0.16516**
		MSE	0.00190	0.00390	0.00349	0.00425	0.00454
	$\hat{\rho}_{T}(l)$	Mean	**0.70043**	0.48396	0.33641	0.23331	0.16348
		MSE	0.00146	0.00225	0.00265	0.00396	0.00407

(1) P = percent of missing data.



(a) (b)

Figure 4: Box Plot estimates of the autocorrelation function obtained for an $ARMA(1,0)$ process (Chi-Square g.f. = 2) with $\phi = 0.7$, using the proposed estimators with sample size $N = 1000$, with (a) 5% and (b) 40% missing data.

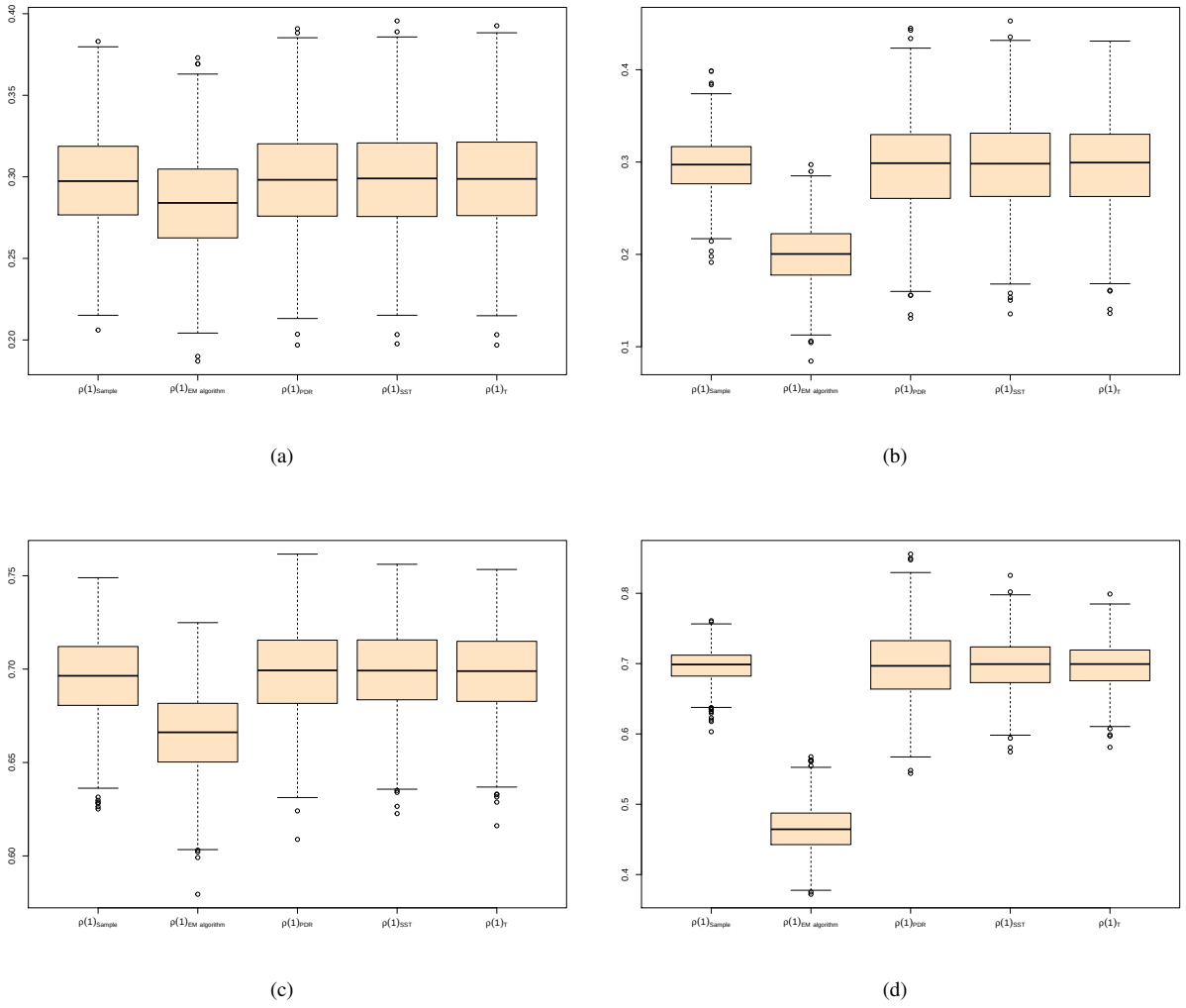
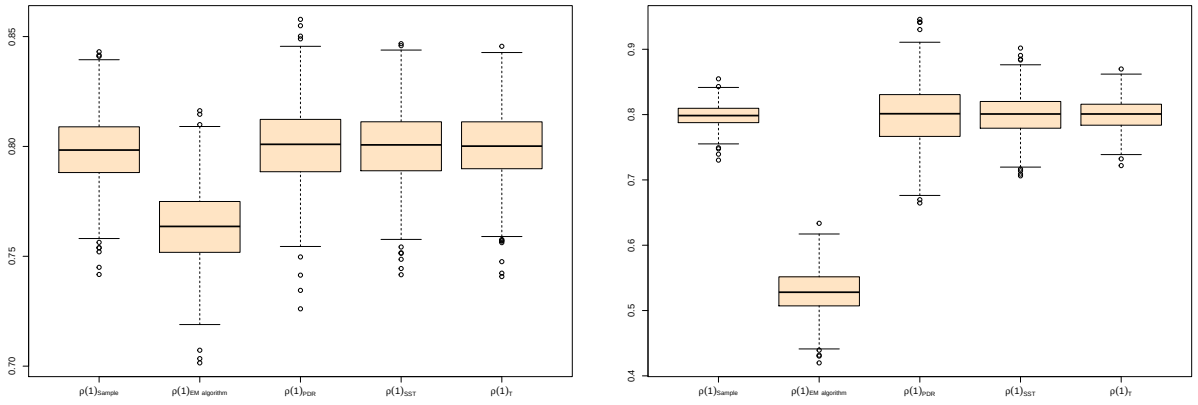
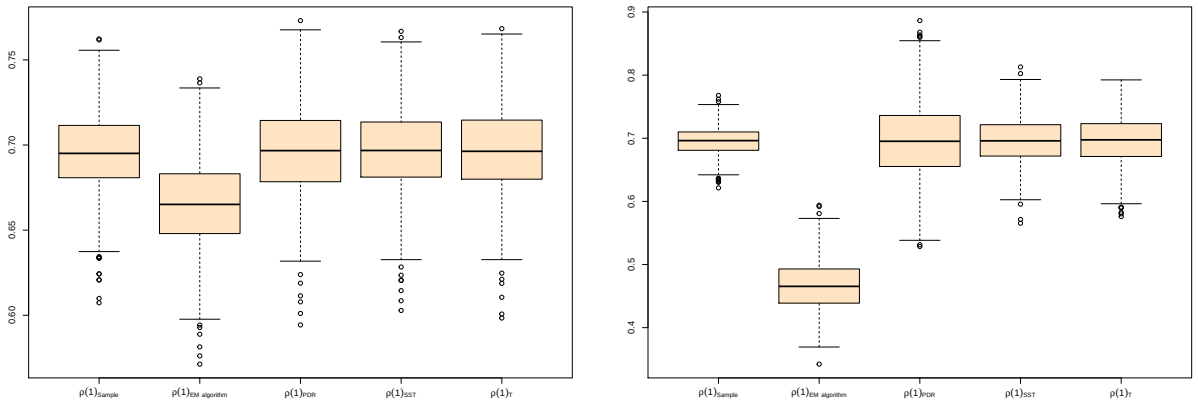


Figure 5: Box Plot estimates of the ACF for an $ARMA(1,0)$ process with $\phi = 0.3$ and $\phi = 0.7$, using the proposed estimators, the EM algorithm and the classical ACF estimator for the complete sample, sample size $N = 1000$, with (a) $\phi = 0.3$ and 5% missing data, (b) $\phi = 0.3$ and 40% missing data, (c) $\phi = 0.7$ and 5% missing data and (d) $\phi = 0.7$ and 40% missing data.



(a) (b)

Figure 6: Box Plot estimates of the ACF for an $ARMA(1,1)$ process with $\phi = 0.7$ and $\theta = 0.3$, the EM algorithm and the classical ACF estimator for the complete sample, sample size $N = 1000$, with (a) 5% and (b) 40% missing data.



(a) (b)

Figure 7: Box Plot estimates of the ACF for an $ARMA(1,0)$ process (Chi-Square g.f. $= 2$) with $\phi = 0.7$, using the proposed estimators, the EM algorithm and the classical ACF estimator for the complete sample, sample size $N = 1000$, with (a) 5% and (b) 40% missing data.

4. An application to PM_{10} pollutant

4.1. Area of study

To apply the studied methodology, this paper used data collected in the Region of Greater Vitória (RGV), Espírito Santo, Brasil. RGV is located on the south coast of the Atlantic Ocean (latitude 20° 19S, longitude 40° 20W). The RGV is constituted by the municipalities of Vitória, Vila Velha, Cariacica, Serra and Viana. Because it is situated in the coastal region, the RGV has a warm tropical climate (Aw), with mild and dry winter, and a hot and rainy summer, with averange temperatures ranging from 24 °C to 30 °C. According to the Brazilian institute of geography and statistics (Instituto Brasileiro de Geografia e Estatística, 2014), the metropolitan region of Vitória has 1,475,332 inhabitants, covers an area of 1,461 square kilometres and is one of the main centers of urban and industrial development in the state. The region suffers from several environmental problems, among them the deterioration of air quality due to atmospheric emissions by industries and the vehicular fleet.

The RGV has an automatic network for air quality monitoring (RAMQAR), owned by the state institute for the environment and water resources, that was put in service in July, 2000. The network is distributed among eight monitoring stations located in the municipalities that compose the RGV, as follows: Serra with three stations (Laranjeiras, Carapina and Cidade Continental); Vitória with three stations (Jardim Camburi, Enseada do Suá, and Central Vitória); Vila Velha with two stations (Ibes and Central Vila Velha); and Cariacica (Cariacica). The locations of the RAMQAR monitoring stations are shown in Figure 8.

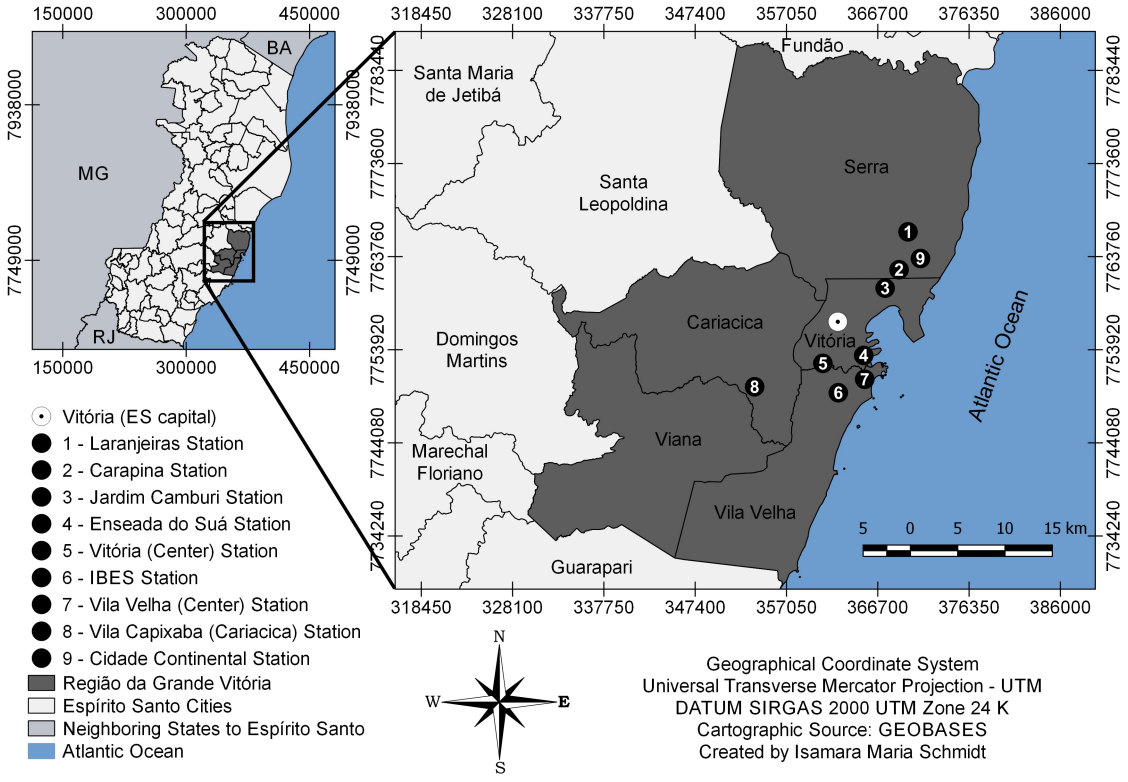


Figure 8: Location of the air quality monitoring stations in the RGV.

The RAMQAR network monitors the following pollutants: PM_{10}, total suspension particles (TSP), Respirable Particulate Material $PM_{2.5}$ Ozone (O_3), Nitrogen Oxide (NO_x), Nitrogen Dioxide (NO_2), Non-Methane Hydrocarbons (HCnM), Methane (CH_4), Nitrogen Monoxide (NO), Carbon Monoxide (CO), Sulphur dioxide (SO_2) and Total

Table 5: Meteorological parameters and pollutants monitored at each RAMQAR station.

Stations	$MP_{2.5}$	PTS	MP_{10}	SO_2	CO	NO	NO_2	NO_X	HCT	O_3	CH_4	HCnM	Meteorological
Laranjeiras		X	X	X	X	X	X	X		X			
Carapina		X	X										DV,VV,UR,PP,P,T,RS,SIGT
Jardim Camburi		X	X	X		X	X	X					
Enseada do Suá	X	X	X	X	X	X	X	X	X	X	X	X	DV,SIGT,VV
Centro (Vitória)		X	X	X	X	X	X	X	X		X	X	
IBES	X	X	X	X	X	X	X	X	X	X	X	X	DV,SIGT,T,UR,VV
Centro (Vila Velha)			X	X									
Vila Capixaba (CEASA)		X	X	X	X	X	X	X		X			DV,SIGT,UR,VV,T
Cidade Continental		X	X	X	X	X		X		X			DV,VV

Source: Instituto Estadual de Meio Ambiente e Recursos Hídricos do Estado do Espírito Santo (2014).

Hydrocarbons (HCT). The following meteorological parameters are also monitored: Scalar wind direction (WD), scalar wind velocity (WV), precipitation (PP), relative humidity (RH), Air temperature (T), atmospheric pressure (P), Standard deviation of wind direction (SIGT) and solar radiation (I). The pollutants and meteorological parameters that RAMQAR monitor at each station are shown in Table 5.

4.2. Analysis of PM_{10} concentrations with missing observations

As previously mentioned, the daily average PM_{10} concentrations (Figure 9) is the data set here analyzed to illustrate the methodology previously discussed. The series is expressed in $\mu g/m^3$ and it was observed in Enseada do Suá, Vitória, Espírito Santo, Brazil. For a preliminary understanding of the variable under study, some statistical descriptive measures are presented in Table 6. An analysis corresponding to the time series, for the period from 2012/01/01 to 2014/12/31, making a total of 1096 observations, and this total 93 are missing data. It is noteworthy that, for the calculation of descriptive statistics, it was considered in the observed data.

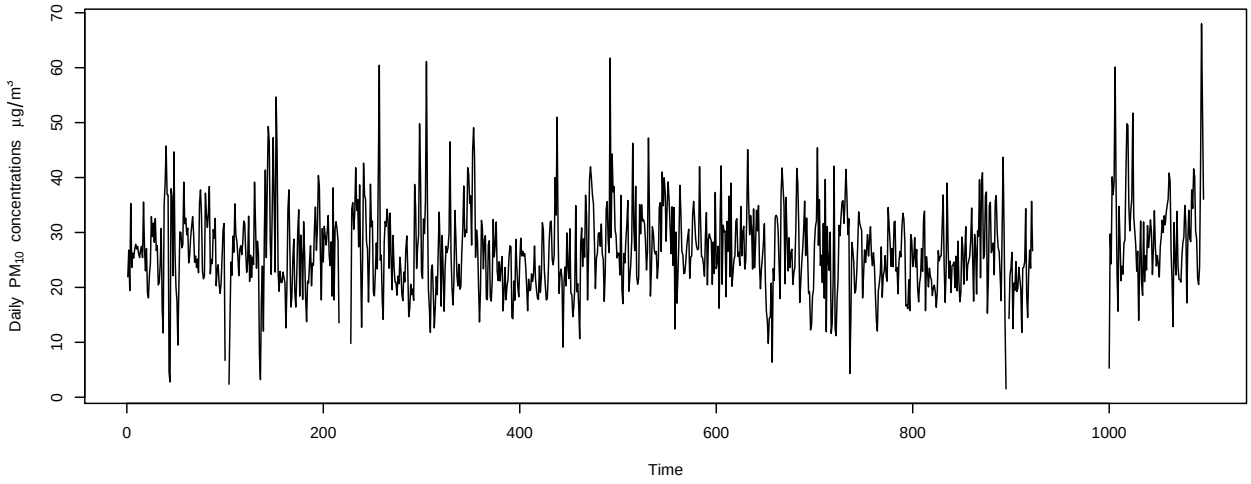


Figure 9: Daily PM_{10} concentrations series with missing data.

The mean PM_{10} concentration was approximately 26.64 $\mu g/m^3$, with a standard deviation of 7.60 $\mu g/m^3$. Note that, on average, the concentrations did not exceed 50 $\mu g/m^3$, but the standard deviation and the high coefficient of variation indicate a poorly representative average. In addition, the results show that the maximum value was more than triple the mean value, demonstrating the great variability of PM_{10} concentrations in RGV. The minimum value was observed in march 2013, in the rainy season of the region, which starts from october until mid-april (Instituto Estadual de Meio Ambiente e Recursos Hídricos do Estado do Espírito Santo, 2014). At this time of year, the performance

Table 6: Descriptive measures of the variable under study.

Descriptive measures	PM_{10} time series
nobs	1096.00
NAs	93.00
Minimum	1.54
Maximum	54.67
1. Quartile	21.62
3. Quartile	31.12
Mean	26.64
Median	26.21
Variance	57.80
Stdev	7.60
Coefficient of variation	42.63
Skewness	0.25
Kurtosis	0.69

of the frontal systems and zones of convergence of humidity favor the increase of precipitation. In this period, the concentrations of PM_{10} are lower, mainly, by the intensification of the efficiency of the removal processes by wet deposition.

Figures 10(a), 10(b) and 10(c), respectively, show the ACF estimates obtained: (a) with the $\hat{\rho}_{PDR}(l)$ estimator, (b) with $\hat{\rho}_{SST}(l)$, and (c) by applying $\hat{\rho}_{T}(l)$. These plots clearly show the presence of the seasonality behavior with period $s = 7$, which is an expected data behavior since the series was observed daily. It is worth mentioning that seasonality in series of PM_{10} is expected, since according to the Instituto Estadual de Meio Ambiente e Recursos Hídricos do Estado do Espírito Santo (2014) a main source emitting particles in the RGV are automotive transports representing more than 60% particle emissions are linked to resuspension of particles in roads, there is a variation between measured concentrations on weekdays and weekends, since the flow of vehicles is higher during the days of the week.

The fact that the time series presents many cycles of different frequencies and amplitudes, that is, series with properties of non-stationarity and seasonality is an opportunity to analyze them in the frequency domain. The use of time series decomposition through spectral analysis emerged as an alternative for the identification of cycle components. The spectrum shows a decomposition of the variance of a data sample through different frequencies. Thus, the spectrum describes as a cyclic good of a given time series. The assays are as varied as the number of impressions of different frequencies, and is a contribution of each cycle and constant in an entire sample. The spectrum then makes a direct contribution through these elementary cycles to a global process variance.

A statistical tool that, for more than a century, has been widely used in the various scientific areas where it is fundamental to find the periodic components of the phenomena under study is the Periodogram. The periodogram was introduced by Schster, in 1898, with the aim of identifying periodic components in time series. For each frequency λ, the ordinate $I(\lambda)$ estimates the contribution of that frequency to the series being studied. Thus, in a superficial way, large values of the periodogram ordinates imply significant frequencies in the series and small values, correspond to little significant frequencies. Usually the frequencies that interest are those that produce "peaks" in the periodogram, especially when working with air quality data. The mathematical concept of the periodogram is formalized following.

Let X_t be a stationary process with zero mean, as defined in equation (1), and consider the autocovariance function $\gamma(l)$. The periodogram for the observed time series $x_1, \cdots, x_n$ is given by

$$I_X^{(n)}(\lambda) = \frac{1}{2\pi n}\left|\sum_{t=1}^{n} X_t e^{-i\lambda t}\right|^2, \quad \lambda \in [0, \pi]. \tag{12}$$

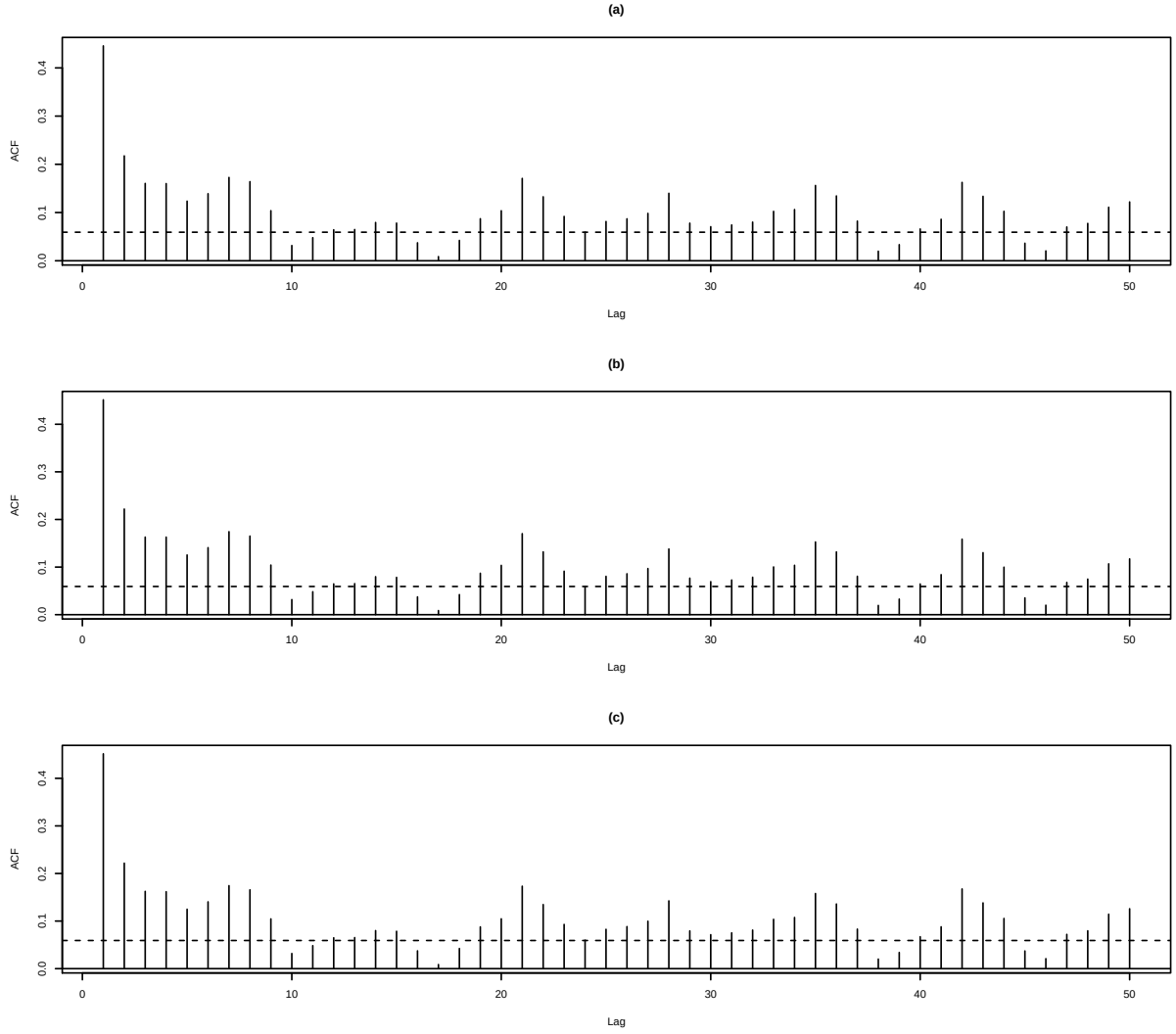


Figure 10: ACF estimates of PM_{10} concentrations with missing data.

It is well-known that the periodogram is an asymptotically uniased estimator of the spectral density function

$$f_Y(\lambda) = \frac{1}{2\pi} \sum_{t=-\infty}^{\infty} \gamma(l) e^{-i\lambda t}, \quad \lambda \in [0, \pi], \tag{13}$$

even though it is inconsistent (Jiang & Hui, 2004). If Xt follows a stationary linear process, then the periodogam $I_Y^{(n)}(\lambda_k)$ at Fourier frequencies $\lambda_k = \frac{2\pi k}{n}$ (for $k = 0, 1, ..., N$, where $N = (n-1)/2$) are asymptotically independent. The asymptotic unbiasedness and independence of the periodogram allow one to construct a consistent estimator of $f_Y(\lambda)$ by locally averaging the periodogram.

Our main tool for deriving estimators for the spectrum of the unobserved process X_t will be the statistic

$$d_Y^{(N)}(\lambda) = \sum_{t=0}^{n-1} Y_t e^{-i\lambda t} = \sum_{t=0}^{n-1} a_t X_t e^{-i\lambda t,} \quad -\pi < \lambda < \pi, \tag{14}$$

which is called the finite Fourier transform of the values $Y_0, \cdots, Y_{n-1}$.

Definition 1. For the time series Y_t with modulated amplitude, its generalized periodogram is defined as

$$GI_Y^{(n)}(\lambda) = \frac{1}{2\pi} \sum_{t=0}^{n-1} \hat{\gamma}_Y(l) e^{-i\lambda t}, \tag{15}$$

where λ are Fourier frequencies, and

$$\hat{\gamma}_Y(l) = C_Y(l) = n^{-1} \sum_{t=1}^{n-l} Y_t Y_{t+l}, \quad 0 \le l \le n-1. \tag{16}$$

The generalized periodogram has properties similar to the common periodograms. It can be used to construct a variety of tests for hidden periodicities following the ideas of common tests based on the periodogram in equation (12) (Jiang & Hui, 2004)

The frequency domain counterpart of the sample ACF is the periodogram which is presented in Figure 11. The sample spectrum has peaks at frequencies very close to zero and also at frequencies which are multiples of 1/7. The periodogram demonstrates that the highest peak is associated with a frequency of 0.14324, implying $s = \frac{1}{0.14324} = 6.98$, that is, a seasonal component with a periodicity of seven days.

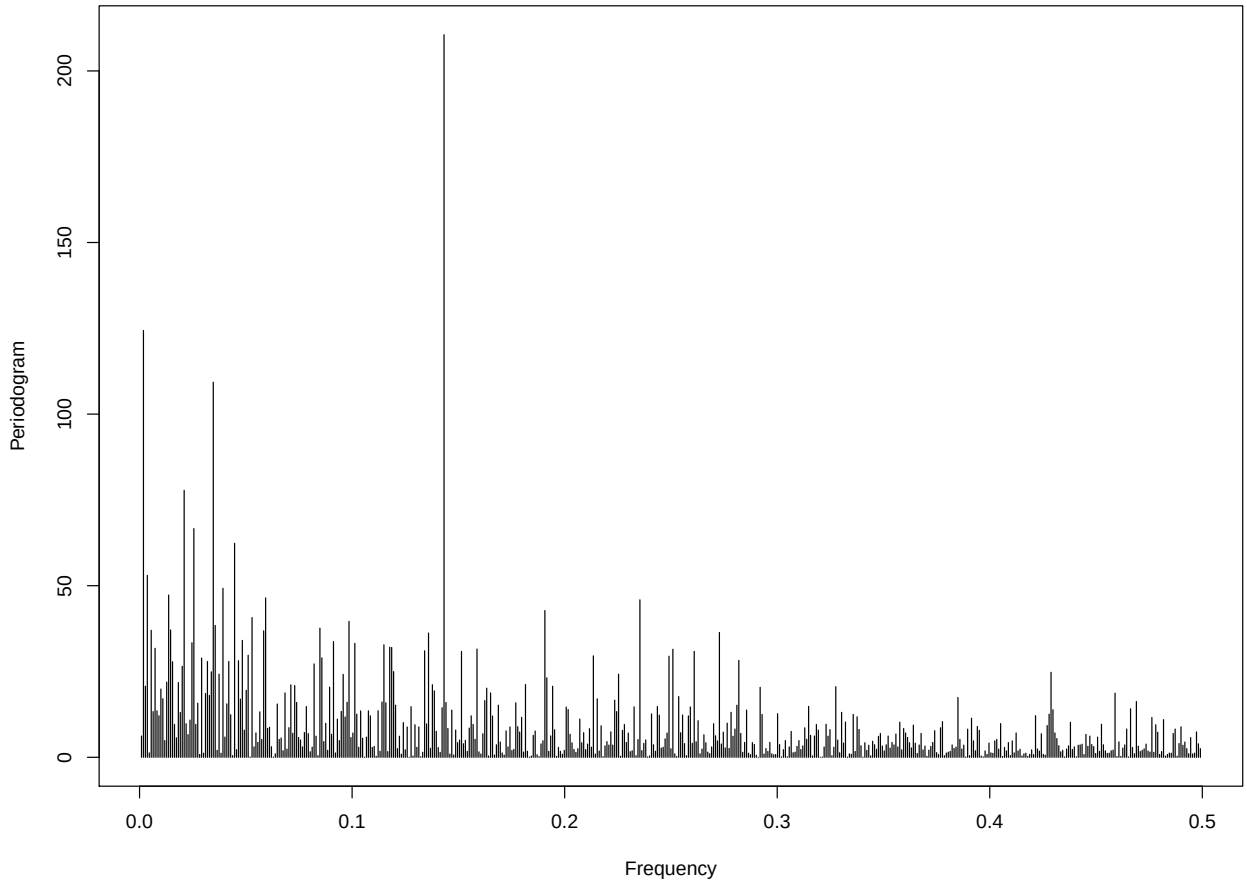


Figure 11: Periodogram of PM_{10} concentrations.

As time series of atmospheric pollutants exhibit periodic variations, which, if identified, may aid in the process of identifying sources of pollution and nature of the phenomenon. In addition, they are analyzes for the process of prediction of peak concentrations of pollutants. A time series methodology (Iordache & Dunea, 2013), but little explored in studies on air pollution (Hies et al., 2000, Sebald et al., 2000, Tchepel & Borrego, 2010, Tchepel et al., 2010), a spectrum analysis is a versatile technique and a critical application in meteorology (Iordache & Dunea, 2013). This may be due to the fact that air quality time series often has periods of consecutive missing values making difficult the use of Spectrum (Fourier) Analysis, which requires equally spaced observations (Iordache & Dunea, 2013), which justifies the use of the methodology of amplitude modulated processes for estimation of the periodogram, as studied in this article.

There is a relationship between the temporal and spatial scales of air pollution. In this context, as short-term fluctuations of concentrations of atmospheric pollutants in RVG are related to local scale phenomena, mainly local meteorological conditions that influence the dispersion patterns. In contrast, the seasonal changes in emissions and the long-term transport of pollution will contribute to the low-frequency spectrum. Typically, time series of atmospheric pollution has a broad spectrum related to the periodicity of atmospheric physical processes and sources in the region (Tchepel et al., 2010).

In addition, a periodogram of the PM_{10} concentrations imputed with the EM algorithm was calculated. It is worth mentioning that for imputation the real series was subdivided into three sub-series of 365 observations. Figure 12 (a) presents the periodogram of the imputed series, estimated by the equation (12). Figure 12 (b) contains the estimated linear regression between the values of the periodograms (Figures 11 and 12 (a)), respectively, estimated with the series in the presence of the missing data and the series imputed by the EM algorithm. It is observed (Figures 11 and 12 (a)) that in both periodograms the property of seasonality is verified, however in the estimation of the periodogram with the imputed series the values are lower than those estimated using the method proposed in this article. Figure 12 (b) shows that for frequencies close to zero and 0.14324. It is important to note that the latter is the frequency that shows the seasonality of the data set, the estimated periodogram, using the imputation method via the EM algorithm, presented a tending to underestimate the values of the periodogram, when compared to the method proposed in this paper, which evidences a possible decrease in the efficiency of this methodology of imputation of missing data in studies in which it is objectified to identify the periodicity of the data. This is due to the fact that the variance tends to be underestimated when using the imputation of missing data, which, consequently, results in the underestimation of the values of the periodogram.

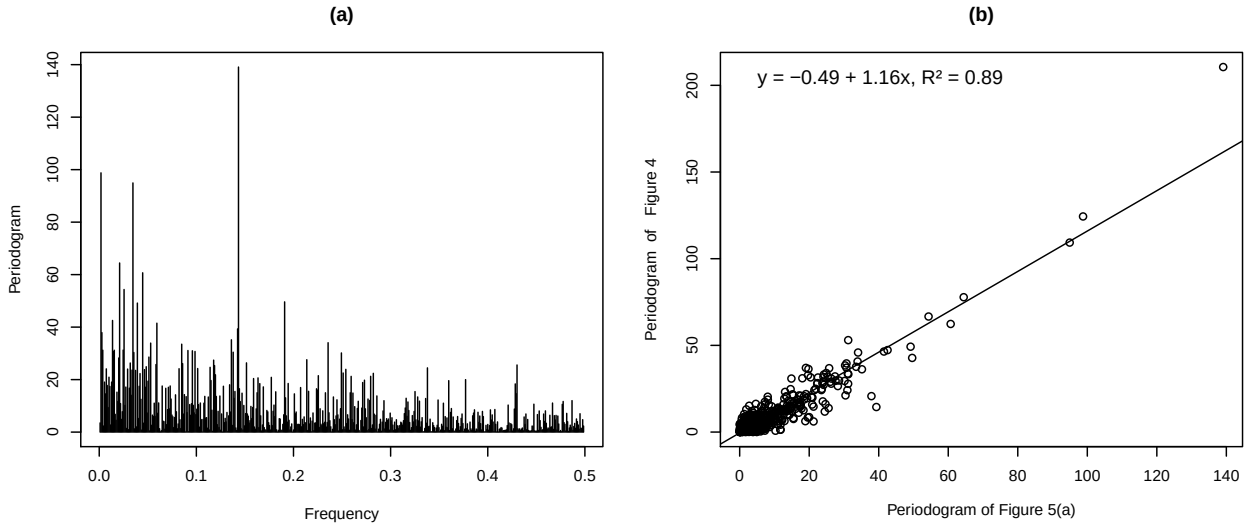


Figure 12: (a) Periodogram of the PM_{10} concentrations series imputed and (b) Regression line between estimated periodograms.

According to Rubin (2004) one of the problems resulting from the imputation of missing data is that a single

imputation can lead to the underestimation of the variances because the variance due to the data imputation was not counted. In the case of the EM algorithm, it provides estimates of maximum likelihood in the presence of missing data. In the first step "E" the missing data are estimated based on the observed values and the initial values of the imputation model parameters. Already, step "M" re-estimates the parameters of the imputation model using the imputed data. The algorithm iterates from steps E to M until the log probability converges to a stationary point (Barzi & Woodward, 2004). In this way, the underestimation of the typical variance of the average imputations is minimized, but this process is affected as the percentage of missing data is high. It should be noted that the simulation study corroborates this result, one that has shown that the efficiency of the imputation methods decreases with increasing percentage of missing data.

5. Concluding remarks

This paper presents a study of methodologies for the treatment of missing data in time series. Such methodologies are based on imputation methods and estimators of the time series autocorrelation function in the presence of missing data. For the processes investigated in this study, the ACF estimates were obtained using the estimators proposed by Parzen (1963) and Yajima & Nishino (1999) and by the algorithm EM (Dempster & Rubin, 1977) have significantly small MSE, compared to the corresponding theoretical values . The results of the numerical experiments suggest that the EM method tends to underestimate the ACF values. The empirical investigation considered different scenarios, emphasizing the effect of the percentage of missing data in the estimators. A data application of PM_{10} was considered and with both estimators it was possible to identify the seasonality property of the series. From the obtained results, it can be observed that the methodologies tested in the work present accurate results, even under extreme conditions of missing data (40%). Therefore, these methodologies can be used as an alternative for the estimation of the autocorrelation function of time series in the presence of missing data, in addition, they can be applied in studies with incomplete database of concentrations of atmospheric pollutants.

Acknowledgements

The authors are grateful to Fundação de Amparo à Pesquisas do Espírito Santo - FAPES, for the financial support.

References

Barzi, F., & Woodward, M. (2004). Imputations of missing values in practice: results from imputations of serum cholesterol in 28 cohort studies. *American Journal of Epidemiology*, *160*, 34–45.

Beale, E. M., & Little, R. J. (1975). Missing values in multivariate analysis. *Journal of the Royal Statistical Society. Series B (Methodological)*, (pp. 129–145).

Bloomfield, P. (1970). Spectral analysis with randomly missing observations. *Journal of the Royal Statistical Society. Series B (Methodological)*, (pp. 369–380).

Bondon, P. (2005). Influence of missing values on the prediction of a stationary time series. *Journal of Time Teries Analysis*, *26*, 519–525.

Bondon, P., & Bahamonde, N. (2012). Least squares estimation of ARCH models with missing observations. *Journal of Time Series Analysis*, *33*, 880–891.

Box, G., Jenkins, G., & Reinsel, G. (2008). *Time Series Analysis: Forecasting and Control*. (4th ed.). Prentice Hall.

Brockwell, P., & Davis, R. (2002). *Introduction to Time Series and Forecasting*. (2nd ed.). Springer Verlag.

Colantonio, A., Di Pietro, R., Ocello, A., & Verde, N. V. (2010). Abba: Adaptive bicluster-based approach to impute missing values in binary matrices. In *Proceedings of the 2010 ACM Symposium on Applied Computing* (pp. 1026–1033). ACM.

Dempster, A., & Rubin, D. (1977). Maximum likelihood from incomplete data via the algorithm em. *Journal of the Royal Statistical Society, B*, *37*, 211–252.

Drake, C., Knapik, O., & Leśkow, J. (2014). Em-based inference for cyclostationary time series with missing observations. In *Cyclostationarity: Theory and Methods* (pp. 23–35). Springer.

Dunsmuir, W., & Robinson, P. M. (1981a). Asymptotic theory for time series containing missing and amplitude modulated observations. *Sankhyā: The Indian Journal of Statistics, Series A*, (pp. 260–281).

Dunsmuir, W., & Robinson, P. M. (1981b). Estimation of times series models in the presence of missing data. *Journal of the American Statistical Association*, *76*, 560–568.

Dunsmuir, W., & Robinson, P. M. (1981c). Parametric estimators for stationary time series with missing observations. *Advances in Applied Probability*, *13*, 129–146.

Efromovich, S. (2014). Efficient non-parametric estimation of the spectral density in the presence of missing observations. *Journal of Time Series Analysis*, *35*, 407–427.

Farhangfar, A., Kurgan, L. A., & Pedrycz, W. (2004). Experimental analysis of methods for imputation of missing values in databases. In *Intelligent Computing: Theory and Applications II* (pp. 172–183). International Society for Optics and Photonics volume 5421.

Farhangfar, A., Kurgan, L. A., & Pedrycz, W. (2007). A novel framework for imputation of missing values in databases. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, *37*, 692–709.

Ghazal, M., & Elhassanein, A. (2006). Periodogram analysis with missing observations. *Journal of Applied Mathematics and Computing*, *22*, 209–222.

Greenland, S., & Rothman, K. J. (1998). *Modern epidemiology*. (2nd ed.). Philadelphia, Lippincott-Raven.

Hartley, H., & Hocking, R. (1971). The analysis of incomplete data. *Biometrics*, (pp. 783–823).

Hies, T., Treffeisen, R., Sebald, L., & Reimer, E. (2000). Spectral analysis of air pollutants. part 1: elemental carbon time series. *Atmospheric Environment*, *34*, 3495–3502.

Iglesias, P., Jorquera, H., & Palma, W. (2005). Data analysis using regression models with missing observations and long memory: an application study. *Computational statistics and Data Analysis*, *50*, 2028–2043.

Instituto Brasileiro de Geografia e Estatística (2014). *Banco de dados. Cidades*. Rio de Janeiro. URL: http://www.cidades.ibge.gov.br/xtras/home.php.

Instituto Estadual de Meio Ambiente e Recursos Hídricos do Estado do Espírito Santo (2014). *Relatório da qualidade do ar da Região da Grande Vitória*. Vitória. URL: http://www.meioambiente.es.gov.br/download/Relat%C3%B3rio_Anual_de_Qualidade_do_Ar_2013.pdf.

Iordache, Ş., & Dunea, D. (2013). Cross-spectrum analysis applied to air pollution time series from several urban areas of romania. *Environmental Engineering & Management Journal (EEMJ)*, *12*.

Jiang, J., & Hui, Y. (2004). Spectral density estimation with amplitude modulation and outlier detection. *Annals of the Institute of Statistical Mathematics*, *56*, 611–630.

Junger, W. L. (2008). *Análise, imputação de dados e interfaces computacionais em estudos de séries temporais epidemiológicas. PhD Dissertation*. Programa de Pós-graduação em Saúde Coletiva, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, Brazil.

Junger, W. L., & Leon, A. P. (2015). Imputation of missing data in time series for air pollutants. *Atmospheric Environment*, *102*, 96–104.

Junninen, H., Niskaa, H., Tuppurainenc, K., Ruuskanena, J., & Koleh-Mainen, M. (2004). Methods for imputation of missing values in air quality data sets. *Atmospheric Environment*, *38*, 2895–2907.

Lakshminarayan, K., Harp, S. A., & Samad, T. (1999). Imputation of missing data in industrial databases. *Applied intelligence*, *11*, 259–275.

Lévy-Leduc, C., Boistard, H., Moulines, E., Taqqu, M. S., & Reisen, V. A. (2011a). Large sample behavior of some well-known robust estimators under long-range. *Statistics*, *45*, 59–71.

Lévy-Leduc, C., Boistard, H., Moulines, E., Taqqu, M. S., & Reisen, V. A. (2011b). Robust estimation of the scale and of the autocovariance function of gaussian short-and long-range dependent processes. *Journal of Time Series Analysis*, *32*, 135–156.

Little, R. J. (1992). Regression with missing x's: a review. *Journal of the American Statistical Association*, *87*, 1227–1237.

Little, R. J., & Rubin, D. B. (2014). *Statistical analysis with missing data* volume 333. John Wiley & Sons.

Mcknight, P. C., Mcknight, K. M., & Figueredo, A. J. (2007). *Missing data: a gentle introduction*. New York. The Guilford Press.

Metaxoglou, K., & Smith, A. (2007). Maximum likelihood estimation of varma models using a state-space em algorithm. *Journal of Time Series Analysis*, *28*, 666–685.

Norazian, M. N., A., S. Y., N., A. R., & M., B. A. M. (2008). Estimation of missing values in air pollution data using single imputation techniques. *ScienceAsia*, *34*, 341–345.

Nunes, L. N., Klück, M. M., & Fachel, J. M. G. (2009). Uso da imputação múltipla de dados faltantes: uma simulação utilizando dados epidemiológicos. *Cadernos de Saúde Pública*, *25*, 268–278.

Parzen, E. (1963). On spectral analysis with missing observations and amplitude modulation. *Sankhya*, *25*, 383–392.

Plaia, A., & Bondì, A. L. (2006). Single imputation method of missing values in environmental pollution data sets. *Atmospheric Environment*, *40*, 7316–7330.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria. URL: https://www.R-project.org/.

Reisen, V. A. (1994). Estimation of the fractional difference parameter in the ARIMA(p,d,q) model using the smoothed periodogram. *Journal of Time Series Analysis*, *15*, 335–350.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581–592.

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys* volume 81. John Wiley & Sons.

Schafer, J. L. (1997a). *Analysis of incomplete multivariante data*. London, Chapman & Hall.

Schafer, J. L. (1997b). *Analysis of incomplete multivariate data*. Chapman and Hall/CRC.

Sebald, L., Treffeisen, R., Reimer, E., & Hies, T. (2000). Spectral analysis of air pollutants. part 2: ozone time series. *Atmospheric Environment*, *34*, 3503–3509.

Shin, D. W., & Sarkar, S. (1995). Estimation for stationary ar(1) models with non-consecutively observed samples. *Sankhyā*, *57*, 287–298.

Takeuchi, K. (1995). A comment on recent development of economic data analysis. *The 63rd anual meeting of Japan Statistical Society*, .

Tchepel, O., & Borrego, C. (2010). Frequency analysis of air quality time series for traffic related pollutants. *Journal of Environmental Monitoring*, *12*, 544–550.

Tchepel, O., Costa, A., Martins, H., Ferreira, J., Monteiro, A., Miranda, A., & Borrego, C. (2010). Determination of background concentrations for air quality models using spectral analysis and filtering of monitoring data. *Atmospheric Environment*, *44*, 106–114.

Toda, H. Y., & Makenzie, C. (1999). Lm tests for unit roots in the presence of missing observations: small sample evidence. *Mathematics and computers in simulation*, *48*, 457–468.

Toloi, C., & Morettin, P. A. (1993). Spectral analysis for amplitude-modulated time series. *Journal of Time Series Analysis*, *14*, 409–432.

Van Buuren, S., van Mulligen, E., & Brand, J. (1994). Routine multiple imputation in statistical databases. In *Scientific and Statistical Database Management, 1994. Proceedings., Seventh International Working Conference on* (pp. 74–78). IEEE.

Wei, W. (2006). *Time Series Analysis: Univariate and Multivariate Methods*. Addison Wesley.

Wu, C.-H., Wun, C.-H., & Chou, H.-J. (2004). Using association rules for completing missing data. In *Hybrid Intelligent Systems, 2004. HIS'04. Fourth International Conference on* (pp. 236–241). IEEE.

Xia, Y., Fabian, P., Stohl, A., & Winterhalter, M. (1999). Forest climatology: estimation of missing values for bavaria, germany. *Agricultural and Forest Meteorology*, *96*, 131–144.

Yajima, Y., & Nishino, H. (1999). Estimation of the autocorrelation function of a stationary time series with missing observations. *The Indian Journal of Statistics*, *61*, 189–207.

Spectral approaches for time series with missing data: an application to air pollution data

Wanderson de Paula Pinto[1], Valdério Aselmo Reisen[1,2]

[1] *Graduate Program in Environmental Engineering (PPGEA) - Federal University of Espírito Santo, Brazil.*

[2] *Department of Statistics, Federal University of Espírito Santo, Brazil.*

Abstract

In this paper, two estimators were proposed for the time series spectrum in the presence of missing data. These estimators extend to complete data cases. The asymptotic properties of the periodogram estimators were established and their empirical properties were investigated for finite samples under different error reporting scenarios. The performance of the estimators was evaluated through Monte Carlo simulations. The numerical results showed that, as expected, under the scenario for time series without missing data, the classical method and those proposed in this work presented similar results. On the other hand, in the presence of missing data, while the classical spectral density estimators may not be applied without a previous treatment of the data, the proposed methods presented good results for missing data percentages smaller or equal to 20%. The efficiency of the proposed methods decreases as the percentage of missing data increases, respectively, to 30% and 40%. As a practical application, the proposed methodology was used to estimate the time series spectrum of the daily average concentrations of the pollutant PM_{10} in the Greater Vitória Region, Espírito Santo, Brazil. This series generally has interesting characteristics such as seasonality and missing data.

Keywords: Lomb-Scargle periodogram, Amplitude modulate, Missing data, PM_{10} pollutant.

1. Introduction

Basically, the time series analysis has two usual approaches: the time and the frequency domain. In both, the objective is to construct models for the series to investigate its generative mechanism, to look for relevant characteristics, such as periodicities and tendency, among others, in order to describe its behavior, what allows, for example, to establish a coherent forecast method. The time domain approach, with methods proposed by the statisticians George Box and Gwilym Jenkins in the '70s, are vastly used in several areas of science, such as in the time series analysis of atmospheric pollutant concentrations. The frequency domain approach is less used but has relevant importance when applied to time series with periodic structure.

For more than a century, a statistical tool has been widely used to find the periodic component phenomena in time series: the periodogram, introduced by Schuster (1898). Its basic functionality plots, for each frequency λ, the ordinate $I(\lambda)$, which estimates the contribution of that frequency to the series. Thus, in a superficial way, large values of the periodogram ordinates imply significant frequencies in the series and small values, correspond to little significant frequencies. Usually, the frequencies of interest are those that produce "peaks" in the periodogram graph. Later, in this paper, the mathematical concept of the periodogram will be formalized.

The theory of this paper was mainly developed motivated by the study and modeling the real time series related to the daily average concentrations of the pollutant PM_{10} in the Greater Vitória Region, Espírito Santo, Brazil. These series present a frequent problem of missing data in its structure, due to fails in the air quality monitoring network. These fails generally occur because the equipment for measuring the concentrations of atmosphere pollutants may have defects that make it impossible to operate for some time, causing data loss. Data analysis, including only the available observations without a statistical treatment for the missing data, can produce a false estimate of the effect measure and underestimate its accuracy Junger (2008). Therefore, several authors have developed alternative methodologies for the analysis of time series with missing data. For example, Toda and Makenzie (1999), Junninen et al. (2004), Iglesias et al. (2005), Plaia and Bondì (2006) and Norazian et al. (2008), use the imputation technique, whilst Metaxoglou and Smith (2007), Drake et al. (2014) and Junger and Leon (2015) suggest the use of Expectation-Maximization (EM) algorithms to mitigate this problem. However, this last approach has the disadvantage of assuming a specific

(usually normal) distribution for the data.

The periodicity is another usual characteristic of the series related to the daily average concentrations of some air pollutants. In the literature, specifically in the analysis of environmental time series, the spectral analyses is mostly used as a tool to investigate the periodicities in the analyzed data. For example, Hies et al. (2000) presented a method, using spectral analysis, to analyze time series of elemental carbon for different sources of atmospheric pollution in Berlin. Sebald et al. (2000) used the spectral analysis to investigate the tropospheric ozone formation and decomposition processes in Germany and Kumar and De Ridder (2010) estimated a generalized autoregressive conditional heteroscedasticity (GARCH) model associated with the FFT-ARIMA (fast moving Fourier-autoregressive integrated moving average) to predict ozone concentrations in two European cities, Brussels and London, as examples of the application of spectral analysis in atmospheric pollution data.

The classic spectral analysis is rarely applied to air pollution data because this data generally present a large quantity of missing data or have an irregular sampling interval or even both. The discrete Fourier transform, used in the classic periodogram, requires equally spaced time series sampling and does not tolerate missing values (Priestley, 1981). Some methodologies use data imputation to fill the missing data before doing the spectral analysis. This, however, is not entirely satisfactory because it modifies the statistical properties of the data by entering artificial data Leroy (2012).

In this paper, to overcome these limitations, two methodological approaches were used to estimate the time series spectrum with missing data. The first, proposed by Parzen (1963), the estimator was constructed based on the autocovariance function calculated using the modulated amplitude process, where the analysis is performed by means of an alternative time series $\{X_t = C_t Y_t\}$, where $\{C_t\}_{t\in\mathbb{Z}}$ represents the censoring process. If $\{Y_t\}$ is observed at the instant t, then $\{C_t\}$ gets one, otherwise $\{C_t\}$ is filled with zero. The asymptotic properties of the estimator of the autocovariance function, proposed by Parzen (1963), were investigated in Dunsmuir and Robinson (1981) under several assumptions about the noise of linear representation $\{\epsilon_t\}_{t\in\mathbb{Z}}$, assumed that $\{C_t\}_{t\in\mathbb{Z}}$ is asymptotically stationary. More recently, Yajima and Nishino (1999); Pinto (2013) compares three estimators of the autocorrelation function for a stationary process with missing observations. They impose the same assumptions on $\{\epsilon_t\}_{t\in\mathbb{N}}$ as those in Dunsmuir and Robinson (1981).

The second statistical methodology proposed, the Lomb-Scargle periodogram, was introduced for the first time in astrophysics, and it allows to estimate the periodogram of the time series with missing data, or unequally spaced sampled, without using an imputation technique to replace the missing values. When studying variables in astronomy, Lomb (1976) proposed a way to find periodicities in data not equally spaced. In an attempt to find an alternative to impute pseudo-data in sinusoidal models, Lomb (1976) proposed to use least squares for sinusoidal curves. Scargle (1982) extended Lomb's work by defining the Lomb-Scargle periodogram and deriving its null distribution. Press and Rybicki (1989) proposed a practical mathematical formulation. It is noteworthy that few studies have applied the Lomb-Scargle periodogram to air pollution data (Hocke and Kämpfer, 2009; Dutton et al., 2010; Bowdalo et al., 2016), and no one of then evaluated the efficiency of this method for different percentages of missing data.

In this context, the objective of this work was to evaluate, through a Monte Carlo simulation study, methodologies for estimation of the time series periodogram with missing data and to apply these methodologies to make the spectral analysis of the data of atmospheric pollutant concentrations monitored in the Greater Vitória Region, Espírito Santo, Brazil.

This paper is organized as follows. The section 2 presents the methodology for the spectral analysis of time series of atmospheric pollutant concentrations with missing data. Section 3 presents the simulations and the empirical studies. Section 4 deals with the statistical analysis of time series of daily PM_{10} concentrations. Finally, section 5 presents the paper conclusions.

2. Material and methods

2.1. Estimation of the periodogram with amplitude modulation

Let $\{Y_t\}_{t\in\mathbb{Z}}$ be a stationary time series with zero mean and autocovariance function $\gamma(l) = \mathbb{E}[Y_t, Y_{t+l}]$, $l = 0, 1, \cdots$, which satisfies

Assumption 1.

$$\sum_{l=-\infty}^{\infty} |\gamma(l)| < \infty. \tag{1}$$

The aim of this article is to study autocovariance and spectral estimation when in the time series $\{Y_t\}_{t\in\mathbb{Z}}$ some observations may be missed. In this

context, the data X_t are not observed for each $k = 1, 2, \cdots, N$ but rather only for $t = t_1, t_2, \cdots, t_m$, where $t_j - t_{j-1} \leq 1$. Following Parzen (1963), we introduce the amplitude modulated observations

$$X_t = C_t Y_t \tag{2}$$

where

$$C_t = \begin{cases} 1 & \text{if } Y_t \text{ is observed} \\ 0 & \text{if } Y_t \text{ is missing.} \end{cases}$$

Throughout the paper, it is considered that the following assumptions are true. These properties are essential to allow the retrieval of the covariance structure of $\{Y_t\}_{t\in\mathbb{Z}}$ (Bahamonde et al., 2010).

Assumption 2. *$\{C_t\}_{t\in\mathbb{Z}}$ is a real, deterministic sequence, asymptotically stationary, satisfying $\sum_t |C_{t+1} - C_t| < \infty$, and $\{X_t\}_{t\in\mathbb{Z}}$ is a strictly stationary process with $\mathbb{E}[|X_t|^k] < \infty$, $k > 0$. (Toloi and Morettin, 1993).*

Assumption 3. *$\{C_t\}_{t\in\mathbb{Z}}$ and $\{X_t\}_{t\in\mathbb{Z}}$ are independent, strictly stationary processes, with $\mathbb{E}[|C_t|^k] < \infty$, $k > 0$ and $\mathbb{E}[|X_t|^k] < \infty$, $k > 0$ (Yajima and Nishino, 1999; Toloi and Morettin, 1993).*

Assumption 4. *$\{C_t\}_{t\in\mathbb{Z}}$ and $\{X_t\}_{t\in\mathbb{Z}}$ is a sequence of independent and identically distributed random variables with a non-zero mean μ_a and strictly positive variance σ_C^2. X_t is as strictly stationary process with $\mathbb{E}[|X_t|^k] < \infty$, $k > 0$, independent of $\{C_t\}$ (Toloi and Morettin, 1993).*

In order that the following convergence holds

$$\overline{C}_N = \frac{1}{N}\sum_{t=1}^{N} C_t \xrightarrow{a.s.} \mu_C \tag{3}$$

we shall assume that the sequence $\{C_t\}_{t\in\mathbb{Z}}$ is ergodic.

We set $\mu_C = \mathbb{E}[C_0]$ and $\gamma_C(l) = Cov[C_0, C_l]$ and for notational convenience, we set

$$v(l) = \mathbb{E}[C_0 C_l] = \gamma_C(l) = \mu_C^2.$$

According to Bahamonde and Doukhan (2016) the following assumption is a condition for estimating the covariance function.

Assumption 5. *(Bahamonde and Doukhan, 2016)* $\mathbb{E}[C_0] \neq 0$ *and* $\mathbb{E}[C_0 C_l] \neq 0$ *for each* $l = 0, 1, \cdots$.

Remark 1. *(Bahamonde and Doukhan, 2016) A Assumption 5 essentially means that we indeed to observe a reasonable amount of data. The condition 5 holds if* $\mathbb{E}[C_0] = \mathbb{P}(C_0 = 1) \neq 0$. *This follows from ergodicity which holds eg. for the case of independently distributed C.*

We denote by $\bar{X}_N$, $\bar{Y}_N$ the sample means of $(X_t)_{t=1}^N$ and $(Y_t)_{t=1}^N$. Usual estimates of the autocovariance coefficients $\gamma_Y(l) = Cov[Y_t, Y_{t+l}]$ are $\hat{\gamma}_{X,N}(l)$ and $\hat{v}_N(l)$ as defined:

$$\hat{\gamma}_{X,N}(l) = \begin{cases} \frac{1}{N-l}\sum_{t=1}^{N-l} X_t X_{t+l} & \text{if} \quad 0 \leq l < N, \\ \frac{1}{N-l}\sum_{t-l}^{N} X_t X_{t+l} & \text{if} \quad -N < l < 0. \end{cases}$$

$$\hat{v}_{C,N}(l) = \begin{cases} \frac{1}{N-l}\sum_{t=1}^{N-l} C_t C_{t+l} & \text{if} \quad 0 \leq l < N, \\ \frac{1}{N-l}\sum_{t-l}^{N} C_t C_{t+l} & \text{if} \quad -N < l < 0. \end{cases}$$

Both estimates are obtained from the observations $\{X_1, \cdots, X_N\} = \{Y_1 C_1, \cdots, Y_N C_N\}$ according to Equation 2 and they are unbiased (Bahamonde et al., 2010). The following estimator of the theoretical ACF of interest is defined under the Equation 2 and was introduced by Parzen (1963),

$$\hat{\gamma}_{Y,N}(l) = \frac{\hat{\gamma}_{X,N}(l)}{\hat{v}_{C,N}(l)}, \quad \text{if} \quad \hat{v}_{C,N}(l) \neq 0. \tag{4}$$

Note that Assumption 5 implies together with any law of large number type that

$$\lim_{N\to\infty} \mathbb{P}(\hat{v}_{C,N}(l) = 0) \neq 0$$

thus, these empirical covariances are well defined.

The correlation function $\rho_Y(h)$ is estimated by

$$\hat{\rho}_{Y,N}(l) = \frac{\hat{\gamma}_{Y,N}(l)}{\hat{v}_{Y,N}(0)}. \tag{5}$$

According to Bahamonde and Doukhan (2016) the construction of the estimator $\hat{\gamma}_{Y,N}(l)$ defined by equation 4 assumes that the observed process $\{X_t\}_{t\in\mathbb{Z}}$is centered. This assumption is convenient to derive asymptotic behavior, but according to Bahamonde and Doukhan (2016) it is not necessarily

the best formulation from the theoretical point of view. The authors propose an alternative estimate representation for the time series sample ACF in the presence of missing data given by:

$$\widetilde{\gamma}_{Y,N}(l) = \hat{\gamma}_{Y,N}(l) - \{\overline{X}_N\}^2. \tag{6}$$

Using the fact that C_t is a non-negative and bounded random variables we derive that $\widetilde{\gamma}_{Y,N}(l) - \hat{\gamma}_{Y,N}(l)$ is arbitrarily small. Therefore $\widetilde{\gamma}_{Y,N}(l) - \hat{\gamma}_{Y,N}(l) = o_{\mathbb{P}}$ so that limiting distributional results proved for $\hat{\gamma}_{Y,N}(l)$ will hold for $\widetilde{\gamma}_{Y,N}(l)$ (Bahamonde and Doukhan, 2016).

Theorem 1. *Let $\{Y_t\}_{t\in\mathbb{Z}}$ be a real valued, stationary sequence time series of square integrable observed, satisfy Assumptions 2 to 5, with censored data. Let the modulating process $\{C_t\}_{t\in\mathbb{Z}}$be a nonnegative bounded stationary process. We assume that $\{C_t\}_{t\in\mathbb{Z}}$ be independent of the process $\{Y_t\}_{t\in\mathbb{Z}}$. Then,*

$$\hat{\gamma}_{Y,N}(l) - \gamma(l) \xrightarrow{a.s.} 0 \quad as \quad N \longrightarrow \infty. \tag{7}$$

See the proof in Bahamonde et al. (2010).

2.1.1. Spectral density function

Let $\{Y_t\}_{t\in\mathbb{Z}}$ a stationary process with zero mean and autocovariance function, $\gamma(l)$, satisfying a condition asymptotic independence, in the sense that time-dependent values of the process are poorly dependent, which can be expressed in the form Assumption 1. For a given time series $\{Y_1 \cdots, Y_N\}$, the classical periodogram, is defined by

$$I_N(\lambda) = \frac{1}{2\pi N}\left|\sum_{t=1}^{N} Y_t e^{-i\lambda t}\right|^2, \quad \lambda \in [-\pi, \pi]. \tag{8}$$

Under Assumption 1, the spectral density of $Y_t, t = 1, 2, \cdots$, is defined as the Fourier transform of $\gamma(l)$, that is,

$$f_Y(\lambda) = \frac{1}{2\pi}\sum_{l=-\infty}^{\infty} \gamma(l)e^{-i\lambda l}, -\pi \leq \lambda \leq \pi, \tag{9}$$

with $e^{i\lambda} = cos\lambda + isen\lambda$ and $i = \sqrt{-1}$. It is noted that, from the equation 9 follows

$$\gamma(l) = \int_{-\pi}^{\pi} e^{i\lambda l} f(\lambda)\mathrm{d}\lambda, l \in \mathbb{Z}, \tag{10}$$

that is, the sequence $\gamma(l)$ can be recovered from $f(\lambda)$ using the inverse Fourier transform.

If we place $l = 0$ in the equation 10, we have

$$\gamma(0) = \mathrm{Var}(Y_t) = \int_{-\pi}^{\pi} f(\lambda)\mathrm{d}\lambda \tag{11}$$

which shows that the spectrum $f(\lambda)$ may be interpreted as the decomposition of the variance of a process. The term $f(\lambda)\mathrm{d}\lambda$ is the contribution to the variance attributable to the component of the process with frequencies in the interval $(\lambda, \lambda + \mathrm{d}\lambda)$. A peak in the spectrum indicates an important contribution to the variance from the components at frequencies in the corresponding interval Wei (2006). A natural estimator of the spectral density function is the periodogram, defined in the next subsection.

2.1.2. Spectral density function under missing data

In this section we consider the estimation of functionals of the spectral density function from time series in presence of missing observations. Using the estimator of the covariance of $\{Y_t\}_{t\in\mathbb{Z}}$ given by $\{\hat{\gamma}_{Y,N}(l)\}$, we introduce estimates of the spectral density.

Definition 1. *For the time series X_t with modulated amplitude, its generalized periodogram is defined as*

$$\hat{I}^{AM}_{Y,N}(\lambda_j) = \frac{1}{2\pi}\sum_{l\in\mathbb{N}} \hat{\gamma}_{Y,N}(l)e^{-i\lambda_j l} = \frac{1}{2\pi}\sum_{|l|<N} \frac{\hat{\gamma}_{X,N}(l)}{\hat{v}_{C,N}(l)}e^{-i\lambda_j l}, \tag{12}$$

where $\lambda_j = \frac{2\pi l}{N}, (l = 0, 1, \cdots, N)$ are Fourier frequencies.

The generalized periodogram has properties similar to the common periodograms. It can be used to construct a variety of tests for hidden periodicities following the ideas of common tests based on the periodogram in equation 8 (Jiang and Hui, 2004)

Proposition 1. *If $\{Y_t\}$ is ergodic and normal with mean zero, then the following results hold for $j = 1, 2, \cdots, N$*

1. *Asymptotic unbiasedness:* $\mathbb{E}[\hat{I}^{AM}_{Y,N}(\lambda_j)] = f_Y(\lambda_j) + O\left(\frac{1}{\sqrt{N}}\right)$.

2. *Variance approximation:* $Var[\hat{I}^{AM}_{Y,N}(\lambda_j)] = f^2_Y(\lambda_j) + \tau^2_Y(\lambda_j) + o(1)$.

3. *Asymptotic independence: for* $\lambda_k \neq \lambda_j$,

$$Cov\left(\hat{I}^{AM}_{Y,N}(\lambda_k), \hat{I}^{AM}_{Y,N}(\lambda_j)\right) = O\left(\frac{1}{N}\right).$$

The detailed proof of Proposition 1 can be found in (Jiang and Hui, 2004), as well as the definition of $\tau^2_Y(\lambda_j)$.

Note that the generalized periodogram is an asymptotically unbiased estimator of 9 at Fourier frequencies. However, it is inconsistent in estimating 9. A consistent estimator of 9, for $\lambda \in [0, \pi]$, can be obtained via directly smoothing the data $\{(\lambda_j, \hat{I}^{AM}_{Y,N}(\lambda_j)), l = 1, 2, \cdots, N\}$ using a locally weighted average (Jiang and Hui, 2004). Following Fan and Kreutzberger (1998); Jiang and Hui (2004), we run the local linear regression smoother to the data, and obtain the estimator

$$\hat{f}_Y(\lambda) = \sum_{l=1}^{N} \left(\frac{\lambda - \lambda_j}{h}\right) \hat{I}^{AM}_{Y,N}(\lambda_j), \tag{13}$$

where

$$K_N(t) = \frac{1}{N} \frac{s_{N,2} - ht \cdot s_{N,1}}{s_{N,0} \cdot s_{N,2} - s^2_{N,1}} K(t)$$

where $K(t)$ is a kernel function and $s_{N,l} = \frac{1}{Nh} \sum_{l=1}^{N} K\left(\frac{\lambda - \lambda_j}{l}\right) (\lambda - \lambda_j)^l$.

Theorem 2. *Let* $\{Y_1, Y_2, \cdots, Y_N\}$ *be a sample observation of a second order stationary time series* $\{Y_t\}_{t\in\mathbb{Z}}$, *with missing data, satisfying Assumptions 2 to 5. Let* $\hat{I}^{AM}_{Y,N}(\cdot)$ *be an estimator of the spectral density of* $\{Y_t\}_{t\in\mathbb{Z}}$, *where* $\gamma(l)$ *satisfies the Assumption 1. Then,*

$$|\hat{I}^{AM}_{Y,N}(\lambda) - f_Y(\lambda)| \xrightarrow{a.s.} 0 \quad as \quad N \longrightarrow \infty, \quad for \quad \lambda \in [-\pi, \pi]. \tag{14}$$

Due to some analytical complexities, the proof of this result will be left for the version of the paper that will submitted.

2.2. The Lomb-Scargle Periodogram

There are several statistical methodologies in the literature for periodicity research in time series. However, in the vast majority, it is assumed as an initial hypothesis that the data are regularly spaced and statistically homogeneous. The periodogram, as described in equation 8, presents a series of problems when applied in time series not equally spaced or with missing data

(Scargle, 1982, 1989). According to Dilmaghani et al. (2007), the main one is the fact that sines and cosines are no more orthogonal on the set of unequally spatialized observations. This invalidates a derivation of equation 1, which is based on the orthogonality of sines and cosines on the set of natural frequencies (Bloomfield, 2004). The solution is to define a time offset to make the sine and cosine orthogonal on the set of unequally spaced observations for an arbitrary frequency (Dilmaghani et al., 2007).

This change results in the application of the Lomb-Scargle periodogram. Lomb (1976) while studying astronomical data, sought a way to find journals in data not equally spaced. Scargle (1982) continued Lomb's work by defining the Lomb-Scargle periodogram.

The Lomb PSD estimation method is based on the general transform theory which shows that the projection of a signal $Y(t)$ onto one element of an orthonormal base $b_i(t)$ is the value c that minimizes the mean squared error ($\epsilon(c)$) defined as the integral, over the definition interval, of the squared differences between $Y(t)$ and $c \times b_i(t)$. The Lomb method implements this minimization over the unevenly distributed sampled values of $Y(t)$ considering that the basis functions are the Fourier kernel $b_i(t) = e^{\omega_i t}$, with $\omega_i = 2\pi f_i$.

The Lomb method for power spectral density estimation is based on the minimization of the squared differences between the projection of the signal onto the basis function and the signal under study (Lomb, 1976). This method can be generalized to any transform estimation on unevenly sampled signals. Let $\{Y(t)\}_{t\in\mathbb{Z}}$ be the continuous signal under study and $b_i(t)$ an orthogonal basis set that defines the transform. It is well known that the coefficients $c(i)$ that represent $\{Y(t)\}$ in the transform domain are

$$c(i) = \int_{-\infty}^{\infty} Y(t) b_i(t) dt \tag{15}$$

and also that these coefficients $c_i(t)$ are those which minimize the squared error $\epsilon(c_i)$ defined as

$$\epsilon(c_i) = \int_{-\infty}^{\infty} \left[Y(t) - c(i) b_i(t)\right]^2 dt. \tag{16}$$

In dealing with uniformly sampled signals, this formalism becomes its discrete counterpart, which in the case of the Fourier domain is well studied as the discrete-time Fourier transform (DTFT), the discrete-valued version (DFT), and the associated fast algorithm (FFT) to calculate it. When the

signal $Y(t)$ is accessible only at unevenly spaced samples at t_n instants, the solution has generally been to reduce it to a uniformly sampled signal by imputation of the data. However, this method brings some distortion in the spectrum we are estimating. To avoid this problem, Lomb (1976) proposed to estimate the Fourier spectra of an unequally sampled signal by adjusting the model

$$Y(t_n) + \epsilon_n = acos(\omega_i t_n) + bsin(\omega_i t_n) \tag{17}$$

in such a way that the mean squared error ϵ_n is minimized with the proper a and b parameters. This expression is a particularization for real signals of the more general

$$Y(t_n) + \epsilon_n = c(i)e^{j\omega_i t_n}, \tag{18}$$

where $Y(t)$ and c can be complex valued. For any transform, not necessarily the Fourier transform, the expression will be

$$Y(t_n) + \epsilon_n = c(i)b_i(t_n). \tag{19}$$

Minimization of ϵ_n variance (mean squared error) leads to minimization of

$$\sum_{n=1}^{N} |Y(t_n) - c(i)b_i(t_n)|^2 \tag{20}$$

which results in a value for $c(i)$

$$c(i) = \frac{1}{k}\sum_{n=1}^{N} |Y(t_n)b_i^*(t_n)|^2 \quad k = \sum_{n=1}^{N} |b_i(t_n)|^2. \tag{21}$$

This result can be referred to as a generalized Lomb method to estimate transforms of unevenly sampled data (Laguna et al., 1998).

The signal power at index of the transformation will be (Lomb, 1976)

$$P_Y(i) = c(i)\sum_{n=1}^{N} Y(t_n)^* b_i(t_n) = k \times |c(i)|^2 = |\tilde{c}(i)|^2, \tag{22}$$

with $\tilde{c}(i) = c(i)\sqrt{k}$. Lomb (1976) introduces a delay at the basis (*sin* and *cos* rather than exponentials), which becomes in a more efficient estimation algorithm, the Lomb normalized periodogram (Laguna et al., 1998).

Be,

$$\overline{Y} = N^{-1} \sum_{t=1}^{N} (Y_t), \tag{23}$$

and

$$\hat{\sigma}^2 = (N-1)^{-1} \sum_{t=1}^{N} (Y_t - \overline{Y})^2, \tag{24}$$

respectively, the mean and the process variance estimator Y_t. The normalized Lomb-Scargle periodogram at the f frequency is defined as (Press and Rybicki, 1989)

$$P_{Y,N}^{L}(\lambda_j) = (2\hat{\sigma}^2)^{-1} \left\{ \frac{[R(\lambda_j)]^2}{\sum_{t=1}^{N} \cos^2 \lambda_j (t_n - \tau)} + \frac{[I(\lambda_j)]^2}{\sum_{t=1}^{N} \sin^2 \lambda_j (t_n - \tau)} \right\}, \tag{25}$$

where λ_j are Fourier frequencies, $\overline{Y}$ and $\hat{\sigma}^2$ are the mean and variance of the measurements, respectively, defined in equations 23 and 24, while the sums

$$R(\lambda_j) = \sum_{t=1}^{N} (Y_t - \overline{Y}) \cos \lambda_j (t_n - \tau), \tag{26}$$

$$I(\lambda_j) = \sum_{t=1}^{N} (Y_t - \overline{Y}) \sin \lambda_j (t_n - \tau). \tag{27}$$

The frequency-dependent time offset τ is evaluated at each λ via

$$\tau = \frac{1}{2\lambda_j} \arctan \left[\frac{\sum_{t=1}^{N} \sin(2\lambda_j t_n)}{\sum_{t=1}^{N} \cos(2\lambda_j t_n)} \right]. \tag{28}$$

According to Scargle (1982), if t_n becomes $t_n + T_0$, then τ becomes $\tau + T_0$. The Lomb-Scargle periodogram can be used for the analysis of equally spaced time series because its statistical properties are equivalent to the classical periodogram, but for Scargle (1982) the calculation of the periodogram by equation 25 is a little more complicated than by the traditional methodology (equation 8). For Scargle (1982) the secret is to impose continuity on τ as

a function of λ, and to use sufficiently high-frequency resolution so that no phase jumps are missed. It is also important to note that

$$\lim_{\lambda \to 0} \tau(\lambda) = \left(\frac{1}{N}\right) \sum_{i=1}^{N} t_n = \langle t \rangle. \tag{29}$$

Equations (25 to 29) describe the Lomb-Scargle periodogram. These equations were defined according to the Lomb (1976), Scargle (1982), Scargle (1989), Press and Rybicki (1989), Glynn et al. (2006), Dilmaghani et al. (2007), Hocke and Kämpfer (2009), Townsend (2010) and Leroy (2012).

2.2.1. Estimation of autocovariance function via Lomb periodogram

In this section, an autocovariance function estimator is proposed for time series with missing data in the frequency domain using the inverse transform of the Lomb-Scargle periodogram.

Proposition 2. *Let $\hat{\gamma}_N(l)$ be an estimator of the autocovariance function of the sample $\{Y_1, Y_2, \cdots, Y_N\}$ of $\{Y_t\}_{t\in\mathbb{Z}}$ satisfying Assumption 1. If the discrete Fourier transform of Y_t is given by $P^L_{Y,N}(\cdot)$, then*

$$\hat{\gamma}_{Y,N}(l) = \mathcal{F}^{-1}[P^L_{Y,N}(\lambda_j)], \quad -\pi < \lambda_j \leq \pi, \quad for \quad l = 0, \pm 1, \pm 2, \cdots, [N/2-1]. \tag{30}$$

where $\mathcal{F}^{-1}$ represents the inverse Fourier transform. The autocovariance $\hat{\gamma}(l)$ is can be calculated by taking the inverse Fourier transform in 25 using the FFT algorithm. If $N \longrightarrow \infty$ for $l = 1, 2, \cdots, N-1$, then $\hat{\gamma}_{Y,N}(l) - \gamma(l) = O_P\left(\frac{1}{\sqrt{N}}\right)$.

Due to some analytical complexities, the proof of this result will be left for the version of the paper that will submitted.

3. Monte Carlo simulation study

This section reports a Monte Carlo simulation study to investigate the performance of the spectral density estimators of the sample with missing data previously discussed. For the numerical experiments, the data generating process of $\{Y_t\}_{t\in\mathbb{Z}}$ is an autoregressive process of order 1 (AR(1)) as follows:

$$Y_t = \phi Y_{t-1} + \varepsilon_t, \tag{31}$$

where $|\phi| < 1$ and ε_t is a zero mean Gaussian white noise process with variance σ^2. For further details about this process, see Brockwell and Davis (2002), Wei (2006) and Box et al. (2008), among others.

In the simulations, $\phi = 0.3$, $\phi = 0.7$ and $p = 5\%, 10\%, 15\%, 20\%, 30\%$ and 40% are set. The sample size are $N = 100$ and $N = 500$. Two scenarios are considered: (i) samples that have no missing data ($p = 0$) and (ii) samples that have missing data ($p \neq 0$). The results of the empirical means of the periodogram estimators are presented in the Figures 1 to 6 and the root of mean square error values (RMSE) and bias in the Tables 1 to 3. Both of the studied scenarios considered 1000 replications. All simulations were performed using Software R (R Core Team, 2019).

In this paper, the percentage of 5% of missing data was considered as reference of the smallest amount of tolerable missing information. The percentage of 40%, on the other hand, was used to evaluate the estimation methods of the periodogram under extreme conditions of missing information.

The empirical RMSE corresponds to the mean over all values of the mean of RMSE_j for the Fourier frequencies at $j = 1, \cdots, [N/2 - 1]$.

$$\text{RMSE} = \frac{1}{m} \sum_{j=1}^{m} \text{RMSE}_j \tag{32}$$

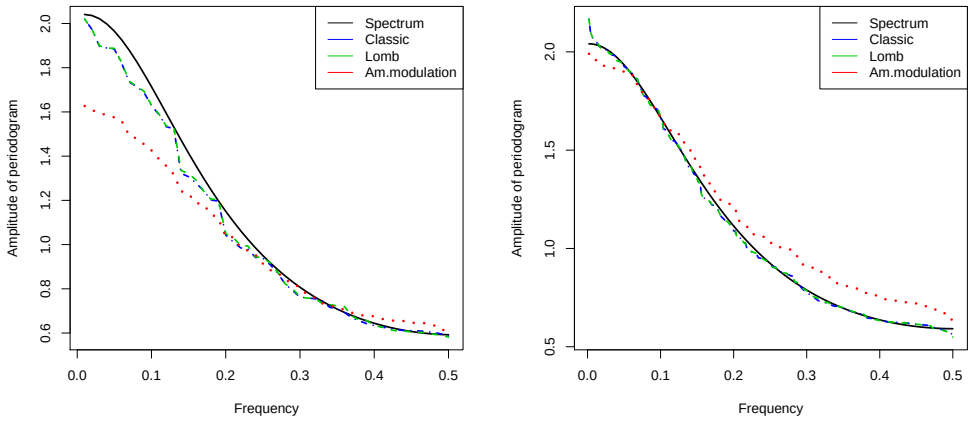
where

$$\text{RMSE}_j = \sqrt{\frac{1}{K} \sum_{k=1}^{K} \left(\hat{f}^{(k)}(\lambda_j) - f(\lambda_j) \right)^2} \tag{33}$$

where $\hat{f}^{(k)}(\lambda_j)$ is the estimated periodogram corresponds to the replication $k (k = 1, \cdots, K)$, $f(.)$ is the spectral density of the $AR(p)$ process, $K = 1000$ and $m = [N/2 - 1]$.

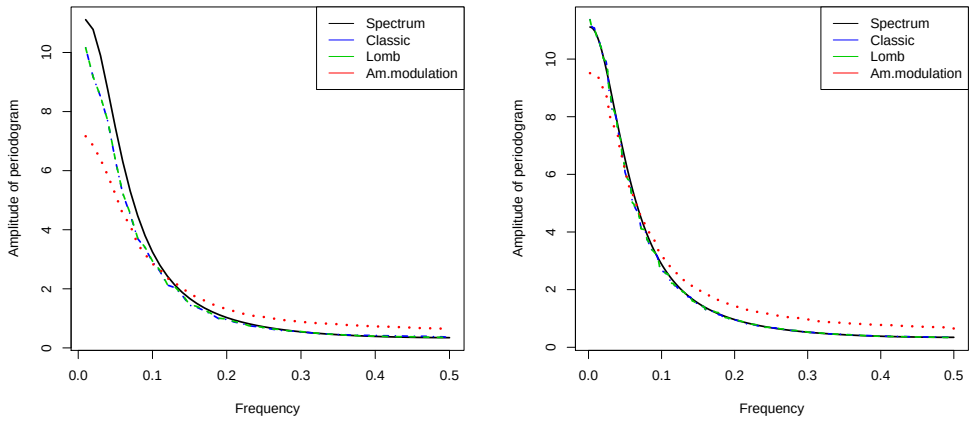
The Figures 1 and 2 show the graphs of the theoretical spectrum of an AR(1) process, with $\phi = 0.3$ (Figure 1) and $\phi = 0.7$ (Figure 2), and the averages of the periodograms estimated by $\hat{I}_N$, $\hat{P}^L_{X,N}$ and $\hat{I}^{AM}_{Y,N}$ for the scenario with no missing data, with $N = 100$ and $N = 500$, respectively. For the case of simulations with samples in the absence of missing data, we observe a similar behavior of the two estimators of the spectrum, indicating that the variability of the series is captured by the estimators. The results suggest that under the hypothesis of $p = 0$ the proposed estimators can be used to estimate the spectrum of stationary processes.

The Table 1 displays the RMSE values defined by Equation 32 and the bias values for the case of the series generated by an AR(1) process, with $\phi = 0.3$ and $\phi = 0.7$, with no missing data. The entries in this table show that, even for a small sample size ($N = 100$), the periodograms $\hat{P}^{L}_{X,N}$ and $\hat{I}^{AM}_{Y,N}$ are fairly well behaved in relation to the classical periodogram (Equation 8). As the sample size increases ($N = 500$), all periodograms presented similar values of RMSE and bias, both for $\phi = 0.3$ and $\phi = 0.7$, as would be expected from the asymptotic results. The periodogram $\hat{I}^{AM}_{Y,N}$ provided smallest estimates for the RMSE, but it is not the best alternative for time series with no missing data, since it tends to underestimate the values when the sample size is small and overestimate when the sample size is large. The estimates obtained with $\hat{P}^{L}_{X,N}$ and $\hat{I}^{AM}_{Y,N}$ were more accurate, which is expected, since that they are closer to the classic periodogram.



(a) (b)

Figure 1: Empirical means of the periodograms, size samples of (a) $N = 100$ and (b) $N = 500$, from the $AR(1)$ process ($\phi = 0.3$) and $p = 0$. The theoretical spectral density is displayed with a solid line.



(a) (b)

Figure 2: Empirical means of the periodograms, size samples of (a) $N = 100$ and (b) $N = 500$, from the $AR(1)$ process ($\phi = 0.7$) and $p = 0$. The theoretical spectral density is displayed with a solid line.

Table 1: RMSE and Bias of the periodograms estimators defined in section 2 for $AR(1)$, with $p = 0$.

Parameter	Estimator		Sample size (N)	
			100	500
$\phi = 0.3$	$I_N(\lambda)$	$\widehat{RMSE}$	0.8223	0.9042
		$\widehat{Bias}$	-0.0421	-0.0064
	$P^L_{Y,N}(\lambda)$	$\widehat{RMSE}$	0.8219	0.9046
		$\widehat{Bias}$	-0.0385	-0.0054
	$\hat{I}^{AM}_{Y,N}(\lambda_j)$	$\widehat{RMSE}$	0.5428	0.5966
		$\widehat{Bias}$	-0.1085	0.0700
$\phi = 0.7$	$I_N(\lambda)$	$\widehat{RMSE}$	1.9729	2.3210
		$\widehat{Bias}$	-0.2181	-0.0329
	$P^L_{Y,N}(\lambda)$	$\widehat{RMSE}$	1.9716	2.3206
		$\widehat{Bias}$	-0.2127	-0.0292
	$\hat{I}^{AM}_{Y,N}(\lambda_j)$	$\widehat{RMSE}$	1.6058	1.551
		$\widehat{Bias}$	-0.2031	0.2450

The discussion of the efficiency of the proposed estimators, in series with different percentages of missing data, is based on the empirical results given

in Tables 2 and 3 and Figures 3, 4, 5 and 6. When compared to the $\hat{I}^{AM}_{Y,N}$, the periodogram $\hat{P}^{L}_{X,N}$ showed the highest values for RMSE in all missing data configurations evaluated. For both estimators, the values of RMSE undergo significant changes when the percentage of missing data exceeds 20%, regardless of sample size and model parameter (see Tables 2 and 3); for example, considering the case presented in Table 2 for $N = 500$, when the percentage is 40% missing data (extreme case of missing information) the RMSE value of the estimates obtained with $\hat{I}^{AM}_{Y,N}$ showed an increase of 108.9% when compared to the estimates obtained with no missing data. The estimates of the RMSE obtained with the estimator $\hat{P}^{L}_{X,N}$, for the scenario of 40% of missing data, showed an increase of 0.32% in relation to the RMSE of series without missing data.

Considering the results presented in Tables 2 and 3 and Figures 3, 4, 5 and 6, even when the periodogram $\hat{P}^{L}_{X,N}$, shows the minimum accurate results of the measurements of accuracy, it is the most suitable for spectral estimation of stationary processes with large percentages of missing data and with expressive serial correlation. On the other hand, the results showed that the estimator $\hat{I}^{AM}_{Y,N}$ is more suitable for time series with low serial correlation. It should be noted (see Figures 3, 4, 5 and 6) that for close to zero frequencies, the estimators showed a slight tendency to underestimate the results, especially for small samples sizes with large percentages of missing data. In series with 40% missing data, a significant decrease in the efficiency of the estimators was evidenced, suggesting the need to develop more efficient alternatives in these scenarios.

The classical periodogram cannot be applied in time series with missing data without a previous treatment of the data, either the imputation or analysis including only the available data. Therefore, its use should be avoided when the series contains missing data, since data analysis including only available data, without statistical treatment for missing data, may produce false estimates of extent of the effect measure and underestimation of its accuracy. On the other hand, the estimators proposed in this work presented accurate estimates, even for a large number of missed observations, so that an alternative methodology for the spectral analysis of series can be used in the presence of missing data.

Table 2: RMSE and Bias of the periodograms estimators defined in section 2 for $AR(1)$, with $\phi = 0.3$ and $p = 1$.

N	Estimator		Percentage of missing data					
			5%	10%	15%	20%	30%	40%
100	$P^L_{Y,N}(\lambda)$	$\widehat{RMSE}$	0.8368	0.8149	0.8509	0.8233	0.8350	0.8778
		$\widehat{Bias}$	-0.0635	-0.0914	-0.1115	-0.1643	-0.2481	-0.3486
	$\hat{I}^{AM}_{Y,N}(\lambda_j)$	$\widehat{RMSE}$	0.5529	0.5748	0.6038	0.6440	0.7383	0.9084
		$\widehat{Bias}$	-0.0577	-0.0538	-0.0209	0.0351	0.1227	0.2382
500	$P^L_{Y,N}(\lambda)$	$\widehat{RMSE}$	0.8942	0.8751	0.8673	0.8612	0.8750	0.9075
		$\widehat{Bias}$	-0.0247	-0.0595	-0.091	-0.1321	-0.2112	-0.3068
	$\hat{I}^{AM}_{Y,N}(\lambda_j)$	$\widehat{RMSE}$	0.6115	0.6788	0.7486	0.7976	1.0010	1.2468
		$\widehat{Bias}$	0.0857	0.1683	0.2370	0.2707	0.4376	0.6237

Table 3: RMSE and Bias of the periodograms estimators defined in section 2 for $AR(1)$, with $\phi = 0.7$ and $p = 1$.

N	Estimator		Percentage of missing data					
			5%	10%	15%	20%	30%	40%
100	$P^L_{Y,N}(\lambda)$	$\widehat{RMSE}$	1.9297	1.9289	1.9355	1.9709	2.0942	2.3358
		$\widehat{Bias}$	-0.3221	-0.4145	-0.5813	-0.6639	-0.9601	-1.3419
	$\hat{I}^{AM}_{Y,N}(\lambda_j)$	$\widehat{RMSE}$	1.6517	1.6900	1.8023	1.8005	1.9498	2.1081
		$\widehat{Bias}$	-0.1693	-0.1850	-0.1291	-0.1099	0.0753	0.1705
500	$P^L_{Y,N}(\lambda)$	$\widehat{RMSE}$	2.1279	2.0613	2.0088	1.9803	1.9894	2.1174
		$\widehat{Bias}$	-0.1292	-0.229	-0.3256	-0.4441	-0.7333	-1.1033
	$\hat{I}^{AM}_{Y,N}(\lambda_j)$	$\widehat{RMSE}$	1.5851	1.6065	1.6802	1.7721	2.0109	2.2798
		$\widehat{Bias}$	0.2987	0.3509	0.4455	0.5465	0.7941	1.0912

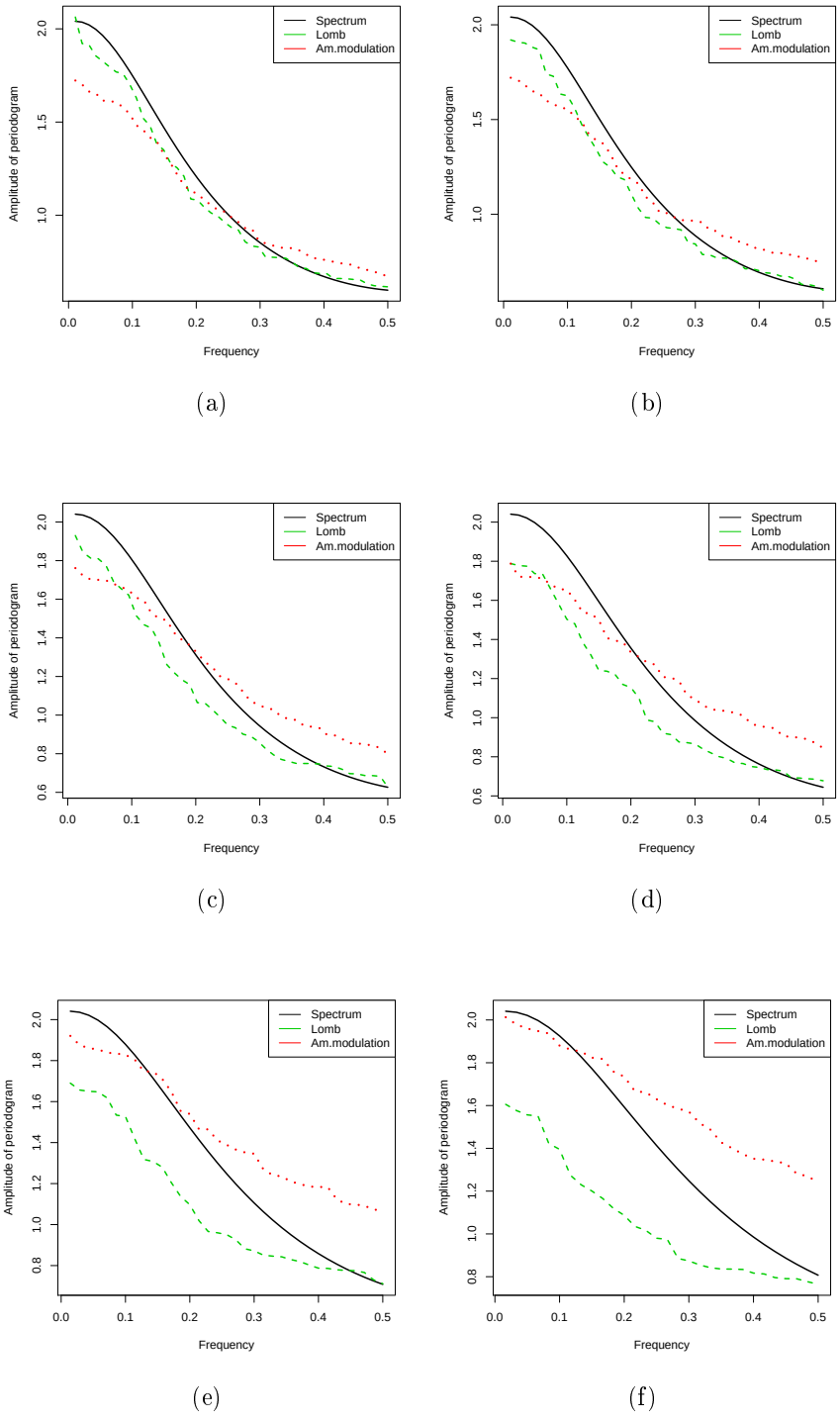


Figure 3: Empirical means of the periodograms, size samples of $N = 100$, from the $AR(1)$ process ($\phi = 0.3$) with (a) $p = 5\%$, (b) $p = 10\%$, (c) $p = 15\%$, (d) $p = 20\%$ (e) $p = 30\%$ and (f) $p = 40\%$. The theoretical spectral density is displayed with a solid line.

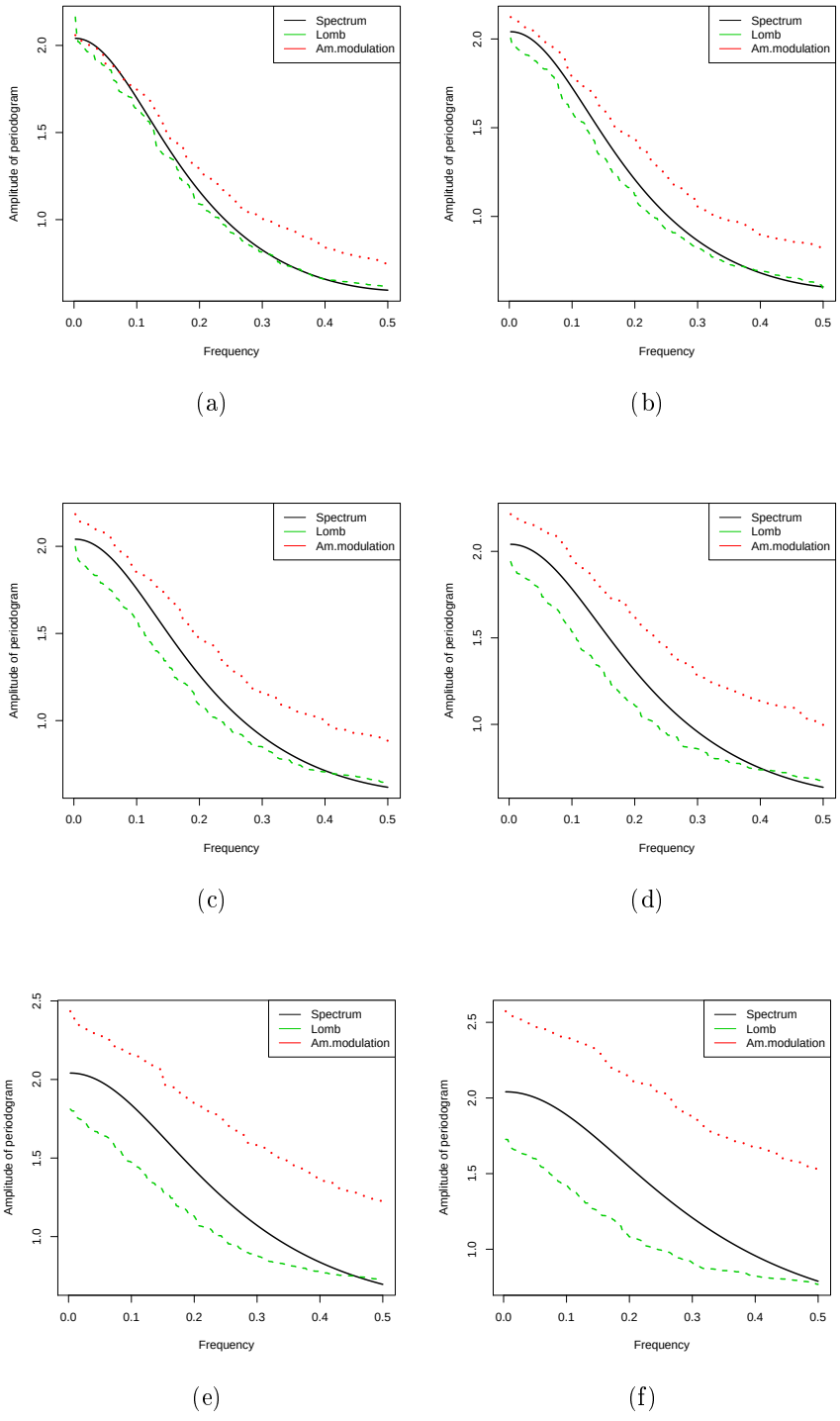


Figure 4: Empirical means of the periodograms, size samples of $N = 500$, from the $AR(1)$ process ($\phi = 0.3$) with (a) $p = 5\%$, (b) $p = 10\%$, (c) $p = 15\%$, (d) $p = 20\%$ (e) $p = 30\%$ and (f) $p = 40\%$. The theoretical spectral density is displayed with a solid line.

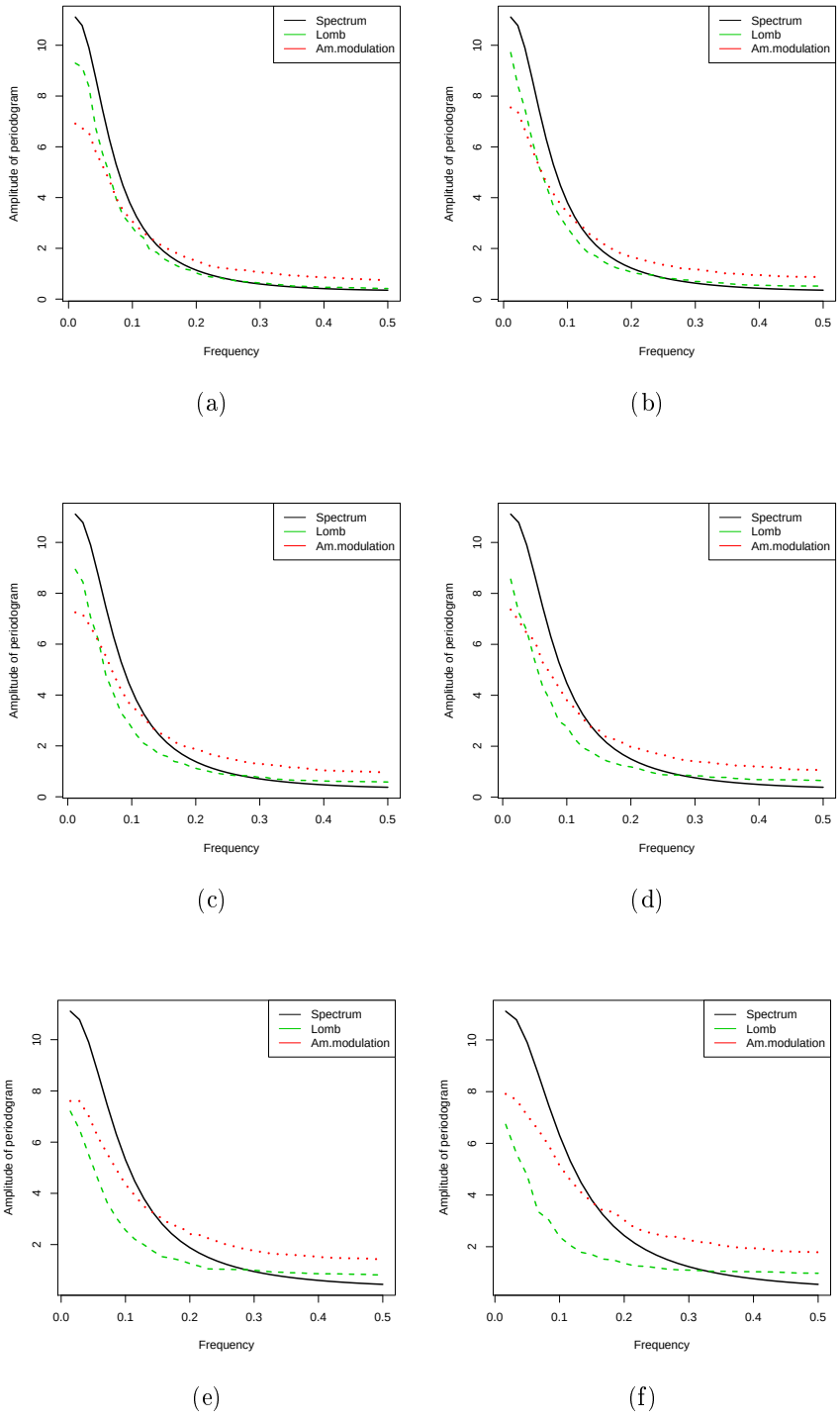


Figure 5: Empirical means of the periodograms, size samples of $N = 100$, from the $AR(1)$ process ($\phi = 0.7$) with (a) $p = 5\%$, (b) $p = 10\%$, (c) $p = 15\%$, (d) $p = 20\%$ (e) $p = 30\%$ and (f) $p = 40\%$. The theoretical spectral density is displayed with a solid line.

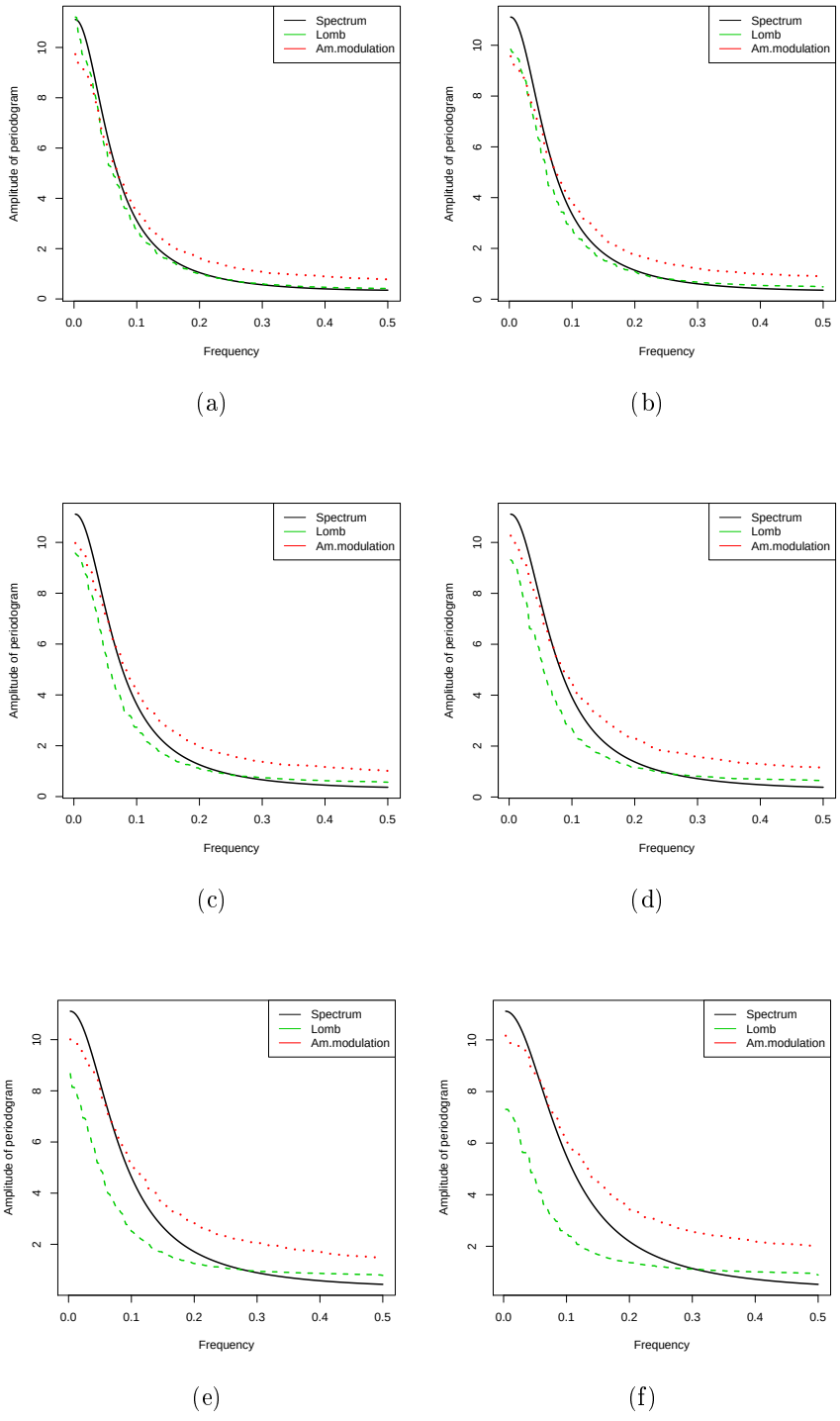


Figure 6: Empirical means of the periodograms, size samples of $N = 500$, from the $AR(1)$ process ($\phi = 0.7$) with (a) $p = 5\%$, (b) $p = 10\%$, (c) $p = 15\%$, (d) $p = 20\%$ (e) $p = 30\%$ and (f) $p = 40\%$. The theoretical spectral density is displayed with a solid line.

4. An application to PM_{10} pollutant

4.1. Area of study

To apply the studied methodology, this paper used data collected in the Region of Greater Vitória (RGV), Espírito Santo, Brasil. RGV is located on the south coast of the Atlantic Ocean (latitude 20° 19S, longitude 40° 20W). The RGV is constituted by the municipalities of Vitória, Vila Velha, Cariacica, Serra and Viana. Because it is situated in the coastal region, the RGV has a warm tropical climate (Aw), with mild and dry winter, and a hot and rainy summer, with average temperatures ranging from 24 °C to 30 °C. According to the Brazilian Institute of Geography and Statistics (Instituto Brasileiro de Geografia e Estatística, 2014), the metropolitan region of Vitória has 1,475,332 inhabitants, covers an area of 1,461 square kilometers and is one of the main centers of urban and industrial development in the state. The region suffers from several environmental problems, among them the deterioration of air quality due to atmospheric emissions by industries and the vehicular fleet.

The RGV has an automatic network for air quality monitoring (RAMQAR), owned by the state institute for the environment and water resources, that was put in service in July, 2000. The network is distributed among nine monitoring stations located in the municipalities that compose the RGV, as follows: Serra with three stations (Laranjeiras, Carapina and Cidade Continental); Vitória with three stations (Jardim Camburi, Enseada do Suá, and Central Vitória); Vila Velha with two stations (Ibes and Central Vila Velha); and Cariacica (Cariacica). The locations of the RAMQAR monitoring stations are shown in Figure 7.

The RAMQAR network monitors the following pollutants: PM_{10}, total suspension particles (TSP), Respirable Particulate Material $PM_{2.5}$ Ozone (O_3), Nitrogen Oxide (NO_x), Nitrogen Dioxide (NO_2), Non-Methane Hydrocarbons (HCnM), Methane (CH_4), Nitrogen Monoxide (NO), Carbon Monoxide (CO), Sulphur dioxide (SO_2) and Total Hydrocarbons (HCT). The following meteorological parameters are also monitored: Scalar wind direction (WD), scalar wind velocity (WV), precipitation (PP), relative humidity (RH), Air temperature (T), atmospheric pressure (P), Standard deviation of wind direction (SIGT) and solar radiation (I). The pollutants and meteorological parameters that RAMQAR monitor at each station are shown in Table 4.

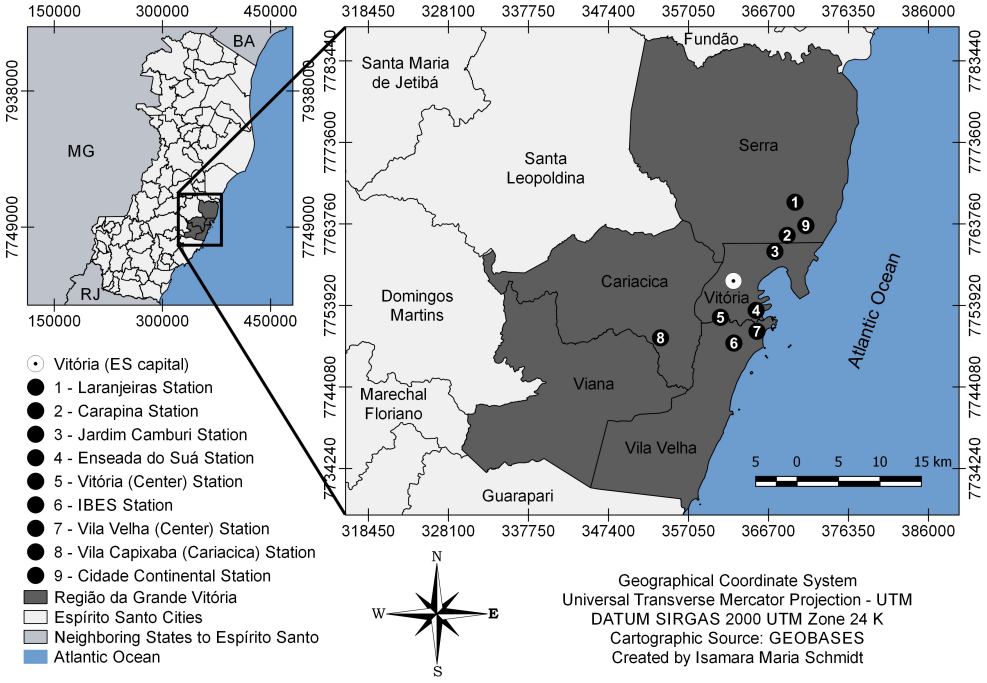


Figure 7: Location of the air quality monitoring stations in the RGV.

Table 4: Meteorological parameters and pollutants monitored at each RAMQAR station.

Stations	$PM_{2.5}$	TSP	PM_{10}	SO_2	CO	NO	NO_2	NO_X	HCT	O_3	CH_4	HCnM	Meteorological
Laranjeiras		X	X	X	X	X	X	X		X			
Carapina		X	X										WD,WV,RH,PP,P,T,I,SIGT
Jardim Camburi		X	X	X		X	X	X					
Enseada do Suá	X	X	X	X	X	X	X	X	X	X	X	X	WD,SIGT,WV
Central (Vitória)		X	X	X	X	X	X	X	X		X	X	
IBES	X	X	X	X	X	X	X	X	X	X	X	X	WD,SIGT,T,RH,WV
Central (Vila Velha)			X	X									
Vila Capixaba (CEASA)		X	X	X	X	X	X	X		X			WD,SIGT,RH,WV,T
Cidade Continental		X	X	X	X	X		X		X			WD,WV

Source: Instituto Estadual de Meio Ambiente e Recursos Hídricos do Estado do Espírito Santo (2014).

4.2. Analysis of PM_{10} concentrations with missing observations

The daily average PM_{10} concentrations (Figure 8) is the data set here analyzed to illustrate the methodology previously discussed. The series is expressed in $\mu g/m^3$ and it was observed in the stations of Carapina (CARA), Laranjeiras (LARA), Jardim Camburi (CAMB), Enseada do Suá (ENSE), Central Vitória (CENVIT), Ibes (IBES), Central Vila Velha (CENVV), Cariacica (CARI) and Cidade Continental (CIDCONT), all of them located in Espírito

Santo, Brazil. These time series were analyzed in the period from January 1st, 2008 to December 31st, 2018.

The choice of this data set was motivated by the fact that, in addition to the importance of modeling and predicting this pollutant in the context of air quality control, all PM_{10} series present missing data (see Figure 8 and Table 4.2), which would require an application of the statistical methodology suggested here for their spectral analysis. The recent works of Reisen et al. (2014a,b, 2017); Fajardo et al. (2018); Pinto et al. (2018); Solci et al. (2019); Reisen et al. (2019), deal with these series for models from different contexts. The main interest of this paper is to show the applicability of the spectral estimators in the presence of missing data in a real data set, that does not meet the standard assumption of the classical methodology of data without missing observations. These observations may be associated with the non-functioning of the contaminant concentration measurement equipment.

It should be mentioned that the PM_{10} series should have a periodic component with seasonal period $s = 7$, since according to Instituto Estadual de Meio Ambiente e Recursos Hídricos do Estado do Espírito Santo (2014), the main source emitting particles in the RGV is vehicles (more than 60% of particulate emissions are linked to particulate resuspension on roads). Thus, there is a variation between the concentrations measured on weekdays and on weekend days. The flow of vehicles is greater during the days of the week. Data may also have other significant periodic components that are generally related to annual deterministic trends in the series, and these components will not be studied in this paper.

Table 4.2 presents the descriptive statistics of the inhalable particulate matter time series with their number and percentage of missing data. It is important to observe the high percentages of missing data that these series have justify the use of the Lomb-Scargle periodogram. In general, by observing the variances, the coefficient of variation and the interquartile distances, it can be seen that the data under study presented a great variability in statistical terms, which is corroborated by Figure 8. It is noteworthy that, for the calculation of descriptive statistics, only the observed data was considered.

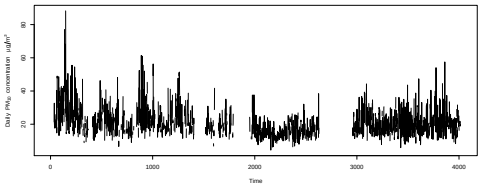
It is worth mentioning that, considering the observed data, the concentrations monitored at the Carapina station exceeded the value of 50 $\mu g.m^(-3)$ on 19 occasions, Cariacica on 500, Enseada do Suá on 51, Jardim Camburi on 14, Laranjeiras on 347, Central Vila Velha on 11, Ibes on 57, Vitória on 27 and Cidade Continental on 61 times. These results meet the guidelines established by WHO and the final air quality standard of CONAMA Res-

	CARA	LARA	CAMB	ENSE	CENVIT	IBES	CENVV	CARI	CIDCONT
nobs	4018.00	4018.00	4018.00	4018.00	4018.00	4018.00	4018.00	4018.00	4018.00
NAs	1531.00	1131.00	1367.00	952.00	821.00	1278.00	2619.00	1711.00	2929.00
NAs/N (%)	38.10	28.14	33.94	23.69	20.43	31.81	65.18	42.26	72.29
Minimum	4.42	6.71	3.54	8.33	6.79	5.00	5.21	5.50	0.00
Maximum	88.25	118.79	66.17	83.58	83.12	88.12	90.75	120.83	208.00
1. Quartile	15.33	21.60	16.25	20.79	19.21	19.50	17.06	26.75	17.21
3. Quartile	23.88	41.21	27.08	32.08	29.42	32.34	28.00	47.79	28.08
Mean	20.52	32.60	22.39	27.02	24.94	26.66	23.31	38.93	27.08
Median	18.88	30.88	21.12	25.92	23.71	25.38	22.04	36.50	21.83
Variance	62.82	203.20	69.31	81.92	67.57	97.50	79.76	288.23	911.65
Stdev	7.93	14.25	8.33	9.05	8.22	9.87	8.93	16.98	30.19
Skewness	1.81	0.82	0.91	0.95	1.01	0.96	1.57	1.00	4.95
Kurtosis	6.27	0.99	1.22	1.99	2.11	2.07	6.63	1.52	26.79
CV (%)	38.65	43.71	37.20	33.93	32.96	37.02	38.31	43.62	111.49

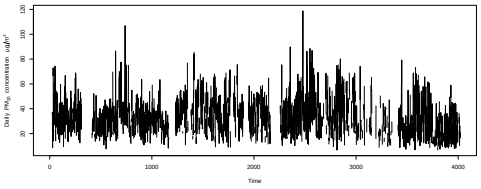
olution nº 491/2018 (not yet in force) (BRASIL, 2018) for this pollutant. Special mention should be made of the Cariacica and Laranjeiras stations, which exceeded the WHO limit in 512 and 380 days, respectively.

Figures 9 and 10 show the periodograms estimated with $\hat{P}^{L}_{X,N}$ and $\hat{I}^{AM}_{Y,N}$ of the natural logarithm of the time series of PM_{10} measured in monitoring stations of the RGV. The estimated periodograms represent the relative contributions of different frequencies to the variance. In general, spectral estimators have large peaks concentrated at frequencies close to zero and also at frequencies that are integer multiples of $\frac{1}{7}$. These plots also indicate seasonality of period 7, corresponding to the seven-day period for all monitoring stations. Period 7 is due to the fact that the series correspond to the average daily observations. An interesting distinction between the graph of the periodogram $\hat{P}^{L}_{X,N}$ (Figure 9) and the graph of the periodogram $\hat{I}^{AM}_{Y,N}$ (Figure 10) is the difference of the vertical scale. However, in all the graphs it is possible to identify the seasonality property, which suggests that the methodology proposed in this work can be used to estimate the time series spectrum of atmospheric pollutant concentrations in the presence of missing data. It is worth noting that, the periodograms for the data of the Cidade Continental station, located in the municipality of Serra, were not estimated, due to the fact that the missing data percentage is greater than 70%.

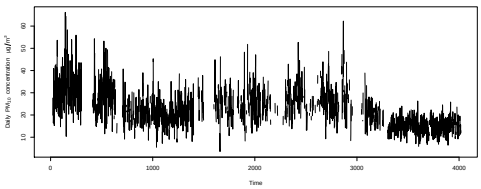
Seasonality in PM_{10} series is expected, since according to the official emission inventory published by Instituto Estadual de Meio Ambiente e Recursos Hídricos do Estado do Espírito Santo (2011), the main source emitting particles in the RGV are linked to dust resuspension of particles in roads (around 60% of PM_{10}). Thus, there is variation between concentrations measured on weekdays and weekend days, since vehicle flow is higher during the days of



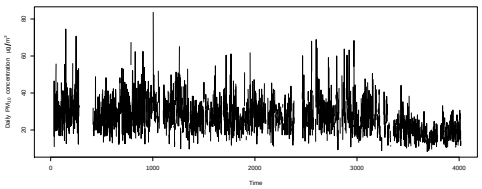
(a) CARA station



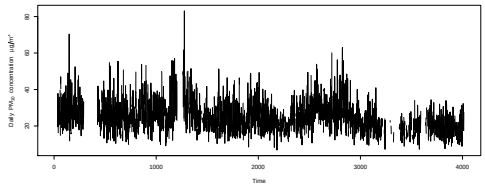
(b) LARA station



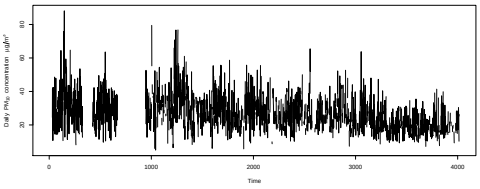
(c) CAMB station



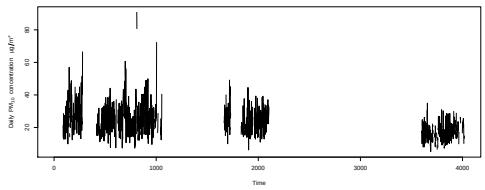
(d) ENSE station



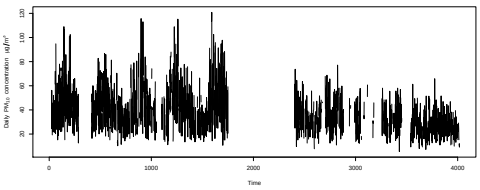
(e) CENVIT station



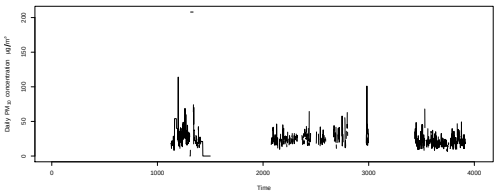
(f) IBES station



(g) CENVV station



(h) CARI station



(i) CIDCONT station

Figure 8: Daily PM_{10} concentrations series with missing data.

the week. However, PM_{10} can also be emitted from other sources, such as soil erosion and building construction. In addition, pollution transported from other sites can make a significant contribution to the observed levels.

It is important to say that the mathematical modeling of time series of atmospheric pollution plays an essential role in the scientific understanding of atmospheric phenomena and in the provision of political strategies to deal with relevant problems such as climate change, air quality and, in general, with degradation of the ecosystem. In mathematical modeling, the model accuracy must be evaluated in relation to the observations to ensure that the errors in model formulations are minimal and that the model limitations are understood. In this context, it is essential to search for models which are able to deal with characteristics of the data that may make its analysis more difficult to perform. This section showed an application to air pollution time series with missing data which, as it was discussed throughout this paper, is a characteristic that causes several problems in frequency analysis.

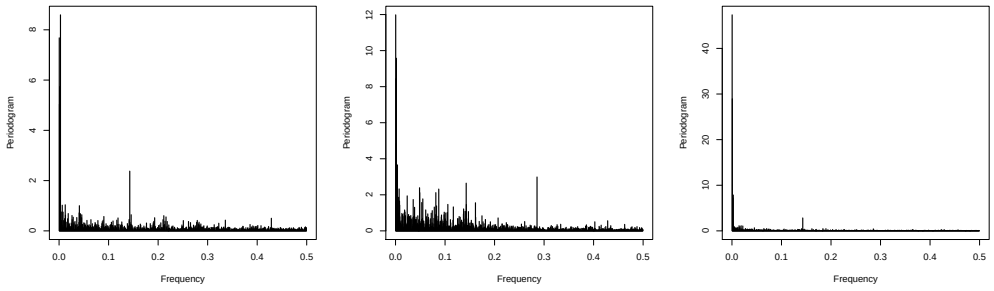
5. Concluding remarks

This paper proposes the use of two estimators for the spectral density of a time series with missing data. Asymptotic properties of the proposed spectral estimators are established and simulations are provided to show the performance of the estimator under different scenarios. The efficiency of the methods, under different percentages of missing data, is also investigated in the simulation study. The results showed that the two methods are suitable for estimation of the spectrum of stationary time series. Therefore, the proposed method becomes a good alternative for the spectral analysis of data sets with missing observations. An application to a real data set related to the daily average concentrations of the pollutant PM_{10} in the Greater Vitória Region, Espírito Santo, Brazil was performed to show the usefulness of the methodology proposed.

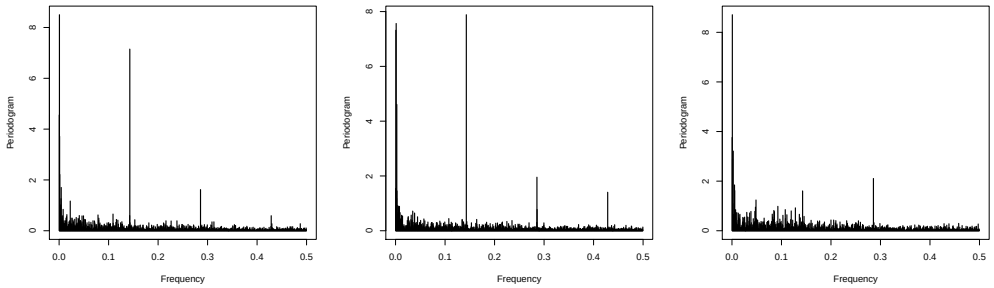
References

Bahamonde, N., Doukhan, P., 2016. Spectral estimation in the presence of missing data. Theory of Probability and Mathematical Statistics 95, 59–79.

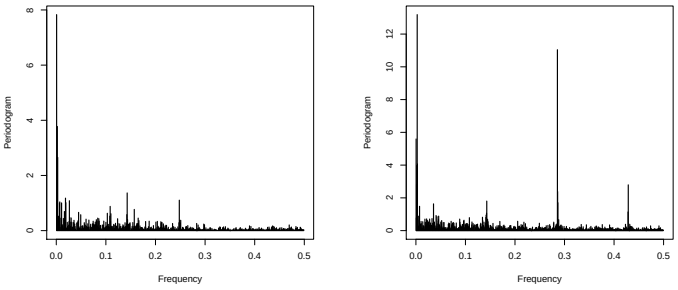
Bahamonde, N., Doukhan, P., Moulines, E., 2010. Estimation of the autocovariance function with missing observations. arXiv preprint arXiv:1004.3717.



(a) CARA station (b) LARA station (c) CAMB station

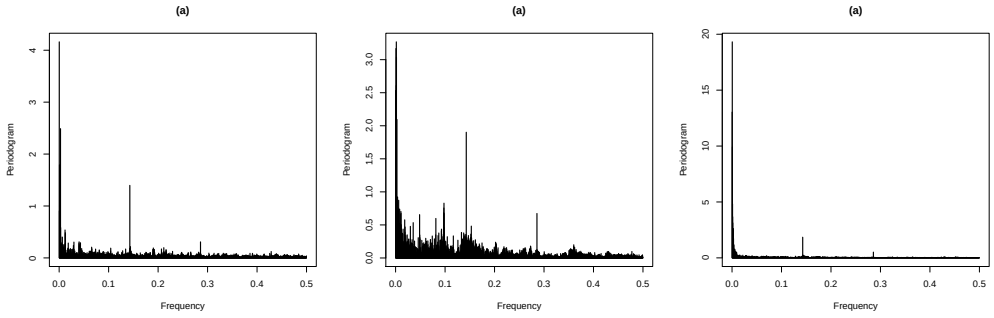


(d) ENSE station (e) CENVIT station (f) IBES station

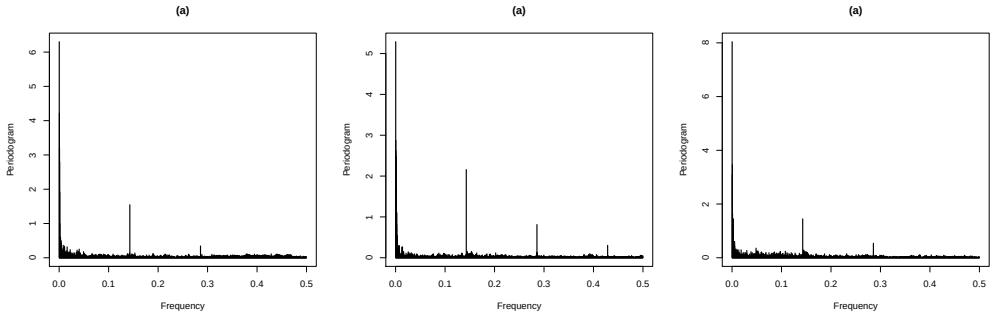


(g) CENVV station (h) CARI station

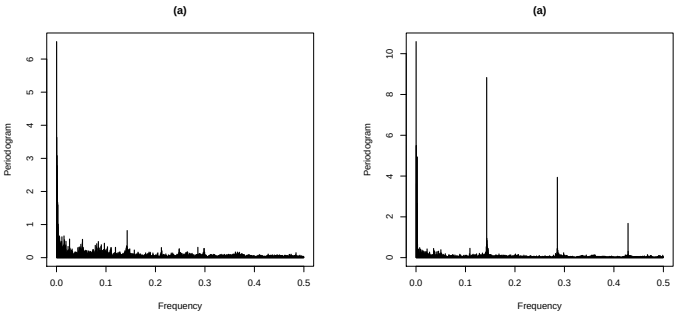
Figure 9: Estimated $\hat{P}^L_{X,N}$ periodograms of daily log PM_{10} concentrations series with missing data.



(a) CARA station (b) LARA station (c) CAMB station



(d) ENSE station (e) CENVIT station (f) IBES station



(g) CENVV station (h) CARI station

Figure 10: Estimated $\hat{I}_{Y,N}^{AM}$ periodograms of daily log PM_{10} concentrations series with missing data.

Bloomfield, P., 2004. Fourier analysis of time series: an introduction. John Wiley & Sons.

Bowdalo, D. R., Evans, M. J., Sofen, E. D., 2016. Spectral analysis of atmospheric composition: application to surface ozone model–measurement comparisons. Atmospheric Chemistry and Physics 16 (13), 8295–8308.

Box, G., Jenkins, G., Reinsel, G., 2008. Time Series Analysis: Forecasting and Control, 4th Edition. Prentice Hall.

BRASIL, 2018. Resolução do CONAMA n. 491 de 19 de novembro de 2018. Dispõe sobre padrães de qualidade do ar. Diário Oficial da República Federativa do Brasil, Publicação DOU nº 223, de 21/11/2018. Brasília, DF.

Brockwell, P., Davis, R., 2002. Introduction to Time Series and Forecasting, 2nd Edition. Springer Verlag.

Dilmaghani, S., Henry, I. C., Soonthornnonda, P., Christensen, E. R., Henry, R. C., 2007. Harmonic analysis of environmental time series with missing data or irregular sample spacing. Environmental science & technology 41 (20), 7030–7038.

Drake, C., Knapik, O., Leśkow, J., 2014. Em-based inference for cyclostationary time series with missing observations. In: Cyclostationarity: Theory and Methods. Springer, pp. 23–35.

Dunsmuir, W., Robinson, P. M., 1981. Asymptotic theory for time series containing missing and amplitude modulated observations. Sankhyā: The Indian Journal of Statistics, Series A, 260–281.

Dutton, S. J., Rajagopalan, B., Vedal, S., Hannigan, M. P., 2010. Temporal patterns in daily measurements of inorganic and organic speciated PM2.5 in Denver. Atmospheric Environment 44 (7), 987–998.

Fajardo, F., Reisen, V. A., Lévy-Leduc, C., Taqqu, M., 2018. M-periodogram for the analysis of long-range-dependent time series. Statistics 52 (3), 665–683.

Fan, J., Kreutzberger, E., 1998. Automatic local smoothing for spectral density estimation. Scandinavian Journal of Statistics 25 (2), 359–369.

Glynn, E. F., Chen, J., Mushegian, A. R., 2006. Detecting periodic patterns in unevenly spaced gene expression time series using Lomb–Scargle periodograms. Bioinformatics 22 (3), 310–316.

Hies, T., Treffeisen, R., Sebald, L., Reimer, E., 2000. Spectral analysis of air pollutants. part 1: elemental carbon time series. Atmospheric Environment 34 (21), 3495–3502.

Hocke, K., Kämpfer, N., 2009. Gap filling and noise reduction of unevenly sampled data by means of the Lomb-Scargle periodogram. Atmospheric chemistry and physics 9 (12), 4197–4206.

Iglesias, P., Jorquera, H., Palma, W., 2005. Data analysis using regression models with missing observations and long memory: an application study. Computational statistics and Data Analysis 50, 2028–2043.

Instituto Brasileiro de Geografia e Estatística, 2014. Banco de dados. Cidades. Rio de Janeiro.
URL http://www.cidades.ibge.gov.br/xtras/home.php

Instituto Estadual de Meio Ambiente e Recursos Hídricos do Estado do Espírito Santo, 2011. Inventário de emissões atmosféricas da Região da Grande Vitória. Vitória.
URL http://www.es.gov.br/Banco%20de%20Documentos/PDF/Maio/100511/RTC10131-R1.pdf

Instituto Estadual de Meio Ambiente e Recursos Hídricos do Estado do Espírito Santo, 2014. Relatório da qualidade do ar da Região da Grande Vitória. Vitória.
URL http://www.meioambiente.es.gov.br/download/Relat%C3%B3rio_Anual_de_Qualidade_do_Ar_2013.pdf

Jiang, J., Hui, Y., 2004. Spectral density estimation with amplitude modulation and outlier detection. Annals of the Institute of Statistical Mathematics 56 (4), 611–630.

Junger, W. L., 2008. Análise, imputação de dados e interfaces computacionais em estudos de séries temporais epidemiológicas. PhD Dissertation. Programa de Pós-graduação em Saúde Coletiva, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, Brazil.

Junger, W. L., Leon, A. P., 2015. Imputation of missing data in time series for air pollutants. Atmospheric Environment 102, 96–104.

Junninen, H., Niskaa, H., Tuppurainenc, K., Ruuskanena, J., Koleh-Mainen, M., 2004. Methods for imputation of missing values in air quality data sets. Atmospheric Environment 38, 2895–2907.

Kumar, U., De Ridder, K., 2010. Garch modelling in association with FFT–ARIMA to forecast ozone episodes. Atmospheric Environment 44 (34), 4252–4265.

Laguna, P., Moody, G. B., Mark, R. G., 1998. Power spectral density of unevenly sampled data by least-square analysis: performance and application to heart rate signals. IEEE Transactions on Biomedical Engineering 45 (6), 698–715.

Leroy, B., 2012. Fast calculation of the Lomb-Scargle periodogram using nonequispaced fast Fourier transforms. Astronomy & Astrophysics 545, A50.

Lomb, N. R., 1976. Least-squares frequency analysis of unequally spaced data. Astrophysics and space science 39 (2), 447–462.

Metaxoglou, K., Smith, A., 2007. Maximum likelihood estimation of varma models using a state-space em algorithm. Journal of Time Series Analysis 28 (5), 666–685.

Norazian, M. N., Shukri, Y. A., Azam, R. N., Al Bakri, A. M. M., 2008. Estimation of missing values in air pollution data using single imputation techniques. ScienceAsia 34 (3), 341–345.

Parzen, E., 1963. On spectral analysis with missing observations and amplitude modulation. Sankhya 25, 383–392.

Pinto, W. P., 2013. O uso da metodologia de dados faltantes em séries temporais com aplicação a dados de concentração de (PM10) observados na Região da Grande Vitória. Master's thesis, Programa de Pós-Graduação em Engenharia Ambiental: Universidade Federal do Espírito Santo, Vitória, Brazil.

Pinto, W. P., Reisen, V. A., Monte, E. Z., 2018. Previsão da concentração de material particulado inalável, na Região da Grande Vitória, ES, Brasil, utilizando o modelo SARIMAX. Engenharia Sanitária e Ambiental 23 (2).

Plaia, A., Bondì, A. L., 2006. Single imputation method of missing values in environmental pollution data sets. Atmospheric Environment 40 (38), 7316–7330.

Press, W. H., Rybicki, G. B., 1989. Fast algorithm for spectral analysis of unevenly sampled data. The Astrophysical Journal 338, 277–280.

Priestley, M. B., 1981. Spectral Analysis and Time Series. Academic Press.

R Core Team, 2019. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
URL https://www.R-project.org/

Reisen, V. A., Leduc, C. L., Cotta, H., 2017. Long-memory models under outliers: an application to air pollution levels.

Reisen, V. A., Sarnaglia, A. J. Q., Reis Jr, N. C., Lévy-Leduc, C., Santos, J. M., 2014a. Modeling and forecasting daily average PM10 concentrations by a seasonal long-memory model with volatility. Environmental Modelling & Software 51, 286–295.

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

Reisen, V. A., Zamprogno, B., Palma, W., Arteche, J., 2014b. A semiparametric approach to estimate two seasonal fractional parameters in the sarfima model. Mathematics and Computers in Simulation 98, 1–17.

Scargle, J. D., 1982. Studies in astronomical time series analysis. ii-statistical aspects of spectral analysis of unevenly spaced data. The Astrophysical Journal 263, 835–853.

Scargle, J. D., 1989. Studies in astronomical time series analysis. iii-fourier transforms, autocorrelation functions, and cross-correlation functions of unevenly spaced data. The Astrophysical Journal 343, 874–887.

Schuster, A., 1898. On the investigation of hidden periodicities with application to a supposed 26 day period of meteorological phenomena. Journal of Geophysical Research 3 (1), 13–41.

Sebald, L., Treffeisen, R., Reimer, E., Hies, T., 2000. Spectral analysis of air pollutants. part 2: ozone time series. Atmospheric Environment 34 (21), 3503–3509.

Solci, C. C., Anselmo Reisen, V., Queiroz Sarnaglia, A. J., Bondon, P., 2019. Empirical study of robust estimation methods for par models with application to the air quality area. Communications in Statistics-Theory and Methods, 1–17.

Toda, H. Y., Makenzie, C., 1999. Lm tests for unit roots in the presence of missing observations: small sample evidence. Mathematics and computers in simulation 48, 457–468.

Toloi, C., Morettin, P. A., 1993. Spectral analysis for amplitude-modulated time series. Journal of Time Series Analysis 14 (4), 409–432.

Townsend, R., 2010. Fast calculation of the Lomb-Scargle periodogram using graphics processing units. The Astrophysical Journal Supplement Series 191 (2), 247.

Wei, W., 2006. Time Series Analysis: Univariate and Multivariate Methods. Addison Wesley.

Yajima, Y., Nishino, H., 1999. Estimation of the autocorrelation function of a stationary time series with missing observations. Sankhyā: The Indian Journal of Statistics, Series A, 189–207.

Yajima, Y., Nishino, H., 1999. Estimation of the autocorrelation function of a stationary time series with missing observations. The Indian Journal of Statistics 61, 189–207.

The application of the spectral decomposition theorem to estimate the ACF function of stationary time series with missing data

Wanderson de Paula Pinto[1], Valdério Aselmo Reisen[1,2]

[1] *Graduate Program in Environmental Engineering (PPGEA) - Federal University of Espírito Santo, Brazil.*

[2] *Department of Statistics, Federal University of Espírito Santo, Brazil.*

Abstract

In this paper, an estimator of $\mathbf{\Gamma}_N$ starting from frequency domain is proposed using the Lomb-Scargle periodogram. The asymptotic properties of the proposed estimator $\{\hat{\rho}^L(h)\}$ are evaluated. In absence of missing data, the finite sample size investigation showed that the proposed method behaves simmilar to the standard one presented estimates close to the true ACF values. Under full sampling, the asymptotic distribution of the proposed estimator is equal to the other two estimators. On the other hand, in the presence of missing data, the estimator $\{\hat{\rho}^L(h)\}$ yields good estimates for some missing data configurations, however, it is suitable once that the classical estimator can not be used. Therefore, the ACF estimator $\{\hat{\rho}^L(h)\}$ proposed is an alternative method to estimate $\mathbf{\Gamma}_N$ in time series with missing data.

Keywords: Autocovariance function, Lomb-Scargle periodogram, Missing data, estimation.

1. Introduction

The estimation of the autocorrelation function (ACF) of the sample $\{Y_1, \cdots, Y_N\}$ is one important step in the process of building time series models (Bahamonde et al., 2010; Yajima and Nishino, 1999). In classical time series analysis, the innovations $\{\epsilon_t\}_{t\in\mathbb{Z}}$ of the linear process $\{Y_t\}$ are often assumed

to be independent and identically distributed (iid), see for example Box et al. (2008) and Brockwell and Davis (2006). In this case, the asymptotic properties of the partial sums, especially the sample ACF and the ratio of the sample covariance have been extensively studied in the literature, see (Bahamonde et al., 2010). The results of the asymptotic theory for the sample ACF of autoregressive processes can be found in Bartlett (1946); Anderson and Walker (1964); Cavazoscadena (1994); He (1996) and Hosking (1996). A summary of properties can be found in Brockwell and Davis (2006).

In practice, a frequent problem in time series is the presence of missing data, usually due to data acquisition failures. Several reasons lead to this scenario, among which we can mention: manual data entry, measurements made incorrectly, equipment with operational failures and high cost of data collection. The analysis of irregularly observed time series is one of the most important problems faced by researchers of diverse areas whose data appear in the form of time series. The study of the asymptotic properties of the ACF function of time series models in the presence of absent observations is more difficult than in the complete case.

Time series analysis in the frequency domain is based on the study of the spectral density function but most of the existing estimation methods are designed for uniformly sampled and complete data (Wang et al., 2005). The estimation of the autocorrelation function can be done in both time domain and frequency domain, see for example cotta.... , however, when observations are missing, the estimation of autocorrelation is still an open problem. Therefore, methods to estimate the spectral density of time series with missing data should be considered.

In this direction, some methodologies have been proposed in the literature, see, for example, Parzen (1963); Bloomfield (1970); Toloi and Morettin (1993); Jiang and Hui (2004); Wang et al. (2005); Bahamonde et al. (2010); Efromovich (2014); Bahamonde and Doukhan (2016), among others. It is worth to mention that the spectral density of a stationary process can be interpreted as the proportion of variance assigned to the oscillations of the time series at a given frequency.

The fast Fourier transform (FFT) is a very efficient technique to calculate the discrete Fourier transform but the limitation of the FFT algorithm is that it requires equally spaced time series (Priestley, 1981). Another limitation of the FFT algorithm is that it does not cope with missing values. In the analysis of time series sampled at irregular time intervals or series with missing data, some methodology is usually employed in order to fill in

the missing data before doing FFT spectral analysis. This, however, is not satisfactory as it modifies the statistical properties of the series by inserting artificial data Leroy (2012). Furthermore, according to Marshall (1980) the imputation will probably smooth the series so that the autocorrelation estimates may not represent the actual autocorrelation structure.

To overcome these limitations, the Lomb-Scargle periodogram was introduced in the astrophysics field to allow the estimation of the spectral density function of time series with missing or unequally spaced data without using imputation techniques. The history behind the Lomb-Scargle periodogram is what follows: In an attempt to find an alternative to impute pseudo-data in sinusoidal models, Lomb (1976) proposed to use least squares for sinusoidal curves. Scargle (1982) extended Lomb's work by defining the Lomb-Scargle periodogram and deriving its distribution.

Therefore, this work considers the estimation of autocovariance function (ACOVF) and ACF with missing data starting from the Lomb-Scargle periodogram. The approach consists of applying the methodology proposed by Lomb (1976) to obtain a discrete Fourier transform for time series with missing data. Next, ACOVF and ACF are obtained from an inverse diagonalization procedure of the matrix containing the estimated spectral density.

The outline of this paper is as follows: Besides the introduction, Section 2 discusses the estimation of the ACOVF and ACF from the Lomb-Scargle periodogram. Section 3 summarizes the simulation experiments. Concluding remarks are given in Section 4.

2. The model and the estimation of the autocovariance function

Let $\{Y_t\}$, $t = 1, 2, ...$, be a stationary process with autocovariance function $\gamma_Y(h) = \mathbb{C}ov[Y_t, Y_{t+h}]$, $h = 0, 1, ...$, which satisfies

(A1) $\sum_{h=-\infty}^{\infty} |\gamma_Y(h)| < \infty$.

Under Assumption (A1), the spectral density of $\{Y_t\}$, $t = 1, 2, ...$, is defined as

$$f(\lambda) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} \gamma_Y(h) e^{-ih\lambda}, \qquad for \quad all \quad \lambda \in [-\pi, \pi]. \tag{1}$$

Let $\{Y_1, Y_2, ..., Y_N\}$ be the first N observations of $\{Y_t\}_{t\in\mathbb{Z}}$. The covariance matrix of $\{Y_1, Y_2, ..., Y_N\}$ is defined by

$$\mathbf{\Gamma}_N = [\gamma(i-j)]_{i,j=1}^{N}. \tag{2}$$

The following properties establishes, from the time to the frequency domain, a very useful analytical connection between functions, that is, this allows to interpret the spectral density as a multiple of the matrix of covariance of the process multiplied by ortogonal vectors. Actually, this has similar interpretation of the classical result of spectral diagonalization of an any positive semi-definite square matrix (Cotta, 2019).

Proposition 1. *Let $\boldsymbol{A}$ be an $N \times N$ regular circulant matrix with first row $[a_0, \ldots, a_{N-1}]$.*

1. *$\boldsymbol{A}$ has eigenvectors g_j with (not necessarily distinct) eigenvalues*

$$\delta_j = \sum_{h=0}^{N-1} a_h \omega^{jh} = p(\omega_j) = \sum_{h=0}^{N-1} a_h \omega_j^h, \quad j = 0, 1, \ldots, N-1, \tag{3}$$

where $\omega_j = \omega^j$, $p(\omega_j) = a_0 + a_1\omega_j + \ldots + a_{N-1}\omega_j^{N-1}$, and

$$g_j = N^{-1/2}[1, \omega^j, \omega^{2j}, \ldots, \omega^{(N-1)j}]'. \tag{4}$$

2. *Setting $j = 0$ in (1), $\delta_0 = a_0 + a_1 + ... + a_{n-1}$ is always an eigenvalue.*
3. *The eigenvectors are mutually orthogonal, that is, $g_j^* g_k = 0$ if $k \neq j$ and $g_j^* g_k = 1$ for $k = j$, where g_j^* is the conjugate transpose of g_j.*
4. *If $\boldsymbol{F}$ is an $N \times N$ Fourier matrix then, it is unitary and $\boldsymbol{F}\boldsymbol{A}\boldsymbol{F}^* = \boldsymbol{\Delta}$, that is $\boldsymbol{A}\boldsymbol{F}^* = \boldsymbol{F}^*\boldsymbol{A}$, where $\Delta = diag[\delta_0, \delta_1, \ldots, \delta_{N-1}]$. Also $\boldsymbol{A} = \boldsymbol{F}^*\boldsymbol{\Delta}F$.*

Proof. This proposition is directly derived from 8.18 to 8.28 in Seber (2008). □

The following Corollaries are straightforward derived from Proposition 1.

Corollary 1. *If $\boldsymbol{A}$ is symmetric regular circulant then:*

1. *$a_h = a_{N-h}, h = 1, \ldots, m$, where*

$$m = \begin{cases} N/2 & N \text{ even} \\ (N-1)/2 & N \text{ odd.} \end{cases} \tag{5}$$

2. *The eigenvectors are* g_j *given by Proposition 1.*
3. *The eigenvalues of* $\boldsymbol{A}$ *are*

$$\delta_j = \sum_{h=0}^{N-1} a_h \cos(2\pi jh/n), \qquad j = 0, \ldots, N-1. \tag{6}$$

4. *A spectral decomposition of* $\boldsymbol{A}$ *is*

$$\boldsymbol{A} = \sum_{j=0}^{N-1} \delta_j g_j g_j^*. \tag{7}$$

Corollary 2. *Let* δ_j *be an eigenvalue of* $\boldsymbol{A}$ *as in Corollary 1. Then, for* N *odd, the pair of real orthogonal eigenvectors corresponding to* δ_j *are*

$$\boldsymbol{c}_j = (\delta_j + \delta_{N-j})/\sqrt{2} = \sqrt{2/N}[1, \cos\lambda_j, \cos 2\lambda_j, \ldots, \cos(N-1)\lambda_j],$$

and

$$\boldsymbol{s}_j = \mathrm{i}(\delta_{N-j} + \delta_j)/\sqrt{2} = \sqrt{2/N}[1, sin\lambda_j, \sin 2\lambda_j, \ldots, \sin(N-1)\lambda_j],$$

for $j = 1, \ldots, [N/2]$ *and setting* $\boldsymbol{c}_0 = \sqrt{1/N}[1, 1, 1, \ldots, 1]$. *Additionally,* $\boldsymbol{P}_N \boldsymbol{A} \boldsymbol{P}'_N = \Delta^s$ *where* $\Delta^s = diag[\delta_0, \delta_1, \delta_1, \ldots, \delta_{[N/2]}, \delta_{[N/2]}]$ *and* $\boldsymbol{P}_N = [\boldsymbol{c}_0, \boldsymbol{c}_1, \boldsymbol{s}_1, \ldots, \boldsymbol{c}_{[N/2]}, \boldsymbol{s}_{[N/2]}]'$. *For the case when* N *is even, both* δ_0 *and* $\delta_{n/2}$ *have multiplicity 1 and* $\boldsymbol{P}_N = [\boldsymbol{c}_0, \boldsymbol{c}_1, \boldsymbol{s}_1, \ldots, 2^{-1/2}\boldsymbol{c}_{N/2}]'$. *Again,* $\boldsymbol{P}_N \boldsymbol{A} \boldsymbol{P}'_N = \Delta^s$.

Proposition 2. *Let* $\boldsymbol{\Gamma}_N$ *be the covariance matrix of the first* N *observations from* $\{Y_t\}_{t\in\mathbb{Z}}$ *which satisfies (A1), and let* $f(.)$ *be its spectral density as given by (1). Let* $\lambda_j = 2\pi j/N, j = 0, \ldots, [N/2]$, *where* $[.]$ *denotes the integer part of* $N/2$. *Let* $\boldsymbol{D}_N$ *be an* $N \times N$ *matrix,*

$$\boldsymbol{D}_N = \begin{cases} \mathrm{diag}\{f(0), \ldots, f(\lambda_{[N/2]}), f(\lambda_{[N/2]})\} & \text{if } N \text{ is odd,} \\ \mathrm{diag}\{f(0), \ldots, f(\lambda_{(N-2)/2}), f(\lambda_{(N-2)/2}), f(\lambda_{N/2})\} & \text{if } N \text{ is even.} \end{cases} \tag{8}$$

and let $\boldsymbol{P}_N$ *be the eigenvectors which will lead to the diagonalization of* $\boldsymbol{\Gamma}_N$ *defined in Corollary 2. Then, the components* x_{ij}^N *of the matrix*

$$\boldsymbol{P}_N \boldsymbol{\Gamma}_N \boldsymbol{P}'_N - 2\pi \boldsymbol{D}_N, \tag{9}$$

converge to zero uniformly as $N \to \infty$ *,i.e.* $\sup_{1\le i,j\le N} |x_{ij}^{(N)}| \to 0$.

Proof. See the proof in Cotta (2019), Proposition 2. □

Let now $\{Y_1, Y_2, ..., Y_N\}$ be a sample of $\{Y_t\}_{t\in\mathbb{Z}}$ and $\hat{f}_N(.)$ be an estimator of the spectral density in (1). Given $\hat{\boldsymbol{D}}_N$ as an estimator of $\boldsymbol{D}_N$ in (8) by replacing $f(.)$ with $\hat{f}_N(.)$, an alternative estimator of $\boldsymbol{\Gamma}_N$ can be obtained by

$$\hat{\boldsymbol{\Gamma}}_N = 2\pi\, \boldsymbol{P}'_N\, \hat{\boldsymbol{D}}_N\, \boldsymbol{P}_N\,. \tag{10}$$

where $\boldsymbol{P}_N$ was defined previously.

The focus is now on the properties of $\hat{f}_N(.)$ related to $f(.)$ that allow the convergence of $\hat{\gamma}_N(.)$ towards $\gamma(.)$. Let $\{Y_t\}_{t\in\mathbb{Z}}$ be a second order stationary process. Given a sample $\{Y_1, Y_2, ..., Y_N\}$, the classical periodogram function, at the Fourier frequency $\lambda_j = 2\pi j/N, j = 1, \ldots, [N/2]$, is defined as

$$I_N(\lambda_j) = \frac{1}{2\pi N}\left|\sum_{k=1}^{N} Y_k \exp(\mathrm{i}k\lambda_j)\right|^2. \tag{11}$$

Although the periodogram is a natural estimator of $f(.)$, It is well-known that $I_N(.)$ is not a consistent estimator of $f(.)$ in the sense that the variance of $I_N(.)$ does not go to zero as N goes to infinite. In addition, $I_N(.)$ has an erratic and wildly fluctuating form. These features make the periodogram to be a poor estimator of the spectral density $f(.)$ see, for example, Priestley (1981). Sine $\{Y_t\}_{t\in\mathbb{Z}}$ is a stationary process with autocovariance function that satisfies Assumption 1, one way to obtain an estimator of the spectral density with reduced variance is simply to omit some terms of $I_N(.)$ which correspond to the tail of the sample autocovariance function. In general, omitting the terms in $I_N(.)$ will increase the bias, however if these correspond to the tails of the sample ACF satisfying Assumption 1, this will not seriously affect the bias of this "new" periodogram. In this context, the new periodogram is given as follows and it is usually called truncated window periodogram (Cotta, 2019).

Before introducing a class of consistent estimators of the spectral density of $\{Y_t\}_{t\in\mathbb{Z}}$, the following assumptions are introduced:

(**A2**) $\{M_N := M\}$ is a sequence of positive integers with $M \to \infty$ and $\frac{M}{N} \to 0$ as $N \to \infty$.

(**A3**) $\{W_N(.)\}$ is a sequence of weight functions with $W_N(k) = W_N(-k)$ and $W_N(k) \geq 0$, for all k.

(**A4**) $\sum_{|k|\leq M} W_N(k) = 1$ and $\sum_{|k|\leq M} W_N^2(k) \to 0$ as $N \to \infty$.

In view of (A1)-(A4), a consistent class of estimators has the form

$$\hat{f}_{T,N}(\lambda_j) = (2\pi)^{-1} \sum_{k=-m}^{m} W_N(k) I_N(\lambda_{j+k}). \tag{12}$$

For more details of the properties of $\hat{f}_{T,N}$ see, for example, Brockwell and Davis (2013), Priestley (1981) and Fuller (1996).

Theorem 1. *Let* $\{Y_1, Y_2, ..., Y_N\}$ *be a sample observation of a second order stationary time series* $\{Y_t\}_{t\in\mathbb{Z}}$ *satisfying* (A1). *Let* $\hat{f}_{T,N}(.)$ *be an estimator of the spectral density of* $\{Y_t\}_{t\in\mathbb{Z}}$ *satisfying* (A2) *to* (A4). *Define* $\hat{\boldsymbol{D}}_N$ *as in* (8) *but replacing* $f(.)$ *with* $\hat{f}_{T,N}(.)$. *Let* $\hat{\boldsymbol{\Gamma}}_N$ *as in* (10) *where* $\hat{\gamma}_N(h)$ *is obtained. Then,*

$$\hat{\boldsymbol{\Gamma}}_N - 2\pi\, \boldsymbol{P}'_N\, \boldsymbol{D}_N\, \boldsymbol{P}_N = o_p\left(\frac{1}{\sqrt{N}}\right) \quad as \quad N \to \infty. \tag{13}$$

Proof. See the proof in Cotta (2019), Theorem 1. □

In this article, an alternative way to derive the $I_N(\lambda_j)$ periodogram function is based on the Lomb-Scargle periodogram (Lomb, 1976; Scargle, 1982).

The Lomb-Scargle periodogram is now introduced and exhibits similar performance (in theoretical and empirical meaning), to $I_N(\lambda)$, $\lambda \in [-\pi, \pi]$, when the sample is equally spaced, i.e, with no missing data. The Lomb-Scargle periodogram reduces to the Fourier transform in case of evenly sampled data. In presence of data gaps, the sine and cosine model functions are orthogonalized by additional phase factors (Lomb, 1976). Scargle (1989) investigated the reconstruction of unevenly sampled time series by application of the Lomb-Scargle periodogram and subsequent, the inverse Fourier transform (Hocke and Kämpfer, 2009).

The Lomb-Scargle periodogram is equal to a linear least-squares fit of sine and cosine model functions to the observed time series Y_t which shall be centered around zero (Lomb, 1976; Press et al., 1992). Writing $\{Y_t\}_{t\in\mathbb{Z}}$ as a sum of sine and cosine, we have:

$$Y_t = Acos(\lambda_j t_j - \tau) + Bsin(\lambda_j t_j - \tau) + \epsilon_j \tag{14}$$

where A and B are constant amplitudes, λ is the Fourier frequency ($\lambda_j = 2\pi j/N, j = 1, \cdots, [N/2]$), ϵ_j is noise at time t_j, and τ is the additional phase which is required for the orthogonalization of the sine and cosine model functions of equation (14) when the data are unevenly spaced.

The power spectral density $P_N^L(\lambda_j)$ of the Lomb-Scargle periodogram is given by

$$P_N^L(\lambda_j) = \frac{1}{2}\left\{\frac{[R(\lambda_j)]^2}{\sum_{t=1}^{N}\cos^2\lambda_j(t_j-\tau)} + \frac{[I(\lambda_j)]^2}{\sum_{t=1}^{N}\sin^2\lambda_j(t_j-\tau)}\right\}, \tag{15}$$

where

$$R(\lambda_j) = \sum_{t=1}^{N}(Y_t - \overline{Y})\cos\lambda_j(t_j-\tau), \tag{16}$$

$$I(\lambda_j) = \sum_{t=1}^{N}(Y_t - \overline{Y})\sin\lambda_j(t_j-\tau), \tag{17}$$

$\overline{Y}$ is the mean of Y series and τ is evaluated at each λ via

$$\tau = \frac{1}{2\lambda_j}\arctan\left[\frac{\sum_{t=1}^{N}\sin(2\lambda_j t_j)}{\sum_{t=1}^{N}\cos(2\lambda_j t_j)}\right]. \tag{18}$$

Scargle (1982) presented τ as an exact solution for the orthogonalization of the sine and cosine functions of (14) in case of unevenly sampled data ($\sum_{j=1}^{N} cos(\lambda_j t_j - \tau)\sum_{j=1}^{N} sin(\lambda_j t_j - \tau) = 0$).

The Lomb-Scargle periodogram can be used for the analysis of equally spaced time series because its statistical properties are equivalent to the classical periodogram. However, for Scargle (1982), the calculation of the periodogram by equation 15 is more complicated than by the traditional methodology (equation 11). For Scargle (1982), the key factor is to impose continuity on τ as a function of λ and to use a sufficiently high-frequency resolution so that no phase jumps are missed. It is also important to note that

$$\lim_{\lambda\to 0}\tau(\omega) = \left(\frac{1}{N}\right)\sum_{i=1}^{N} t_j = \langle t\rangle. \tag{19}$$

Equations (14 to 19) describe the Lomb-Scargle periodogram. These equations were defined according to the Lomb (1976), Scargle (1982), Scargle (1989), Press and Rybicki (1989), Glynn et al. (2006), Dilmaghani et al. (2007), Hocke and Kämpfer (2009), Townsend (2010) and Leroy (2012).

As previously mentioned, the main objective of this work is to obtain an ACF estimator with missing data which satisfies all assumptions of the definition of ACF presented in Proposition 1.5.1 and Definition 1.5.1 (non-negative definiteness) (Theorem 1.5.1) in Brockwell and Davis (2006). In order to obtain such estimator, in 2, the diagonal elements of matrix $\boldsymbol{D}_N$ is replaced by the Lomb-Scargle spectral estimator $P_N^L(.)$. This leads to the following equation

$$\hat{\boldsymbol{\Gamma}}_N^L = 2\pi\, \boldsymbol{P}'_N\, \hat{\boldsymbol{D}}_n^L\, \boldsymbol{P}'_N, \tag{20}$$

where $\hat{\boldsymbol{D}}_N^L$ is defined similarly as (8) but replacing $f(.)$ with $P_N^L(.)$.

Proposition 3. *Let $\{Y_1, Y_2, ..., Y_N\}$ be a sample observation of a second order stationary time series $\{Y_t\}_{t\in\mathbb{Z}}$ satisfying (A1). Let $I_N^L(.)$ be an estimator of the spectral density of $\{Y_t\}_{t\in\mathbb{Z}}$ and define $\hat{\boldsymbol{D}}_N$ as in (8) but replacing $f(.)$ with $I_N^L(.)$. Let $\hat{\boldsymbol{\Gamma}}_N$ as in (10) where $\hat{\gamma}_N^L(h)$ is obtained. Suppose that (A1) to (A4) hold. Then, $\hat{\gamma}_N^L(h) - \gamma(h) = o_p\left(\frac{1}{\sqrt{N}}\right)$ as $N \to \infty$ for $h = 0, \ldots, N-1$.*

Now, given a sample $\{Y_1, Y_2, ..., Y_N\}$ of $\{Y_t\}_{t\in\mathbb{Z}}$, the $N \times N$ autocorrelation matrix $\hat{\boldsymbol{\rho}}_N$ and its version $\hat{\boldsymbol{\rho}}_N^L$ are, respectively,

$$\hat{\boldsymbol{\rho}}_N = \frac{\hat{\boldsymbol{\Gamma}}_N}{\hat{\gamma}_{11}}, \tag{21}$$

and

$$\hat{\boldsymbol{\rho}}_N^L = \frac{\hat{\boldsymbol{\Gamma}}_N^L}{\hat{\gamma}_{11}^L}, \tag{22}$$

where $\hat{\gamma}_{11} = \hat{\gamma}_N(0)$ and $\hat{\gamma}_{11}^L = \hat{\gamma}_N^L(0)$. Finally, the ACOVF and ACF, $\hat{\gamma}_N(.)$ and $\hat{\rho}_N(.)$, and their counterparts, namely $\hat{\gamma}_N^M$, $\hat{\rho}_N^M$, are extracted from the rows of the circulant matrices $\hat{\boldsymbol{\Gamma}}_N$, $\hat{\boldsymbol{\rho}}_N$, $\hat{\boldsymbol{\Gamma}}_N^L$ and $\hat{\boldsymbol{\rho}}_N^L$, respectively.

3. Monte Carlo simulation study

This section reports a Monte Carlo simulation study to investigate the performance of the robust sample ACOVF and ACF estimators discussed

previously. For the numerical experiments, the data generating process of $\{Y_t\}_{t\in\mathbb{N}}$ is an autoregressive process of order 1 ($AR(1)$) as follows:

$$Y_t = \phi Y_{t-1} + \varepsilon_t, \tag{23}$$

where $|\phi| < 1$ and ε_t is a zero mean Gaussian white noise process with variance σ^2.

Let $\{Y_1, Y_2, ..., Y_N\}$ be a realization of $\{Y_t\}_{t\in\mathbb{N}}$, the standard biased ACF estimator is

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}, \quad -N < h < N \tag{24}$$

where

$$\hat{\gamma}(h) = N^{-1} \sum_{t=1}^{N-|h|} (Y_{t+|h|} - \bar{Y})(Y_t - \bar{Y}), \quad -N < h < N \tag{25}$$

where $\bar{Y} = N^{-1}\sum_{t=1}^{N} Y_t$.

For the comparison between the proposed ACF estimator of samples with missing data and the proposed estimator of Parzen (1963), the latter is summarized in the sequel. Let $\{Y_t\}_{t\in\mathbb{N}}$ be a discrete-time second-order stationary time series with zero-mean. Following Parzen (1963), we assume that the observations are given by

$$Z_t = C_t Y_t \tag{26}$$

where $\{C_t\}_{t\in\mathbb{N}}$ is a non-negative modulating process taking values in $[0, 1]$. When C_t takes values in 0, 1, the observations are censured, more general modulations can be considered as well. Throughout the paper, this process is assumed to be independent of $\{Y_t\}_{t\in\mathbb{N}}$. This property is essential in order to allow recovery of the covariance structure of $\{Y_t\}_{t\in\mathbb{N}}$ (Bahamonde et al., 2010).

We denote by $\bar{Z}$, $\bar{Y}$ the sample means of $(Z_t)_{t=1}^N$ and $(Y_t)_{t=1}^N$. Usual estimates of the autocovariance coefficients $\gamma_Z(h) = Cov[Z_t, Z_{t+h}]$ are the empirical estimates $\hat{\gamma}_{Z,N}(h)$ and $\hat{v}_N(h)$ as follow:

$$\hat{\gamma}_{Y,N}(h) = \begin{cases} \frac{1}{N-h}\sum_{t=1}^{N-h}(Z_t - \bar{Z})(Z_{t+h} - \bar{Z}) & \text{if} \quad 0 \le h < N, \\ \frac{1}{N-h}\sum_{t-h}^{N}(Z_t - \bar{Z})(Z_{t+h} - \bar{Z}) & \text{if} \quad -N < h < 0. \end{cases}$$

$$\hat{v}_{C,N}(h) = \begin{cases} \frac{1}{N-h}\sum_{t=1}^{N-h} C_t C_{t+h} & \text{if} \quad 0 \leq h < N, \\ \frac{1}{N-h}\sum_{t-h}^{N} C_t C_{t+h} & \text{if} \quad -N < h < 0. \end{cases}$$

Both estimates are obtained from the observations $\{Z_1, \cdots, Z_N\} = \{Y_1 C_1, \cdots, Y_N C_N\}$ according to Equation (26) and they are unbiased (Bahamonde et al., 2010). The following estimator is defined considering Equation (26) and was introduced by Parzen (1963),

$$\hat{\gamma}_{Y,N}(h) = \frac{\hat{\gamma}_{Z,N}(h)}{\hat{v}_{C,N}(h)}, \quad \text{if} \quad \hat{v}_{C,N}(h) \neq 0. \tag{27}$$

The correlation function $\rho_Y(h)$ is estimated by $\hat{\rho}_N^{AM}(h) = \frac{\hat{\gamma}_{Z,N}(h)}{\hat{v}_{Z,N}(0)}$. Dunsmuir and Robinson (1981) investigated the asymptotic properties of (27) under various assumptions about $\{\epsilon_t\}_{t\in\mathbb{N}}$ and assuming that $\{C_t\}_{t\in\mathbb{N}}$ is asymptotically stationary. More recently, Yajima and Nishino (1999); Pinto (2013) compared three estimators of the autocorrelation function of stationary processes with missing observations. They impose the same assumptions on $\{\epsilon_t\}_{t\in\mathbb{N}}$ as those in Dunsmuir and Robinson (1981).

In the simulations, $\phi = 0.3$, $\phi = 0.7$ and $p = 5\%, 10\%, 15\%, 20\%$ and 40% are set. The sample size are $N = 100, 300, 600$ and 1000, and each experiment is replicated 1000 times. Two scenarios are considered: (i) samples have no missing data ($p = 0$) and (ii) samples have missing data ($p \neq 0$). Under both scenarios, the comparison between the estimators is done by contrasting the plots and the empirical root mean square error (RMSE) and bias of the theoretical $\rho(h) = \phi^h$ with $\hat{\rho}(h)$, $\hat{\rho}_N^L(h)$ and $\hat{\rho}_N^{AM}(h)$, for $h = 0, 1, \ldots, 8$.

It is worth saying that, the percentage of 5% of missing data is considered by the literature as reference of the smallest amount of tolerable missing information that any method can provide reliable estimates. Above this value, missind data tecniques must be considered and the percentage of 40% is used to evaluate the estimation methods of the autocorrelation function under extreme conditions of missing information.

In this direction, Figure 1 displays the plots of $\rho(h)$ and the means of $\hat{\rho}(h)$, $\hat{\rho}_N^L(h)$ and $\hat{\rho}_N^{AM}(h)$ for the scenario with no missing data, with $\phi = 0.3$ and $N = 100, 300, 600$ and 1000. The case of $\phi = 0.7$ and $p = 0$ is shown in Figure 2. For the case of simulations with samples in the absence of missing data, we observed similar behavior in all graphs, indicating that all the estimators capture the correlation structure of the series. Therefore, these estimators have the same asymptotic properties under complete sampling

but as will be shown below, they can behave asymptotically different in the presence of missing data.

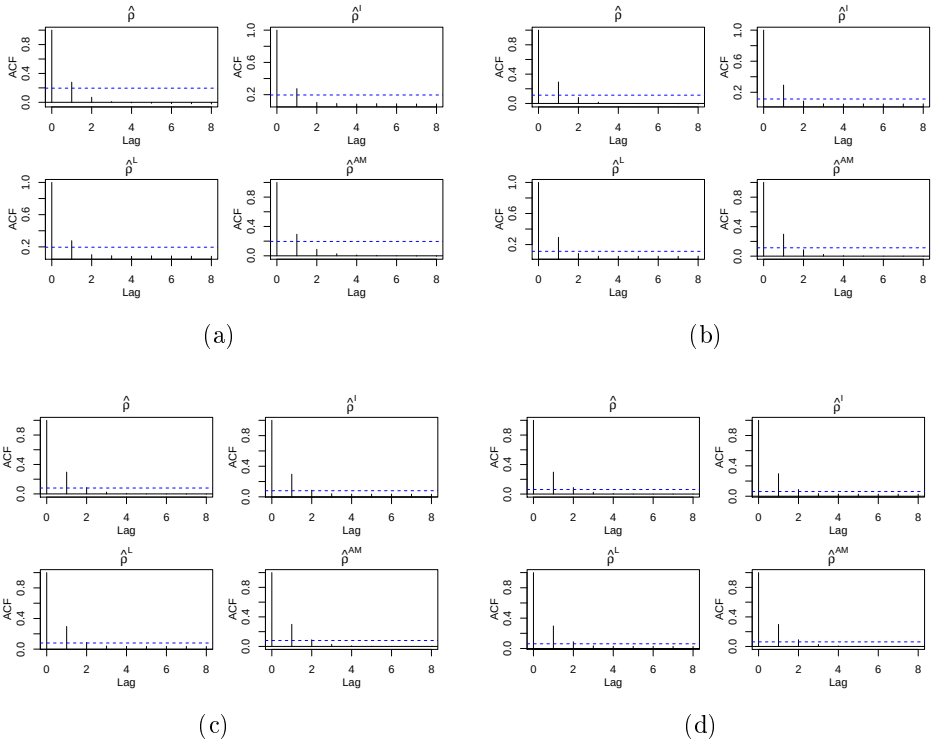


Figure 1: Autocorrelation function of z_t. From left to right and top to bottom, plots are $\rho(h)$, $\hat{\rho}(h)$, $\hat{\rho}^L(h)$, and $\hat{\rho}^{AM}(h)$ when $p = 0$, $\phi = 0.3$ and (a) $N = 100$, (b) $N = 300$, (c) $N = 600$ and (d) $N = 1000$.

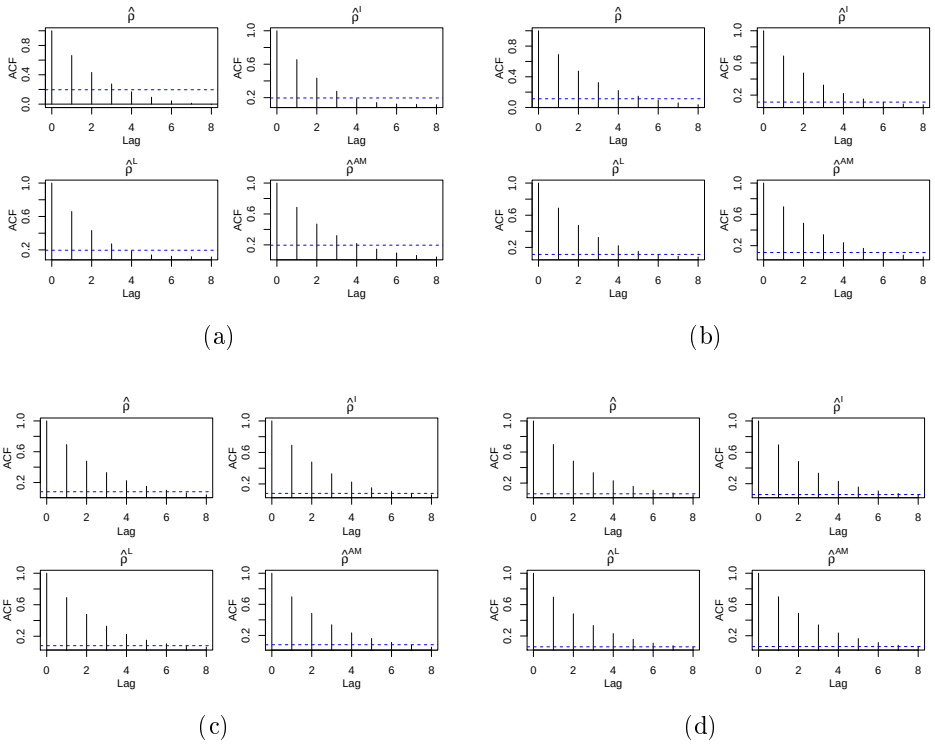


Figure 2: Autocorrelation function of z_t. From left to right and top to bottom, plots are $\rho(h)$, $\hat{\rho}(h)$, $\hat{\rho}^{L}(h)$, and $\hat{\rho}^{AM}(h)$ when $p = 0$, $\phi = 0.7$ and (a) $N = 100$, (b) $N = 300$, (c) $N = 600$ and (d) $N = 1000$.

Related to Figures 1 and 2, Figure 3 and 4 show the boxplot of the simulated ACF values for both scenarios, respectively. It is observed that the estimator of the autocorrelation function $\{\hat{\rho}^{L}(h)\}$, in the frequency domain, proposed in this work presents a pattern of variability similar to the classical estimator $\{\hat{\rho}(h)\}$ and to the proposed Parzen (1963) estimator $\{\hat{\rho}^{AM}(h)\}$ of the idea of amplitude modulated.

In the tables 1 and 3 , we present the root mean squared errors (RMSE) of $\hat{\rho}(h)$, $\hat{\rho}^{L}_{N}(h)$ and $\hat{\rho}^{AM}_{N}(h)$ as N increases, for $h = 0, 1, \ldots, 8$, $p = 0$, $\phi = 0.3$ and $\phi = 0.7$. Tables 2 and 4 present the estimated results for BIAS. It is observed that estimators have RMSE close to each other in the absence of missing data, depending on the parameter evaluated. Besides, as expected as the sample size increases, the estimated RMSE values decrease. By observing

the Tables 2 and 4 one can conclude that the proposed estimator $\hat{\rho}_N^L(h)$ underestimates the theoretical value of the ACF, which is also observed in the results of the estimators $\hat{\rho}(h)$ and $\hat{\rho}_N^{AM}(h)$. Thus, $\hat{\rho}_N^L(h)$ is suitable even in the context where the presence of missing data is uncertain.

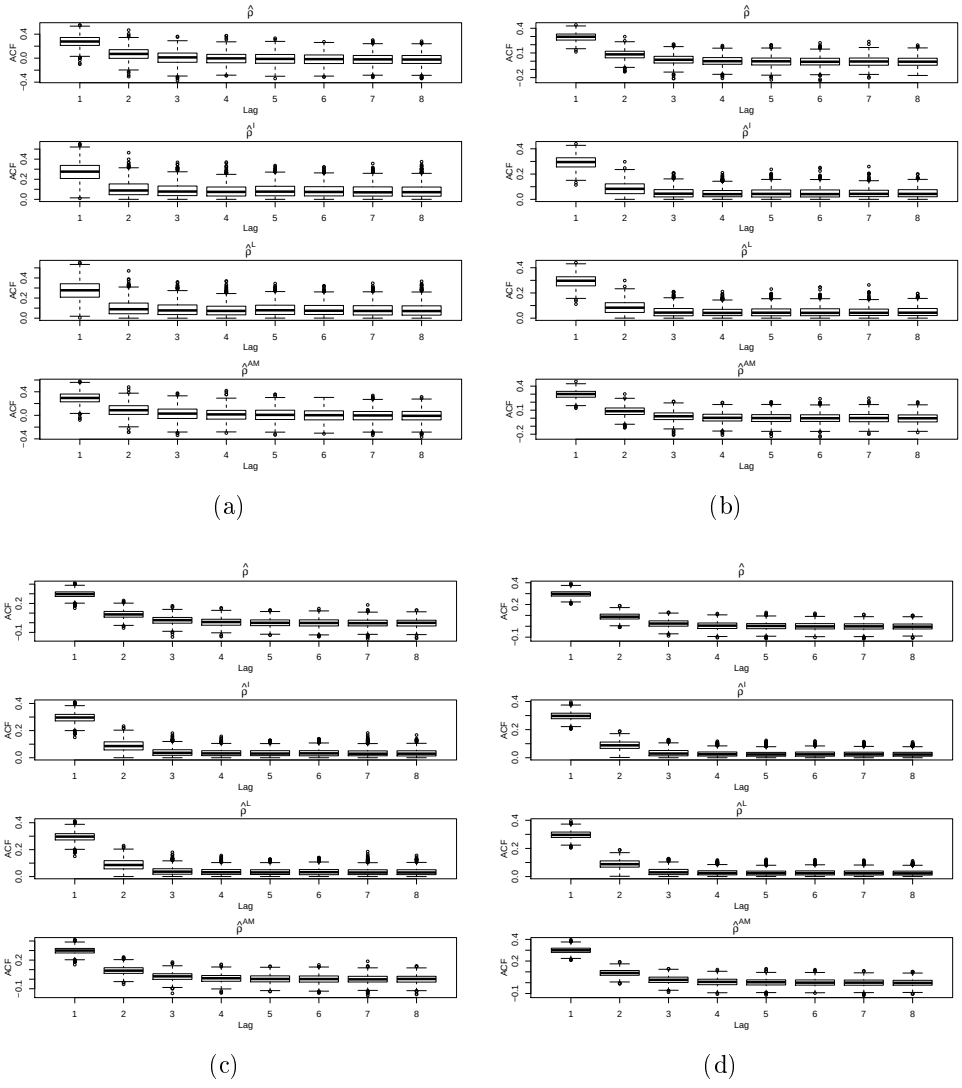


Figure 3: Boxplots of estimated ACF of z_t when $p = 0$, $\phi = 0.3$ and (a) $N = 100$, (b) $N = 300$, (c) $N = 600$ and (d) $N = 1000$.

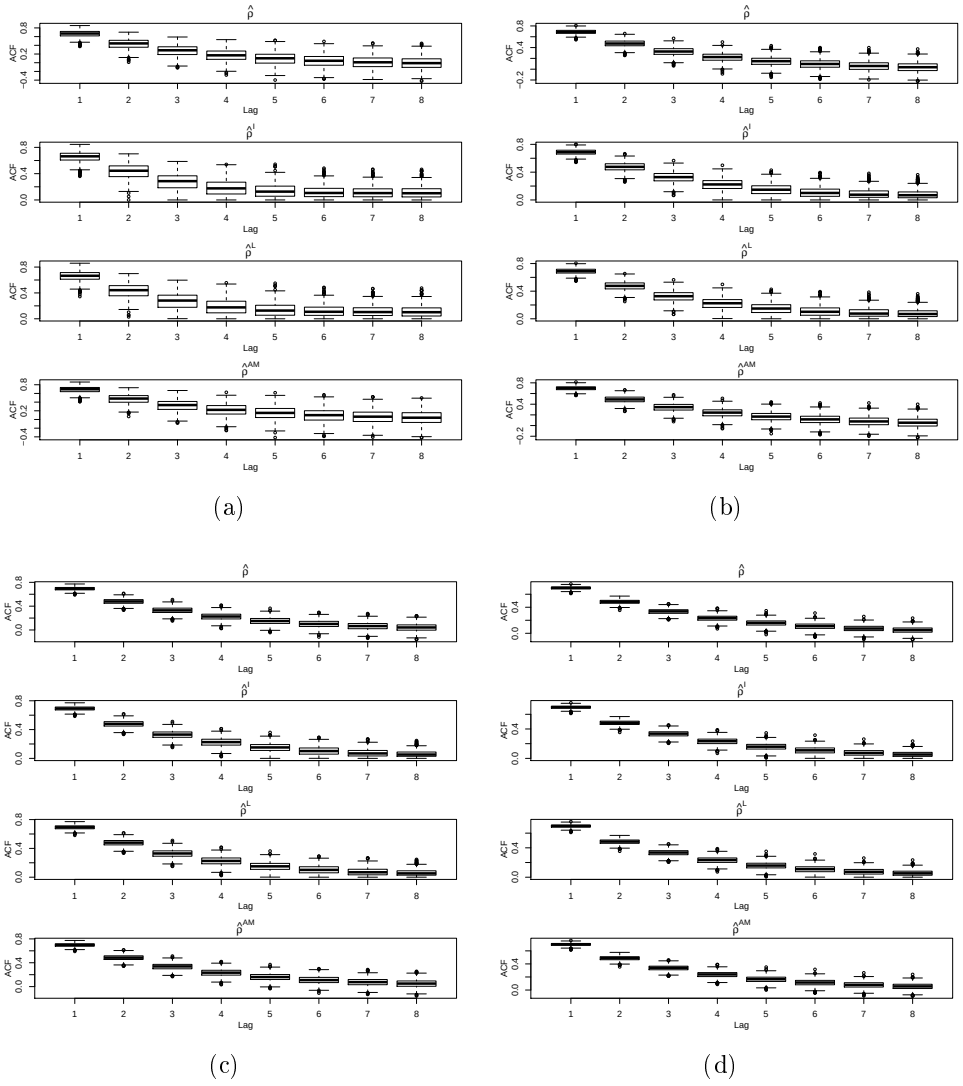


Figure 4: Boxplots of estimated ACF of z_t when $p = 0$, $\phi = 0.7$ and (a) $N = 100$, (b) $N = 300$, (c) $N = 600$ and (d) $N = 1000$.

The results for the scenario with missing data $p \neq 0$ are displayed in tables 5, 6, 7 and 8 for $p = 5\%$, $p = 10\%$, $p = 15\%$ and $p = 20\%$, respectively. Comparing the tables, it is observed that as the missing data percentage is increased the performance of both estimators reduces. However, as expected,

$\hat{\rho}_N^{AM}(h)$ still exhibits a somewhat better performance than $\hat{\rho}_N^L(h)$. Therefore, the comparison is not only about the advantage of the use $\hat{\rho}_N^L(h)$ for missing data, but also in the case of unequally spaced observations and dimension reduction techniques.

Table 1: RMSE of $\hat{\rho}(h)$, $\hat{\rho}_N^L(h)$ and $\hat{\rho}_N^{AM}(h)$, for $h = 0, 1, \ldots, 8$, and $p = 0$ ($\phi = 0.3$).

	N	1	2	3	4	5	6	7	8
$\hat{\rho}(h)$	100	0.0948	0.1062	0.1056	0.1046	0.1072	0.1053	0.1063	0.1040
	300	0.0552	0.0602	0.0625	0.0629	0.0625	0.0605	0.0615	0.0619
	600	0.0415	0.0442	0.0444	0.0444	0.0448	0.0453	0.0443	0.0437
	1000	0.0301	0.0351	0.0344	0.0356	0.0348	0.0348	0.0336	0.0338
$\hat{\rho}_N^I(h)$	100	0.0994	0.0781	0.0923	0.0994	0.1083	0.1086	0.1069	0.1089
	300	0.0582	0.0557	0.0480	0.0559	0.0623	0.0629	0.0634	0.0639
	600	0.0385	0.0427	0.0331	0.0386	0.0428	0.0455	0.0456	0.0437
	1000	0.0302	0.0339	0.0265	0.0298	0.0328	0.0344	0.0352	0.0348
$\hat{\rho}_N^L(h)$	100	0.0954	0.0747	0.0855	0.0995	0.1074	0.1079	0.1091	0.1087
	300	0.0551	0.0524	0.0462	0.0563	0.0612	0.0606	0.0622	0.0631
	600	0.0415	0.0426	0.0318	0.0381	0.0430	0.0448	0.0444	0.0438
	1000	0.0301	0.0344	0.0256	0.0296	0.0330	0.0343	0.0336	0.0339
$\hat{\rho}_N^{AM}(h)$	100	0.0937	0.1068	0.1076	0.1082	0.1135	0.1130	0.1137	0.1128
	300	0.0546	0.0600	0.0623	0.0630	0.0633	0.0613	0.0626	0.0641
	600	0.0414	0.0444	0.0447	0.0444	0.0450	0.0454	0.0445	0.0442
	1000	0.0299	0.0349	0.0343	0.0355	0.0349	0.0350	0.0338	0.0340

Table 2: BIAS of $\hat{\rho}(h)$, $\hat{\rho}_N^L(h)$ and $\hat{\rho}_N^{AM}(h)$, for $h = 0, 1, \ldots, 8$ and $p = 0$ ($\phi = 0.3$).

	N	1	2	3	4	5	6	7	8
$\hat{\rho}(h)$	100	-0.0186	-0.0193	-0.0222	-0.0191	-0.0182	-0.0185	-0.0233	-0.0223
	300	-0.0089	-0.0079	-0.0090	-0.0090	-0.0057	-0.0065	-0.0050	-0.0028
	600	-0.0047	-0.0051	-0.0051	-0.0060	-0.0041	-0.0047	-0.0041	-0.0016
	1000	-0.0044	-0.0041	-0.0029	-0.0039	-0.0020	-0.0017	-0.0014	-0.0017
$\hat{\rho}_N^I(h)$	100	-0.0249	0.0145	0.0635	0.0764	0.0864	0.0863	0.0848	0.0854
	300	-0.0094	0.0008	0.0277	0.0413	0.0493	0.0504	0.0503	0.0512
	600	-0.0040	-0.0023	0.0142	0.0282	0.0336	0.0360	0.0360	0.0349
	1000	-0.0019	-0.0018	0.0075	0.0205	0.0251	0.0271	0.0283	0.0278
$\hat{\rho}_N^L(h)$	100	-0.0197	0.0140	0.0566	0.0769	0.0844	0.0858	0.0884	0.0877
	300	-0.0092	-0.0025	0.0246	0.0416	0.0475	0.0481	0.0496	0.0506
	600	-0.0052	-0.0040	0.0128	0.0272	0.0337	0.0353	0.0356	0.0349
	1000	-0.0046	-0.0037	0.0067	0.0202	0.0253	0.0269	0.0266	0.0272
$\hat{\rho}_N^{AM}(h)$	100	-0.0036	-0.0020	-0.0050	-0.0022	-0.0017	-0.0023	-0.0074	-0.0066
	300	-0.0033	-0.0014	-0.0024	-0.0025	0.0008	-0.0001	0.0014	0.0037
	600	-0.0020	-0.0020	-0.0019	-0.0029	-0.0010	-0.0016	-0.0010	0.0015
	1000	-0.0027	-0.0021	-0.0009	-0.0018	-0.0000	0.0003	0.0006	0.0004

Table 3: RMSE of $\hat{\rho}(h)$, $\hat{\rho}_N^L(h)$ and $\hat{\rho}_N^{AM}(h)$, for $h = 0, 1, \ldots, 8$, and $p = 0$ ($\phi = 0.7$).

	N	1	2	3	4	5	6	7	8
$\hat{\rho}(h)$	100	0.0845	0.1276	0.1527	0.1614	0.1638	0.1632	0.1581	0.1550
	300	0.0456	0.0710	0.0847	0.0912	0.0942	0.0953	0.0960	0.0958
	600	0.0294	0.0454	0.0554	0.0629	0.0670	0.0689	0.0705	0.0697
	1000	0.0241	0.0375	0.0451	0.0494	0.0508	0.0505	0.0504	0.0506
$\hat{\rho}_N^I(h)$	100	0.0889	0.1253	0.1424	0.1264	0.1035	0.0891	0.0940	0.1048
	300	0.0439	0.0675	0.0819	0.0879	0.0842	0.0747	0.0686	0.0679
	600	0.0304	0.0464	0.0563	0.0610	0.0620	0.0596	0.0522	0.0451
	1000	0.0219	0.0344	0.0425	0.0469	0.0501	0.0500	0.0456	0.0395
$\hat{\rho}_N^L(h)$	100	0.0862	0.1280	0.1485	0.1364	0.1067	0.0934	0.0957	0.1072
	300	0.0459	0.0714	0.0853	0.0896	0.0875	0.0757	0.0675	0.0646
	600	0.0298	0.0454	0.0557	0.0629	0.0668	0.0628	0.0557	0.0508
	1000	0.0242	0.0375	0.0452	0.0493	0.0508	0.0495	0.0456	0.0403
$\hat{\rho}_N^{AM}(h)$	100	0.0749	0.1160	0.1421	0.1527	0.1569	0.1602	0.1598	0.1611
	300	0.0432	0.0673	0.0812	0.0894	0.0937	0.0956	0.0971	0.0974
	600	0.0287	0.0445	0.0545	0.0620	0.0661	0.0683	0.0702	0.0698
	1000	0.0237	0.0371	0.0446	0.0489	0.0502	0.0501	0.0503	0.0507

Table 4: BIAS of $\hat{\rho}(h)$, $\hat{\rho}_N^L(h)$ and $\hat{\rho}_N^{AM}(h)$, for $h = 0, 1, \ldots, 8$ and $p = 0$ ($\phi = 0.7$).

	N	1	2	3	4	5	6	7	8
$\hat{\rho}(h)$	100	-0.0363	-0.0547	-0.0667	-0.0703	-0.0720	-0.0709	-0.0643	-0.0606
	300	-0.0149	-0.0235	-0.0265	-0.0259	-0.0244	-0.0248	-0.0238	-0.0239
	600	-0.0058	-0.0088	-0.0106	-0.0116	-0.0124	-0.0121	-0.0115	-0.0102
	1000	-0.0037	-0.0059	-0.0080	-0.0100	-0.0103	-0.0088	-0.0071	-0.0059
$\hat{\rho}_N^I(h)$	100	-0.0452	-0.0565	-0.0665	-0.0542	-0.0264	0.0060	0.0365	0.0588
	300	-0.0132	-0.0151	-0.0177	-0.0188	-0.0162	-0.0064	0.0082	0.0242
	600	-0.0083	-0.0115	-0.0135	-0.0155	-0.0161	-0.0140	-0.0063	0.0045
	1000	-0.0045	-0.0065	-0.0085	-0.0092	-0.0098	-0.0087	-0.0053	0.0011
$\hat{\rho}_N^L(h)$	100	-0.0389	-0.0542	-0.0667	-0.0535	-0.0262	0.0069	0.0360	0.0592
	300	-0.0155	-0.0235	-0.0273	-0.0252	-0.0209	-0.0103	0.0045	0.0212
	600	-0.0072	-0.0084	-0.0118	-0.0113	-0.0133	-0.0081	-0.0006	0.0104
	1000	-0.0042	-0.0058	-0.0083	-0.0096	-0.0105	-0.0080	-0.0043	0.0027
$\hat{\rho}_N^{AM}(h)$	100	-0.0120	-0.0164	-0.0205	-0.0199	-0.0196	-0.0181	-0.0115	-0.0083
	300	-0.0066	-0.0101	-0.0103	-0.0081	-0.0057	-0.0058	-0.0046	-0.0047
	600	-0.0018	-0.0023	-0.0027	-0.0028	-0.0031	-0.0025	-0.0019	-0.0006
	1000	-0.0014	-0.0022	-0.0034	-0.0049	-0.0050	-0.0034	-0.0016	-0.0004

Table 5: RMSE of $\hat{\rho}(h)$, $\hat{\rho}_N^L(h)$ and $\hat{\rho}_N^{AM}(h)$, for $h = 0, 1, \ldots, 8$, and $\phi = 0.3$.

p		N	1	2	3	4	5	6	7	8
5%	$\hat{\rho}_N^L(h)$	100	0.1031	0.0749	0.0875	0.1003	0.1097	0.1106	0.1083	0.1073
		300	0.0595	0.0529	0.0482	0.0592	0.0629	0.0631	0.0648	0.0667
		600	0.0449	0.0435	0.0332	0.0388	0.0441	0.0433	0.0447	0.0455
	$\hat{\rho}_N^{AM}(h)$	100	0.0982	0.1075	0.1131	0.1139	0.1168	0.1196	0.1150	0.1153
		300	0.0553	0.0634	0.0646	0.0674	0.0669	0.0662	0.0679	0.0693
		600	0.0415	0.0453	0.0462	0.0457	0.0468	0.0442	0.0462	0.0463
10%	$\hat{\rho}_N^L(h)$	100	0.1184	0.0760	0.0930	0.1052	0.1127	0.1128	0.1151	0.1108
		300	0.0727	0.0550	0.0489	0.0606	0.0645	0.0676	0.0668	0.0668
		600	0.0562	0.0452	0.0342	0.0410	0.0466	0.0460	0.0463	0.0484
	$\hat{\rho}_N^{AM}(h)$	100	0.1103	0.1175	0.1201	0.1217	0.1207	0.1221	0.1254	0.1232
		300	0.0615	0.0684	0.0674	0.0696	0.0683	0.0706	0.0687	0.0698
		600	0.0434	0.0476	0.0490	0.0488	0.0502	0.0484	0.0488	0.0480
15%	$\hat{\rho}_N^L(h)$	100	0.1231	0.0749	0.0951	0.1086	0.1167	0.1201	0.1154	0.1133
		300	0.0834	0.0538	0.0539	0.0631	0.0703	0.0691	0.0711	0.0683
		600	0.0685	0.0471	0.0334	0.0425	0.0467	0.0481	0.0492	0.0500
	$\hat{\rho}_N^{AM}(h)$	100	0.1126	0.1232	0.1244	0.1248	0.1312	0.1322	0.1296	0.1268
		300	0.0642	0.0714	0.0729	0.0728	0.0742	0.0732	0.0744	0.0725
		600	0.0471	0.0502	0.0506	0.0499	0.0507	0.0510	0.0504	0.0526
20%	$\hat{\rho}_N^L(h)$	100	0.1348	0.0741	0.1019	0.1094	0.1184	0.1196	0.1191	0.1163
		300	0.0965	0.0565	0.0525	0.0638	0.0684	0.0697	0.0696	0.0704
		600	0.0788	0.0469	0.0343	0.0450	0.0468	0.0495	0.0499	0.0501
	$\hat{\rho}_N^{AM}(h)$	100	0.1190	0.1262	0.1325	0.1330	0.1314	0.1289	0.1333	0.1349
		300	0.0698	0.0759	0.0764	0.0765	0.0777	0.0777	0.0751	0.0792
		600	0.0476	0.0527	0.0515	0.0546	0.0533	0.0522	0.0539	0.0538
40%	$\hat{\rho}_N^L(h)$	100	0.1718	0.0820	0.1096	0.1268	0.1318	0.1318	0.1323	0.1289
		300	0.1481	0.0558	0.0593	0.0727	0.0769	0.0753	0.0783	0.0773
		600	0.1368	0.0525	0.0394	0.0494	0.0541	0.0555	0.0563	0.0539
	$\hat{\rho}_N^{AM}(h)$	100	0.1684	0.1706	0.1712	0.1818	0.1747	0.1798	0.1798	0.1751
		300	0.0914	0.0969	0.1008	0.0998	0.0997	0.0972	0.1004	0.1000
		600	0.0684	0.0698	0.0687	0.0715	0.0711	0.0714	0.0694	0.0701

Table 6: BIAS of $\hat{\rho}(h)$, $\hat{\rho}_N^L(h)$ and $\hat{\rho}_N^{AM}(h)$, for $h = 0, 1, \ldots, 8$ and $\phi = 0.3$.

		N	1	2	3	4	5	6	7	8
5%	$\hat{\rho}_N^L(h)$	100	-0.0345	0.0104	0.0609	0.0789	0.0868	0.0866	0.0870	0.0849
		300	-0.0217	-0.0044	0.0269	0.0444	0.0493	0.0499	0.0513	0.0523
		600	-0.0184	-0.0082	0.0135	0.0273	0.0341	0.0343	0.0359	0.0358
	$\hat{\rho}_N^{AM}(h)$	100	-0.0017	0.0010	0.0029	-0.0010	-0.0003	0.0034	0.0021	0.0044
		300	-0.0010	-0.0004	0.0013	0.0002	0.0021	0.0017	0.0006	-0.0030
		600	-0.0002	-0.0014	-0.0011	-0.0008	-0.0010	0.0007	0.0004	-0.0023
10%	$\hat{\rho}_N^L(h)$	100	-0.0580	0.0106	0.0619	0.0820	0.0892	0.0895	0.0917	0.0902
		300	-0.0415	-0.0051	0.0274	0.0465	0.0511	0.0545	0.0532	0.0534
		600	-0.0368	-0.0121	0.0133	0.0302	0.0363	0.0362	0.0371	0.0388
	$\hat{\rho}_N^{AM}(h)$	100	-0.0096	-0.0058	-0.0070	-0.0065	-0.0048	-0.0036	-0.0091	-0.0035
		300	-0.0042	0.0024	0.0014	0.0005	0.0001	-0.0002	-0.0016	0.0002
		600	-0.0028	-0.0016	-0.0002	0.0008	0.0001	0.0026	0.0021	0.0019
15%	$\hat{\rho}_N^L(h)$	100	-0.0673	0.0118	0.0658	0.0830	0.0934	0.0954	0.0925	0.0913
		300	-0.0555	-0.0109	0.0310	0.0475	0.0547	0.0557	0.0567	0.0541
		600	-0.0514	-0.0166	0.0136	0.0307	0.0361	0.0380	0.0397	0.0399
	$\hat{\rho}_N^{AM}(h)$	100	-0.0002	-0.0019	-0.0021	-0.0013	0.0001	0.0035	0.0000	0.0044
		300	-0.0040	-0.0021	-0.0017	-0.0040	-0.0039	-0.0007	-0.0017	0.0003
		600	-0.0040	-0.0021	-0.0021	-0.0009	-0.0015	-0.0021	-0.0012	0.0003
20%	$\hat{\rho}_N^L(h)$	100	-0.0851	0.0120	0.0724	0.0848	0.0933	0.0947	0.0952	0.0935
		300	-0.0694	-0.0148	0.0309	0.0486	0.0535	0.0555	0.0555	0.0550
		600	-0.0641	-0.0185	0.0147	0.0326	0.0361	0.0389	0.0407	0.0396
	$\hat{\rho}_N^{AM}(h)$	100	-0.0082	-0.0025	0.0064	0.0044	0.0025	0.0001	-0.0048	-0.0081
		300	-0.0035	-0.0067	-0.0032	-0.0030	-0.0006	0.0009	-0.0043	-0.0041
		600	0.0016	0.0005	0.0013	0.0003	0.0001	0.0022	0.0036	-0.0001
40%	$\hat{\rho}_N^L(h)$	100	-0.1338	0.0189	0.0771	0.0986	0.1031	0.1061	0.1051	0.1023
		300	-0.1285	-0.0169	0.0359	0.0546	0.0604	0.0596	0.0620	0.0625
		600	-0.1250	-0.0297	0.0187	0.0365	0.0422	0.0433	0.0450	0.0427
	$\hat{\rho}_N^{AM}(h)$	100	-0.0040	-0.0019	-0.0057	-0.0018	-0.0115	-0.0029	0.0023	-0.0064
		300	-0.0032	-0.0006	0.0054	0.0037	0.0028	0.0002	-0.0012	-0.0058
		600	0.0022	-0.0010	-0.0005	-0.0023	0.0004	-0.0041	-0.0066	-0.0061

Table 7: RMSE of $\hat{\rho}(h)$, $\hat{\rho}_N^L(h)$ and $\hat{\rho}_N^{AM}(h)$, for $h = 0, 1, \ldots, 8$, and $\phi = 0.7$.

p		N	1	2	3	4	5	6	7	8
5%	$\hat{\rho}_N^L(h)$	100	0.1114	0.1412	0.1512	0.1338	0.1085	0.0981	0.1032	0.1091
		300	0.0677	0.0831	0.0903	0.0919	0.0857	0.0722	0.0654	0.0631
		600	0.0548	0.0589	0.0639	0.0657	0.0669	0.0607	0.0521	0.0471
	$\hat{\rho}_N^{AM}(h)$	100	0.0833	0.1210	0.1437	0.1548	0.1622	0.1679	0.1714	0.1688
		300	0.0474	0.0721	0.0837	0.0905	0.0950	0.0966	0.0970	0.0963
		600	0.0327	0.0476	0.0564	0.0621	0.0651	0.0672	0.0671	0.0682
10%	$\hat{\rho}_N^L(h)$	100	0.1422	0.1575	0.1548	0.1291	0.1039	0.0875	0.0883	0.1011
		300	0.0999	0.1007	0.1034	0.1022	0.0908	0.0765	0.0662	0.0632
		600	0.0871	0.0770	0.0734	0.0707	0.0699	0.0625	0.0533	0.0479
	$\hat{\rho}_N^{AM}(h)$	100	0.0868	0.1196	0.1390	0.1526	0.1604	0.1654	0.1674	0.1689
		300	0.0511	0.0718	0.0847	0.0938	0.0978	0.1003	0.1006	0.0991
		600	0.0355	0.0487	0.0569	0.0630	0.0670	0.0681	0.0687	0.0688
15%	$\hat{\rho}_N^L(h)$	100	0.1808	0.1887	0.1712	0.1364	0.0986	0.0865	0.0928	0.1090
		300	0.1345	0.1237	0.1160	0.1073	0.0904	0.0730	0.0610	0.0617
		600	0.1208	0.0993	0.0887	0.0800	0.0747	0.0654	0.0540	0.0465
	$\hat{\rho}_N^{AM}(h)$	100	0.0994	0.1330	0.1506	0.1603	0.1655	0.1700	0.1735	0.1778
		300	0.0559	0.0759	0.0877	0.0941	0.0967	0.1004	0.1012	0.1032
		600	0.0395	0.0524	0.0616	0.0662	0.0697	0.0701	0.0720	0.0720
20%	$\hat{\rho}_N^L(h)$	100	0.2200	0.2072	0.1830	0.1407	0.1039	0.0881	0.0912	0.1004
		300	0.1689	0.1421	0.1246	0.1110	0.0926	0.0735	0.0621	0.0610
		600	0.1550	0.1207	0.1013	0.0873	0.0776	0.0645	0.0517	0.0454
	$\hat{\rho}_N^{AM}(h)$	100	0.1052	0.1367	0.1569	0.1676	0.1739	0.1788	0.1802	0.1795
			0.0621	0.0792	0.0903	0.0979	0.1020	0.1063	0.1067	0.1029
		600	0.0427	0.0540	0.0638	0.0671	0.0710	0.0719	0.0726	0.0743
40%	$\hat{\rho}_N^L(h)$	100	0.3717	0.2871	0.2107	0.1470	0.1035	0.0840	0.0917	0.1044
		300	0.3138	0.2400	0.1853	0.1377	0.0989	0.0711	0.0600	0.0576
		600	0.2975	0.2178	0.1663	0.1283	0.0969	0.0700	0.0527	0.0424
	$\hat{\rho}_N^{AM}(h)$	100	0.1526	0.1834	0.1831	0.2011	0.2015	0.2047	0.2082	0.2139
		300	0.0905	0.1049	0.1106	0.1135	0.1182	0.1225	0.1233	0.1237
		600	0.0614	0.0694	0.0797	0.0820	0.0814	0.0845	0.0870	0.0866

Table 8: BIAS of $\hat{\rho}(h)$, $\hat{\rho}_N^L(h)$ and $\hat{\rho}_N^{AM}(h)$, for $h = 0, 1, \ldots, 8$ and $\phi = 0.7$.

		N	1	2	3	4	5	6	7	8
5%	$\hat{\rho}_N^L(h)$	100	-0.0734	-0.0786	-0.0750	-0.0580	-0.0244	0.0102	0.0432	0.0646
		300	-0.0490	-0.0449	-0.0393	-0.0334	-0.0236	-0.0087	0.0071	0.0217
		600	-0.0445	-0.0363	-0.0329	-0.0270	-0.0243	-0.0154	-0.0073	0.0057
	$\hat{\rho}_N^{AM}(h)$	100	-0.0124	-0.0137	-0.0124	-0.0118	-0.0096	-0.0077	-0.0057	-0.0058
		300	-0.0048	-0.0072	-0.0064	-0.0043	-0.0021	-0.0002	0.0025	0.0033
		600	-0.0042	-0.0059	-0.0065	-0.0067	-0.0059	-0.0046	-0.0040	-0.0029
10%	$\hat{\rho}_N^L(h)$	100	-0.1122	-0.1068	-0.0951	-0.0689	-0.0338	0.0030	0.0318	0.0543
		300	-0.0861	-0.0731	-0.0634	-0.0527	-0.0397	-0.0201	0.0018	0.0214
		600	-0.0794	-0.0598	-0.0459	-0.0329	-0.0269	-0.0169	-0.0070	0.0065
	$\hat{\rho}_N^{AM}(h)$	100	-0.0112	-0.0163	-0.0148	-0.0146	-0.0112	-0.0093	-0.0091	-0.0108
		300	-0.0054	-0.0084	-0.0108	-0.0102	-0.0102	-0.0086	-0.0055	-0.0030
		600	-0.0031	-0.0042	-0.0022	-0.0005	-0.0005	-0.0002	-0.0003	0.0001
15%	$\hat{\rho}_N^L(h)$	100	-0.1524	-0.1424	-0.1181	-0.0834	-0.0366	0.0031	0.0345	0.0624
		300	-0.1226	-0.1004	-0.0817	-0.0664	-0.0472	-0.0238	-0.0015	0.0213
		600	-0.1143	-0.0856	-0.0653	-0.0487	-0.0379	-0.0245	-0.0094	0.0052
	$\hat{\rho}_N^{AM}(h)$	100	-0.0156	-0.0255	-0.0239	-0.0246	-0.0231	-0.0166	-0.0141	-0.0119
		300	-0.0058	-0.0111	-0.0110	-0.0117	-0.0110	-0.0114	-0.0105	-0.0070
		600	-0.0017	-0.0034	-0.0024	-0.0026	-0.0016	-0.0007	-0.0008	-0.0012
20%	$\hat{\rho}_N^L(h)$	100	-0.1935	-0.1661	-0.1367	-0.0918	-0.0448	0.0004	0.0335	0.0568
		300	-0.1574	-0.1195	-0.0924	-0.0688	-0.0487	-0.0226	0.0006	0.0196
		600	-0.1491	-0.1083	-0.0807	-0.0606	-0.0462	-0.0307	-0.0142	0.0023
	$\hat{\rho}_N^{AM}(h)$	100	-0.0133	-0.0225	-0.0226	-0.0226	-0.0163	-0.0126	-0.0098	-0.0058
		300	-0.0082	-0.0081	-0.0066	-0.0048	-0.0065	-0.0060	-0.0056	-0.0035
		600	-0.0030	-0.0025	-0.0031	-0.0036	-0.0048	-0.0037	-0.0026	-0.0010
40%	$\hat{\rho}_N^L(h)$	100	-0.3470	-0.2575	-0.1798	-0.1107	-0.0516	-0.0049	0.0333	0.0582
		300	-0.3041	-0.2241	-0.1643	-0.1141	-0.0715	-0.0338	-0.0061	0.0170
		600	-0.2925	-0.2099	-0.1531	-0.1114	-0.0784	-0.0482	-0.0236	-0.0035
	$\hat{\rho}_N^{AM}(h)$	100	-0.0222	-0.0244	-0.0123	-0.0082	-0.0088	-0.0082	-0.0065	-0.0064
		300	-0.0102	-0.0093	-0.0096	-0.0094	-0.0063	-0.0029	-0.0028	-0.0006
		600	-0.0071	-0.0073	-0.0064	-0.0067	-0.0059	-0.0062	-0.0051	-0.0066

4. Conclusion

This paper presents an estimation method for the autocovariance and autocorrelation functions of stationary univariate processes. The procedure consists in replacing the traditional periodogram with the Lomb-Sacargle periodogram in the inverse diagonalization procedure of the matrix containing the estimated spectral density. According to the results, it can be concluded that there is no loss of information, so the proposed method can be used to estimate the autocovariance and autocorrelation functions, even in the presence of missing data. Therefore, the authors suggest the use of the proposed method in time series in which there are occurrences of missing observations.

References

Anderson, T., Walker, A., 1964. On the asymptotic distribution of the autocorrelations of a sample from a linear stochastic process. The Annals of Mathematical Statistics 35 (3), 1296–1303.

Bahamonde, N., Doukhan, P., 2016. Spectral estimation in the presence of missing data. Theory of Probability and Mathematical Statistics 95, 59–79.

Bahamonde, N., Doukhan, P., Moulines, E., 2010. Estimation of the autocovariance function with missing observations. arXiv preprint arXiv:1004.3717.

Bartlett, M. S., 1946. On the theoretical specification and sampling properties of autocorrelated time-series. Supplement to the Journal of the Royal Statistical Society 8 (1), 27–41.

Bloomfield, P., 1970. Spectral analysis with randomly missing observations. Journal of the Royal Statistical Society. Series B (Methodological), 369–380.

Box, G., Jenkins, G., Reinsel, G., 2008. Time Series Analysis: Forecasting and Control, 4th Edition. Prentice Hall.

Brockwell, P. J., Davis, R. A., 2006. Time Series: Theory and Methods, 2nd Edition. Springer Series in Statistics.

Brockwell, P. J., Davis, R. A., 2013. Time series: theory and methods. Springer Science & Business Media.

Cavazoscadena, R., 1994. The asymptotic distribution of sample autocorrelations for a class of linear filters. Journal of multivariate analysis 48 (2), 249–274.

Cotta, H. H. A., 2019. Robust methods in multivariate time series with high dimension data. Doutorado em engenharia ambiental, Programa de Pós-Graduação em Engenharia Ambiental: Universidade Federal do Espírito Santo, Vitória, Espírito Santo, Brazil.

Dilmaghani, S., Henry, I. C., Soonthornnonda, P., Christensen, E. R., Henry, R. C., 2007. Harmonic analysis of environmental time series with missing data or irregular sample spacing. Environmental science & technology 41 (20), 7030–7038.

Dunsmuir, W., Robinson, P. M., 1981. Asymptotic theory for time series containing missing and amplitude modulated observations. Sankhyā: The Indian Journal of Statistics, Series A, 260–281.

Efromovich, S., 2014. Efficient non-parametric estimation of the spectral density in the presence of missing observations. Journal of Time Series Analysis 35 (5), 407–427.

Fuller, W. A., 1996. Introduction to statistical time series. Vol. 428. John Wiley & Sons.

Glynn, E. F., Chen, J., Mushegian, A. R., 2006. Detecting periodic patterns in unevenly spaced gene expression time series using lomb–scargle periodograms. Bioinformatics 22 (3), 310–316.

He, S., 1996. A note on the asymptotic normality of sample autocorrelations for a linear stationary sequence. journal of multivariate analysis 58 (2), 182–188.

Hocke, K., Kämpfer, N., 2009. Gap filling and noise reduction of unevenly sampled data by means of the lomb-scargle periodogram. Atmospheric chemistry and physics 9 (12), 4197–4206.

Hosking, J. R., 1996. Asymptotic distributions of the sample mean, autocovariances, and autocorrelations of long-memory time series. Journal of Econometrics 73 (1), 261–284.

Jiang, J., Hui, Y., 2004. Spectral density estimation with amplitude modulation and outlier detection. Annals of the Institute of Statistical Mathematics 56 (4), 611–630.

Leroy, B., 2012. Fast calculation of the lomb-scargle periodogram using nonequispaced fast fourier transforms. Astronomy & Astrophysics 545, A50.

Lomb, N. R., 1976. Least-squares frequency analysis of unequally spaced data. Astrophysics and space science 39 (2), 447–462.

Marshall, R., 1980. Autocorrelation estimation of time series with randomly missing observations. Biometrika 67 (3), 567–570.

Parzen, E., 1963. On spectral analysis with missing observations and amplitude modulation. Sankhya 25, 383–392.

Pinto, W. P., 2013. O uso da metodologia de dados faltantes em séries temporais com aplicação a dados de concentração de (PM10) observados na Região da Grande Vitória. Master's thesis, Programa de Pós-Graduação em Engenharia Ambiental: Universidade Federal do Espírito Santo, Vitória, Brazil.

Press, W. H., Rybicki, G. B., 1989. Fast algorithm for spectral analysis of unevenly sampled data. The Astrophysical Journal 338, 277–280.

Press, W. H., Teukolsky, S., Vetterling, W., Flannery, B., 1992. Numerical recipes in Fortran. Cambridge University Press, Cambridge, USA, 2 edn.

Priestley, M. B., 1981. Spectral Analysis and Time Series. Academic Press.

Scargle, J. D., 1982. Studies in astronomical time series analysis. ii-statistical aspects of spectral analysis of unevenly spaced data. The Astrophysical Journal 263, 835–853.

Scargle, J. D., 1989. Studies in astronomical time series analysis. iii-fourier transforms, autocorrelation functions, and cross-correlation functions of unevenly spaced data. The Astrophysical Journal 343, 874–887.

Seber, G. A., 2008. A matrix handbook for statisticians. Vol. 15. John Wiley & Sons.

Toloi, C., Morettin, P. A., 1993. Spectral analysis for amplitude-modulated time series. Journal of Time Series Analysis 14 (4), 409–432.

Townsend, R., 2010. Fast calculation of the lomb-scargle periodogram using graphics processing units. The Astrophysical Journal Supplement Series 191 (2), 247.

Wang, Y., Stoica, P., Li, J., Marzetta, T. L., 2005. Nonparametric spectral analysis with missing data via the em algorithm. Digital signal processing 15 (2), 191–206.

Yajima, Y., Nishino, H., 1999. Estimation of the autocorrelation function of a stationary time series with missing observations. Sankhyā: The Indian Journal of Statistics, Series A, 189–207.

6 CONCLUSÕES GERAIS

A poluição atmosférica tem afetado de forma significativa os seres vivos, mesmo quando seus valores estão abaixo do permitido pelas entidades regulamentadoras. Neste sentido, as questões relativas à qualidade do ar têm se tornado cada vez mais importantes, uma vez que vários problemas de saúde decorrem da poluição atmosférica. Dessa forma, diversos estudos aplicando técnicas de análise de séries temporais têm sido realizados, com o intuito de contribuir como ferramentas na tomada de decisões dos agentes públicos e privados no que diz respeito à prevenção de concentrações elevadas, ao controle da poluição atmosférica e à formulação de legislações para esse fim. Uma das metodologias estatísticas adotadas é a análise espectral, sendo a mesma utilizada para identificar propriedades do conjunto de dados, como por exemplo a sazonalidade. No entanto, observa-se que, entre os estudos que têm adotado esta técnica, uma característica comum é negligenciar a presença de dados faltantes *(missing data)*, que pode levar à subestimar a precisão dos resultados. Nota-se que nas séries temporais relacionadas à poluição atmosférica um problema frequente é a presença de dados faltantes, geralmente devido a falhas dos equipamentos de monitoramento.

Assim, este documento concentra-se no estudo de metodologias usadas para estimar a função de autocorrelação e a densidade espectral de séries temporais univariadas na presença ou sem dados faltantes. Os estimadores sugeridos são baseadas na metodologia de Amplitude Modulada, proposta por Parzen (1963), e no periodograma de Lomb-Scargle (LOMB, 1976; SCARGLE, 1982). Além disso, foi proposto estimadores das funções de autocovariância e autocorrelação de séries temporais, considerando a conexão entre o domínio do tempo e da frequência por meio da relação entre a função de autocovariância e a densidade espectral.

Assim, este trabalho objetivou propor métodos para a estimação da função de autocorrelação e da densidade espectral de séries temporais univariadas estacionárias na presença de dados faltantes, com aplicação para a poluição do ar, na Região da Grande Vitória, Espírito Santo. Três linhas de pesquisa foram propostas: i) avaliar diferentes estimadores da função de autocorrelação de séries temporais estacionárias na presença de dados faltantes, com aplicação de dados de poluição atmosférica; ii) avaliar, por meio de estudo de simulação de Monte Carlo, a utilização do periodograma de Lomb-Scargle (LOMB, 1976; SCARGLE, 1982) e da metodologia de Amplitude Modulada, proposta por Parzen (1963), para estimar a densidade espectral de séries temporais na presença de dados faltantes. Além disso, aplicar estas metodologias em dados de PM_{10} monitorados na RGV, ES, Brasil; e, iii) propor estimadores das funções de autocovariância e autocorrelação de séries temporais, considerando a conexão entre o domínio do tempo e da frequência por meio da relação entre a função de autocovariância e a densidade espectral. As três linhas de pesquisa deram origem à três artigos principais, estando suas conclusões descritas a seguir.

O primeiro artigo apresenta um estudo de metodologias para tratamento de dados faltantes em séries temporais. Tais metodologias baseiam-se em métodos de imputação e estimadores da

função de autocorrelação de séries temporais na presença de dados faltantes. Para os processos investigados neste estudo, as estimativas de ACF obtidas usando-se os estimadores propostos por (PARZEN, 1963) e (YAJIMA; NISHINO, 1999) e pelo algoritmo EM (DEMPSTER; RUBIN, 1977; JUNGER; LEON, 2015) têm MSE significativamente pequenos, comparado aos valores teóricos correspondentes. Os resultados dos experimentos numéricos sugerem que o método EM tende a subestimar os valores de ACF. A investigação empírica considerou diferentes cenários, enfatizando o efeito da porcentagem de dados faltantes nos estimadores. Uma aplicação a dados de PM_{10} foi considerada e com ambos os estimadores foi possível identificar a propriedade de sazonalidade da série. A partir dos resultados obtidos, pode-se observar que as metodologias testadas no trabalho apresentam resultados acurados, mesmo sob condições extremas de dados faltantes (40%). Portanto, estas metodologias podem ser utilizadas como uma alternativa para estimação da função de autocorrelação de séries temporais na presença de dados faltantes, além disso, podem ser aplicadas em estudos com banco de dados incompletos de concentrações de poluentes atmosféricos.

Em relação ao segundo artigo, foi proposto a utilização de dois estimadores para a densidade espectral de uma série temporal com dados faltantes. As propriedades assintóticas dos estimadores espectrais propostos são estabelecidas e simulações são fornecidas para mostrar o desempenho do estimador sob diferentes cenários. A eficiência dos métodos, sub diferentes porcentagens de dados faltantes, também é investigada no estudo de simulação. Os resultados evidenciaram que os dois métodos são adequados para estimação do espectro de séries temporais estacionárias. Portanto, o método proposto torna-se uma boa alternativa para a análise espectral de conjuntos de dados com observações faltantes. Um conjunto de dados relacionados à variável poluição do ar é usado para mostrar a utilidade da metodologia proposta em aplicações reais.

Já no terceiro artigo, foi apresentado um método de estimação para as funções de autocovariância e autocorrelação de processos univariados estacionários. O procedimento consiste em substituir o periodograma tradicional pelo periodograma de Lomb-Sacargle no procedimento de diagonalização inversa da matriz contendo a densidade espectral estimada. De acordo com os resultados obtidos, pode-se concluir que não há perda de informação assintoticamente, assim o método proposto pode ser utilizando para estimar as funções de autocovariância e autocorrelação, mesmo na presença de dados faltantes. Portanto, os autores sugerem o uso do método proposto em uma série temporal em que há ocorrências de observações faltantes.

Para trabalhos futuros sugere-se: i) que as técnicas de análise espectral discutidas neste trabalho sejam utilizadas para reconstruir, por meio de algum procedimento matemático, as séries temporais de concentrações de poluentes com dados faltantes; ii) os métodos de estimação da densidade espectral propostos neste trabalho foram concentrados em investigações empíricas. Portanto, as propriedades assintóticas dos estimadores e as provas relativas permanecem como problemas abertos para novas linhas de pesquisa; e, iii) estender o estudo de simulação das técnicas propostas nesta Tese para os casos em que as séries temporais além de apresentarem dados faltantes, apresentam comportamento de memória longa e volatilidade.

7 REFERÊNCIAS

ADAMKIEWICZ, G. et al. Nitrogen dioxide concentrations in neighborhoods adjacent to a commercial airport: a land use regression modeling study. **Environmental Health**, BioMed Central, v. 9, n. 1, p. 73, 2010.

AGIRRE-BASURKO, E.; IBARRA-BERASTEGI, G.; MADARIAGA, I. Regression and multilayer perceptron-based models to forecast hourly O_3 and NO_2 levels in the Bilbao area. **Environmental Modelling & Software**, Elsevier, v. 21, n. 4, p. 430–446, 2006.

ALLEN, R. W. et al. The transferability of no and no2 land use regression models between cities and pollutants. **Atmospheric Environment**, Elsevier, v. 45, n. 2, p. 369–378, 2011.

ALLISON, P. D. Estimation of linear models with incomplete data. **Sociological methodology**, JSTOR, p. 71–103, 1987.

ARAIN, M. et al. The use of wind fields in a land use regression model to predict air pollution concentrations for health exposure studies. **Atmospheric Environment**, Elsevier, v. 41, n. 16, p. 3453–3464, 2007.

ARBEX, M. A. et al. A poluição do ar e o sistema respiratório. **Jornal Brasileiro de Pneumologia**, Sociedade Brasileira de Pneumologia e Tisiologia, v. 38, n. 5, p. 643–655, 2012.

ARHAMI, M.; KAMALI, N.; RAJABI, M. M. Predicting hourly air pollutant levels using artificial neural networks coupled with uncertainty analysis by monte carlo simulations. **Environmental Science and Pollution Research**, Springer, v. 20, n. 7, p. 4777–4789, 2013.

ATARI, D. O. et al. Spatial variability of ambient nitrogen dioxide and sulfur dioxide in sarnia,chemical valley, ontario, canada. **Journal of Toxicology and Environmental Health, Part A**, Taylor & Francis, v. 71, n. 24, p. 1572–1581, 2008.

BAIRD, C. **Química Ambiental; trad. Maria Angeles Lobo Recio e Luiz Carlos Marques Carrera**. [S.l.]: Porto Alegre: Bookman, 2002.

BEALE, E. M.; LITTLE, R. J. Missing values in multivariate analysis. **Journal of the Royal Statistical Society. Series B (Methodological)**, JSTOR, p. 129–145, 1975.

BÉJOT, Y. et al. A review of epidemiological research on stroke and dementia and exposure to air pollution. **International Journal of Stroke**, SAGE Publications Sage UK: London, England, v. 13, n. 7, p. 687–695, 2018.

BERGIN, M. S. et al. Regional atmospheric pollution and transboundary air quality management. **Annu. Rev. Environ. Resour.**, Annual Reviews, v. 30, p. 1–37, 2005.

BLOOMFIELD, P. Spectral analysis with randomly missing observations. **Journal of the Royal Statistical Society. Series B (Methodological)**, JSTOR, p. 369–380, 1970.

BONDON, P. Influence of missing values on the prediction of a stationary time series. **Journal of time series analysis**, Wiley Online Library, v. 26, n. 4, p. 519–525, 2005.

BONDON, P.; BAHAMONDE, N. Least squares estimation of arch models with missing observations. **Journal of Time Series Analysis**, Wiley Online Library, v. 33, n. 6, p. 880–891, 2012.

BOWDALO, D. R.; EVANS, M. J.; SOFEN, E. D. Spectral analysis of atmospheric composition: application to surface ozone model–measurement comparisons. **Atmospheric Chemistry and Physics**, Copernicus GmbH, v. 16, n. 13, p. 8295–8308, 2016.

BOWE, B. et al. Associations of ambient coarse particulate matter, nitrogen dioxide, and carbon monoxide with the risk of kidney disease: a cohort study. **The Lancet Planetary Health**, Elsevier, v. 1, n. 7, p. e267–e276, 2017.

BOWE, B. et al. The 2016 global and national burden of diabetes mellitus attributable to pm 2· 5 air pollution. **The Lancet Planetary Health**, Elsevier, v. 2, n. 7, p. e301–e312, 2018.

BOX, G.; JENKINS, G.; REINSEL, G. **Time series analysis: forecasting and control**. 4th. ed. [S.l.]: Prentice Hall, 2008.

BRAGA, A. L. F. et al. Associação entre poluição atmosférica e doenças respiratórias e cardiovasculares na cidade de Itabira, Minas Gerais,Brasil. **Cadernos de Saúde Pública**, SciELO Public Health, v. 23, p. S570–S578, 2007.
BRASSEUR, G. et al. **Atmospheric chemistry and global change**. [S.l.]: Oxford University Press, 1999.
BRAVO, M. A. et al. Air pollution and mortality in sao paulo, brazil: Effects of multiple pollutants and analysis of susceptible populations. **Journal of Exposure Science and Environmental Epidemiology**, Nature Publishing Group, v. 26, n. 2, p. 150, 2016.
BROWN, R. J.; HARRIS, P. M.; COX, M. G. Improved strategies for calculating annual averages of ambient air pollutants in cases of incomplete data coverage. **Environmental Science: Processes & Impacts**, Royal Society of Chemistry, v. 15, n. 5, p. 904–911, 2013.
BRUNEKREEF, B.; HOLGATE, S. T. Air pollution and health. **The lancet**, Elsevier, v. 360, n. 9341, p. 1233–1242, 2002.
BUUREN, S. V.; MULLIGEN, E. van; BRAND, J. Routine multiple imputation in statistical databases. In: IEEE. **Scientific and Statistical Database Management, 1994. Proceedings., Seventh International Working Conference on**. [S.l.], 1994. p. 74–78.
CALDERÓN-GARCIDUEÑAS, L. et al. Air pollution, a rising environmental risk factor for cognition, neuroinflammation and neurodegeneration: the clinical impact on children and beyond. **Revue neurologique**, Elsevier, v. 172, n. 1, p. 69–80, 2016.
CANÇADO, J. E. D. et al. Repercussões clínicas da exposição à poluição atmosférica. **J bras pneumol**, SciELO Brasil, v. 32, n. Supl 1, p. S5–S11, 2006.
CASTRO, H. A. d. et al. Efeitos da poluição do ar na função respiratória de escolares, Rio de Janeiro, RJ. **Revista de Saúde Pública**, SciELO Public Health, v. 43, p. 26–34, 2009.
CHALOULAKOU, A. et al. Measurements of pm10 and pm2. 5 particle concentrations in athens, greece. **Atmospheric Environment**, Elsevier, v. 37, n. 5, p. 649–660, 2003.
CHANG, C.-C.; CHEN, P.-S.; YANG, C.-Y. Short-term effects of fine particulate air pollution on hospital admissions for cardiovascular diseases: A case-crossover study in a tropical city. **Journal of Toxicology and Environmental Health, Part A**, Taylor & Francis, v. 78, n. 4, p. 267–277, 2015.
CHOI, Y.-S. et al. Spectral analysis of weekly variation in pm10 mass concentration and meteorological conditions over china. **Atmospheric Environment**, Elsevier, v. 42, n. 4, p. 655–666, 2008.
COLANTONIO, A. et al. Abba: Adaptive bicluster-based approach to impute missing values in binary matrices. In: ACM. **Proceedings of the 2010 ACM Symposium on Applied Computing**. [S.l.], 2010. p. 1026–1033.
CONAMA. Dispõe sobre padrães de qualidade do ar, previstos no PRONAR. **Diário Oficial da República Federativa do Brasil**, Brasília, DF, p. 15937–15939, junho 1990.
CONAMA. Dispõe sobre padrães de qualidade do ar. **Diário Oficial da República Federativa do Brasil, PublicaÃ§Ã£o DOU nÂº 223, de 21/11/2018**, Brasília, DF, p. 155–156, novembro 2018.
CVITAŠ, T. et al. Spectral analysis of boundary layer ozone data from the eurotrac tor network. **Journal of Geophysical Research: Atmospheres**, Wiley Online Library, v. 109, n. D2, 2004.
DALLAROSA, J. B. **Estudo da formação e dispersão de ozônio troposférico em áreas de atividade de processamento de carvão aplicando modelos numéricos**. 127 f. Dissertação (Mestrado em Sensoriamento Remoto) — Programa de Pós-Gradução em Sensoriamento Remoto, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2005.
DAVISON, A.; HEMPHILL, M. On the statistical analysis of ambient ozone data when measurements are missing. **Atmospheric Environment (1967)**, Elsevier, v. 21, n. 3, p. 629–639, 1987.

DEMPSTER, A.; RUBIN, D. Maximum likelihood from incomplete data via the algorithm em. **Journal of the Royal Statistical Society, B**, v. 37, p. 211–252, 1977.
DERISIO, J. C. **Introdução ao control de poluição ambiental**. [S.l.]: Oficina de Textos, 2016.
DILMAGHANI, S. et al. Harmonic analysis of environmental time series with missing data or irregular sample spacing. **Environmental science & technology**, ACS Publications, v. 41, n. 20, p. 7030–7038, 2007.
DRAKE, C.; KNAPIK, O.; LEŚKOW, J. Em-based inference for cyclostationary time series with missing observations. In: **Cyclostationarity: Theory and Methods**. [S.l.]: Springer, 2014. p. 23–35.
DUNSMUIR, W.; ROBINSON, P. M. Asymptotic theory for time series containing missing and amplitude modulated observations. **Sankhyā: The Indian Journal of Statistics, Series A**, JSTOR, p. 260–281, 1981.
DUNSMUIR, W.; ROBINSON, P. M. Estimation of times series models in the presence of missing data. **Journal of the American Statistical Association**, v. 76, n. 375, p. 560–568, 1981.
DUNSMUIR, W.; ROBINSON, P. M. Parametric estimators for stationary time series with missing observations. **Advances in Applied Probability**, Cambridge Univ Press, v. 13, n. 01, p. 129–146, 1981.
DUTTON, S. J. et al. Temporal patterns in daily measurements of inorganic and organic speciated PM2.5 in Denver. **Atmospheric Environment**, Elsevier, v. 44, n. 7, p. 987–998, 2010.
ECOSOFT CONSULTORIA E SOFTWARES AMBIENTAIS. **Inventário de emissões atmosféricas da Região da Grande Vitória**. Vitória, 2011. Disponível em: ⟨http://www.es.gov.br/Banco%20de%20Documentos/PDF/Maio/100511/RTC10131-R1.pdf⟩. Acesso em: 20 de nov. de 2016.
EFROMOVICH, S. Efficient non-parametric estimation of the spectral density in the presence of missing observations. **Journal of Time Series Analysis**, Wiley Online Library, v. 35, n. 5, p. 407–427, 2014.
ESTRELLA, B. et al. Air pollution control and the occurrence of acute respiratory illness in school children of quito, ecuador. **Journal of public health policy**, Springer, p. 1–18, 2018.
FAJARDO, F. et al. M-periodogram for the analysis of long-range-dependent time series. **Statistics**, Taylor & Francis, v. 52, n. 3, p. 665–683, 2018.
FARHANGFAR, A.; KURGAN, L. A.; PEDRYCZ, W. Experimental analysis of methods for imputation of missing values in databases. In: INTERNATIONAL SOCIETY FOR OPTICS AND PHOTONICS. **Intelligent Computing: Theory and Applications II**. [S.l.], 2004. v. 5421, p. 172–183.
FARHANGFAR, A.; KURGAN, L. A.; PEDRYCZ, W. A novel framework for imputation of missing values in databases. **IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans**, IEEE, v. 37, n. 5, p. 692–709, 2007.
FRAUCHES, D. de O. et al. Doenças respiratórias em crianças e adolescentes: um perfil dos atendimentos na atenção primária em vitória/es. **Revista Brasileira de Medicina de Família e Comunidade**, v. 12, n. 39, p. 1–11, 2017.
FREITAS, C. U. d. et al. Poluição do ar e impactos na saúde em Vitória, Espírito Santo. **Revista de Saúde Pública**, REVISTA DE SAUDE PUBLICA, v. 50, 2016.
FUENTES, M. Interpolation of nonstationary air pollution processes: a spatial spectral approach. **Statistical Modelling**, Sage Publications Sage CA: Thousand Oaks, CA, v. 2, n. 4, p. 281–298, 2002.
GHAZAL, M.; ELHASSANEIN, A. Periodogram analysis with missing observations. **Journal**

of Applied Mathematics and Computing, Springer, v. 22, n. 1, p. 209–222, 2006.
GILL, M. K. et al. Effect of missing data on performance of learning algorithms for hydrologic predictions: Implications to an imputation technique. **Water resources research**, Wiley Online Library, v. 43, n. 7, 2007.
GODISH, T. **Air quality**. 2. ed. Chelsea, Michigan: Lewis, 1991. 422 p.
GOMES, K. **Modelagem INAR (p) para previsão de índices de qualidade do ar**. Dissertação (Mestrado) — Universidade Federal do Espírito Santo, 2009.
GÓMEZ-CARRACEDO, M. et al. A practical comparison of single and multiple imputation methods to handle complex missing data in air quality datasets. **Chemometrics and Intelligent Laboratory Systems**, Elsevier, v. 134, p. 23–33, 2014.
GOUVEIA, N. et al. Poluição do ar e efeitos na saúde nas populações de duas grandes metrópoles brasileiras. **Epidemiologia e Serviços de Saúde**, Coordenação-Geral de Desenvolvimento da Epidemiologia em Serviços/Secretaria de Vigilância em Saúde/Ministério da Saúde, v. 12, n. 1, p. 29–40, 2003.
GOVERNO DO ESTADO DO ESPÍRITO SANTO. Estabelece novos padrões de qualidade do ar e dá providências correlatas. **Diário Oficial do Estado do Espírito Santo**, Vitória, ES, dezembro 2013.
GOYAL, P.; CHAN, A. T.; JAISWAL, N. Statistical models for the prediction of respirable suspended particulate matter in urban cities. **Atmospheric Environment**, Elsevier, v. 40, n. 11, p. 2068–2077, 2006.
GRIPA, W. R. et al. Análise de predição e previsão das concentrações de material particulado inalável (PM_{10}) na cidade de Carapina, ES. **Revista Brasileira de Estatística**, v. 73, n. 237, p. 37–57, 2012.
GULLIVER, J. et al. Land use regression modeling to estimate historic (1962- 1991) concentrations of black smoke and sulfur dioxide for great britain. **Environmental science & technology**, ACS Publications, v. 45, n. 8, p. 3526–3532, 2011.
HARTLEY, H.; HOCKING, R. The analysis of incomplete data. **Biometrics**, JSTOR, p. 783–823, 1971.
HE, H.-D.; LU, W.-Z. Spectral analysis of vehicle pollutants at traffic intersection in hong kong. **Stochastic environmental research and risk assessment**, Springer, v. 26, n. 8, p. 1053–1061, 2012.
HIES, T. et al. Spectral analysis of air pollutants. part 1: elemental carbon time series. **Atmospheric Environment**, Elsevier, v. 34, n. 21, p. 3495–3502, 2000.
HINRICHS, R. A.; KLEINBACH, M.; REIS, L. B. dos. Energia e meio ambiente, tradução da ed 4 americana. **Cengage Learning Edições Ltda, São Paulo, SP**, 2011.
HOCKE, K.; KÄMPFER, N. Gap filling and noise reduction of unevenly sampled data by means of the Lomb-Scargle periodogram. **Atmospheric chemistry and physics**, Copernicus GmbH, v. 9, n. 12, p. 4197–4206, 2009.
HOEK, G. et al. A review of land-use regression models to assess spatial variation of outdoor air pollution. **Atmospheric environment**, Elsevier, v. 42, n. 33, p. 7561–7578, 2008.
HOEK, G. et al. Long-term air pollution exposure and cardio-respiratory mortality: a review. **Environmental health**, BioMed Central, v. 12, n. 1, p. 43, 2013.
HOLGATE, S. T. et al. **Air pollution and health**. [S.l.]: Academic Press, 1999.
IGLESIAS, P.; JORQUERA, H.; PALMA, W. Data analysis using regression models with missing observations and long-memory: an application study. **Computational Statistics & Data Analysis**, v. 50, n. 8, p. 2028–2043, 2006.
INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Banco de dados. Cidades**. Rio de Janeiro, 2014. Disponível em: ⟨http://www.cidades.ibge.gov.br/xtras/home.php⟩. Acesso em: 20 de mar. de 2014.

INSTITUTO ESTADUAL DE MEIO AMBIENTE E RECURSOS HÍDRICOS DO ESTADO DO ESPÍRITO SANTO. **Relatório da qualidade do ar da Região da Grande Vitória**. Vitória, 2014. Disponível em: ⟨http://www.meioambiente.es.gov.br/download/Relat%C3%B3rio_Anual_de_Qualidade_do_Ar_2013.pdf⟩. Acesso em: 27 de jun. de 2016.

INSTITUTO ESTADUAL DE MEIO AMBIENTE E RECURSOS HÍDRICOS DO ESTADO DO ESPÍRITO SANTO. **Norma de Procedimento - IEMA NÂº 003: Procedimento Operacional do Monitoramento da Qualidade do Ar**. Vitória, 2018. Disponível em: ⟨https://iema.es.gov.br/Media/iema/CQAI/Normas\%20e\%20procedimentos/IEMA\%20003\%20-\%20Normatiza\%C3\%A7\%C3\%A3o\%20Monitoramento\%20CQAI\%20(1).pdf⟩. Acesso em: 21 de jan. de 2019.

IORDACHE, Ş.; DUNEA, D. Cross-spectrum analysis applied to air pollution time series from several urban areas of Romania. **Environmental Engineering & Management Journal (EEMJ)**, v. 12, n. 4, 2013.

JACOBS, E. T.; BURGESS, J. L.; ABBOTT, M. B. The donora smog revisited: 70 years after the event that inspired the clean air act. **American journal of public health**, American Public Health Association, v. 108, n. S2, p. S85–S88, 2018.

JERRETT, M. et al. A review and evaluation of intraurban air pollution exposure models. **Journal of Exposure Science and Environmental Epidemiology**, Nature Publishing Group, v. 15, n. 2, p. 185, 2005.

JIAN, L. et al. An application of arima model to predict submicron particle concentrations from meteorological factors at a busy roadside in hangzhou, china. **Science of the Total Environment**, Elsevier, v. 426, p. 336–345, 2012.

JORQUERA, H. et al. Forecasting ozone daily maximum levels at Santiago, Chile. **Atmospheric Environment**, Elsevier, v. 32, n. 20, p. 3415–3424, 1998.

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

JUNGER, W. L.; LEON, A. P. Imputation of missing data in time series for air pollutants. **Atmospheric Environment**, Elsevier, v. 102, p. 96–104, 2015.

JUNNINEN, H. et al. Methods for imputation of missing values in air quality data sets. **Atmospheric Environment**, v. 38, n. 18, p. 2895–2907, 2004.

KAI, S. et al. Using three methods to investigate time-scaling properties in air pollution indexes time series. **Nonlinear Analysis: Real World Applications**, Elsevier, v. 9, n. 2, p. 693–707, 2008.

KANAROGLOU, P. S. et al. Estimation of sulfur dioxide air pollution concentrations with a spatial autoregressive model. **Atmospheric Environment**, Elsevier, v. 79, p. 421–427, 2013.

KANDLIKAR, M. Air pollution at a hotspot location in delhi: detecting trends, seasonal cycles and oscillations. **Atmospheric environment**, Elsevier, v. 41, n. 28, p. 5934–5947, 2007.

KELSALL, J. E.; ZEGER, S. L.; SAMET, J. M. Frequency domain log-linear models; air pollution and mortality. **Journal of the Royal Statistical Society: Series C (Applied Statistics)**, Wiley Online Library, v. 48, n. 3, p. 331–344, 1999.

KÖPPEN, W. Versuch einer klassifikation der klimate, vorzugsweise nach ihren beziehungen zur pflanzenwelt. **Geographische Zeitschrift**, JSTOR, v. 6, n. 11. H, p. 593–611, 1900.

KUMAR, U.; RIDDER, K. D. Garch modelling in association with FFT–ARIMA to forecast ozone episodes. **Atmospheric Environment**, Elsevier, v. 44, n. 34, p. 4252–4265, 2010.

LAKSHMINARAYAN, K.; HARP, S. A.; SAMAD, T. Imputation of missing data in industrial databases. **Applied intelligence**, Springer, v. 11, n. 3, p. 259–275, 1999.

LANDRIGAN, P. J. et al. The lancet commission on pollution and health. **The Lancet**,

Elsevier, v. 391, n. 10119, p. 462–512, 2018.
LATORRE, M. R. D. O.; CARDOSO, M. R. A. Análise de séries temporais em epidemiologia: uma introdução sobre os aspectos metodológicos. **Rev. bras. epidemiol**, v. 4, n. 3, p. 145–152, 2001.
LEITE, R. C. M. et al. Utilização de regressão logística simples na verificação da qualidade do ar atmosférico de uberlândia. **Engenharia Sanitária e Ambiental**, SciELO Brasil, v. 16, n. 1, 2011.
LEROY, B. Fast calculation of the lomb-scargle periodogram using nonequispaced fast fourier transforms. **Astronomy & Astrophysics**, EDP Sciences, v. 545, p. A50, 2012.
LIRA, R. **Modelagem e previsão da qualidade do ar na cidade de Uberlândia–MG**. 152 f. Tese (Doutorado em Engenharia Química) — Programa de Pós-Gradução em Engenharia Química, Universidade Federal de Uberlândia, Uberlândia, Minas Gerais, 2009.
LITTLE, R. J. Regression with missing xś: a review. **Journal of the American Statistical Association**, Taylor & Francis Group, v. 87, n. 420, p. 1227–1237, 1992.
LITTLE, R. J.; RUBIN, D. B. **Statistical analysis with missing data**. [S.l.]: John Wiley & Sons, 2014. v. 333.
LIU, H. et al. Association of short-term exposure to ambient carbon monoxide with hospital admissions in china. **Scientific reports**, Nature Publishing Group, v. 8, n. 1, p. 13336, 2018.
LIU, P.-W. G. Establishment of a Box-Jenkins multivariate time-series model to simulate ground-level peak daily one-hour ozone concentrations at Ta-Liao in Taiwan. **Journal of the Air & Waste Management Association**, v. 57, n. 9, p. 1078–1090, 2007.
LIU, P.-W. G. Simulation of the daily average PM_{10} concentrations at Ta-Liao with Box-Jenkins time series models and multivariate analysis. **Atmospheric Environment**, v. 43, n. 13, p. 2104–2113, 2009.
LIU, P.-W. G.; JOHNSON, R. Forecasting peak daily ozone levels-I. A regression with time series errors model having a principal component trigger to fit 1991 ozone levels. **Journal of the Air & Waste Management Association**, v. 52, n. 9, p. 1064–1074, 2002.
LIU, P.-W. G.; JOHNSON, R. Forecasting peak daily ozone levels: part 2- A regression with time series errors model having a principal component trigger to forecast 1999 and 2002 ozone levels. **Journal of the Air & Waste Management Association**, v. 53, n. 12, p. 1472–1489, 2003.
LIU, P.-W. G. et al. Establishing multiple regression models for ozone sensitivity analysis to temperature variation in Taiwan. **Atmospheric Environment**, v. 79, p. 225–235, 2013.
LOGAN, W. et al. Mortality in the london fog incident, 1952. **Lancet**, p. 336–8, 1953.
LOMB, N. R. Least-squares frequency analysis of unequally spaced data. **Astrophysics and space science**, Springer, v. 39, n. 2, p. 447–462, 1976.
LYRA, G. B.; ODA-SOUZA, M.; VIOLA, D. N. Modelos lineares aplicados à estimativa da concentração do material particulado (pm10) na cidade do rio de janeiro, rj. **Revista Brasileira de Meteorologia**, SciELO Brasil, v. 26, n. 3, p. 392–400, 2011.
MARR, L. C.; HARLEY, R. A. Spectral analysis of weekday–weekend differences in ambient ozone, nitrogen oxide, and non-methane hydrocarbon time series in california. **Atmospheric Environment**, Elsevier, v. 36, n. 14, p. 2327–2335, 2002.
MARTINS, L. C. et al. Poluição atmosférica e atendimentos por pneumonia e gripe em São Paulo, Brasil. **Revista de Saúde Pública**, SciELO Public Health, v. 36, n. 1, p. 88–94, 2002.
MATOS, E. P. **Estudo epidemiológico espacial e temporal na análise da associação entre poluição do ar e o número de atendimentos hospitalares por causas respiratórias em crianças**. 231 f. Dissertação (Mestrado) — Programa de Pós-Gradução em Engenharia Ambiental, Universidade Federal do Espírito Santo, Vitória, ES, 2012.
MAYNARD, R. Key airborne pollutants-the impact on health. **Science of the Total**

Environment, Elsevier, v. 334, p. 9–13, 2004.
MCKNIGHT, P. E. et al. **Missing data: A gentle introduction**. [S.l.]: Guilford Press, 2007.
METAXOGLOU, K.; SMITH, A. Maximum likelihood estimation of varma models using a state-space em algorithm. **Journal of Time Series Analysis**, Wiley Online Library, v. 28, n. 5, p. 666–685, 2007.
MIHELCIC, J. R.; ZIMMERMAN, J. B.; AUER, M. T. **Environmental engineering: Fundamentals, sustainability, design**. [S.l.]: Wiley Hoboken, 2014. v. 1.
MILLER, L. et al. Evaluation of missing value methods for predicting ambient btex concentrations in two neighbouring cities in southwestern ontario canada. **Atmospheric Environment**, Elsevier, v. 181, p. 126–134, 2018.
MONTE, E. Z. **Análise de componentes principais em séries temporais multivariadas com heteroscedasticidade condicional e outliers: uma aplicação para a poluição do ar, na Região da Grande Vitória, Espírito Santo, Brasil**. 167 f. Tese (Doutorado em Engenharia Ambiental) — Programa de Pós-Gradução em Engenharia Ambiental, Universidade Federal do Espírito Santo, Vitória, ES, 2016.
MONTE, E. Z.; ALBUQUERQUE, T. T. D. A.; REISEN, V. A. Ozone concentration forecast in the Region of Grande Vitória, Espírito Santo, Brazil, using the armax-garch model. **Revista Brasileira de Meteorologia**, SciELO Brasil, v. 30, n. 3, p. 285–294, 2015.
MONTE, E. Z.; ALBUQUERQUE, T. T. d. A.; REISEN, V. A. Impactos das variáveis meteorológicas na qualidade do ar da região da grande vitórria, espírito santo, brasil. **Revista Brasileira de Meteorologia**, scielo, v. 31, p. 546–554, 12 2016. ISSN 0102-7786. Disponível em: ⟨http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-77862016000500546&nrm=iso⟩.
MONTE, E. Z.; ALBUQUERQUE, T. T. d. A.; REISEN, V. A. Inter-relações entre as concentrações de ozônio e de dióxido de nitrogênio na região da grande vitória, espírito santo, brasil. **Engenharia Sanitaria e Ambiental**, scielo, v. 22, p. 679–690, 08 2017. ISSN 1413-4152. Disponível em: ⟨http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-41522017000400679&nrm=iso⟩.
MORETTIN, P. A.; TOLOI, C. **Análise de Séries Temporais**. [S.l.]: ABE: Projeto Fisher, 2006.
MOSHENBERG, S.; LERNER, U.; FISHBAIN, B. Spectral methods for imputation of missing air quality data. **Environmental Systems Research**, SpringerOpen, v. 4, n. 1, p. 26, 2015.
NASCIMENTO, A. P. et al. Association between the concentration of fine particles in the atmosphere and acute respiratory diseases in children. **Revista de saude publica**, SciELO Public Health, v. 51, p. 3, 2017.
NASCIMENTO, L. et al. Efeitos da poluição atmosférica na saúde infantil em São José dos Campos, SP. **Revista de Saúde Pública**, v. 40, p. 77–82, 2006.
NEGRETE, B. R. et al. Poluição atmosférica e internações por insuficiência cardíaca congestiva em adultos e idosos em Santo André (SP). **Arquivos Brasileiros de Ciências da Saúde**, v. 35, n. 3, 2010.
NORAZIAN, M. N. et al. Estimation of missing values in air pollution data using single imputation techniques. **ScienceAsia**, v. 34, p. 341–345, 2008.
NUNES, L. N.; KLÜCK, M. M.; FACHEL, J. M. G. Uso da imputação múltipla de dados faltantes: uma simulação utilizando dados epidemiológicos. **Cadernos de Saúde Pública**, SciELO Public Health, v. 25, p. 268–278, 2009.
OLTMANS, S. et al. Long-term changes in tropospheric ozone. **Atmospheric Environment**, Elsevier, v. 40, n. 17, p. 3156–3173, 2006.
OSTRO, B. et al. Air pollution and mortality: results from a study of santiago, chile. **Journal of exposure analysis and environmental epidemiology**, v. 6, n. 1, p. 97–114, 1995.

OSTRO, B. et al. Air pollution and mortality: results from a study of santiago, chile. **Journal of Exposure Analysis and Environmental Epidemiology**, v. 6, p. 97–114, 1996.
OSTRO, B. et al. Air pollution and mortality: results from a study of santiago, chile. **Journal of Exposure Analysis and Environmental Epidemiology**, v. 6, p. 97–114, 1996.
OSTRO, B. D. et al. Air pollution and health effects: A study of medical visits among children in Santiago, Chile. **Environmental Health Perspectives**, v. 107, p. 69–73, 1999.
OSTRO, B. D.; HURLEY, S.; LIPSETT, M. J. Air pollution and daily mortality in the coachella valley, california: A study of PM_{10} dominated by coarse particles. **Environmental Research**, v. 81, p. 231–238, 1999.
PARZEN, E. On spectral analysis with missing observations and amplitude modulation. **Sankhyā: The Indian Journal of Statistics, Series A**, JSTOR, n. 4, p. 383–392, 1963.
PASCHALIDOU, A. K. et al. Forecasting hourly pm 10 concentration in cyprus through artificial neural networks and multiple regression models: implications to local environmental management. **Environmental Science and Pollution Research**, Springer, v. 18, n. 2, p. 316–327, 2011.
PIRES, J. et al. Identification of redundant air quality measurements through the use of principal component analysis. **Atmospheric environment**, Elsevier, v. 43, n. 25, p. 3837–3842, 2009.
PIRES, J. et al. Management of air quality monitoring using principal component and cluster analysis-part i: So2 and pm10. **Atmospheric Environment**, Elsevier, v. 42, n. 6, p. 1249–1260, 2008.
PLAIA, A.; BONDÌ, A. L. Single imputation method of missing values in environmental pollution data sets. **Atmospheric Environment**, Elsevier, v. 40, n. 38, p. 7316–7330, 2006.
PLAIA, A.; BONDÌ, A. L. Regression imputation for space-time datasets with missing values. In: **Data Analysis and Classification**. [S.l.]: Springer, 2010. p. 465–472.
Pope III, C. A. et al. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. **Jama**, American Medical Association, v. 287, n. 9, p. 1132–1141, 2002.
PRESS, W. H.; RYBICKI, G. B. Fast algorithm for spectral analysis of unevenly sampled data. **The Astrophysical Journal**, v. 338, p. 277–280, 1989.
PRIESTLEY, M. B. **Spectral Analysis and Time Series**. [S.l.]: Academic Press, 1981.
R Development Core Team. **R: A language and environment for statistical computing**. Vienna, Austria, 2016. Disponível em: ⟨http://www.r-project.org/⟩.
RAO, S. T.; SAMSON, P.; PEDDADA, A. Spectral analysis approach to the dynamics of air pollutants. **Atmospheric Environment (1967)**, Elsevier, v. 10, n. 5, p. 375–379, 1976.
RAO, S. T.; ZURBENKO, I. G. Detecting and tracking changes in ozone air quality. **Air & waste**, Taylor & Francis, v. 44, n. 9, p. 1089–1092, 1994.
RAZ, R. et al. Autism spectrum disorder and particulate matter air pollution before, during, and after pregnancy: a nested case-control analysis within the nurseśhealth study ii cohort. **Environmental health perspectives**, NLM-Export, v. 123, n. 3, p. 264–270, 2014.
REISEN, V. A. et al. Robust estimation of fractional seasonal processes: Modeling and forecasting daily average so2 concentrations. **Mathematics and Computers in Simulation**, Elsevier, v. 146, p. 27–43, 2018.
REISEN, V. A. et al. Modeling and forecasting daily average PM_{10} concentrations by a seasonal long-memory model with volatility. **Environmental Modelling & Software**, Elsevier, v. 51, p. 286–295, 2014.
REISEN, V. A. et al. Robust factor modelling for high-dimensional time series: An application to air pollution data. **Applied Mathematics and Computation**, Elsevier, v. 346, p. 842–852, 2019.

REISEN, V. A.; SILVA, A. N. **O uso da linguagem R para cálculos de Estatística Básica**. [S.l.]: Editora da Universidade Federal do Espírito Santo, 2011.
RUBIN, D. B. Inference and missing data. **Biometrika**, Oxford University Press, v. 63, n. 3, p. 581–592, 1976.
SALCEDO, R. et al. Time-series analysis of air pollution data. **Atmospheric Environment**, Elsevier, v. 33, n. 15, p. 2361–2372, 1999.
SARNAGLIA, A. J. Q.; MONROY, N. A. J.; VITÓRIA, A. G. da. Modeling and forecasting daily maximum hourly ozone concentrations using the regar model with skewed and heavy-tailed innovations. **Environmental and Ecological Statistics**, Springer, v. 25, n. 4, p. 443–469, 2018.
SCARGLE, J. D. Studies in astronomical time series analysis. ii-statistical aspects of spectral analysis of unevenly spaced data. **The Astrophysical Journal**, v. 263, p. 835–853, 1982.
SCHAFER, J. L. **Analysis of incomplete multivariate data**. [S.l.]: Chapman and Hall/CRC, 1997.
SCHLINK, U.; HERBARTH, O.; TETZLAFF, G. A component time-series model for so2 data: Forecasting, interpretation and modification. **Atmospheric Environment**, Elsevier, v. 31, n. 9, p. 1285–1295, 1997.
SEBALD, L. et al. Spectral analysis of air pollutants. part 2: ozone time series. **Atmospheric Environment**, Elsevier, v. 34, n. 21, p. 3503–3509, 2000.
SOUSA, S. et al. Prediction of ozone concentrations in oporto city with statistical approaches. **Chemosphere**, Elsevier, v. 64, n. 7, p. 1141–1149, 2006.
SOUZA, J. B. d. et al. Componentes principais e modelagem linear generalizada na associação entre atendimento hospitalar e poluição do ar. **Revista de Saúde Pública**, v. 48, n. 3, p. 451–458, 2014.
TAKEUCHI, K. A comment on"recent development of economic data analysis"at the 63rd annual meeting of japan statistical society. 1995.
TCHEPEL, O.; BORREGO, C. Frequency analysis of air quality time series for traffic related pollutants. **Journal of Environmental Monitoring**, Royal Society of Chemistry, v. 12, n. 2, p. 544–550, 2010.
TCHEPEL, O. et al. Determination of background concentrations for air quality models using spectral analysis and filtering of monitoring data. **Atmospheric Environment**, Elsevier, v. 44, n. 1, p. 106–114, 2010.
TIBUAKUU, M. et al. Air pollution and cardiovascular disease: a focus on vulnerable populations worldwide. **Current Epidemiology Reports**, Springer, v. 5, n. 4, p. 370–378, 2018.
TO, T. et al. Progression from asthma to chronic obstructive pulmonary disease. is air pollution a risk factor? **American journal of respiratory and critical care medicine**, American Thoracic Society, v. 194, n. 4, p. 429–438, 2016.
TODA, H. Y.; MCKENZIE, C. R. Lm tests for unit roots in the presence of missing observations: small sample evidence. **Mathematics and computers in simulation**, Elsevier, v. 48, n. 4, p. 457–468, 1999.
TUFIK, S. et al. Revisão sistemática sobre a epidemiologia das doenças cardiovasculares e respiratórias e suas associações com a poluição do ar em vitória/es. **Clinical & Biomedical Research**, v. 37, n. 2, 2017.
VARDOULAKIS, S.; KASSOMENOS, P. Sources and factors affecting pm10 levels in two european cities: Implications for local air quality management. **Atmospheric Environment**, Elsevier, v. 42, n. 17, p. 3949–3963, 2008.
VERONEZE, R. **Tratamento de dados faltantes empregando biclusterização com imputação múltipla**. 238 f. Dissertação (Mestrado em Engenharia Elétrica) — Programa de

Pós-Gradução em Engenharia Elétrica, Universidade Estadual de Campinas, Campinas, SP, 2011.
VINGARZAN, R. A review of surface ozone background levels and trends. **Atmospheric environment**, Elsevier, v. 38, n. 21, p. 3431–3442, 2004.
WANG, H. et al. An urban-rural and sex differences in cancer incidence and mortality and the relationship with pm2. 5 exposure: An ecological study in the southeastern side of hu line. **Chemosphere**, Elsevier, v. 216, p. 766–773, 2019.
WHEELER, A. J. et al. Intra-urban variability of air pollution in windsor, ontario-measurement and modeling for human exposure assessment. **Environmental Research**, Elsevier, v. 106, n. 1, p. 7–16, 2008.
WORLD HEALTH ORGANIZATION. **WHO air quality guidelines global update 2005. Report on a working group meeting, Bonn/Germany**. Copenhagen, 2005. Disponível em: ⟨http://www.euro.who.int/_data/assets/pdf_file/0008/147851/E87950.pdf⟩. Acesso em: 18 de jan. de 2019.
WORLD HEALTH ORGANIZATION. **Air pollution estimates**. Copenhagen, 2014. Disponível em: ⟨http://www.who.int/phe/health_topics/outdoorair/databases/FINAL_HAP_AAP_BoD_24March2014.pdf?ua=1⟩. Acesso em: 18 de jun. de 2018.
WU, C.-H.; WUN, C.-H.; CHOU, H.-J. Using association rules for completing missing data. In: IEEE. **Hybrid Intelligent Systems, 2004. HIS'04. Fourth International Conference on**. [S.l.], 2004. p. 236–241.
XIA, Y. et al. Forest climatology: estimation of missing values for bavaria, germany. **Agricultural and Forest Meteorology**, Elsevier, v. 96, n. 1-3, p. 131–144, 1999.
XIONG, Q. et al. Fine particulate matter pollution and hospital admissions for respiratory diseases in beijing, china. **International journal of environmental research and public health**, Multidisciplinary Digital Publishing Institute, v. 12, n. 9, p. 11880–11892, 2015.
YAJIMA, Y.; NISHINO, H. Estimation of the autocorrelation function of a stationary time series with missing observations. **Sankhyā: The Indian Journal of Statistics, Series A**, JSTOR, p. 189–207, 1999.
ZAINURI, N. A.; JEMAIN, A. A.; MUDA, N. A comparison of various imputation methods for missing values in air quality data. **Sains Malaysiana**, Penerbit Universiti Kebangsaan Malaysia, v. 44, n. 3, p. 449–456, 2015.
ZAKARIA, N. A.; NOOR, N. M. Imputation methods for filling missing data in urban air pollution data formalaysia. **Urbanism. Arhitectura. Constructii**, Institutul National de Cercetare-Dezvoltare in Constructii, Urbanism. Arhitecturã. Constructii, v. 9, n. 2, p. 159, 2018.
ZHANG, P. Multiple imputation: theory and method. **International Statistical Review**, Wiley Online Library, v. 71, n. 3, p. 581–592, 2003.
ZHANG, X.; CHEN, X.; ZHANG, X. The impact of exposure to air pollution on cognitive performance. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 115, n. 37, p. 9193–9197, 2018.
ZÚÑIGA, J. et al. Assessment of the possible association of air pollutants pm10, o3, no2 with an increase in cardiovascular, respiratory, and diabetes mortality in panama city: a 2003 to 2013 data analysis. **Medicine**, Wolters Kluwer Health, v. 95, n. 2, 2016.

8 APÊNDICE: ESTUDOS ADICIONAIS

Neste apêndice encontram-se os quatro artigos adicionais desta tese, que estão diretamente ligados ao tema central da pesquisa. Os artigos estão formatados de acordo com as normas das revistas as quais os artigos foram aceitos ou submetidos.

DOI: 10.1590/S1413-41522018168758

Artigo Técnico

Previsão da concentração de material particulado inalável, na Região da Grande Vitória, ES, Brasil, utilizando o modelo SARIMAX

Inhalable particulate matter concentration forecast, in the Greater Vitória Region, ES, Brazil, using the SARIMAX model

Wanderson de Paula Pinto[1], Valdério Anselmo Reisen[2], Edson Zambon Monte[3]

RESUMO

Este trabalho objetivou modelar e prever a concentração média diária de material particulado inalável (MP_{10}), na Região da Grande Vitória (RGV), Espírito Santo, Brasil, utilizando o modelo SARIMAX para o período de 01/01/2012 a 30/04/2015. Os dados deste estudo foram do tipo séries temporais de concentrações de MP_{10} e de variáveis meteorológicas (velocidade do vento, umidade relativa, precipitação pluvial e temperatura), obtidas junto ao Instituto Estadual de Meio Ambiente e Recursos Hídricos (IEMA), sendo escolhida a estação da Enseada do Suá para fazer o estudo de predição e previsão. Baseando-se em indicadores de desempenho de modelagem, verificou-se que o modelo SARIMAX $(1,0,2)\ (0,1,1)_7$ é o mais acurado entre os estudados, objetivando fazer predições e previsões da qualidade do ar na RGV. Em comparação com os modelos ARMA, o desempenho estatístico do modelo SARIMAX foi superior, no que diz respeito à predição de eventos de qualidade do ar regular. Dentre as variáveis meteorológicas avaliadas, a velocidade do vento e a precipitação pluvial foram significativas e melhoraram o ajuste do modelo. Em termos de previsão da qualidade do ar, os modelos de séries temporais mostraram resultados satisfatórios.

Palavras-chave: poluição do ar; MP_{10}; séries temporais; SARIMAX.

ABSTRACT

This study aimed to model and forecast the average daily concentration of inhalable particulate matter (PM_{10}), in the Greater Vitoria Region (GVR), Espírito Santo, Brazil, using the SARIMAX model, for the period from January 1st, 2012 to April 30th, 2015. Data set from the State Environmental Institute was used. The Enseada do Suá station was chosen for purposes of prediction and forecasting. Some meteorological parameters (wind speed, relative humidity, rainfall and temperature) measured at the GVR were taken as explanatory variables of PM_{10} concentrations. Based on modelling performance indicators, it was verified the SARIMAX model $(1.0.2)\ (0.1.1)_7$ is the most accurate between the ones studied, purposing to predict and forecast the air quality in the GVR. The statistical performance of the SARIMAX model was better than the ARMA model, with regard to prediction of regular air quality events. Among the evaluated meteorological variables, wind speed and rainfall were significant and improved the model estimated. Regarding to air quality forecasting, the time series models showed satisfactory results.

Keywords: air pollution; PM_{10}; time series; SARIMAX.

INTRODUÇÃO

A poluição atmosférica caracteriza-se basicamente pela presença de gases tóxicos e partículas sólidas no ar (SEINFELD & PANDIS, 2006). Suas fontes são classificadas em antropogênicas, como, por exemplo, emissões provenientes de indústrias e escapamentos de veículos; e naturais, como as decorrentes de erupções vulcânicas. Segundo Holgate *et al.* (1999), um nível elevado dos poluentes pode ocasionar desde irritação nos olhos, nariz e garganta, bronquite e pneumonia, até doenças respiratórias crônicas, câncer de pulmão, problemas cardíacos etc. Diversos estudos epidemiológicos têm demonstrado associações significativas entre a exposição às concentrações elevadas de poluentes atmosféricos e problemas de saúde (OSTRO *et al.*, 1996; MARTINS *et al.*, 2002; GOUVEIA *et al.*, 2003;

[1]Doutorando em Engenharia Ambiental pela Universidade Federal do Espírito Santo (UFES) – Vitória (ES), Brasil.
[2]Professor do Departamento de Estatística, do Programa de Pós-Graduação em Engenharia Ambiental e do Programa de Pós-Graduação em Economia da UFES – Vitória (ES), Brasil.
[3]Professor do Departamento de Economia e do Programa de Pós-Graduação em Economia e membro do Grupo de Pesquisa em Econometria da UFES – Vitória (ES), Brasil.
***Autor correspondente:** wandersondpp@gmail.com
Recebido: 02/09/16 – **Aceito:** 23/02/17 – **Reg. ABES:** 168758

ALMEIDA, 2006; NASCIMENTO *et al.*, 2006; CURTIS *et al.*, 2006; BRAGA *et al.*, 2007; SOUZA *et al.*, 2014).

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 de grupos de risco, como crianças, idosos e pessoas com doenças respiratórias. Além disso, também pode servir para as autoridades competentes como informação para a preparação de planos para redução de emissões e gerenciamento da qualidade do ar (GOMES, 2009).

Na literatura, existem vários trabalhos que utilizaram modelos estatísticos para modelar e fazer previsão da qualidade do ar. Agirre-Basurko, Ibarra-Berastegui e Madariaga (2006) utilizaram três modelos, um de regressão linear múltipla e dois de rede neural para modelar e prever a qualidade do ar da cidade de Bilbao, Espanha. O fluxo de veículos e as variáveis meteorológicas foram usados como modelos de dados de entrada – temperatura, umidade relativa, pressão, radiação, gradiente de temperatura, direção do vento e velocidade do vento – no período de 1993 a 1994. Como saída prevista para modelos, adotou-se as concentrações de O_3 e NO_2, com horizonte de previsão de oito horas à frente. Os resultados mostraram que os modelos de rede neural obtiveram resultados melhores para a previsão das concentrações de O_3 e NO_2 quando comparados ao modelo de regressão linear múltipla. Quanto ao desempenho dos modelos de rede neural, o que mais se destacou foi o modelo que considerou a sazonalidade da série das concentrações de O_3 e NO_2.

Goyal, Chan e Jaiswal (2006) realizaram um estudo com três modelos estatísticos aplicados à média diária de concentração de MP_{10}, medido nas cidades de Delhi e Hong Kong. O trabalho objetivou desenvolver um modelo estatístico de previsão das concentrações de MP_{10} e promover um estudo comparativo através do desempenho dos modelos, a saber:

1. modelo de regressão linear múltipla (modelo 1);
2. modelo de séries temporais ARIMA (modelo 2);
3. combinação entre os modelos 1 e 2 (modelo 3).

Além do MP_{10}, alguns parâmetros meteorológicos foram adotados, como a velocidade do vento, a temperatura, a radiação solar e a umidade relativa do ar, medidos no período de junho de 2000 a junho de 2001. 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 a junho de 2002. O modelo 3 foi utilizado, e os resultados da previsão foram satisfatórios.

Gomes (2009) realizou um estudo de previsão de índices de qualidade do ar da Região da Grande Vitória (RGV), Espírito Santo, Brasil, utilizando o modelo autorregressivo de valores inteiros INAR(p). O período de análise foi de 01/01/07 a 19/03/07, sendo as previsões datadas de 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, conforme a Resolução do Conselho Nacional do Meio Ambiente (CONAMA, 1990) nº 03. Porém, baseados nas diretrizes da World Health Organization (WHO, 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, ou seja, mesmo estando dentro o limite do padrão nacional essa concentração é prejudicial para a saúde humana.

Gripa *et al.* (2012) compararam dois modelos: um de séries temporais e um de regressão linear múltipla, para modelagem e previsão das concentrações médias de MP_{10}, monitorados na RGV, com a incorporação de fatores meteorológicos. Ambos os modelos evidenciaram resultados semelhantes. No entanto, o modelo de regressão apresentou medidas de previsão das concentrações médias de MP_{10} um pouco melhores do que o modelo de séries temporais. Reisen *et al.* (2014) modelaram a média diária de concentração de MP_{10}, na cidade de Cariacica, Espírito Santo, Brasil, utilizando um processo integrado fracionado sazonal, com volatilidade. Os autores concluíram que o modelo, ajustado com erros heterocedásticos, captou bem a dinâmica da série e foi capaz de prever os períodos de maior volatilidade.

Monte, Albuquerque e Reisen (2015) realizaram um estudo para estimar e prever a concentração horária de ozônio na RGV, utilizando um modelo ARMAX-GARCH no período de 01/01/2011 a 31/12/2011. O estudo utilizou dados cedidos pelo Instituto Estadual de Meio Ambiente e Recursos Hídricos (IEMA). De acordo com os resultados apresentados, as previsões horárias para o dia 31/12/2011 foram muito próximas dos valores observados. Observou-se também que as estimativas seguiram a trajetória diária da concentração de ozônio. Os autores concluíram que o modelo ARMAX-GARCH é mais eficaz na predição de episódios de poluição de ozônio, em comparação aos modelos autorregressivos e de médias móveis (ARMA) e autorregressivos e de médias móveis com variáveis explicativas (ARMAX).

Vale ressaltar que uma série temporal é composta, em geral, por três componentes não observáveis, a saber: tendência, sazonalidade e aleatoriedade. A sazonalidade é uma componente difícil de ser modelada, pois é necessário compatibilizar a questão física do problema em estudo com a questão estatística. Define-se um fenômeno sazonal como aquele que ocorre regularmente em períodos fixos de tempo (LATORRE & CARDOSO, 2001). Na análise de séries temporais, os

modelos ARMA podem ser empregados quando a série em estudo está livre de tendência e de sazonalidade, os modelos ARIMA são utilizados quando há tendência e, para incorporar o componente de sazonalidade, utilizam-se os modelos autorregressivos integrados e de médias móveis sazonal multiplicativo (SARIMA) (LATORRE & CARDOSO, 2001; MORETTIN & TOLOI, 2006). Conforme Reisen *et al.* (2014), os modelos que descrevem de forma adequada o comportamento físico do dados são essenciais para a previsão precisa em qualquer área de aplicação, pois a sazonalidade é um fenômeno característico do poluente MP_{10}.

Nesse contexto, este trabalho objetivou modelar e prever a concentração média diária de MP_{10}, na RGV, utilizando o modelo SARIMAX, no período 01/01/2012 a 30/04/2015. Mesmo tendo ultrapassado apenas 3 vezes os padrões primário e secundário (150 $\mu g.m^{-3}$), média de 24 horas, estabelecidos pela Resolução CONAMA nº 03, de 28/06/1990 (CONAMA, 1990), no período de estudo, em 144 dias, de um total de 1.216, as concentrações monitoradas na estação da Enseada do Suá excederam o valor de 50 $\mu g.m^{-3}$, o que vai de encontro com as diretrizes estabelecidas pela WHO (2005) para esse poluente. O IEMA estabelece uma qualidade do ar boa, para concentrações de MP_{10} entre 0 e 45 $\mu g.m^{-3}$; regular para aquelas entre 46 e 120 $\mu g.m^{-3}$; inadequada, entre 121 e 250 $\mu g.m^{-3}$ (IEMA, 2013). No período de estudo, em 450 ocasiões a qualidade do ar foi classificada como regular; e em 6, como inadequadas. Logo, a importância desta pesquisa é justificada, principalmente no que diz respeito à formulação de medidas preventivas por parte das autoridades competentes, uma vez que a concentração de MP_{10} vem atingindo níveis que são prejudiciais à saúde na região de estudo.

MATERIAL E MÉTODOS

Área de estudo e variáveis analisadas

Para a realização desse estudo, utilizou-se séries temporais de concentrações de poluentes atmosféricos e de variáveis meteorológicas monitorados na RGV, que é constituída pelos municípios de Vitória, Vila Velha, Cariacica, Serra e Viana. Segundo o Instituto Brasileiro de Geografia e Estatística (IBGE, 2010), a RGV abrange uma área de 1.461 km^2, com aproximadamente 1.475.332 habitantes, sendo um dos principais polos de desenvolvimento urbano e industrial do estado. A região sofre com diversos tipos de problemas ambientais, entre os quais está a deterioração da qualidade do ar, devido às emissões atmosféricas por indústrias, pela frota veicular e ressuspensão do solo causada pelo vento e tráfego veicular.

Vale ressaltar que a RGV possui uma Rede Automática de Monitoramento da Qualidade do Ar (RAMQAR), inaugurada em julho de 2000, de propriedade do IEMA. Essa rede é distribuída em oito estações, a saber:

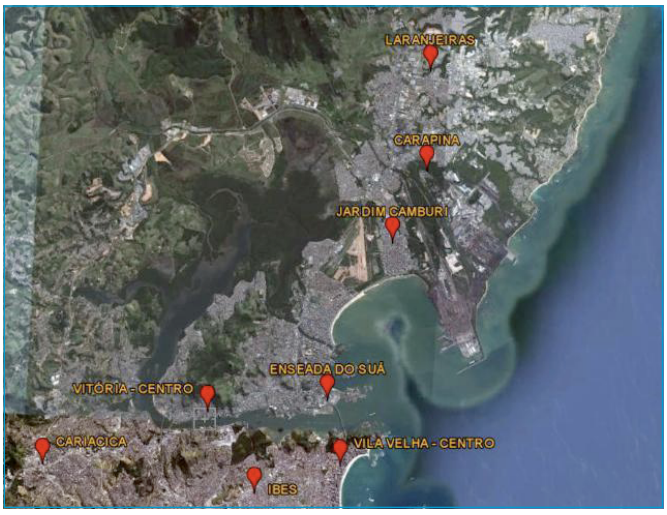
1. o município Serra; com duas estações localizadas nas regiões de Laranjeiras e Carapina;
2. o município Vitória, com três estações localizadas nas regiões de Jardim Camburi, Enseada do Suá e Centro de Vitória;
3. o município de Vila Velha apresenta duas estações localizadas nas regiões do Ibes e Centro de Vila Velha;
4. o município de Cariacica, com uma estação em Cariacica.

A localização espacial das estações de monitoramento da RAMQAR está ilustrada na Figura 1.

A RAMQAR monitora os seguintes poluentes: partículas totais em suspensão (PTS); MP_{10}; O_3; NO_x; CO e hidrocarbonetos (HC). E realiza, ainda, o monitoramento dos seguintes parâmetros meteorológicos: direção dos ventos (DV); velocidade dos ventos (VV); precipitação pluviométrica (PP); umidade relativa do ar (UR); precipitação pluviométrica (PP); temperatura (T); pressão atmosférica (P); e radiação solar (I). Os poluentes e parâmetros meteorológicos monitorados em cada estação RAMQAR encontram-se no Quadro 1. Na análise deste trabalho, as variáveis MP_{10}, VV, UR, PP e T foram utilizadas conforme descrição do Quadro 2.

Modelo SARIMAX

Uma série temporal é um conjunto de observações de qualquer fenômeno aleatório, ordenadas no tempo. A análise de séries temporais consiste em procurar alguma relação de dependência existente temporalmente nos dados, identificando o mecanismo gerador da série com objetivo de extrair periodicidades relevantes nas observações, descrever o seu comportamento e fazer previsões (MORETTIN & TOLOI, 2006; BAYER & SOUZA, 2010).



Fonte: Google Earth.

Figura 1 – Localização espacial das estações de monitoramento da qualidade do ar da RGV.

Seja Y_t (t = 1, 2, 3, ...) um processo linear com representação dada pela Equação 1:

$$(1 - B)^d(1 - B^s)^D\Phi(B^s)\Phi(B)Y_t = \Theta(B^s)\theta(B)\varepsilon_t, \quad (1)$$

Em que

s é chamado período sazonal do processo e,

ε_t, é ruído branco (RB), definido como uma sequência de variáveis aleatórias não correlacionadas com média zero e variância constante ao longo do tempo (WEI, 2006), ou seja, $\varepsilon_t \sim RB(0, \sigma_\varepsilon^2)$.

Quadro 1 - Parâmetros meteorológicos e poluentes monitorados em cada estação da Rede Automática de Monitoramento da Qualidade do Ar.

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

Fonte: adaptado de IEMA, 2013.

PTS: partículas totais em suspensão; MP_{10}: material particulado inalável; SO_2: dióxido de enxofre; CO: monóxido de carbono; NO_x: dióxido de nitrogênio; HC: hidrocarbonetos; O_3: ozônio; DV: direção dos ventos; VV: velocidade dos ventos; UR: umidade relativa do ar; PP: precipitação pluviométrica; P: pressão atmosférica; T: temperatura; I: radiação solar.

Quadro 2 - Descrição das variáveis material particulado inalável, velocidade do vento, umidade, precipitação e temperatura.

Variáveis	Unidades	Descrição
MP_{10}	$\mu g.m^{-3}$	Existem medições para todas as oito estações da RAMQAR. Entretanto, como existem muitos dados faltantes para as demais estações, optou-se por trabalhar apenas com as médias diárias da estação da Enseada do Suá. Além disso, essa estação ultrapassou algumas vezes o padrão do CONAMA para este poluente.
Velocidade do vento	$m.s^{-1}$	Valores medidos na estação de Carapina. Foi escolhida essa estação por apresentar a menor porcentagem de dados faltantes no período de estudo.
Umidade relativa	%	Como existem muitos dados faltantes para a estação de Cariacica, adotou-se o valor de Carapina.
Precipitação	mm	Valores medidos na estação de Carapina, única que possui medição para tal variável.
Temperatura	ºC	Média aritmética entre as estações de Carapina e Cariacica, únicas que possuem medições para tal variável.

MP_{10}: material particulado inalável; RAMQAR: Rede Automática de Monitoramento da Qualidade do Ar; CONAMA: Conselho Nacional do Meio Ambiente.

Em $(1 - B)^d(1 - B^s)^D$, d e D são números inteiros não negativos e representam o número de diferenças simples e sazonais, respectivamente, aplicadas sobre o processo Y_t.

Tem-se que, B é o operador de defasagem definido como $B^kY_t = Y_{t-k}$, $k \in \mathbb{N}$, $\phi(z) = 1 - \Sigma_{k=1}^{p} \phi_k z^k$, $\Phi(z^s) = 1 - \Sigma_{k=1}^{p} \Phi_k z^{sk}$, $\theta(z) = 1 - \Sigma_{k=1}^{q} \theta_k z^k$ e $\Theta(z^s) = 1 - \Sigma_{k=1}^{Q} \Theta_k z^{sk}$ são polinômios de ordem P, p, Q, $q \in \mathbb{N}$ respectivamente, com $z \in \mathbb{C}$, em que $\mathbb{C}$ representa o conjunto dos números complexos e $\{\phi_k\}$, $\{\theta_k\}$, $\{\Phi_k\}$, $\{\Theta_k\}$ são sequências de números reais. O processo Y_t, com representação dada na Equação 1 é denominado modelo autorregressivo integrado e de médias móveis sazonal multiplicativo (SARIMA), de ordem (p, d, q) × (P, D, Q)$_s$. O processo Y_t é estacionário e invertível se d = D = 0 e as raízes de $\phi(z)$, $\Phi(z^s)$, $\theta(z)$ e $\Theta(z^s)$ são não comuns e encontram-se fora do círculo unitário (WEI, 2006). As ordens p, d, q, P, D, Q e *s* devem ser identificadas seguindo a metodologia de Box e Jenkins (1970).

Nesse estudo, foi considerado o modelo SARIMAX (modelo autorregressivo integrado e de médias móveis sazonal multiplicativo com variáveis explicativas), que é uma extensão do modelo SARIMA, utilizando outras séries temporais como variáveis explicativas. O modelo SARIMAX explica a variável dependente por meio de: variáveis explicativas; defasagens das variáveis explicativas; e defasagens da variável dependente (MOURA; MONTINI; CASTRO, 2011).

Metodologia de modelagem

A comprovação da sazonalidade foi verificada pela análise espectral e pelo teste G de Fisher. As hipóteses testadas foram as seguintes:

1. H_0: não existe sazonalidade;
2. H_1: existe sazonalidade.

A estatística do teste é dada pela Equação 2:

$$G = \frac{\max[I_p(f_i)]}{\Sigma_{i=1}^{\left(\frac{N}{2}\right)} I_p(f_i)} \quad (2)$$

Em que:

I_p = periodograma no período p; e

N = número de observações da série.

Segundo Barbosa *et al.* (2015), o periodograma consiste na decomposição da série temporal em uma série de Fourier. No eixo das ordenadas ficam localizadas as frequências da série (f_i) e no eixo das abscissas as respectivas intensidades de cada frequência $I_p(f_i)$, definidas pela Equação 3:

$$I_p(f_i) = \frac{2}{\left(\frac{N}{2}\right)} \left\{ \sum_{t=1}^{n} \in_t \cos\left[\frac{2\pi i}{\left(\frac{N}{2}\right)} t\right]^2 + \sum_{t=1}^{n} \in_t \sin\left[\frac{2\pi i}{\left(\frac{N}{2}\right)} t\right]^2 \right\} \quad (3)$$

Em que

$\in_t$ representa o componente estocástico da série temporal associado ao tempo t e N é o número de observações da série. A distribuição exata de *G* é dada por $Z_\alpha = 1 - (\frac{\alpha}{c})^{\frac{1}{c-1}}$ sendo $c = (\frac{N}{2})$ e α o nível de significância adotado. Se G>Z, a hipótese H_0 é rejeitada e a série apresenta periodicidade no período i. Para determinar o período sazonal (s) verifica-se a qual frequência está associado o maior valor $I_p(f_i)$ e, então, divide 1 por esse valor de frequência, isto é, $s = (\frac{1}{f_i})$. Para detalhes, consultar Morettin e Toloi (2006).

A metodologia de Box e Jenkins aplicada neste trabalho está dividida nas seguintes etapas:

1. identificação;
2. estimação;
3. diagnóstico;
4. previsão.

A identificação do modelo a ser ajustado aos dados é uma das fases mais complicadas. Para isso, foram utilizados os critérios de seleção de modelos.

O *Akaike Information Criterion* (AIC) (AKAIKE, 1973) é um critério de seleção comumente utilizado. Utilizando os estimadores de máxima verossimilhança para os parâmetros do modelo, em que $(\hat{\xi})$ é a função de log-verossimilhança maximizada e k é o número de parâmetros do modelo, o AIC é dado por: $AIC = -2l(\hat{\xi}) + 2k$. Em uma perspectiva bayesiana, Schwarz (1978) e Akaike (1978) introduziram o *Bayesian Information Criterion* (BIC), dado por: $BIC = -2l(\hat{\xi}) + k\log(n)$, em que n é o número de observações da amostra. Com base nesses critérios de informação, pode-se ajustar diversos modelos e escolher aquele que obtiver o menor valor para o critério de informação.

Além dos critérios de seleção de modelos, outras duas medidas de qualidade de ajuste foram utilizadas para auxiliar na seleção do modelo que melhor se ajusta à série temporal em estudo e, principalmente, na avaliação da qualidade do ajuste. As medidas utilizadas neste trabalho foram a raiz do erro quadrático médio (REQM) e erro absoluto médio (EAM), definidas pelas Equações 4 e 5:

$$REQM = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2} \quad (4)$$

$$EAM = \frac{1}{N}\sum_{i=1}^{N}|Y_i - \hat{Y}_i| \quad (5)$$

Em que

Y_i e $\hat{Y}_i$ são, respectivamente, os valores observados e previstos no instante *i*. Um dos principais objetivos deste trabalho foi 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.

RESULTADOS E DISCUSSÕES

Análise dos dados e ajuste dos modelos

Para um entendimento preliminar das variáveis em estudo são apresentadas algumas medidas descritivas (Tabela 1). A análise corresponde às variáveis apresentadas no Quadro 2, para o período de 01/01/2012 a 30/04/2015, perfazendo um total de 1.216 observações. Todas as análises foram feitas utilizando o *software* livre R (R CORE TEAM, 2016). A presença de dados faltantes nas séries motivam o uso da metodologia de imputação via algoritmo *expectation-maximisation* (EM), proposta por Junger e Leon (2015) e implementada na biblioteca R *multivariante time-series data imputation* (MTSDI).

A concentração média de MP_{10} foi de, aproximadamente, 43,0 $\mu g.m^{-3}$ com desvio padrão de 15,3 $\mu g.m^{-3}$. Nota-se que, em média, as concentrações não ultrapassaram o valor de 50,0 $\mu g.m^{-3}$, porém, o desvio padrão e o coeficiente de variação alto indicam uma média pouco representativa. Além disso, os resultados mostram que o valor máximo foi mais do que o triplo do valor médio, demonstrando a grande variabilidade das concentrações de MP_{10} na RGV. O valor mínimo foi observado em março de 2013, na estação chuvosa da região, que se inicia a partir do mês de outubro e vai até meados de abril (IEMA, 2013). Nessa época do ano, a atuação dos sistemas frontais e de zonas de convergência de umidade favorecem o aumento de precipitação. Nesse período, as concentrações de MP_{10} são menores, principalmente pela intensificação da eficiência dos processos de remoção por deposição úmida. Em relação às variáveis

Tabela 1 – Medidas descritivas das variáveis sob estudo.

Medidas descritivas	MP_{10}	Velocidade	Umidade	Precipitação	Temperatura
Média	43,8853	2,4918	76,9636	0,1217	23,1228
Mediana	41,7917	2,3683	76,7184	0,0000	23,7253
Desvio padrão	15,3020	0,7980	6,6742	0,5198	3,9298
Coeficiente de variação	34,8681	32,0233	8,6719	427,0784	16,9954
Valor máximo	159,1250	5,5412	98,2983	12,7750	30,7085
Valor mínimo	8,9167	1,0392	29,7162	0,0000	5,5700

MP_{10}: material particulado inalável.

meteorológicas, nota-se que as mesmas mostraram grande variabilidade em termos estatísticos.

A Tabela 2 mostra as correlações calculadas entre as concentrações de médias diárias de MP_{10}, de VV, da UR, de PP e T. Observa-se que as variáveis meteorológicas apresentam relação linear com as concentrações de MP_{10}, na estação de Enseada do Suá. Conforme esperado, os índices de MP_{10} estão associados às mudanças dessas variáveis, ou seja, infere-se dos resultados (Tabela 2) que o aumento da VV e da T na RGV acarreta aumento nas concentrações de MP_{10} e que o aumento da PP e UR acarreta diminuição das concentrações de MP_{10}.

A Figura 2 apresenta a série de MP_{10} e sua decomposição em componentes de tendência, sazonalidade e aleatoriedade. Nota-se a presença da propriedade de sazonalidade e a inexistência de tendência nos dados, uma vez que o gráfico da componente tendência não apresenta um comportamento de crescimento ou decrescimento ao longo do tempo. Segundo Wei (2006), o primeiro passo na análise de séries temporais é verificar se as mesmas apresentam média, variância e covariância constantes ao longo do tempo, ou seja, se são estacionárias. Caso elas não forem estacionárias, aplica-se a primeira diferença na mesma para tentar estacionarizá-las. O resultado para o teste de Dickey-Fuller Aumentado (ADF) (DICKEY & FULLER, 1981) revela que as séries em estudo são estacionárias na média, o que confirma que a série não apresenta tendência.

Como mencionado anteriormente, há indícios de sazonalidade nos dados. Portanto, para confirmar tal hipótese e verificar a necessidade de aplicação de diferenças sazonais, foi realizada a decomposição espectral da série MP_{10}, conforme mostra a Figura 3.

O periodograma demonstra que o maior pico está associado à frequência 0,14333; o que implica em $s = \frac{1}{0,14333} = 6,97$, ou seja, uma componente sazonal com periodicidade de sete dias. De acordo com

Tabela 2 – Matriz de correlação entre as variáveis sob estudo.

	MP_{10}	Velocidade	Umidade	Precipitação	Temperatura
MP_{10}	1,0000				
Velocidade do vento	0,1149	1,0000			
Umidade	-0,1377	-0,4653	1,0000		
Precipitação	-0,1332	-0,0979	0,2680	1,0000	
Temperatura	0,1794	0,2871	-0,2125	-0,0653	1,0000

MP_{10}: material particulado inalável.

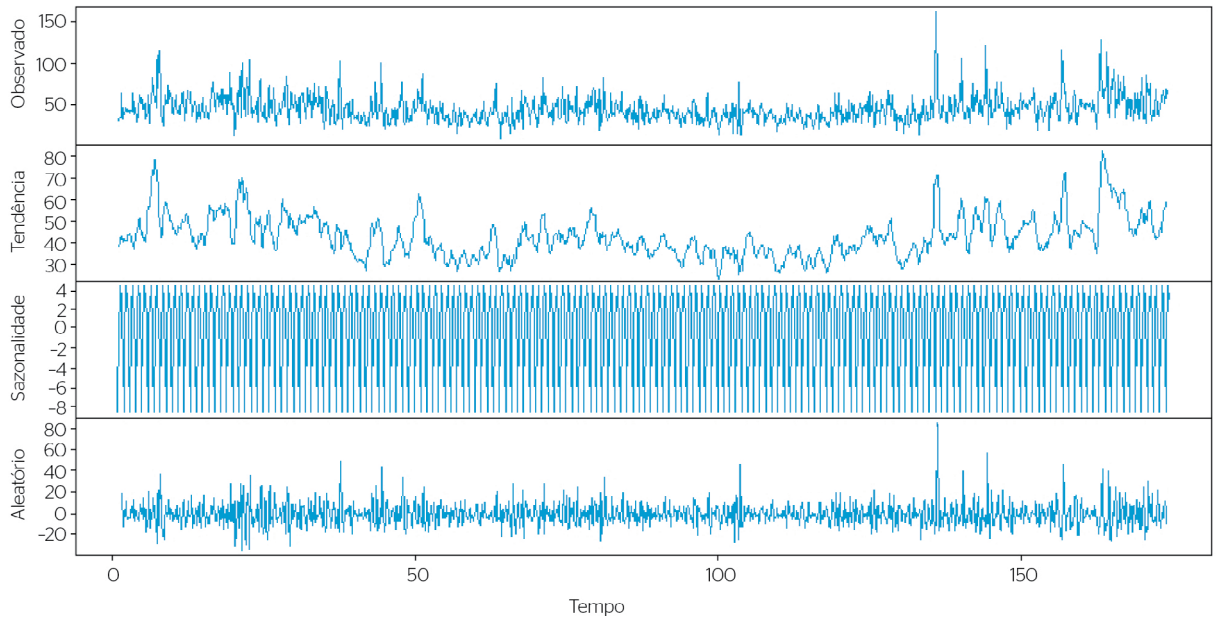


Figura 2 – Decomposição da série temporal em componentes de sazonalidade, de tendência e de aleatoriedade.

o teste de Fisher, foram obtidos os valores das estatísticas G=0,1167 e Z=0,0156. Como o valor de G é superior ao de Z, rejeita-se a hipótese H_0, confirmando a existência da componente sazonalidade para períodos de 7 dias ao nível nominal de significância de 5%. A sazonalidade em séries de MP_{10} é esperada, visto que segundo o IEMA (2013), a principal fonte emissora de partículas na RGV são veículos automotores, representando mais de 60% das emissões de partículas que estão ligados à ressuspensão de partículas em vias. Dessa forma, existe variação entre as concentrações medidas nos dias úteis e finais de semana, uma vez que o fluxo de veículos é maior durante os dias da semana. Logo, uma diferença sazonal de ordem s=7 foi aplicada para eliminar a sazonalidade presente na série. A Figura 4 mostra a função de autocorrelação (FAC) (a) da série e a FAC (b) da série diferenciada de ordem 7.

A identificação preliminar das ordens, p e q e P e Q, do modelo SARIMAX a ser estimado ocorreu pela análise do comportamento da FAC e da função de autocorrelação parcial (FACP), para detalhes ver Wei (2006). Dessa forma, com os valores de p e q variando de 1 a 3, e P e Q de 0 a 3, intervalos escolhidos pela indicações do correlograma, testaram-se alguns modelos com essas combinações. Como a

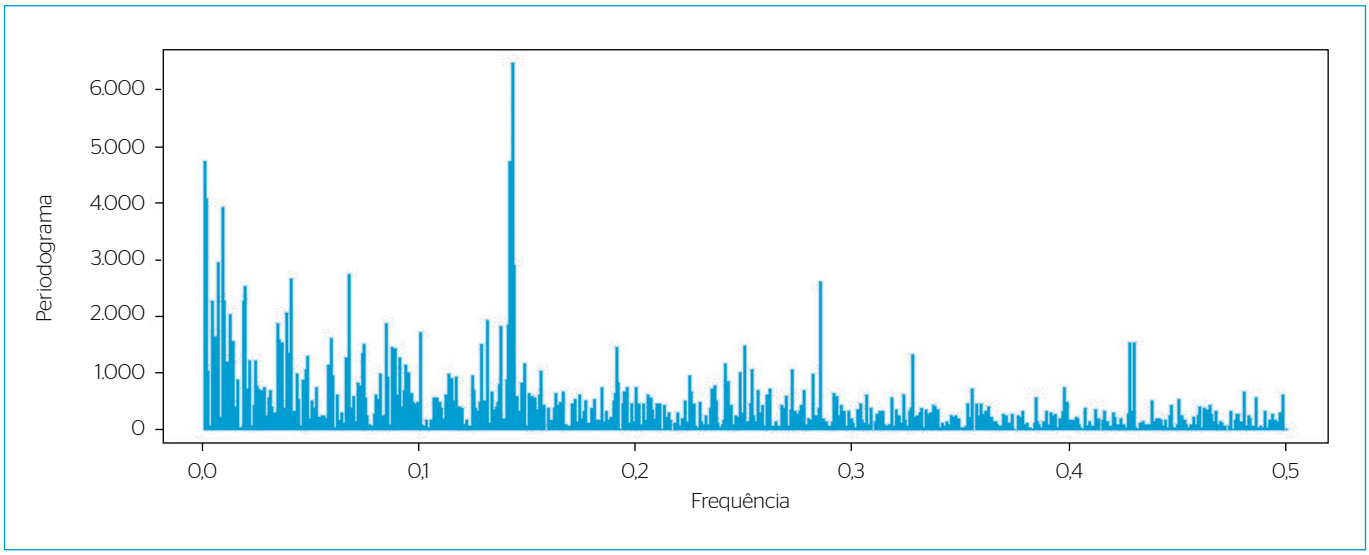
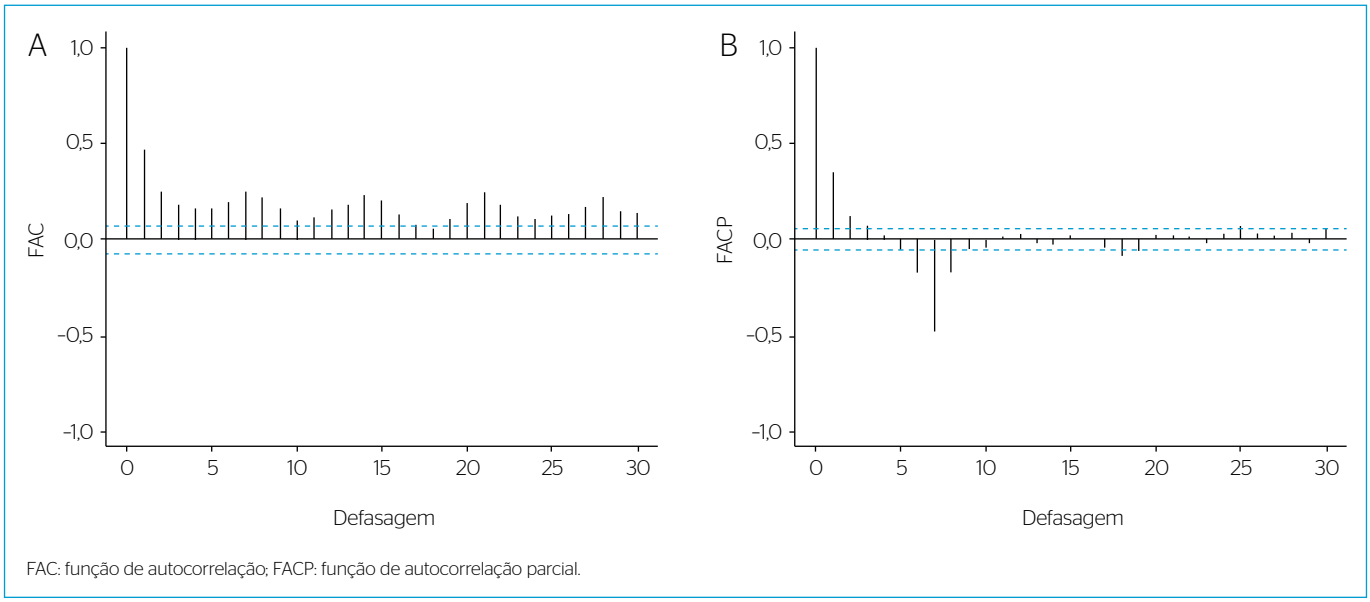


Figura 3 – Periodograma da série de material particulado inalável.



FAC: função de autocorrelação; FACP: função de autocorrelação parcial.

Figura 4 – Função de autocorrelação da série (A); e função de autocorrelação da série (B) após aplicação da diferença sazonal.

série é estacionária, não houve a necessidade de aplicar diferenças de ordem 1, então, tem-se que d=0. Da mesma forma, como uma diferença sazonal foi aplicada, tem-se D=1. Feito isso, a identificação das ordens autorregressivas e de médias móveis, do melhor modelo para representar os dados, baseou-se nos critérios de informação AIC, BIC e nos valores da REQM e EAM. Dentre todos o modelos estimados, o modelo com melhor ajuste foi o SARIMAX$(1,0,2)(0,1,1)_7$, por ter apresentado os menores valores calculados para os critérios de informação e medidas de qualidade. Sendo assim, esse torna-se o modelo de interesse.

Para verificar a adequação do modelo, a teoria estatística estabelece suposições das propriedades do mesmo, tais como não autocorrelação e normalidade dos resíduos (WEI, 2006). Na Tabela 3 são apresentadas as estatísticas para os testes Shapiro e Wilk (1965), Bera e Jarque (1981), Ljung e Box (1978) e Box e Pierce (1970). Observa-se que os resíduos não são normalmente distribuídos, resultado que já era esperado por se tratar de uma variável ambiental. Porém, assumiu-se, pela teoria assintótica sobre a média das distribuições de probabilidade, a suposição de que os resíduos são normalmente distribuídos. O valor p dos testes de correlação são superiores ao nível de significância de 5%, indicando que não se rejeita a hipótese nula de erros não autocorrelacionados. Nenhuma autocorrelação apresentou-se significativamente diferente de zero no correlograma residual apresentado na Figura 5, o que vai ao encontro dos resultados apresentados na Tabela 3 e com a hipótese de homogeneidade dos resíduos.

A Tabela 4 contém as estimativas dos parâmetros do modelo ajustado, considerando como variáveis explicativas a VV e a PP. Observa-se que todos os coeficientes são estatisticamente significativos de acordo com o teste z. Uma vez que as variáveis umidade e temperatura não foram estatisticamente significativas, as mesmas não foram incorporadas no ajuste do modelo final. Chaloulakou, Grivas, Spyrellis (2003) estimaram a média diária de MP_{10} em Atenas por meio da análise de regressão. Os autores observaram que a variável VV influencia na concentração de MP_{10}. Esses resultados vão ao encontro dos apresentados no presente trabalho. Lyra, Oda-Souza e Viola (2011) observaram que as variáveis UR e a PP foram os elementos meteorológicos significativos para explicar a variabilidade do MP_{10}, na cidade do Rio de Janeiro. Os resultados para umidade encontrados pelos autores contrariam as estimativas deste trabalho, uma vez que a umidade não foi

Tabela 3 – Testes estatísticos de normalidade* e correlação** dos resíduos do modelo escolhido.

Parâmetro	Estimativa
Jarque-Bera*	<0,0001
Shapiro-Wilk*	<0,0001
Ljung-Box**	0,9095
Box-Pierce**	0,9114

Tabela 4 – Estimativa dos parâmetros do modelo ajustado SARIMAX$(1,0,2)(0,1,1)_7$

Parâmetro	Estimativa	Erro padrão	Valor de z	Valor p
Veloc	2,4419	0,9544	2,5586	0,0105
precp	-2,4995	0,5955	-4,1971	<0,0001
ϕ_1	0,7168	0,0763	9,3974	<0,0001
θ_1	-0,3444	0,0823	-4,1831	<0,0001
θ_2	-0,1261	0,0418	-3,0155	0,0026
Θ_1	-0,9058	0,0124	-72,8414	<0,0001

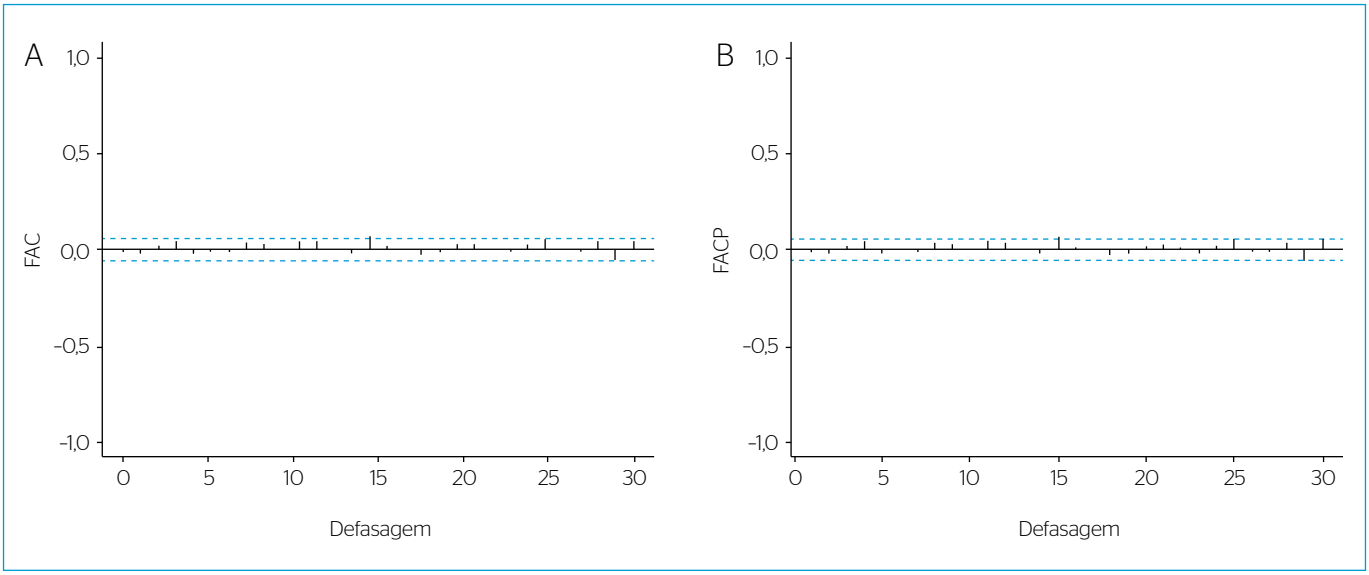


Figura 5 – Funções amostrais de autocorrelação (A) e autocorrelação parcial (B) dos resíduos do modelo SARIMAX$(1,0,2)(0,1,1)_7$.

um elemento significativo para explicar as concentrações de MP_{10} na RGV. Entretanto, no trabalho de Lyra, Oda-Souza e Viola (2011) a precipitação foi estatisticamente significativa, similar ao observado para o modelo SARIMAX ajustado.

O sinal positivo do coeficiente estimado da variável VV indica a existência de uma relação direta entre a concentração de MP_{10} e essa variável, demostrando que o aumento da VV eleva a concentração de MP_{10}. Segundo o Relatório da Qualidade do Ar da Grande Vitória (IEMA, 2013), 69,3% das emissões de MP_{10} para a atmosfera da RGV estão ligadas à ressuspensão causada pelo vento e tráfego veicular, corroborando o resultado encontrado, uma vez que o aumento da VV tende a elevar a ressuspensão do solo. Já a PP apresentou relação negativa com a concentração de MP_{10}. O coeficiente negativo no modelo se deve ao processo de deposição úmida e atenuação da ressuspensão do solo (LYRA; ODA-SOUZA; VIOLA, 2011). Durante eventos de precipitação ocorre diminuição do MP_{10}, pois o solo úmido dificulta a ressuspensão do particulado no solo (VARDOULAKIS & KASSOMENOS, 2008).

Estudo de previsões

Para corroborar a utilização do modelo SARIMAX e para fins de comparação entre modelos, realizou-se estimativas para os modelos ARMA, SARMA, ARMAX e SARMAX e verificou-se qual é o melhor método para fazer previsões das concentrações de MP_{10}. Para isso, o conjunto de dados sob estudo foi dividido em dois subconjuntos; adotou-se o período de 01/01/2012 a 14/04/2015 para fazer as estimações e reservou-se os dados compreendidos entre 15/04/2015 a 30/04/2015 para se fazer as previsões, ou seja, para que o cálculo das estatísticas de avaliação das previsões dê um passo à frente.

Na Tabela 5 são apresentados os resultados estimados para REQM e EAM para os modelos ajustados. Observa-se que o modelo SARIMAX apresentou um aumento na precisão das previsões nos horizontes (h) calculados. Os resultados da comparação das medidas de qualidade mostram que para h=0 e h=1 o modelo SARIMAX foi o que obteve melhor resultado, sendo o mais adequado para realizar as previsões.

Tabela 5 – Medidas de avaliação da qualidade de previsão obtidas a partir dos modelos ajustados.

Horizonte	Medida	Modelos				
		ARMA	SARMA	ARMAX	SARMAX	SARIMAX
h=0	REQM	13,2885	13,2540	13,0881	13,0504	12,8507
	EAM	9,6660	9,6456	9,4331	9,4220	9,3400
h=1	REQM	13,3421	13,3047	13,1487	13,1089	12,9212
	EAM	9,6610	9,6386	9,4356	9,4223	9,3564

REQM: raiz do erro quadrático médio; EAM: erro absoluto médio.

Avaliação do modelo SARIMAX para estimativa da qualidade do ar

Conforme apresentado na introdução, o IEMA classifica a qualidade do ar da RGV como boa, para concentrações de MP_{10} entre 0 e 45 $\mu g.m^{-3}$, regular entre 46 e 120 $\mu g.m^{-3}$, e inadequada entre 121 e 250 $\mu g.m^{-3}$ (IEMA, 2013). A qualidade do ar pode, ainda, ser classificada como má, péssima ou crítica. Como no período de estudo a qualidade do ar, monitorada na estação da Enseada do Suá, foi classificada como regular em 450 ocasiões, avaliou-se apenas a eficiência do modelo em predizer eventos classificados como regular. Para isso, baseando-se na metodologia de Ryan (1995) e Liu e Johnson (2003), calculou-se as seguintes estatísticas de avaliação: a probabilidade de detecção (POD), que mensura a probabilidade do modelo predizer corretamente as concentrações maiores ou iguais a 46 $\mu g.m^{-3}$ (qualidade do ar regular), quando realmente esse valor ocorreu; a razão de alarme falso (FAR), que mede a tendência da predição superestimar o valor observado da concentração de MP_{10}; o escore de ameaças (EA), que representa a relação entre as predições corretas de eventos pelo total de eventos regular observado; e a taxa de perda (MISS), que mensura a taxa a qual os eventos de concentrações maiores ou iguais a 46 $\mu g.m^{-3}$ observados não são preditos.

Para fins de comparação entre o pior modelo (ARMA) e o melhor modelo (SARIMAX) para fazer a predição de eventos regulares de qualidade do ar, a Tabela 6 contém a descrição e os resultados das estatísticas de avaliação. Observa-se que, das vezes que foram observados eventos de qualidade regular, o modelo SARIMAX estimou corretamente 64,39% desses eventos. Além disso, apresentou uma FAR de 0,3792, o que significa que em torno de 37,00% das predições do modelo os eventos de qualidade do ar foram classificados como regular, quando o observado foi boa; enquanto no modelo ARMA esse resultado foi em torno de 38,00%. O maior valor de EA mostra que o modelo SARIMAX foi superior ao modelo ARMA na predição correta de eventos de qualidade do ar regular. Além disso, a taxa

Tabela 6 – Descrição e resultados das estatísticas de avaliação da eficiência do modelo em predizer eventos classificados como regular.

Estatísticas/equação[1]	Modelos	
	ARMA	SARIMAX
POD = A/(A + B)	0,6100	0,6439
FAR = C/(C + A)	0,3888	0,3792
EA = A/(A + B + C)	0,4395	0,4621
MISS = 1 – POD	0,3900	0,3561

[1]A: número de eventos observados e preditos; B: número de eventos observados mas não preditos; C: número de eventos estimado mas não observados; POD: probabilidade de detecção; FAR: razão de alarme falso; EA: escore de ameaças; MISS: taxa de perda.

(MISS) referente à ocorrência de eventos que não foram detectados foi menor para o modelo SARIMAX, corroborando, novamente, que o desempenho estatístico do modelo SARIMAX foi superior ao modelo ARMA, no que diz respeito à predição de eventos de qualidades do ar regular.

De maneira geral, os resultados obtidos com a modelagem dos dados de concentrações de MP_{10} com o modelo SARIMAX, tanto no ajuste quanto na previsão e predição de eventos de qualidade do ar, podem ser considerados bons, uma vez que tais resultados foram satisfatórios em termos estatísticos.

CONCLUSÕES

Este trabalho objetivou modelar e prever a concentração média diária de MP_{10} na RGV, Espírito Santo, Brasil, utilizando o modelo SARIMAX, no período de 01/01/2012 a 30/04/2015. Baseando-se em indicadores de desempenho de modelagem, verificou-se que o modelo SARIMAX $(1,0,2)\ (0,1,1)_7$ é o mais acurado, entre os estudados, para fazer predições e previsões da qualidade ar da RGV. Em comparação com os modelos ARMA, o desempenho estatístico do modelo SARIMAX foi superior, no que diz respeito à predição de eventos de qualidade do ar regular.

Entre as variáveis meteorológicas avaliadas, a VV e a PP foram significativas e melhoraram o ajuste do modelo. A UR e a T não foram estatisticamente significativas para explicar a variação das concentrações de MP_{10}. Em termos de previsão da qualidade do ar, os modelos de séries temporais mostraram resultados satisfatórios.

Como parte de um estudo posterior, sugere-se a utilização de um modelo de heterocedasticidade condicional autorregressivo generalizado (GARCH) juntamente aos modelos da classe SARIMAX.

REFERÊNCIAS

AGIRRE-BASURKO, E.; IBARRA-BERASTEGUI, 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 and Software*, v. 21, p. 430-446. http://dx.doi.org/10.1016/j.envsoft.2004.07

AKAIKE, H. (1973) Information theory and an extension of the maximum likelihood principle. *In*: PETROV, B.N.; CSAKI, F. (Eds.). *Proceedings of the Second International Symposium on Information Theory*. Budapeste: Akademiai Kiado. p. 267-281.

_____. (1978) A Bayesian analysis of the minimum AIC procedure. *Annals of the Institute of Statistical Mathematics*, v. 30, n. 1, p. 9-14.

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 (Mestrado em Engenharia Ambiental) – Programa de Pós-Graduação em Engenharia Ambiental, Centro Tecnológico, Universidade Federal do Espírito Santo, Vitória.

BARBOSA, E.C.; SÁFADI, T.; NASCIMENTO, M.; NASCIMENTO, A.C.C.; SILVA, C.H.O.; MANULI, R.C. (2015) Metodologia Box & Jenkins para previsão de temperatura média mensal da cidade de Bauru (SP). *Revista Brasileira de Biometria*, v. 33, n. 1, p. 104-117.

BAYER, F.M.; SOUZA, A.M. (2010) Wavelets e modelos tradicionais de previsão: um estudo comparativo. *Revista Brasileira de Biometria*, v. 28, n. 2, p. 40-61.

BERA, A.K.; JARQUE, C.M. (1981) Efficient tests for normality, homoscedasticity and serial independence of regression residuals: Monte Carlo evidence. *Economics Letters*, v. 7, n. 4, p. 313-318. https://doi.org/10.1016/0165-1765(81)90035-5

BOX, G.; JENKINS, G. (1970) *Time Series Analysis:* Forecasting and Control. São Francisco: Holden-Day.

BOX, G.; PIERCE, D.A. (1970) Distribution of residual autocorrelations in autorregressive integrated moving average time series models. *Journal of the American Statistical Association*, v. 65, n. 332, p. 1509-1526. DOI: 10.2307/2284333

BRAGA, A.L.F.; PEREIRA, L.A.A.; PROCÓPIO, M.; ANDRÉ, P.A.D.; SALDIVA, P.H.D.N. (2007) Associação entre poluição atmosférica e doenças respiratórias e cardiovasculares na cidade de Itabira, Minas Gerais, Brasil. *Cadernos de Saúde Pública*, v. 23, supl. 4, p. S570-S578. http://dx.doi.org/10.1590/S0102-311X2007001600017

CHALOULAKOU, A.; GRIVAS, G.; SPYRELLIS, N. (2003) Neural network and multiple regression models for PM_{10} prediction in Athens: A comparative assessment. *Journal of the Air & Waste Management Association*, v. 53, n. 10, p. 1183-1190. https://doi.org/10.1080/10473289.2003.10466276

CONAMA – Conselho Nacional de Meio Ambiente. (1990) Resolução nº 03, de 28 de junho de 1990. Dispõe sobre padrões de qualidade do ar, previstos no PRONAR. *Diário Oficial da República Federativa do Brasil*, Brasília, Seção 1, p. 15937-15939.

CURTIS, L.; REA, W.; SMITH-WILLIS, P.; FENYVES, E.; PAN, Y. (2006) Adverse health effects of outdoor air pollutants. *Environment International*, v. 32, n. 6, p. 815-830. https://doi.org/10.1016/j.envint.2006.03.012

DICKEY, D.A.; FULLER, W.A. (1981) Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica*, v. 49, n. 4, p. 1057-1072. DOI: 10.2307/1912517

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

GOOGLE EARTH. (2014) *Informações geográficas*. Disponível em: <http://www.google.com.br/intl/pt-PT/earth/>. Acesso em: 10 jun. 2016.

GOUVEIA, N.; MENDONÇA, G.A.S.; LEON, A.P.; CORREIA, J.E.M.; JUNGER, W.L.; FREITAS, C.U.; DAUMAS, R.P.; MARTINS, L.C.; GIUSSEPE, L.; CONCEIÇÃO, G.M.S.; MANERICH, A.; CUNHA-CRUZ, J. (2003) Poluição do ar e efeitos na saúde nas populações de duas grandes metrópoles brasileiras. *Epidemiologia e Serviços de Saúde*, v. 12, p. 29-40. http://dx.doi.org/10.5123/S1679-49742003000100004

GOYAL, P.; CHAN, A.T.; JAISWAL, N. (2006) Statistical models for the prediction of respirable suspended particulate matter in urban cities. *Atmospheric Environment*, v. 40, n. 11, p. 2068-2077. https://doi.org/10.1016/j.atmosenv.2005.11.041

GRIPA, W.R.; REISEN, V.A.; FAJARDO, F.A.; REIS JUNIOR, N.C. (2012) Análise de predição e previsão das concentrações de material particulado inalável (PM_{10}) na cidade de Carapina, ES. *Revista Brasileira de Estatística*, v. 73, n. 237, p. 37-57.

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

IBGE – Instituto Brasileiro de Geografia e Estatística. (2010) *Censo 2010*. Disponível em: <http://www.ibge.gov.br/home/estatistica/populacao/censo2010/resultados_dou/ES2010.pdf>. Acesso em: 01 dez. 2015.

IEMA – Instituto Estadual de Meio Ambiente e Recursos Hídricos do Estado do Espírito Santo. (2013) *Relatório da qualidade do ar da Região da Grande Vitória*. Vitória: IEMA. Disponível em: <https://iema.es.gov.br/Media/iema/Downloads/RAMQAR/Relat%C3%B3rio_Anual_de_Qualidade_do_Ar_2013.pdf>. Acesso em: 10 jun. 2016.

JUNGER, W.L.; LEON, A.P. (2015) Imputation of missing data in time series for air pollutants. *Atmospheric Environment*, v. 102, p. 96-104. https://doi.org/10.1016/j.atmosenv.2014.11.049

LATORRE, M.R.D.O.; CARDOSO, M.R.A. (2001) Análise de séries temporais em epidemiologia: uma introdução sobre os aspectos metodológicos. *Revista Brasileira de Epidemiologia*, v. 4, n. 3, p. 145-152. http://dx.doi.org/10.1590/S1415-790X2001000300002

LIU, P.W.G.; JOHNSON, R. (2003) Forecasting peak daily ozone levels: part 2. A regression with time series errors model having a principal component trigger to forecast 1999 and 2002 ozone levels. *Journal of the Air & Waste Management Association*, v. 53, n. 12, p. 1472-1489. https://doi.org/10.1080/10473289.2003.10466321

LJUNG, G.M.; BOX, G.E.P. (1978) On a measure of lack of fit in time series models. *Biometrika*, v. 65, n. 2, p. 297-303.

LYRA, G.B.; ODA-SOUZA, M.; VIOLA, D.N. (2011) Modelos lineares aplicados à estimativa da concentração do material particulado (PM_{10}) na cidade do Rio de Janeiro, RJ. *Revista Brasileira de Meteorologia*, v. 26, n. 3, p. 392-400. DOI: 10.1590/s0102-77862011000300006

MARTINS, L.C.; LATORRE, M.R.D.O.; CARDOSO, M.R.A.; GONÇALVES, F.L.T.; SALDIVA, P.H.N.; BRAGA, A.L.F. (2002) Poluição atmosférica e atendimentos por pneumonia e gripe em São Paulo, Brasil. *Revista de Saúde Pública*, v. 36, n. 1, p. 88-94. http://dx.doi.org/10.1590/S0034-89102002000100014

MONTE, E.Z.; ALBUQUERQUE, T.T.A.; REISEN, V.A. (2015) Previsão da concentração de ozônio na Região da Grande Vitória, Espírito Santo, Brasil, utilizando o modelo ARMAX-GARCH. *Revista Brasileira de Meteorologia*, v. 30, n. 3, p. 285-294. http://dx.doi.org/10.1590/0102-/10.1590/0102-778620140060

MORETTIN, P.A.; TOLOI, C.M.C. (2006) *Análise de séries temporais*. 2. ed. São Paulo: Blucher.

MOURA, F.A.; MONTINI, A.A.; CASTRO, J.B.B. (2011) Modelagem do Consumo de Energia Elétrica Residencial no Brasil Através de Modelos ARMAX. *In*: SEMINÁRIO DE ADMINISTRAÇÃO, 14., 2011. *Anais*...

NASCIMENTO, L.F.C.; PEREIRA, L.A.A.; BRAGA, A.L.F.; MÓDOLO, M.C.C.; CARVALHO JR., J.A.C. (2006) Efeitos da poluição atmosférica na saúde infantil em São José dos Campos, SP. *Revista de Saúde Pública*, v. 40, p. 77-82. http://dx.doi.org/10.1590/S0034-89102006000100013

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

R CORE TEAM (2016). *R: A language and environment for statistical computing*. Viena: R Foundation for Statistical Computing. Disponível em: <http://www.R-project.org/>.

REISEN, V.A.; SARNAGLIA, A.J.Q.; REIS JUNIOR, N.C.; LÉVY-LEDUC, C.; SANTOS, J.M. (2014) Modeling and forecasting daily average PM_{10} concentrations by a seasonal long-memory model with volatility. *Environmental Modelling & Software*, v. 51, p. 286-295. https://doi.org/10.1016/j.envsoft.2013.09.027

RYAN, W.F. (1995) Forecasting severe ozone episodes in the Baltimore metropolitan area. *Atmospheric Environment*, v. 29, n. 17, p. 2387-2398. https://doi.org/10.1016/1352-2310(94)00302-2

SCHWARZ, G. (1978) Estimating the dimensional of a model. *The Annals of Statistics*, Hayward, v. 6, n. 2, p. 461-464.

SEINFELD, J.H.; PANDIS, S.N. (2006) *Atmospheric Chemistry and Physics*. 2. ed. Nova Jersey: John Wiley & Sons.

SHAPIRO, S.S.; WILK, M.B. (1965) An analysis of variance test for normality (complete samples). *Biometrika*, v. 52, n. 3-4, p. 591-611. DOI: 10.2307/2333709

SOUZA, J.B.; REISEN, V.A.; SANTOS, J.M.; FRANCO, G.C. (2014). Componentes principais e modelagem linear generalizada na associação entre atendimento hospitalar e poluição do ar. *Revista de Saúde Pública*, v. 48, n. 3, p. 451-458. DOI: 10.1590/S0034-8910.2014048005078

VARDOULAKIS, S.; KASSOMENOS, P. (2008) Sources and factors affecting PM_{10} levels in two European cities: implications for local air quality management. *Atmospheric Environment*, v. 42, n. 17, p. 3949-3963. https://doi.org/10.1016/j.atmosenv.2006.12.021

WEI, W. (2006) *Time Series Analysis:* Univariate and Multivariate Methods. Nova York: Addison Wesley.

WORLD HEALTH ORGANIZATION (WHO). (2005) *Air quality guidelines global update 2005*. Report on a working group meeting, Bonn/Alemanha: WHO. Disponível em: <http://www.euro.who.int/_data/assets/pdf_file/0008/147851/E87950.pdf>. Acesso em: 20 jan. 2016.

sociedade & natureza

DOI: 10.14393/SN-v30n3-2018-9

Picos de concentração de poluição atmosférica na Região da Grande Vitória, ES, Brasil: uma aplicação da regressão logística

Atmospheric pollution concentration peaks in the Região da Grande Vitória, ES, Brazil: an application of logistic regression

Wanderson de Paula Pinto[1]
Valdério Anselmo Reisen[2]
Edson Zambon Monte[3]

Resumo
Este trabalho objetivou avaliar os impactos das variáveis meteorológicas temperatura, umidade relativa, velocidade do vento e precipitação na probabilidade de ocorrência de picos/episódios de concentração de poluentes na Região da Grande Vitória (RGV), Espírito Santo, Brasil, por meio do modelo Logit, para o período de 01/01/2012 a 31/12/2014. Os dados deste estudo foram do tipo séries temporais relativos às concentrações de PTS, MP_{10}, SO_2, CO, NO_2, O_3 e às variáveis meteorológicas (velocidade do vento, umidade relativa, precipitação pluvial e temperatura) obtidas junto ao Instituto Estadual de Meio Ambiente e Recursos Hídricos (IEMA). Conforme esperado, os resultados mostraram que os índices de concentrações na RGV estão associados às mudanças das variáveis meteorológicas. Entre as variáveis meteorológicas avaliadas, a velocidade do vento e a precipitação foram mais significativas na redução da probabilidade de ocorrência de classificação do ar como "não boa". O modelo Logit mostrou ser uma ferramenta estatística satisfatória para avaliar a qualidade do ar "não boa" da região de estudo.
Palavras-chave: Poluição atmosférica; modelo Logit; variáveis meteorológicas.

Abstract
This study aimed to evaluate the impacts of meteorological variables temperature, relative humidity, wind speed and precipitation in the probability of occurrence of pollutants concentration peaks/episodes in the Região da Grande Vitória (RGV), Espírito Santo, Brazil, using the Logit model, for period from 2012/01/01 to 2014/12/31. It were adopted time series of the pollutants PTS, PM_{10}, SO_2, CO, NO_2, O_3 and meteorological variables (wind speed, relative humidity, rainfall and temperature)

[1] Doutorando em Engenharia Ambiental pela Universidade Federal do Espírito Santo – UFES, Vitória, ES, Brasil. wandersondpp@gmail.com
[2] Professor do Departamento de Economia e do Programa de Pós-Graduação em Economia e membro do Grupo de Pesquisa em Econometria da UFES, Vitória, ES, Brasil. edsonzambon@yahoo.com.br
[3] Programa de Pós-Graduação em Engenharia Ambiental e Programa de Pós-Graduação em Economia, Departamento de Estatística, Universidade Federal do Espírito Santo, Vitória, ES, Brasil. valderioanselmoreisen@gmail.com
Artigo recebido em: 01/05/2017. Aceito para publicação em: 30/10/2018.

obtained from the State Institute of Environment and Resources Water (IEMA). As expected, the results showed that the contents of the RGV concentrations are associated with changes of the meteorological variables. The results showed that the precipitation and the wind speed contributed significantly to the reduction of the probability of air quality "not good". The Logit model proved to be a satisfactory statistical tool for evaluating the air quality "not good" in the study region.
Keywords: Atmospheric pollution; Logit model; meteorological variables.

Introdução

A preocupação com os efeitos da poluição do ar veio com o crescimento industrial iniciado no período da Revolução Industrial que teve início na Inglaterra, em meados do século XVIII, devido à alguns episódios de alta concentração de poluentes. Episódios esses que causaram aumento do número de mortes em algumas cidades da Europa e dos Estados Unidos.

O primeiro episódio ocorreu em 1930, no vale do Meuse, na porção Belga, que provocou a morte de 60 pessoas. Anos mais tarde, em 1948, um episódio semelhante ocorreu durante os últimos cinco dias do mês de outubro, na cidade de Donora, Pensilvânia. Porém, um dos mais graves episódios acerca dos efeitos deletérios dos poluentes do ar de que se tem notícia ocorreu em Londres durante o inverno de 1952. Em dezembro 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 quatro dias, deixando cerca de quatro mil pessoas mortas durante as duas semanas seguintes devido à poluição do ar. Segundo Logan (1953) o aumento da mortalidade afetou pessoas de todas as idades, mas particularmente aquelas com 45 anos ou mais. As mortes atribuídas a bronquite e a pneumonia aumentaram oito vezes e três vezes, respectivamente, em uma semana. Um aumento considerável no número de mortes ocorreu mesmo no primeiro dia do nevoeiro. Quatro nevoeiros anteriores em Londres, que resultaram em um aumento súbito de mortes,

foram observados; mas o incidente de 1952 causou de longe o maior aumento. Esses episódios acarretaram em um aumento do número de óbitos em relação à média de óbitos em períodos semelhantes (BRAGA *et al.* 2005).

Nossos ancestrais já conviviam com a poluição do ar natural, oriunda das erupções vulcânicas e da decomposição da matéria orgânica. Mais tarde, com a expansão da unbanização e da industrialização as atividades antropogências foram intensificadas, o que contribuiu com a aceleração da deterioração da qualidade do ar. Com o crescimento populacional, o desenvolvimento econômico e o crescimento da frota motorizada, as fontes de poluição multiplicaram-se, agravando o problema, mesmo em áreas não industrializadas (LIRA, 2009).

Segundo Holgate *et al.* (1999), um nível elevado dos poluentes pode ocasionar desde irritação dos olhos, nariz e garganta, bronquite e pneumonia, até doenças respiratórias crônicas, câncer de pulmão, problemas cardíacos, entre outros. Segundo Cançado *et al*. (2006), isso ocorre,

> Pois a poluição do ar causa uma resposta inflamatória no aparelho respiratório induzida pela ação de substâncias oxidantes, as quais acarretam aumento da produção, da acidez, da viscosidade e da consistência do muco produzido pelas vias aéreas, levando, consequentemente, à diminuição da resposta e/ou eficácia do sistema mucociliar. O aumento da poluição do ar tem sido associado ao aumento da viscosidade sanguínea, de marcadores inflamatórios (proteína C reativa, fibrinogênio) e da progressão de arteriosclerose, a alterações da coagulação, à redução da variabilidade da frequência cardíaca (indicador de risco para arritmia e morte súbita), à vasoconstricção e ao aumento da pressão arterial, todos fatores de risco para doenças cardiovasculares.

Diversos estudos epidemiológicos têm demonstrado associações significativas entre a exposição a concentrações elevadas de poluentes atmosféricos e os problemas de saúde (OSTRO *et al*., 1996; MARTINS *et al.*, 2002; GOUVEIA *et al.*, 2003; ALMEIDA, 2006; NASCIMENTO *et al.*, 2006; CURTIS, *et al.*, 2006; BRAGA *et al.*, 2007; SOUZA *et al.*, 2014). Para detalhes sobre os principais poluentes, suas fontes e efeitos sobre a saúde humana consultar Arbex *et al.* (2012).

Nascimento *et al.* (2006) investigaram dados diários do número de internações por pneumonia na cidade de São José dos Campos - SP, dados de concentrações médias diárias dos poluentes dióxido de enxofre (SO_2), ozônio troposférico (O_3) e material particulado inalável (MP_{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 $\mu g.m^{-3}$ na concentração de MP_{10}, houve um acréscimo de 9,8 % nas internações.

Vale dizer que os modelos aditivos generalizados (MAG) são uma extensão dos modelos lineares generalizados. Neste grupo de modelos estatísticos a variável dependente ou resposta (Y) é um processo de contagem e as varáveis independentes são variáveis canditadas a explicar o comportamento da série ao longo do tempo. No MAG cada variável independente analisada não entra no modelo com o seu valor, mas, sim, adotando uma função não paramétrica de forma não especificada, que é estimada a partir de curvas de alisamento. Sendo assim, não é necessário assumir uma relação linear e/ou aditiva entre a variável dependente e a variável independete em estudo. Maiores detalhes desta classe de modelos podem ser vistos nos trabalhos de Latore e Cardoso (2001), Souza *et al.* (2014), Freitas *et al.* (2016) e Souza *et al.* (2018).

Braga *et al.* (2007) avaliaram os efeitos agudos do MP_{10} sobre os atendimentos em pronto-socorro por doenças cardiovasculares e respiratórias no Município de Itabira – MG. Os resultados evidenciam que elevações de 10 $\mu g.m^{-3}$ de MP_{10} foram associadas 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, houve um efeito agudo, principalmente para os indivíduos com idade entre 45 e 64 anos.

Ikefuti, Barrozo e Braga (2018) avaliaram associações entre acidente vascular cerebral (AVC) e temperatura média do ar, usando dados de mortalidade registrados e dados de estações meteorológicas de 2002 a 2011 na cidade de São Paulo. Uma análise de séries temporais foi aplicada a 55.633 casos de mortalidade. De acordo com os resultados apresentados a temperatura média do ar está associada à mortalidade por acidente vascular cerebral na cidade de São Paulo para homens e mulheres. Vale ressaltar que, segundo os autores no Brasil, as doenças crônicas são responsáveis pela maior porcentagem de todas as mortes entre homens e mulheres. Entre as doenças cardiovasculares, o AVC é a principal causa de morte, sendo responsável por 10% de todas as mortes.

Na Região da Grande Vitória (RGV), Souza *et al.* (2014) realizaram um estudo cujo objetivo foi investigar a associação entre concentrações dos poluentes atmosféricos e atendimentos diários por causas respiratórias em crianças. Foram analisadas as contagens diárias de admissões hospitalares de crianças menores de 6 anos e as concentrações diárias de poluentes atmosféricos, MP_{10}, SO_2, dióxido de nitrogénio (NO_2), O_3 e monóxido de carbono (CO), de janeiro de 2005 a dezembro de 2010. Os autores combinaram duas técnicas para a análise estatística: modelo de regressão de Poisson em modelos aditivos generalizados e análise de componentes principais. Os resultados mostraram que o aumento de 10.49 $\mu g.m^{-3}$ (intervalo interquartílico) nos níveis do poluente MP_{10} levou a um aumento de 3,0% do valor do risco relativo estimado por meio do modelo aditivo generalizado, enquanto no modelo aditivo generalizado usual a estimativa foi de 2,0%. Ainda, segundo os resultados, existe uma relação significativa, ao nível de 95% de confiabilidade, entre os níveis de concentração dos

poluentes e o número de atendimentos hospitalares em crianças menores de 6 anos, mesmo em um ambiente com níveis abaixo dos padrões recomendados pelo Conselho Nacional do Meio Ambiente (CONAMA) e pela *World Health Organization* (WHO).

Freitas *et al.* (2016) analisaram o impacto da poluição atmosférica na morbidade respiratória e cardiovascular de crianças e adultos em Vitória, ES. Foi realizado um estudo utilizando modelos de séries temporais via regressão de Poisson a partir de dados de hospitalizações e poluentes em Vitória, de 2001 a 2006. Foram testadas como variáveis independentes MP_{10}, o SO_2 e o O_3, em defasagem simples e acumulada até cinco dias. Introduziram-se temperatura, umidade e variáveis indicadoras dos dias da semana e feriados da cidade como variáveis de controle nos modelos. Os autores observaram que, para cada incremento de 10 $\mu g.m^{-3}$ dos poluentes MP_{10}, SO_2 e O_3, houve aumentos no risco relativo percentual (RR%) para as hospitalizações por doenças respiratórias, totais de 9,67; 6,98 e 1,93, respectivamente. De acordo com os resultados da pesquisa, as doenças respiratórias apresentaram relação forte e consistente com os poluentes pesquisados em Vitória. O conjunto de dado utilizado como variáveis dependenetes nos modelos foi cotagens diárias de internações pelas causas invertigadas: doenças respiratórias (CID10:J00-J99) como por exemplo, infecções agudas das vias aéreas superiores, doenças crônicas das vias aéreas inferiores, doenças pulmonares devidas a agentes externos, entre outras e doenças cardiovasculares (CID10: I00-I99) em maiores de 39 anos como, doenças hipertensivas, doenças isquêmicas do coração, doenças das artérias, das arteríolas e dos capilares.

Os efeitos causados ao meio ambiente e à saúde da população por causa da emissão de poluentes atmosféricos podem não ser apenas locais, pois dependem de fatores da região, como o relevo do entorno da fonte de emissão, as condições meteorológicas e a natureza dos poluentes. Isso significa que esses poluentes podem viajar milhares de quilômetros pela

atmosfera, atingindo, assim, comunidades distantes do ponto de emissão (LEITE, *et al.*, 2011).

Vale frisar que, de acordo com Moreira, Tirabassi e Moraes (2008), as condições meteorológicas desempenham um papel importantíssimo na dispersão ou acumulação de poluentes, uma vez que estas condições atuam através de dois fenômenos fundamentais: o transporte e a difusão. Além disso, as condições meteorológicas influenciam a deposição no solo dos poluentes. Liu e Johnson (2002) descreveram que a poluição do ar está associada, geralmente, a fatores como temperatura, umidade relativa, velocidade e direção do vento, entre outros. Como exemplo, tem-se que a baixa umidade relativa e a reduzida velocidade do vento tendem a elevar os níveis de poluentes. Já a ocorrência de precipitação pluviométrica e o aumento da velocidade do vento contribuem para a dispersão dos poluentes e, consequentemente, para a redução da concentração dos mesmos.

Como grandes projetos industriais foram implantados na RGV desde o início da década de 1970, a indústria era apontada como a principal fonte de poluição da área. De acordo com o IEMA (2013), porém, o crescimento da frota veicular e os empreendimentos imobiliários têm alterado o perfil da região nos últimos anos. O inventário de emissões atmosféricas da Região da Grande Vitória (ECOSOFT, 2011) estimou que os veículos e a indústria minero-siderúrgica foram os principais responsáveis pelas emissões de CO e hidrocarbonetos, a indústria minero-siderúrgica foi responsável por mais de 70% da emissão de SO_2 para a atmosfera e por mais de 45% dos óxidos de nitrogênio (NO_x), e indicou como principais responsáveis pelos níveis de MP_{10} a ressuspensão do solo (69,3%), as indústrias (19,6%) e os veículos (Escapamento e Evaporativa) (3,9%). Com o aumento da poluição do ar na região, aumentou também a preocupação com os efeitos adversos à saúde humana, à fauna e à flora. Dessa forma, este trabalho objetivou avaliar os impactos das variáveis meteorológicas temperatura, a umidade relativa, a velocidade do vento e a precipitação na probabilidade de ocorrência de

picos/episódios de concentração de poluentes na RGV, por meio do modelo Logit.

O modelo Logit é uma ferramenta estatística que permite o ajuste de um conjunto de variáveis independentes a uma variável dependente dicotômica, ou seja, é possível estimar probabilidades de ocorrências em variáveis dependentes do tipo binário. O modelo Logit aplicado na modelagem de dados ambientais tem sido utilizado por Kuchenho e Thamerus (1996), Mendes e Vega (2011), Silva *et al.* (2011), Gehring *et al.* (2013), entre outros autores. Com objetivo semelhante, cabe ressaltar o trabalho de Leite *et al.* (2011), que analisaram a qualidade do ar atmosférico de Uberlândia, em Minas Gerais, por meio de modelos de regressão logística simples. Os autores utilizaram dados do período de 2003 a 2008 de concentrações de MP_{10} e de variáveis meteorológicas. De acordo com os resultados, os modelos logísticos simples mostraram que as variáveis meteorológicas precipitação e umidade relativa contribuem significativamente para a redução das concentrações de MP_{10}.

Materiais e métodos

Área de estudo e apresentação das variáveis

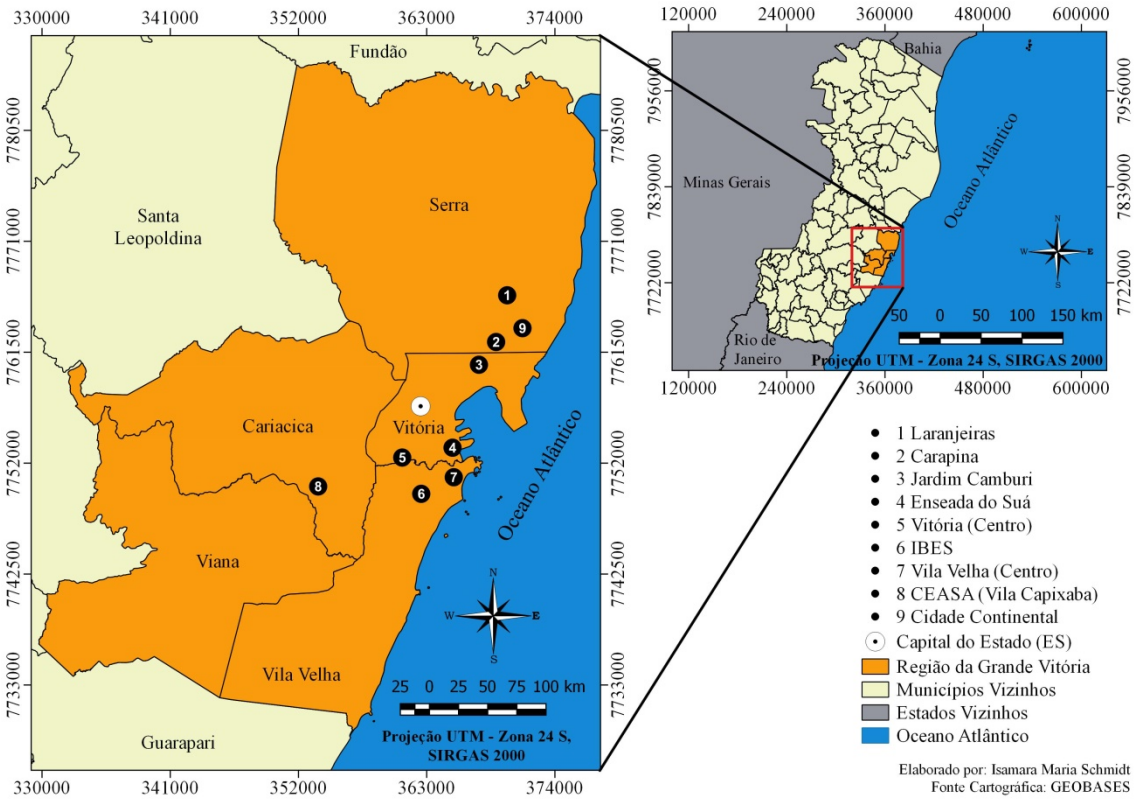
Para a realização deste estudo, utilizaram-se séries temporais de concentrações de poluentes atmosféricos e de variáveis meteorológicas monitorados na RGV (conforme apresentado na Tabela 1), Espírito Santo, Brasil, região que é constituída pelos municípios de Vitória, Vila Velha, Cariacica, Serra e Viana. Segundo o Instituto Brasileiro de Geografia e Estatística (IBGE, 2010), a RGV abrange uma área de 1.461 Km^2, com aproximadamente 1.475.332 habitantes, sendo um dos principais polos de desenvolvimento urbano e industrial do Estado. A região sofre com diversos tipos de problemas ambientais, entre os quais está a deterioração da qualidade do ar, devido às emissões atmosféricas por indústrias, pela frota

veicular e pela ressuspensão do solo causada pelo vento e pelo tráfego veicular.

Vale ressaltar que a RGV possui uma Rede Automática de Monitoramento da Qualidade do Ar (RAMQAR) inaugurada em julho de 2000, de propriedade do IEMA. Essa rede é distribuída em oito estações, a saber: no município da Serra, duas estações localizadas nas regiões de Laranjeiras e de Carapina; no município de Vitória, três estações localizadas nas regiões de Jardim Camburi, Enseada do Suá e Centro de Vitória; no município de Vila Velha, duas estações localizadas nas regiões do Ibes e Centro de Vila Velha; por fim, no município de Cariacica, uma estação em Cariacica. A localização espacial das estações de monitoramento da RAMQAR está ilustrada na Figura 1.

A RAMQAR monitora os seguintes poluentes: Partículas Totais em Suspensão (PTS), MP_{10}, O_3, NO_x, CO e Hidrocarbonetos (HC). Ainda, realiza 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). Na análise deste trabalho, as variáveis PTS, MP_{10}, SO_2, CO, NO_2, O_3, velocidade do vento, umidade relativa, precipitação e temperatura foram utilizadas conforme descrição da Tabela 2.

Figura 1: Localização espacial das estações de monitoramento da qualidade do ar da RGV.



Fonte: Elaborado para os autores por Isamara Maria Schmidt, 2018.

Tabela 1: Parâmetros meteorológicos e poluentes monitorados em cada estação RAMQAR

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

Fonte: Adaptado de Relatório da Qualidade do Ar da Região da Grande Vitória, 2013.

Tabela 2: Descrição das variáveis PTS, MP_{10}, SO_2, CO, NO_2, O_3, velocidade do vento, umidade, precipitação e temperatura

Variáveis	Unidades	Descrição
PTS, MP_{10}, SO_2, CO, NO_2 e O_3	$\mu g.m^{-3}$	Valor máximo medido entre as estações que fazem o monitoramento do poluente, para cada dia da amostra de dados.
Velocidade do Vento	$m.s^{-1}$	Valor máximo medido entre as estações de Carapina, Enseada do Suá, Ibes e Cariacica.
Umidade relativa	%	Como existem muitos dados faltantes para a estação de Cariacica, adotou-se o valor de Carapina.
Precipitação	mm	Valores medidos na estação de Carapina, única que possui medição para tal variável.
Temperatura	ºC	Média aritmética entre as estações de Carapina e Cariacica, únicas que possuem medições para tal variável.

Fonte: Org. dos autores, 2017.

O IEMA classifica diariamente a qualidade do ar da RGV como "boa" e "não boa". Nesse último caso, são englobadas as classificações "regular", "inadequada", "má", "péssima" e "crítica". A qualidade é considerada "boa" se a concentração de cada poluente for igual ou inferior aos valores apresentados na Tabela 3. Já a classificação de "não boa" ocorre quando a concentração é superior aos valores apresentados na Tabela 3.

Este trabalho preocupou-se com os picos de concentração. As concentrações de cada poluente foram representadas pelo valor máximo medido dentre as estações que fazem o monitoramento desses poluentes (Tabela 1), para cada dia da amostra de dados. No mais, como a regressão logística considera a variável dependente como dicotômica, a concentração de cada poluente foi transformada em uma variável *dummy,* apresentando a seguinte classificação: um (1) para classificação "não boa" (picos de concentração) e zero (0) para "boa".

Tabela 3: Limites de concentração para classificação da qualidade do ar como "boa" e "não boa"

Classificação	PTS $\mu g/m^3$ Média 24h	PM_{10} $\mu g/m^3$ Média 24h	SO_2 $\mu g/m^3$ Média 24h	NO_2 $\mu g/m^3$ Média 1h	O_3 $\mu g/m^3$ Média 8h	CO $\mu g/m^3$ Média 1h
Boa	0 - 65	0 - 45	0 - 40	0 - 50	0 - 70	0 - 5.000
Não boa	> 66	> 46	> 41	> 51	> 71	> 5.001

Fonte: IEMA, 2017.

Modelo Logit

Para verificar a influência das variáveis meteorológicas (variáveis preditoras) temperatura, umidade relativa, velocidade do vento e precipitação pluviométrica na probabilidade de ocorrência de picos de concentração de poluentes atmosféricos (classificação "não boa" na qualidade do ar) na RGV, foi utilizado o modelo Logit (GUJARATI; PORTER 2008), que admite valores discretos, zero e um (variável binária) para a variável dependente. Um dos principais objetivos dos modelos de resposta binária é calcular a probabilidade de um dado evento, com determinado conjunto de atributos, de fato acontece (MONTE, ALBUQUERQUE e REISEN, 2016).

No modelo Logit utiliza-se uma função de distribuição acumulada logística, dada por

$$L(X_t\beta) = \frac{1}{1 + e^{-(\beta_0 + \Sigma_{i=1}^{k} \beta_i X_{ti})}}, \quad 1)$$

em que L representa a função de distribuição logística; X_t, variáveis independentes; $\beta = \{\beta_0, \beta_1, \cdots, \beta_k\}$, vetor de parâmetros a serem estimados; k, o número de variáveis explicativas; e, base do logaritmo natural; e, $t = 1, \cdots, n$ (número de observações).

A ocorrência ou não de uma classificação "não boa" da qualidade do ar depende de vários fatores. Como os parâmetros dessa ocorrência não são observáveis para cada ponto do tempo t, pode-se definir uma variável latente ou não observada, Y_t^*, como

$$Y_t^* = \beta_0 + \sum_{i=1}^{k} \beta_i X_{ti} + \mu_t, \quad 2)$$

em que Y_t^* é variável latente e μ_t, erro aleatório. Para detalhes ver Gujarati e Porter (2008).

A ocorrência de uma determinada classificação pode ser descrita pela variável binária Y_t, tal que $Y_t = 1$, se a classificação é "não boa", e $Y_t = 0$, se é "boa". Esses valores observados de Y_t estão relacionados com Y_t^*, como segue:

$Y_t = 1$, se $Y_t^* > 0$; e, $Y_t = 0$, se $Y_t^* = 0$,

$$\text{Prob}(Y_t = 1) = \text{Prob}(Y_t^* > 0) = \text{Prob}\left(\mu_t > -\left(\beta_0 + \sum_{l=1}^{k} \beta_l X_{tl}\right)\right),$$

$$\text{Prob}(Y_t = 0) = \text{Prob}(Y_t^* = 0) = \text{Prob}\left(\mu_t \leq -\left(\beta_0 + \sum_{l=1}^{k} \beta_l X_{tl}\right)\right).$$

Os parâmetros do modelo são estimados pelo Método de Máxima Verossimilhança. A probabilidade de ocorrência da classificação "não boa" (a) e a probabilidade de ocorrência da classificação "boa" (b) pode ser calculada pelas seguintes expressões:

$$\text{a) } P_t = \frac{1}{1+e^{-(\beta_0+\sum_{l=1}^{k}\beta_l X_{tl})}} \text{ e b) } 1 - P_t = \frac{e^{-(\beta_0+\sum_{l=1}^{k}\beta_l X_{tl})}}{1+e^{-(\beta_0+\sum_{l=1}^{k}\beta_l X_{tl})}}, \qquad 3)$$

sendo P_t igual a probabilidade de ocorrência da classificação "não boa", e $1 - P_t$, probabilidade de ocorrência da classificação "boa".

Para determinar o efeito marginal de cada variável preditora, sobre a probabilidade de ocorrência da classificação "não boa", é necessário usar os valores médios das variáveis explicativas. O efeito marginal da variável X_{tl} sobre a variável dependente é descrito pela expressão

$$\frac{\partial P_t}{\partial X_{tl}} = \beta_l \times \frac{1}{1+e^{-(\beta_0+\sum_{l=1}^{k}\beta_l X_{tl})}} \times \frac{e^{-(\beta_0+\sum_{l=1}^{k}\beta_l X_{tl})}}{1+e^{-(\beta_0+\sum_{l=1}^{k}\beta_l X_{tl})}} = \beta_l \frac{e^{-(\beta_0+\sum_{l=1}^{k}\beta_l X_{tl}}}{\left(1+e^{-(\beta_0+\sum_{l=1}^{k}\beta_l X\right.} \quad 4)$$

A Equação 4 mostra o efeito marginal sobre P_t, de um aumento em X_{tl}. Observa-se que o efeito marginal de cada variável explicativa sobre a probabilidade não é constante, visto que depende do valor médio de cada

variável X_{tl}. Para detalhes, ver em Gujarati e Porter (2008) e Hill, Judge e Griffiths (2010).

Resultados e discussão

Para um entendimento preliminar das variáveis em estudo, são apresentadas algumas medidas descritivas (Tabelas 4 e 5). Para detalhes sobre as medidas apresentadas e suas formulações matemáticas, consultar em Gujarati e Porter (2008) e Morettin e Bussab (2017).A análise corresponde às variáveis apresentadas na Tabela 2, para o período de 01/01/2012 a 31/12/2014, perfazendo um total de 1096 observações. Todas as análises foram feitas utilizando o *software* livre R (R DEVELOPMENT CORE TEAM, 2016). A presença de dados faltantes nas séries motivou o uso da metodologia de imputação via algoritmo EM (*expectation-maximisation*), proposta por Junger e Leon (2015) e implementada na biblioteca R *mtsdi* (*multivariante time-series data imputation*).

Tabela 4: Estatísticas descritivas das variáveis poluentes atmosféricos

Estatísticas	Variáveis					
	MP_{10}	PTS	SO2	NO_2	CO	O_3
Média	48,3911	87,1433	21,2563	74,3990	844,5010	35,7936
Mediana	46,3750	79,5833	19,7050	70,1688	805,4279	35,5504
Desvio padrão	16,9144	36,7769	9,9869	26,8003	285,4049	9,1888
Coeficiente de variação	34,9535	42,2028	46,9832	36,0224	33,7957	25,6717
Valor máximo	125,0000	279,1667	73,1713	197,2292	1893,5408	67,1804
Valor mínimo	12,2500	15,0000	2,4221	19,8208	232,7021	0,0000

Fonte: Org. dos autores, 2017.

Nota-se que (Tabela 4), em média, as concentrações de MP_{10}, PTS e NO_2 ultrapassaram os valores apresentados na Tabela 3 para a classificação de qualidade do ar "boa", já as concentrações de SO_2, O_3 e CO não ultrapassaram. Porém, o desvio padrão e o coeficiente de variação alto

indicam uma média pouco representativa e uma grande variabilidade dos dados. Por exemplo, os resultados mostram que a concentração máxima de PTS foi mais do que o triplo do valor médio, demostrando a grande variabilidade desse poluente na RGV.

A Tabela 5 apresenta as estatísticas descritivas para as variáveis meteorológicas. Em geral, observando-se os desvios padrão, o coeficiente de variação e as diferenças entre os máximos e mínimos, percebe-se que as variáveis meteorológicas apresentaram grande variabilidade em termos estatísticos, exceção feita à temperatura. Segundo o IEMA (2013), não há muita variabilidade climatológica na RGV.

Tabela 5: Estatísticas descritivas das variáveis meteorológicas

Estatísticas	Variáveis			
	VV	PP	UR	T
Média	2,4256	0,1289	77,5826	25,0952
Mediana	2,3087	0,0000	77,0229	25,0392
Desvio padrão	0,7904	0,5421	6,0259	2,4129
Coeficiente de variação	32,5846	420,5203	7,7671	9,6150
Valor máximo	5,5412	12,7750	98,2983	31,4000
Valor mínimo	1,0392	0,0000	56,8638	18,1758

Fonte: Org. dos autores, 2017.

A Tabela 6 mostra as correlações calculadas entre as concentrações de médias diárias de PTS, MP_{10}, SO_2, CO, NO_2, O_3, de velocidade do vento, da umidade, de precipitação e de temperatura. Observa-se que as variáveis meteorológicas apresentam relação linear com as concentrações de poluentes. Conforme esperado, os índices de concentrações máximas na RGV estão associados às mudanças dessas variáveis, ou seja, infere-se dos resultados (Tabela 6) que o aumento da velocidade do vento acarreta aumento nas concentrações de MP_{10}, SO_2 e O_3 e diminuição das concentrações de PTS, NO_2 e CO; o aumento da precipitação acarreta uma diminuição nas concentrações de MP_{10}, PTS e SO_2; o aumento da umidade implica em um aumento de NO_2 e CO e em uma diminuição das concentrações de MP_{10}, PTS e SO_2. Já o aumento da temperatura acarreta

um aumento nas contrações de MP_{10}, PTS, SO_2 e O_3 e diminuição de NO_2 e CO.

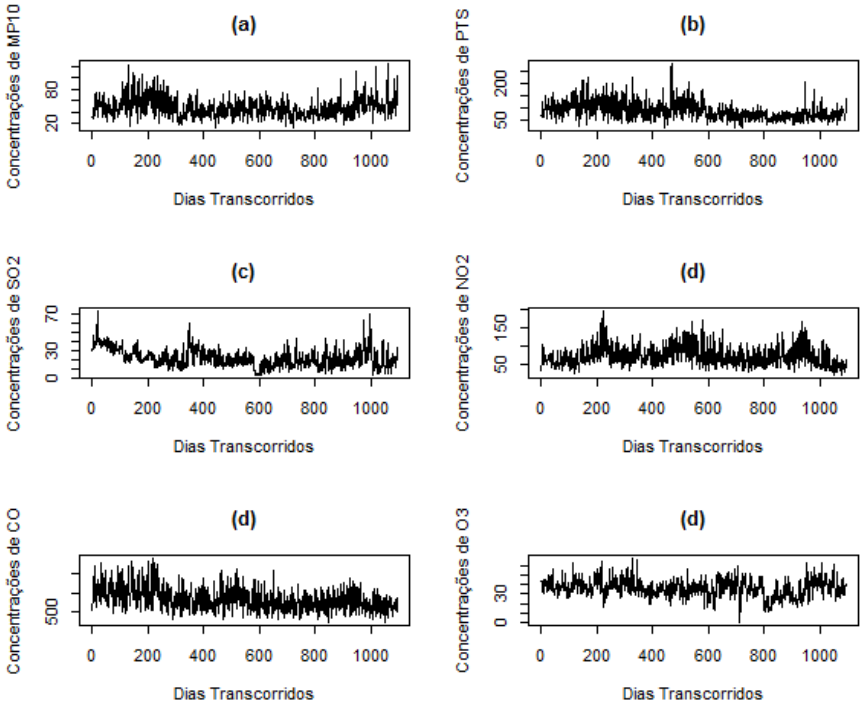
Tabela 6: Matriz de correlação entre as variáveis sob estudo

	MP_{10}	PTS	SO2	NO_2	CO	O_3	VV	PP	UR	T
MP_{10}	1,0000									
PTS	0,6453	1,0000								
SO2	0,2947	0,1960	1,0000							
NO_2	0,2101	0,4105	-0,1060	1,0000						
CO	0,3488	0,4688	0,1825	0,6742	1,0000					
O_3	0,0769	0,0755	0,1697	-0,3567	-0,3248	1,0000				
VV	0,1823	-0,0645	0,3023	-0,5862	-0,3797	0,3739	1,0000			
PP	-0,1801	-0,1631	-0,0539	0,0098	0,0385	0,0019	-0,0535	1,0000		
UR	-0,1772	-0,0755	-0,0843	0,2178	0,3107	-0,2146	-0,3671	0,3240	1,0000	
T	0,0906	0,0966	0,4326	-0,3405	-0,0786	0,0256	0,3222	-0,1425	-0,3406	1,0000

Fonte: Org. dos autores, 2017.

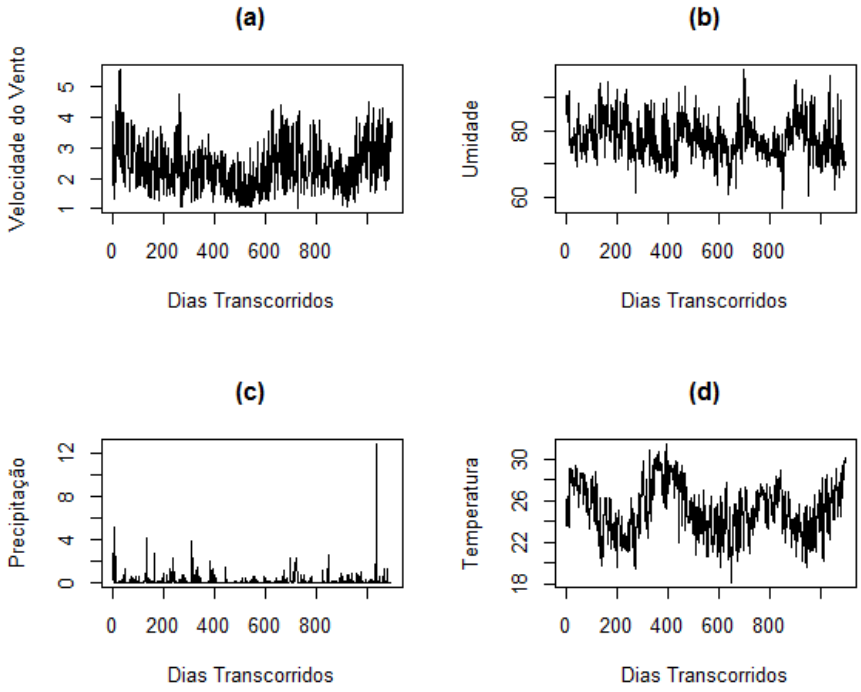
A Figura 2 apresenta as séries temporais das concentrações de poluentes atmosféricos, e a Figura 3 as das variáveis meteorológicas no período de análise desse estudo. De um total de 1096 observações (dias) para o poluente MP_{10}, em cerca de 51% dos dias (médias de 24 horas) a qualidade do ar foi classificação como "não boa"; para o PTS, aproximadamente em 68% dos dias (médias de 24 horas); para o SO_2, a porcentagem foi de 95% dos dias (médias de 24 horas); para o NO_2, 64% das médias horárias foram classificadas como "não boas"; para o O_3, 0,63% das médias de oito horas foram classificadas como "não boas". Já para o poluente CO, a qualidade do ar foi classificada como "não boa" uma única vez.

Figura 2: Séries temporais das concentrações dos poluentes atmosféricos em estudo.



Fonte: Org. dos autores, 2017.

Figura 3: Séries temporais das variáveis meteorológicas em estudo.



Fonte: Org. dos autores, 2017.

Na Tabela 7 são apresentados os resultados para a regressão logística estimada para cada poluente, considerando como variáveis explicativas a umidade relativa, a velocidade do vento, a precipitação e a temperatura. Também são apresentados os respectivos efeitos marginais de cada variável sobre a probabilidade de ocorrência de qualidade do ar "não boa" para todos os poluentes estudados. Verificou-se que as regressões como um todo foram estatisticamente significativas. Uma vez que para o poluente CO a qualidade do ar foi classificada como "não boa" uma única vez, não foi estimada uma regressão para ele, pois o efeito das variáveis meteorológicas não seria estatisticamente significativo.

Para o Poluente MP_{10}, os resultados mostraram que as variáveis explicativas velocidade do vento, precipitação e umidade interferem significantemente na qualidade do ar "não boa". Como a variável temperatura não foi estatisticamente significativa, os resultados para ela

não se encontram na Tabela 6. Analisando os efeitos marginais (Tabela 6), verifica-se que, para cada unidade de acréscimo na velocidade do vento na região, ocorreu um aumento de 5,9 pontos percentuais na probabilidade de qualidade do ar "não boa"; para a precipitação um aumento de uma unidade, ocasionou uma redução na probabilidade de qualidade do ar "não boa" de 27,11 pontos percentuais. Já para a variável umidade, o efeito marginal foi muito pequeno.

É importante ressalvar que, para o poluente MP_{10}, os sinais estimados são coerentes com o esperado. O sinal positivo do coeficiente estimado da variável velocidade do vento indica a existência de uma relação direta entre a concentração de MP_{10} e essa variável, demostrando que o aumento da velocidade do vento eleva a concentração de MP_{10}. Segundo o Relatório da Qualidade do Ar da Grande Vitória (IEMA, 2013), 69,3% das emissões de MP_{10} para a atmosfera da RGV estão ligadas à ressuspensão causada pelo vento e pelo tráfego veicular, corroborando com o resultado encontrado, uma vez que o aumento da velocidade do vento tende à elevar a ressuspensão do solo. Já a precipitação pluvial e a umidade apresentaram relação negativa com a concentração de MP_{10}. O coeficiente negativo no modelo se deve ao processo de deposição úmida e atenuação da ressuspensão do solo (LYRA; ODA-SOUZA; VIOLA, 2011). Durante eventos de precipitação ocorre diminuição do MP_{10}, pois o solo úmido dificulta a ressuspensão do particulado no solo (VARDOULAKIS; KASSOMENOS, 2008).

Para o poluente PTS, os resultados mostraram que a variável precipitação interfere significantemente na qualidade do ar "não boa". Analisando os efeitos marginais (Tabela 7), verifica-se que para cada unidade de acréscimo na precipitação ocasionou uma redução na probabilidade de qualidade do ar "não boa" de 15,69 pontos percentuais; para a temperatura, um aumento de uma unidade ocasionou um aumento na probabilidade de qualidade do ar "não boa" de 2,83 pontos percentuais.

Como as variáveis velocidade do vento e umidade não foram estatisticamente significativas, os resultados para as mesmas não se encontram na Tabela 7.

Os resultados mostraram que as variáveis meteorológicas velocidade do vento e precipitação interferem nas concentrações de SO_2 e NO_2 presentes no ar da RGV, favorecendo na redução da poluição atmosférica, como mostra a Tabela 6. Observou-se que um aumento de uma unidade na velocidade do vento acarreta uma redução de 1,16 pontos percentuais na probabilidade de ocorrência de qualidade do ar "não boa" para o poluente SO_2 e de 7,79 pontos percentuais para o poluente NO_2. Já para a precipitação, um aumento de uma unidade implica em uma redução de 3,95 pontos percentuais na probabilidade de ocorrência de qualidade do ar "não boa" para o poluente SO_2. Ressalta-se, ainda, que o efeito marginal da umidade foi muito pequeno, tanto para o SO_2 quanto para o NO_2; da precipitação e da temperatura foi pequeno para o NO_2. Os resultados revelaram, também, que as variações das variáveis meteorológicas não são estatisticamente significantes para explicar as concentrações de O_3 na RGV.

Vale ressaltar que, as condições meteorológicas influenciam as condições de mistura vertical do ar, e as concentrações de O_3 aumentam, conforme diminuem as condições de mistura. Porém, sabe-se que as concentrações de NO_x aumentam conforme aumeta a estabilidade vertical da atmosfera, o que afeta diretamente as concentrações de O_3. Além disso, as concentrações de ozônio troposférico num determinado local são afetadas pela constituição atmosférica, e esta é notadamente modificada em grandes centros urbanos onde há grandes emissões de precursores devido às atividades antrópicas (CHIQUETTO, 2008).

Para corroborar com os resultados apresentados na Tabela 7, é válido ressaltar que os sinais dos coeficientes estimados para as variáveis meteorológicas estão de acordo com o esperado. O sinal negativo dos coeficientes estimados das variáveis velocidade do vento e precipitação

indica a existência de uma relação inversa entre as concentrações de SO_2 e essas variáveis, demostrando que o aumento da velocidade do vento e da precipitação diminuem as concentrações de SO_2. Segundo Derisio (2012), a poluição numa região ocorre em função das atividades da comunidade aí presentes e das condições meteorológicas. A ocorrência de precipitação indica instabilidade atmosférica, ou seja, favorece a dispersão de poluentes pelos movimentos de ar e as gotas de chuva também removem quantidades consideráveis desses poluentes. O aumento da velocidade do vento favorece a dispersão do SO_2 e do NO_2 na atmosfera. Para Derisio (2012), o vento é o primeiro mecanismo atmosférico de transporte de contaminantes.

Tabela 7: Equações logísticas considerando as variáveis explicativas e seus efeitos marginais para cada poluente

Modelo	Variáveis	Coeficientes	Erro-padrão	Valor de Z	Valor-p	Efeito marginal
ModeloPM_{10}	Constante	0,8654	1,0603	0,8162	0,4144	-
	VV	0,3771*	0,0880	4,2863	0,0000	0,0598
	PREC	-1,7099*	0,3815	-4,4819	0,0000	-0,2711
	UMID	-0,0204	0,0126	-1,6130	0,1067	-0,0032
	TEMP	-	-	-	-	-
	Constante	-2,4256	0,7212	-3,3631	0,0008	-
	VV	-	-	-	-	-
ModeloPTS	PREC	-0,7391*	0,1966	-3,7585	0,0002	-0,1569
	UMID	-	-	-	-	-
	TEMP	0,1331*	0,0288	4,6238	0,0000	0,0283
	Constante	-0,4062	1,4096	-0,2882	0,7732	-
	VV	-0,1854**	0,0937	-1,9774	0,0480	-0,0116
ModeloSO_2	PREC	-0,6310*	0,2087	-3,0242	0,0025	-0,0395
	UMID	-0,0221***	0,0131	-1,6907	0,0909	-0,0014
	TEMP	0,1387*	0,0310	4,4748	0,0000	0,0087
	Constante	3,0386	0,2321	13,0904	0,0000	-
	VV	-0,3185*	0,0136	-23,4487	0,0000	-0,0779
ModeloNO_2	PREC	0,0363*	0,0122	2,9869	0,0028	0,0089
	UMID	-0,0332*	0,0017	-19,9910	0,0000	-0,0081
	TEMP	0,0373*	0,0052	7,2506	0,0000	0,0091
	Constante	0,5211	1,2799	0,4071	0,6839	-
	VV	-	-	-	-	-
ModeloO_3	PREC	-	-	-	-	-
	UMID	-0,0770*	0,0188	-4,0875	0,0000	-0,0003
	TEMP	-	-	-	-	-

Nota: 1) *Significativo a 1%, **Significativo a 5%, ***Significativo a 10%; e, 2) As estimativas foram realizadas utilizando o método de covariâncias robusta GLM (Modelo Linear Generalizado).
Fonte: Org. dos autores, 2017.

Referente ao coeficiente da variável temperatura, para o poluente SO_2, embora pequeno, ele foi estatisticamente significativo, indicando que tal variável contribuiu para a elevação das concentrações de SO_2 na RGV. Como mencionado por Derisio (2012), quando a temperatura aumenta na superfície pode ocorrer a chamada inversão térmica, geralmente a camada de inversão dificulta a dispersão de poluentes, ocasionando o aumento da concentração destes junto ao solo, já que nesse período de inversão os ventos horizontais geralmente têm baixa velocidade.

Considerações Finais

Este trabalho objetivou avaliar os impactos das variáveis meteorológicas temperatura, a umidade relativa, a velocidade do vento e a precipitação na probabilidade de ocorrência de picos/episódios de concentração de poluentes na RGV, Espírito Santo, Brasil, por meio do modelo Logit, no período de 01/01/2012 a 31/12/2014. Para avaliar os efeitos marginais, a qualidade do ar, no que se refere aos poluentes PTS, MP_{10}, SO_2, CO, NO_2 e O_3, foi classificada como "boa" e "não boa".

A análise dos resultados mostrou que as variáveis meteorológicas apresentam relação linear com as concentrações de poluentes atmosféricos e que os índices de concentrações na RGV estão associados às mudanças dessas variáveis. Constatou-se que a ocorrência de precipitação está associada à redução da probabilidade de ocorrência de qualidade do ar "não boa" de três dos cinco poluentes avaliados. Adicionalmente notou-se que a velocidade do vento interfere significativamente na concentração de PM_{10}, SO_2 e NO_2 presente na atmosfera, favorecendo ou não a poluição do ar.

Conclui-se que entre as variáveis meteorológicas avaliadas, a velocidade do vento e a precipitação foram mais significativas na redução da probabilidade de ocorrência de classificação do ar como "não boa", uma vez que apresentou os maiores efeitos marginais. Desta maneira, pode-se ainda

inferir que maiores velocidades do vento e altos volumes de chuvas contribuíram fortemente para a qualidade do ar, na região de estudo, por intensificaram o processo de dispersão e diluição de poluentes. Desta forma, o modelo Logit mostrou ser uma ferramenta estatística satisfatória para avaliar a qualidade do ar "não boa" da RGV.

Por fim, torna-se importante destacar que os efeitos da poluição do ar na saúde são detectados em diversas cidades no mundo. Porém, segundo Arbex *et al.* (2012), apesar de os efeitos da poluição terem sido descritos desde a antiguidade, somente com o advento da revolução industrial a poluição passou a atingir a população em grandes proporções. Dessa forma, este estudo visa contribuir como uma ferramenta para os órgãos regulamentadores na discussão de atividades de prevenção e de controle da poluição atmosférica na RGV.

Referências

ALMEIDA, M. A. I. **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, Programa de Pós-Graduação em Engenharia Ambiental do Centro Tecnológico, Universidade Federal do Espírito Santo, Vitória, 2006.

ARBEX, M. A.; SANTOS, U. P.; MARTINS, L. C.; SALDIVA, P. H. N.; PEREIRA, L. A. A.; BRAGA, A. L. F. A poluição do ar e o sistema respiratório. **Jornal Brasileiro de Pneumologia**, v. 38, n. 5, p. 643-655, 2012. https://doi.org/10.1590/S1806-37132012000500015

BRAGA, A. L. F.; PEREIRA, L. A. A.; PROCÓPIO, M.; ANDRÉ, P. A.; SALDIVA, P. H. N. Associação entre poluição atmosférica e doenças respiratórias e cardiovasculares na cidade de Itabira, Minas Gerais, Brasil. **Caderno de Saúde Pública**, Rio de Janeiro, v. 23, Sup 4, p.S570-S578, 2007.

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

CANÇADO, J. E. D. et al. Repercussões clínicas da exposição à poluição atmosférica. **J bras pneumol**, v. 32, n. Supl 1, p. S5-S11, 2006.

CHIQUETTO, J. B.. **Padrões atmosféricos associados a concentrações de ozônio troposférico na região metropolitana de São Paulo**. Tese (Doutorado em Geografia Física) – São Paulo: Universidade de São Paulo, 2008.

CURTIS, L.; REA, W.; SMITH-WILLIS, P.; FENYVES, E.; PAN, YAQIN. Adverse health effects of outdoor air pollutants. **Environment International**, v. 32, n. 6, p. 815-830, 2006. https://doi.org/10.1016/j.envint.2006.03.012

DERISIO, J. C. **Introdução ao controle de poluição ambiental**. 4 ed. São Paulo: Oficina de Textos, 2012.

ECOSOFT CONSULTORIA E SOFTWARES AMBIENTAIS (ECOSOFT). **Inventário de emissões atmosféricas da região da grande vitória**. Acordo de Cooperação Técnica IEMA & EcoSoft nº 010/2009. Vitória, 2011. Disponível em: < http://www.meioambiente.es.gov.br/download/RTC10131_R1.pdf>. Acesso em: 15 de julho de 2016.

FREITAS, C. U.; LEON, A. P.; JUNGER, W.; GOUVEIA, N. Poluição do ar e impactos na saúde em Vitória, Espírito Santo. **Revista de Saúde Pública**, v. 50, n. 4, p. 1-9, 2016.

GEHRING, U., et al. Air pollution exposure and lung function in children: the escape project. **Environmental Health Perspectives**, v. 121, n. 11-12, p. 1357-1364, 2013. https://doi.org/10.1289/ehp.1306770

GOUVEIA. N.; MENDONÇA, G. A. S.; LEON, A. P.; CORREIA, J. E. M.; JUNGER, W. L.; FREITAS, C. U.; DAUMAS, R. P.; MARTINS, L. C.; GIUSSEPE, L.; CONCEIÇÃO, G. M. S.; MANERICH, A.; CUNHA-CRUZ, J.; Poluição do ar e efeitos na saúde nas populações de duas grandes metrópoles brasileiras. **Epidemial Serv. Saúde**, v. 12, p. 29-40, 2003.

GUJARATI, D. N.; PORTER, D. C. **Basic Econometrics**. 5 ed. New York: McGraw-Hill/Irwin, 2008.

HILL, R. C.; JUDGE, G. G.; GRIFFITHS, W. E. **Econometria**. 3 ed. São Paulo: Saraiva, 2010.

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 agosto de 2016.

IKEFUTI, Priscilla V.; BARROZO, Ligia V.; BRAGA, Alfésio LF. Mean air temperature as a risk factor for stroke mortality in São Paulo, Brazil. **International journal of biometeorology**, p. 1-8, 2018.

INSTITUTO ESTADUAL DE MEIO AMBIENTE E RECURSOS HÍDRICOS DO ESTADO DO ESPÍRITO SANTO. **Relatório da qualidade do ar da Região da Grande Vitória**. Vitória, 2013. Disponível em: <http://www.meioambiente.es.gov.br>. Acesso em: 10 de junho de 2016.

JUNGER, W. L.; LEON, A. P. Imputation of missing data in time series for air pollutants. **Atmospheric Environment**, v. 102, p. 96-104, 2015. https://doi.org/10.1016/j.atmosenv.2014.11.049

KUCHENHO, H.; THAMERUS, M. Extreme value analysis of Munich air pollution data. **Environmental and Ecological Statistics**, v. 3, n. 2, p. 127-141, 1996. https://doi.org/10.1007/BF02427858

LEITE, R. C. M.; GUIMARÃES, E. C.; LIMA, E. A. R. L.; BARROZO, M. A. S. B.; TAVARES, M. Utilização de regressão logística simples na verificação da qualidade do ar atmosférico de Uberlândia. **Engenharia Sanitária e Ambiental**, v. 16, n. 1, p. 175-180, 2011. https://doi.org/10.1590/S1413-41522011000200011

LIRA, R. S. **Modelagem e previsão da qualidade do ar na cidade de Uberlândia – MG**. 152f. Tese (Doutorado em Engenharia Química) – Faculdade de Engenharia Química, Universidade Federal de Uberlândia, Uberlândia, 2009.

LIU, P. W. G.; JOHNSON, R. Forecasting peak daily ozone levels: part 2. A regression with time series errors model having a principal component trigger to forecast 1999 and 2002 ozone levels. **Journal of the Air & Waste Management Association**, v. 53, n. 12, p. 1472-1489, 2002. https://doi.org/10.1080/10473289.2003.10466321

LOGAN, W. P. D. et al. Mortality in the London fog incident, 1952. **Lancet**, p. 336-338, 1953.

LYRA, G. B.; ODA-SOUZA, M.; VIOLA, D. N. Modelos lineares aplicados à estimativa da concentração do material particulado (PM_{10}) na cidade do Rio de Janeiro, RJ. **Revista Brasileira de Meteorologia**, v. 26, n. 3, p. 392-400, 2011.

MARTINS, L. C.; LATORRE, M. R. D. O.; CARDOSO, M. R. A.; GONÇALVES, F. L. T.; SALDIVA, P. H. N.; BRAGA, A. L. F. Poluição atmosférica e atendimentos por pneumonia e gripe em São Paulo, Brasil. **Revista de Saúde Pública**, v. 36, n. 1, p. 88-94, 2002. https://doi.org/10.1590/S0034-89102002000100014

MENDES, C. A. B.; VEGA, F. A. C. Técnicas de regressão logística aplicada à análise ambiental. **Revista Geografia** (Londrina), v. 20, n. 1, p. 5-30, 2011.

MONTE, E. Z.; ALBUQUERQUE, T. T. A.; REISEN, V. A.. Impactos das Variáveis Meteorológicas na Qualidade do Ar da Região da Grande Vitória, Espírito Santo, Brasil. **Rev. bras. meteorol.**, São Paulo, v. 31, n. 4, supl. 1, p. 546-554, dez. 2016. http://dx.doi.org/10.1590/0102-7786312314b20150100.

MOREIRA, D. M.; TIRABASSI, T.; MORAES, M. R. Meteorologia e poluição atmosférica. **Ambiente e Sociedade**. v. 11, n. 1, p. 1-13, 2008. https://doi.org/10.1590/S1414-753X2008000100002

MORETTIN, P. A.; BUSSAB, W. O. **Estatística básica**. Saraiva, 2017.

NASCIMENTO, L. F. C.; PEREIRA, L. A. A.; BRAGA, A. L. F.; MÓDOLOA, M. C. C.; CARVALHO, J. A. C. Efeitos da poluição atmosférica na saúde infantil em São José dos Campos, SP. **Revista de Saúde Pública**, v. 40, p. 77-82, 2006. https://doi.org/10.1590/S0034-89102006000100013

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

R CORE TEAM. **R**: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2016. ISBN 3-900051-07-0, URL http://www.R-project.org/

SILVA, W. S.; PAIXÃO, A. N.; ARAÚJO, A. F. V.; PICANÇO, A. P. Avaliação dos benefícios da coleta de lixo em Palmas, Tocantins: uma aplicação do método de avaliação contingente. **Engenharia Sanitária e Ambiental**, v. 16, n. 2, p. 141-148, 2011. https://doi.org/10.1590/S1413-41522011000200007

SOUZA, J. B.; REISEN, V. A.; SANTOS, J. M.; FRANCO, G. C. Componentes principais e modelagem linear generalizada na associação entre atendimento hospitalar e poluição do ar. **Revista de Saúde Pública**, v. 48, n. 3, p. 451-458, 2014. https://doi.org/10.1590/S0034-8910.2014048005078

SOUZA, Juliana B. et al. Generalized additive models with principal component analysis: an application to time series of respiratory disease and air pollution data. **Journal**

of the Royal Statistical Society: Series C (Applied Statistics), v. 67, n. 2, p. 453-480, 2018. https://doi.org/10.1111/rssc.12239

VARDOULAKIS, S.; KASSOMENOS, P. Sources and factors affecting PM_{10} levels in two European cities: implications for local air quality management. **Atmospheric Environment**, v. 42, n. 17, p. 3949–3963, 2008. https://doi.org/10.1016/j.atmosenv.2006.12.021

Identification of periodic components in time series with missing data: an application to air pollution data

Wanderson de Paula Pinto[a], Valdério Anselmo Reisen[a,b], Edson Zambon Monte[c] and Carlo Corrêa Solci[a]

[a]Post-Graduate Program in Environmental Engineering (PPGEA) - Federal University of Espírito Santo, Brazil; [b]Department of Statistics, Federal University of Espírito Santo, Brazil
[c]Department of Economics, Federal University of Espírito Santo, Brazil

ARTICLE HISTORY
Compiled January 2, 2019

ABSTRACT
This paper deals with the estimation of the periodogram of time series models in the presence of missing data. It is known that the presence of missing data makes it impossible to apply the classical methodology to estimate the periodogram. The statistical methodology used in this article was introduced for the first time in Astrophysics and provides means to estimate the periodogram of time series with missing data, without using an imputation technique to replace the missing values: the Lomb-Scargle periodogram. Monte Carlo tests were performed to compare, in different scenarios, the root mean square errors (RMSE) of the proposed estimator with those of the traditional estimation method. The empirical study evidenced that the suggested estimation method presents good performance in terms of RMSE for different percentages of missing data. It is demonstrated by simulations that in the time series scenario with missing data, the standard methodology provides misleading results, while the proposed method is not affected. The methodology is applied to identify periodic components of a daily time series of inhalable particulate matter (PM_{10}) concentrations that has seasonal characteristics and occasional large peaks of pollutant concentrations.

KEYWORDS
Lomb-Scargle periodogram; missing data; PM_{10} pollutant

1. Introduction

The analysis of time series is a topic well discussed in the academic environment. There are basically two approaches used in the analysis of time series [26]. In both, the objective is to construct models for the series, with the following purposes: investigate its generative mechanism, predict its future values, describe its behavior, look for relevant periodicities in it, among others. In several areas of science, the first approach can be highlighted, its analysis is carried out in the time domain, especially the methodology proposed by the statisticians George Box and Gwilym Jenkins in the 70's. The second approach, which is less often applied in analyzing time series atmospheric pollutants, does it through the frequency domain.

Several time series modelling and prediction methods are available in the literature,

CONTACT Wanderson de Paula Pinto. Email: wandersondpp@gmail.com

such as moving averages (MA), linear regression with time, exponential smoothing of Holt-Winters and ARIMA (Autoregressive Integrated Moving Average) Models. However, a frequent problem in time series from air quality monitoring is the presence of missing data. These data generally occur because the equipment for measuring the concentrations of contaminants in the atmosphere may have defects that make it impossible to operate for some time, causing data loss. Data analysis, including only the available observations without a statistical treatment for the missing data, can produce a false estimate of the effect measure and underestimate its accuracy [18]. Therefore, several authors have been developing alternative methodologies for the analysis of time series with missing data. For example, [36], [20], [15], [28] and [27], use the imputation technique [25], [7] and [19] suggest the use of Expectation-Maximization (EM) algorithms to address this problem. However, this approach has the disadvantage of assuming a specific (usually normal) distribution for the data.

The fact that time series present many cycles of different frequencies and amplitudes, that is, series with properties of non-stationarity and seasonality, is an opportunity to analyze them in the frequency domain. The use of time series decomposition via spectral analysis emerged as an alternative for identification of the frequency components of these cycles. The spectrum shows the variance decomposition of a data sample through different frequencies. Thus, the spectrum describes the cyclic properties of a given time series. It is assumed that the fluctuations of the underlying process are produced by a large number of elementary cycles of different frequencies, and that the contribution of each cycle is constant throughout the sample. The spectrum then gives the relative contribution made by these elementary cycles to the overall process variance.

Spectral data analysis has been gaining popularity, especially in the areas of economics and finance, as it can be seen in [40], [12], [5] and [11]. However, its application can be extended to other areas of knowledge in which spectral analysis utilization is still poorly reported, such as the air quality evaluation of a given region. In this context, the following works can be highlighted as applications examples of spectral analysis in atmospheric pollution data: [13] presented a method, using spectral analysis, to analyse time series of elemental carbon for different sources of atmospheric pollution in Berlin; [35] used the spectral analysis to investigate the tropospheric ozone formation and decomposition processes in Germany; and [21] estimated a generalized autoregressive conditional heteroscedasticity (GARCH) model associated with the FFT-ARIMA (fast Fourier transform-Autoregressive Integrated Moving Average), to predict ozone concentrations in two European cities, Brussels and London.

In the literature, spectral analysis of environmental time series usually have the drawback of requiring the evaluation of hourly, daily or monthly means of the data to investigate its periodicities. However, the fact that air pollution data usually have considerable percentages of the missing data which often causes irregular sampling intervals, may be the main reason that spectral analysis is rarely appended to this area. The fast Fourier transform (FFT) is a very efficient manner to calculate the discrete Fourier transform, but a limitation of the FFT algorithm and the common periodograms is that it requires equally spaced time series [30]. Another limitation of the FFT algorithm is that it does not tolerate missing values. In the analysis of time series with irregular sampling intervals or series with missing data, some methodology is applied to fill in the missing data before doing FFT spectral analysis. This, however, is not entirely satisfactory because the inclusion of artificial data modifies the statistical properties of a time series [23].

In this article, to overcome these limitations, the Lomb-Scargle periodogram, which

is a statistical methodology introduced for the first time in astrophysics will be used. This methodology is capable of estimating the periodogram of time series in the presence of missing or unequally spaced data without using an imputation technique to replace the missing values. When studying variables in Astronomy, [24] proposed a way to find periodicities in data not equally spaced. In an attempt to find an alternative to impute pseudo-data in sinusoidal models, [24] proposed to use least squares for sinusoidal curves. [32] extended Lomb's work by defining the Lomb-Scargle periodogram and deriving its null distribution. [29] proposed a practical mathematical formulation. It is noteworthy that few studies have applied the Lomb-Scargle periodogram to air pollution data [2, 6, 8, 14], and none of them evaluated the efficiency of this method for different percentages of missing data. In this context, the objective of this work is use the Lomb-Scargle periodogram to modify Fisher's test in a way that enables it to identify periodic components in time series with missing data. The modified version of Fisher's test will also be applied in the identification of periodic components in time series of atmospheric pollution data which present missing observations.

This paper is organized as follows: Section 2 presents the methodology for spectral analysis of atmospheric pollution time series with missing data; Section 3 presents simulation and a empirical studies; Section 4 deals with the analysis of PM_{10} concentrations; and Section 5 draws some conclusions.

2. Material and methods

2.1. Spectral density function

Let $\{Y_t\}_{t\in\mathbb{Z}}$ a stationary process with zero mean and autocovariance function, $\gamma(l) = \mathbb{E}[Y_t, Y_{t+l}]$, satisfying the asymptotic independence condition, in the sense that values of the process which are highly separated in time are poorly dependent, such condition can be expressed as

$$\sum_{l=-\infty}^{\infty} |\gamma(l)| < \infty. \tag{1}$$

Under these conditions the spectral density function Y_t is defined as the Fourier transform of $\gamma(l)$, that is,

$$f_Y(\omega) = \frac{1}{2\pi} \sum_{l=-\infty}^{\infty} \gamma(l) e^{-i\omega l}, -\pi \leq \omega \leq \pi, \tag{2}$$

with $e^{i\omega} = cos\omega + isen\omega$ and $i = \sqrt{-1}$. It is noted that, from the Equation (2), it follows that

$$\gamma(l) = \int_{-\pi}^{\pi} e^{i\omega l} f(\omega) \mathrm{d}\omega, l \in \mathbb{Z}, \tag{3}$$

that is, the sequence $\gamma(l)$ can be recovered from $f(\omega)$ using the inverse Fourier transform.

If we place $l = 0$ in the Equation (3), then

$$\gamma(0) = \mathrm{Var}(Y_t) = \int_{-\pi}^{\pi} f(\omega)\mathrm{d}\omega \tag{4}$$

which shows that the spectrum $f(\omega)$ may be interpreted as the decomposition of the variance of a process. The term $f(\omega)\mathrm{d}\omega$ is the contribution to the variance attributed to the component of the process with frequencies in the interval $(\omega, \omega + \mathrm{d}\omega)$. A peak in the spectrum indicates an important contribution to the variance from the components at frequencies in the corresponding interval [38]. A natural estimator of the spectral density function is the periodogram, defined in the Subsection 2.2.

2.2. The Periodogram

The periodogram is a statistical tool that, for more than a century, has been widely used in the various scientific areas in which it is fundamental to find the periodic components of the phenomena under study. This tool was introduced by [34] with the aim of identifying periodic components in time series. For each frequency ω, the ordinate $I(\omega)$ estimates the contribution of that frequency to the series in study. Thus, in a superficial way, larger values of the periodogram ordinates imply in more significant frequencies in the series and smaller values, correspond to less significant frequencies. Usually, the frequencies of interest are those that produce "peaks" in the periodogram, especially when working with air quality data. In the following paragraph, the mathematical concept of periodogram is formalized.

Let $\{Y_t\}_{t\in\mathbb{Z}}$ be a stationary process with zero mean. For a given time series $\{Y_1 \cdots, Y_N\}$, the classical periodogram at Fourier frequencies, $\omega_j = \frac{2\pi j}{N}, j = 1, \cdots, [N/2]$, is defined by

$$I_N(\omega_j) = \frac{1}{2\pi N}\left|\sum_{t=1}^{N} Y_t e^{-i\omega t}\right|^2. \tag{5}$$

There are several statistical methodologies in the literature for periodicity research in time series. However, in the vast majority, it is assumed as an initial hypothesis that the data are regularly spaced and statistically homogeneous. The periodogram, as described in Equation (5), presents a series of problems when applied in time series not equally spaced or with missing data [32, 33]. According to [6], the main one is the fact that sines and cosines are no more orthogonal on the set of unequally spaced observations. This invalidates a derivation of Equation (5), which is based on the orthogonality of sines and cosines on the set of natural frequencies [1]. The solution is to define a time offset to make the sine and cosine orthogonal on the set of unequally spaced observations for an arbitrary frequency [6].

This change results in the application of the Lomb-Scargle periodogram. [24], while studying astronomical data, sought a way to find periodicities in not equally spaced data. [32] continued Lomb's work by defining the Lomb-Scargle periodogram.

2.3. The Lomb-Scargle Periodogram

The Lomb method for power spectral density estimation is based on the minimization of the squared differences between the projection of the signal onto the basis function and the signal under study [24]. This method can be generalized to any transform estimation on unevenly sampled signals.

Let $\{Y_t\}_{t\in\mathbb{Z}}$ be a time series whose observations have been sampled at times $t_i (i = 1, \cdots, N)$. The Lomb-Scargle periodogram is constructed as follows.

Let,

$$\overline{Y} = N^{-1}\sum_{t=1}^{N}(Y_t), \tag{6}$$

and

$$\widehat{\sigma}^2 = (N-1)^{-1}\sum_{t=1}^{N}(Y_t - \overline{Y}), \tag{7}$$

be respectively, the mean and variance estimators of Y_t. The normalized Lomb-Scargle periodogram at the angular frequency ω_j is defined as [29]

$$P_Y(\omega) = (2\widehat{\sigma}^2)^{-1}\left\{\frac{R_j^2}{\sum_{t=1}^{N}\cos^2\omega_j(t_i-\tau)} + \frac{I_j^2}{\sum_{t=1}^{N}\sin^2\omega_j(t_i-\tau)}\right\}, \tag{8}$$

where, $\overline{Y}$ and $\hat{\sigma}^2$ are the mean and variance of the measurements, respectively, defined in Equations (6) and (7), while R_j and I_j are, respectively, the sums

$$R_j = \sum_{j=1}^{N}(Y_j - \overline{Y})\cos\omega(t_j-\tau), \tag{9}$$

and

$$I_j = \sum_{j=1}^{N}(Y_j - \overline{Y})\sin\omega(t_j-\tau). \tag{10}$$

The frequency-dependent time offset τ is evaluated at each ω_j via

$$\tau = \frac{1}{2\omega_j}\arctan\left[\frac{\sum_{t=1}^{N}\sin(2\omega_j t_i)}{\sum_{t=1}^{N}\cos(2\omega_j t_i)}\right]. \tag{11}$$

According to [32], if t_i becomes $t_i + T_0$, then τ becomes $\tau + T_0$ and hence the periodogram of Equation (8) also possesses the useful property of invariance to time

translation that its classical version owns. The Lomb-Scargle periodogram can be used for the analysis of equally spaced time series because its statistical properties are equivalent to the classical periodogram. But, for [32], the calculation of the periodogram by Equation (8) is a little more complicated than by the traditional methodology given by Equation (5), since there is the 2π ambiguity in the arctan function in Equation (11). For [32], the secret is to impose continuity on τ as a function of ω, and to use a sufficiently high frequency resolution to ensure that no phase jumps are missed. It is also important to note that

$$\lim_{\omega \to 0} \tau(\omega) = \left(\frac{1}{N}\right) \sum_{i=1}^{N} t_i = \langle t \rangle. \tag{12}$$

Equations (8) to (12) describe the Lomb-Scargle periodogram. These equations were defined according to the [24], [32], [33], [29], [22] [10], [6], [14], [37] and [23].

2.4. Test for periodic components

In practice, it is known that one of the objectives periodogram analysis is to look for "hidden periodicities". If the time series actually contains a single periodic component in the frequency ω, it is expected that the periodogram $I^N(\omega_j)$ reaches its maximum value at the Fourier frequency ω_j nearest to ω. Thus, we can search out the maximum periodogram ordinate and test whether this ordinate can be reasonably considered as the maximum in a random sample of $N/2$ random variables [38].

To test for the periodicity formally, some kind of a test statistic must be chosen. An exact test for $\max\{I_N(\omega_j)\}$ was derived by [9], based on the following statistic

$$T = \frac{\max\{I_N(\omega_j)\}}{\sum_{j=1}^{N/2} I_N(\omega_j)} = \frac{I_N^{(1)}(\omega_{(1)})}{\sum_{j=1}^{N/2} I_N(\omega_j)}, \tag{13}$$

where $I_N^{(1)}(\omega_{(1)})$ is the largest periodogram ordinate at Fourier frequency $\omega_{(1)}$. Under the null hypothesis that $Y_t \sim N(0, \sigma^2)$, [9] showed that

$$P(T > g) = \sum_{j=1}^{m} (-1)^{(j-1)} \binom{n}{j} (1 - jg)^{n-1}, \tag{14}$$

where $n = N/2, g > 0$, and m is the largest integer less than $1/g$. Thus, for any given significance level α, we can use Equation (14) to find the critical value g_α such that $P(T > g_\alpha) = \alpha$. If the T value calculated from the series is larger than g_α, then we reject the null hypothesis and conclude that the series Y_t contains a periodic component. This test procedure is known as Fisher's test.

A good approximation of Equation (14) is obtained using only the first term of the expansion, that is,

$$P(T > g) \simeq n(1 - g)^{n-1}. \tag{15}$$

[39] suggested extending Fisher's test for this second largest ordinate based on the test statistic

$$T_2 = \frac{I_N^{(2)}(\omega_2)}{\{\sum_{j=1}^{N/2} I_N(\omega_j)\} - I_N^{(1)}(\omega_1)}, \tag{16}$$

where $I_N^{(2)}(\omega_{(2)})$ is the second largest periodogram ordinate at Fourier frequency $\omega_{(2)}$ and the distribution in Equation (14) is taken as the distribution of T_2 with N replaced by $(N-1)$. The procedure can be continued until an insignificant result is obtained [38].

In this work, the Lomb-Scargle periodogram is used to make a modification in Fisher's test, with the objective of applying it to identify periodic components in time series with missing data. Thus, the statistic of Fisher's modified test can be obtained by the equation

$$T_{Modified} = \frac{\max\{P_Y(\omega_j)\}}{\sum_{j=1}^{N/2} P_Y(\omega_j)}. \tag{17}$$

3. Monte Carlo simulation study

The empirical behavior of the Lomb-Scargle periodogram was analyzed using Monte Carlo simulations. The $\{Y_t\}_{t\in\mathbb{Z}}$ time series was generated from a SARMA (Seasonal Autoregressive Moving Average) model. A process $\{Y_t\}_{t\in\mathbb{Z}}$ is defined as a zero-mean SARMA$(p,q)\times(P,Q)_s$ model with non-seasonal orders p and q, seasonal orders P and Q, and seasonal period $s \in \mathbb{N}^* = \mathbb{N} - \{0\}$ if it satisfies

$$\Phi(B^s)\phi(B)Y_t = \Theta(B^s)\theta(B)\epsilon_t, \tag{18}$$

where $\{\epsilon_t\}_{t\in\mathbb{Z}}$ is a white noise process with $\mathbb{E}(\epsilon_t) = 0$ and $\mathrm{Var}(\epsilon_t) = \sigma_\epsilon^2$, and B is the backshift operator which satisfies $BY_t = Y_{t-1}$ for any process $\{Y_t\}_{t\in\mathbb{Z}}$.

If we let z be any complex number, then, in Equation (18), $\Phi(\cdot), \Theta(\cdot), \phi(\cdot)$ and $\theta(\cdot)$ are the Psth, Qsth, pth and qth-degree polynomials

$$\Phi(z^s) = 1 - \Phi_1 z^s - \Phi_2 z^{2s} - \cdots - \phi_P z^{Ps},$$

$$\Theta(z^s) = 1 - \Theta_1 z^s - \Theta_2 z^{2s} - \cdots - \theta_Q z^{Qs},$$

$$\phi(z) = 1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p,$$

$$\theta(z) = 1 - \theta_1 z - \theta_2 z^2 - \cdots - \theta_q z^q.$$

It is assumed that these polynomials have no common zeros and satisfy the conditions $\Phi(z^s)\phi(z) \neq 0$ and $\Theta(z^s)\theta(z) \neq 0$ for $|z| = 1$. Furthermore, in the above equations, $(\Phi_i)_{1\leq i\leq P}$, $(\Theta_j)_{1\leq j\leq Q}$, $(\phi_k)_{1\leq k\leq p}$ and $(\theta_l)_{1\leq l\leq q}$. For further details, see [4], [38] and [3], among others.

It was performed 1000 replications with sample sizes $N = 100$ and 500. The generated data comes from a SARMA$(1,0)(1,0)_7$ process with $\phi = 0.3$, $\Phi = 0.9$ and $\epsilon_t \sim N(0,1)$. The generated series were investigated with percentages of missing data equal to 5%, 10%, 15%, 20%, 30% and 40%. The simulations were performed in the programming language R [31]. The empirical $RMSE$ corresponds to the mean over all values of $RMSE_j$ for the Fourier frequencies

$$RMSE = \frac{1}{m}\sum_{j=1}^{m} RMSE_j, \tag{19}$$

where

$$RMSE_j = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(f(\omega_j) - \tilde{f}^{(k)}(\omega_j)\right)^2}, \tag{20}$$

and $\tilde{f}^{(k)}(\omega_j)$ is the estimated periodogram of replication k $(k = 1, \cdots, K)$, $f(\cdot)$ is the spectral density of the SARMA$(p,q)(P,D)$ process, $K = 1000$ and $m = [N/2 - 1]$.

Table 1 shows the results of the statistical accuracy measures evaluated under different percentages of missing data. These results were obtained by comparing the theoretical spectrum, i.e., the classical periodogram obtained by Equation (5) with 0% of missing data, with the spectrum estimated by the Lomb-Scargle periodogram (Equations (8) to (12)) for percentages of missing data equal to 5%, 10%, 15%, 20%, 30% and 40%. For series with missing data, it was verified that the Lomb-Sacargle periodogram presents lower values for both RMSE and Bias than the classical one, regardless of the sample size and the percentage of missing data. In series with zero percent of missing data it was found that the Lomb-Scargle periodogram shows a slight tendency to suppress the theoretical value of the spectrum.

The performance of the estimator was also evaluated from the perspective of identifying periodic components in the time series with missing data. To perform this identification the modified Fisher test presented in the previous section was used. Since a SARMA$(1,0)(1,0)_7$ process was simulated, the periodogram should be dominated by a very large peak in the frequency $\omega_0 = 0.1433$. This frequency corresponds to a seasonal period of $s = 1/0.1433 \approx 7$, which indicates that the data displays an approximate cycle of seven days.

In addition, Table 1 presents the results for the modified Fisher test with a significance level of $\alpha = 0.05$. For $N = 100$, if we use the first term approximation given in Equation (15), then $n(1-g_{0.05})^{n-1} = 50(1-g_{0.05})^{49} = 0.05$, implying $g_{0.05} = 0.13149$. If a sample size of $N = 500$ is chosen and the approximation of Equation (15) is adopted, we have $n(1-g_{0.05})^{n-1} = 250(1-g_{0.05})^{249} = 0.05$ which gives $g_{0.05} = 0.03363$. Since in both cases $T_{Modified} > g_{0.05}$, the result is highly significant and we conclude that the series contains a periodic component in the frequency $\omega_0 = 0.1433$. It is noteworthy that the modified Fisher test was able to capture the periodic component indepen-

dent of the percentage of missing data tested. Thus, through Monte Carlo experiments it was shown that the methodology tested in this work allows the estimation of the periodogram of a time series with missing data with certain statistical accuracy.

Table 1. Precision measurements of estimations of the classical (evaluated with 0% of missing data) and the Lomb-Scargle (calculated for 5%, 10%, 15%, 20%, 30% and 40% of missing data) periodograms for SARMA$(1,0)(1,0)_7$, with $\phi = 0.3$ and $\Phi = 0.9$ processes.

N		Percentage of missing data						
		0%	5%	10%	15%	20%	30%	40%
100	$\widehat{Mean}$	48.96520	48.63572	48.11165	45.07622	39.63127	34.46129	34.70265
	$\widehat{RMSE}$	2.02053	2.00681	1.96315	1.93141	1.73006	1.69661	1.70754
	$\widehat{Bias}$	0.11352	0.09705	0.06587	0.07327	0.11359	0.10900	0.11667
	$\widehat{Peak}$	15.66587	15.16224	14.37756	14.02002	12.70213	10.95739	9.51175
	$T_{Modified}$	0.31966†	0.31139	0.29828	0.30895	0.31552	0.28872	0.27256
	Result	Periodic	Periodic	Periodic	Periodic	Periodic	Periodic	Periodic
500	$\widehat{Mean}$	248.9218	247.9154	245.0968	217.1088	205.7863	196.2442	181.181
	$\widehat{RMSE}$	2.49311	2.44588	2.3524	2.15522	2.06967	2.05553	2.10212
	$\widehat{Bias}$	0.33642	0.31252	0.31471	0.35715	0.34378	0.38697	0.4312
	$\widehat{Peak}$	41.81718	39.72839	37.02617	35.65959	33.60837	29.30517	25.31138
	$T_{Modified}$	0.16799	0.16024†	0.15103	0.16344	0.16250	0.14851	0.13886
	Result	Periodic	Periodic	Periodic	Periodic	Periodic	Periodic	Periodic

†These values refer to the T statistic.

4. Application

4.1. Study area

To apply the studied methodology, this paper used data collected in the Region of Greater Vitória (RGV), Espírito Santo, Brazil. RGV is located on the south coast of the Atlantic Ocean (latitude 20 °19 ′S, longitude 40 °20 ′W). The RGV is constituted by the municipalities of Vitória, Vila Velha, Cariacica, Serra and Viana. Because it is situated in the coastal region, the RGV has a warm tropical climate (Aw), with mild and dry winter, and a hot and rainy summer, with averange temperatures ranging from 24 °C to 30 °C. According to [16], the metropolitan region of Vitória has 1,475,332 inhabitants, covers an area of 1,461 square kilometres and is one of the main centers of urban and industrial development in the state. The region suffers from several environmental problems, among them the deterioration of air quality due to atmospheric emissions by industries and the vehicular fleet.

The RGV has an automatic network of air quality monitoring (RAMQAr), owned by the State Institute for the Environment and Water Resources (IEMA), that was put in service in July, 2000. The network is distributed among eight monitoring stations located in the municipalities that compose the RGV, as follows: Serra, with two stations (Carapina and Laranjeiras); Vitória, with three stations (Jardim Camburi, Enseada do Suá, and Central Vitória); Vila Velha, with two stations (Ibes and Central Vila Velha); and Cariacica (Cariacica). The locations of the RAMQAR monitoring stations are shown in Figure 1.

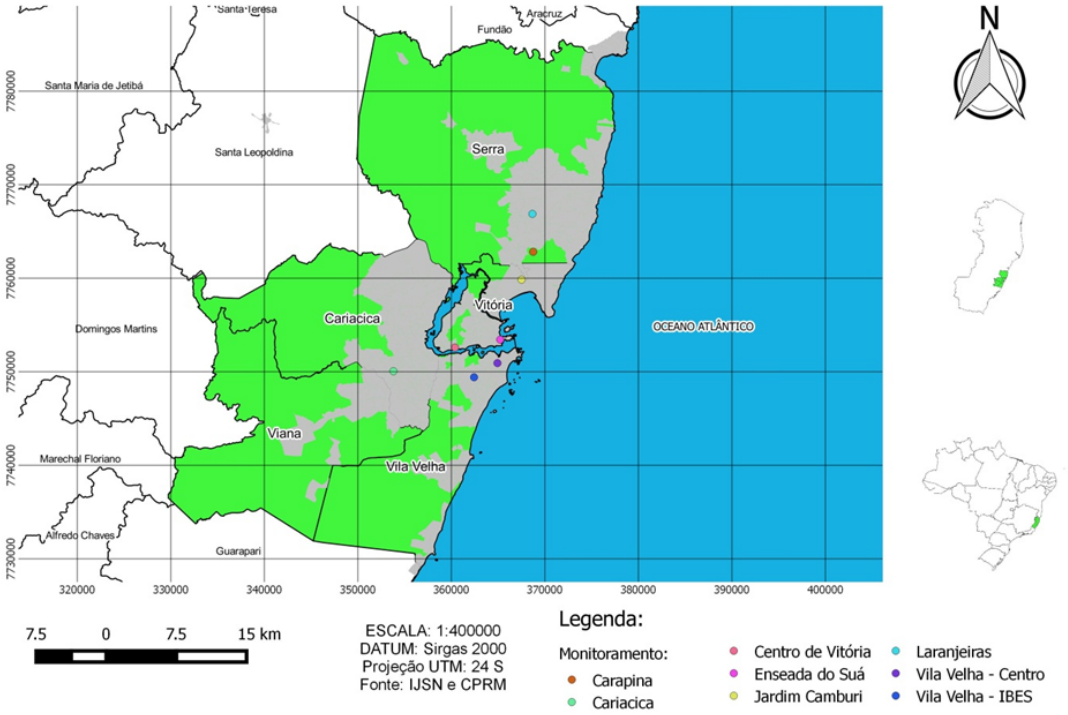


Figure 1. Location of the air quality monitoring stations in the RGV.

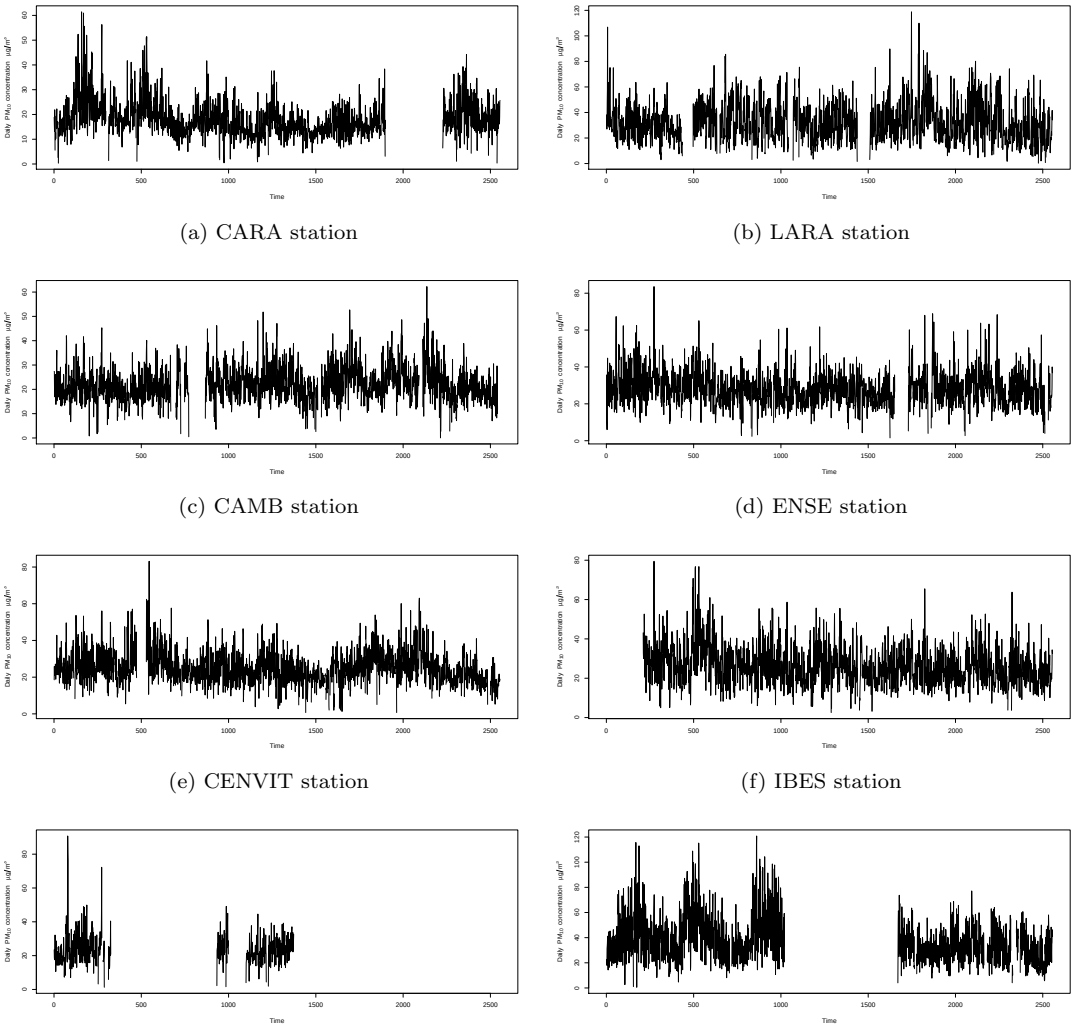
4.2. An application to PM_{10} pollutant

As previously mentioned, the daily average PM_{10} concentrations (Figure 2) is the data set here analyzed to illustrate the methodology previously discussed. The series is expressed in $\mu g/m^3$ and it was observed in Carapina (CARA), Laranjeiras (LARA), Jardim Camburi (CAMB), Enseada do Suá (ENSE), Central Vitória (CENVIT), Ibes (IBES), Central Vila Velha (CENVV) and Cariacica (CARI), all of them located in Espírito Santo, Brazil. These time series were analysed in the period from January 1st, 2010 to December 31st, 2016.

PM_{10} series is expected to have a periodic component with have seasonal period $s = 7$, since according to [17], the main source emitting particles in the RGV is automotive vehicles (more than 60% of particulate emissions is connected to the resuspension of particles in roads). Thus, there is a variation between the concentrations measured on weekdays and the days of weekend. The flow of vehicles is greater during the days of the week. The data may also have other significant periodic components which are usually related to annual deterministic trends in the series, these components will not be studied in this paper.

Table 2 presents the descriptive statistics of the inhalable particulate matter time series with their number and percentage of missing data. It is important to observe the high percentages of missing data that these series have justify the use of Lomb-Scargle periodogram. In general, by observing the variances, the coefficient of variation and the interquartile distances, it can be seen that the data under study presented a great variability in statistical terms, which is corroborated by Figure 3. It is noteworthy that, for the calculation of descriptive statistics, only the observed data was considered.

It is worth mentioning that the concentrations monitored at the stations of Cara-



(a) CARA station

(b) LARA station

(c) CAMB station

(d) ENSE station

(e) CENVIT station

(f) IBES station

(g) CENVV station

(h) CARI station

Figure 2. Daily PM_{10} concentrations series with missing data.

Table 2. Descriptive statistics of the time series of PM_{10} monitored in RGV.

	CARA	LARA	CAMB	ENSE	CENVIT	IBES	CENVV	CARI
N	2555.00	2555.00	2555.00	2555.00	2555.00	2555.00	2555.00	2555.00
NAs	362.00	207.00	210.00	140.00	121.00	227.00	1938.00	756.00
NAs/N (%)	14.17	8.10	8.22	5.48	4.74	8.88	75.85	29.59
Minimum	0.33	0.29	0.04	1.54	0.71	2.54	1.29	0.54
Maximum	61.42	118.79	62.25	83.58	83.12	79.46	90.75	120.83
1. Quartile	13.54	21.25	17.62	22.21	19.50	19.29	18.50	26.94
3. Quartile	21.21	41.27	26.08	33.29	29.75	32.04	28.58	48.31
Mean	18.10	32.30	22.15	28.24	25.12	26.27	24.20	39.16
Median	16.88	30.46	21.50	27.38	24.00	25.06	23.04	37.00
Variance	49.21	215.48	48.66	81.23	74.35	94.21	89.85	314.11
Stdev	7.02	14.68	6.98	9.01	8.62	9.71	9.48	17.72
Skewness	1.43	0.80	0.60	0.82	0.84	0.89	2.08	0.95
Kurtosis	4.19	1.21	1.39	2.25	1.99	1.93	11.12	1.39
CV (%)	38.78	45.45	31.51	31.91	34.32	36.96	39.17	45.25

CV = Coefficient of variation; N = Sample size; NAs = Number of missing data.

pina, Laranjeiras, Jardim Camburi, Enseada do Suá, Central Vitória, Ibes, Central Vila Velha and Cariacica exceeded the limit value of $50\mu g/m^3$ on 12, 380, 10, 84, 56, 76, 10 and 512 occasions, respectively. These results meet the guidelines established by the World Health Organization (WHO) for this pollutant. Despite the fact that the guidelines of WHO where met by all monitoring stations, special mention should be made to Cariacica and Laranjeiras ones, in which the WHO limit was exceeded in, respectively, 20.04% and 14.87% of the total days. Therefore, the importance of this research is justified, especially with regard to the formulation of preventive measures by competent authorities, since the concentration of PM_{10} in the study region is frequently reaching levels that are harmful to health.

Figure 4 shows the Lomb-Scargle periodograms of the PM_{10} time series observed in the RGV. The graphs represent the relative contributions of different frequencies to the variance. The spectra obtained for the PM_{10} time series clearly identified peaks at frequencies close to 0.143 which corresponds to a seven day period for all monitoring stations. It would not be possible to correctly identify the time series with missing data if the classical periodogram was used instead of the Lomb-Scargle one. It is worth noting that the periodogram of Central Vila Velha station data was not estimated due to its very large percentage of missing data.

The results shown in Figure 4 are quantified in Table 3. It can be observed that the periodograms show that in all stations the highest peak is associated to the frequency 0.1429, implying $s = 1/0.1429 = 6.97$, that is, a seasonal component of seven days length. According to the results presented in Table 3 for the modified Fisher test, since $T_{Modified} > g_{0.05}$, the null hypothesis was rejected, confirming the existence of the periodic component for periods of seven days at the nominal level of significance of 5%.

Table 3. Tests for hidden periodic components in the PM_{10} time series monitored in RGV.

	Peak	ω_0	$T_{Modified}$	$g_{0.05}$	Result	s
CARA	27.30675	0.14291	0.02594	0.00792	Periodic	7
LARA	33.89355	0.14291	0.02703	0.00792	Periodic	7
CAMB	51.30669	0.14291	0.04175	0.00792	Periodic	7
ENSE	52.61087	0.14291	0.04196	0.00792	Periodic	7
CENVIT	61.83433	0.14291	0.05252	0.00792	Periodic	7
IBES	37.27977	0.14292	0.03185	0.00792	Periodic	7
CARI	91.67547	0.14291	0.08840	0.00792	Periodic	7

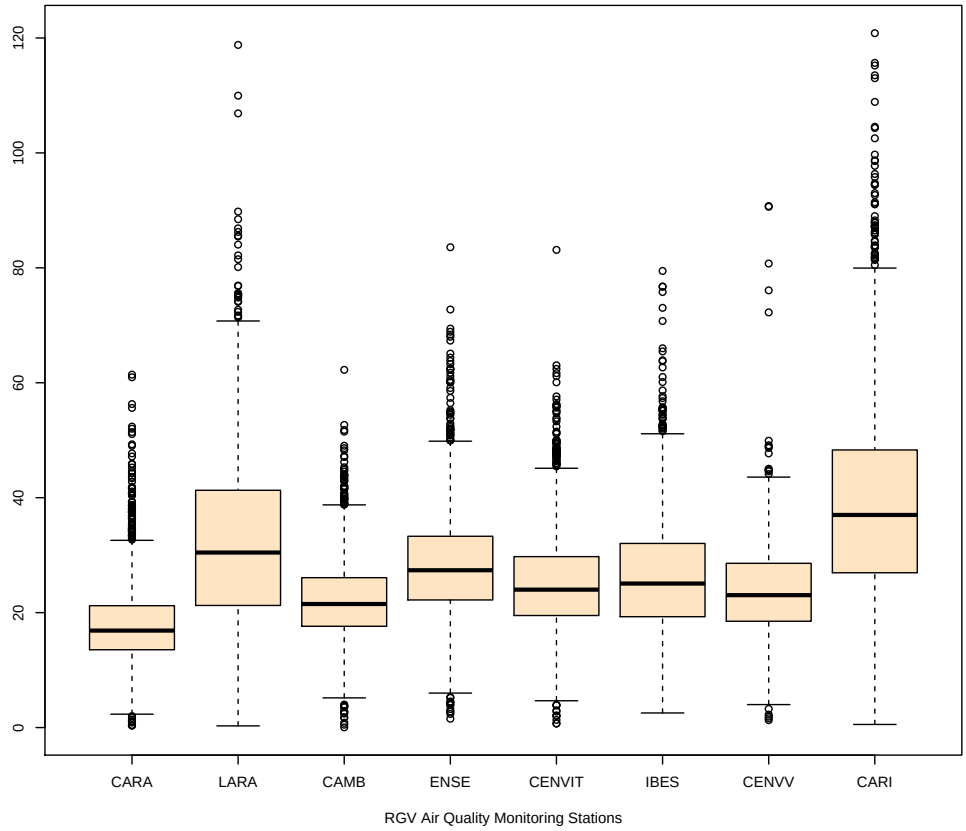
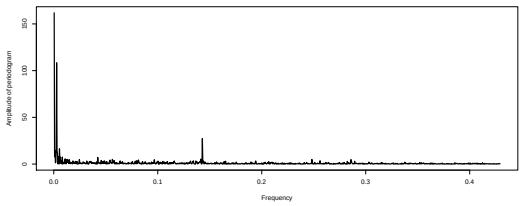
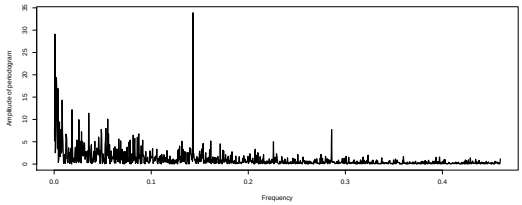


Figure 3. Location of the air quality monitoring stations in the RGV.

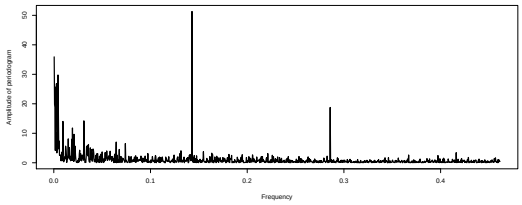
It is important to say that the mathematical modeling of time series of atmospheric pollution plays an essential role in the scientific understanding of atmospheric phenomena and in the provision of political strategies to deal with relevant problems such as climate change, air quality and, in general, with degradation of the ecosystem. In mathematical modeling the model accuracy must be evaluated in relation to the observations to ensure that the errors in model formulations are minimal and that the model limitations are understood. In this context, it is essential to search for models which are able to deal with characteristics of the data that may make its analysis more difficult to perform. This section is an application to air pollution time series with missing data which, as it was discussed throughout this paper, is a characteristic that causes several problems in frequency analysis. In this work, the Lomb-Scargle periodogram was used to make a modification in Fisher's test, to identify periodic components in time series with missing data. In this context, modified version of Fisher's test (which uses the Lomb-Scargle periodogram) was proposed and applied to incomplete PM_{10} time series to identify a hidden periodic component with seasonal period $s = 7$. In mathematical modelling, searching for periodic components is a fundamental step in defining the model that will be adopted.



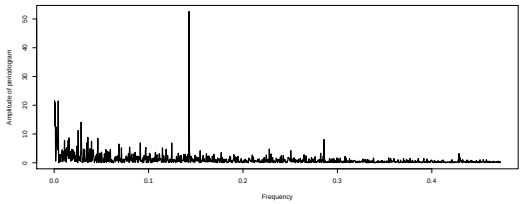
(a) CARA station



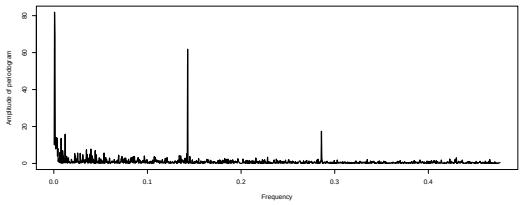
(b) LARA station



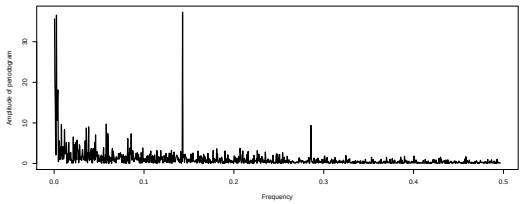
(c) CAMB station



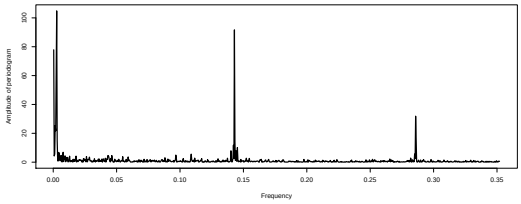
(d) ENSE station



(e) CENVIT station



(f) IBES station



(g) CARI station

Figure 4. Periodogram of daily PM_{10} concentrations series with missing data.

5. Conclusions

The objective of this work was use the Lomb-Scargle periodogram to modify Fisher's test in a way that enables it to identify periodic components in time series with missing data. The behavior of the Lomb-Scargle periodogram estimator was compared with the theoretical spectral density estimator using Monte Carlo simulations and accurate result was obtained.

By means of simulations it was shown that the modified Fisher test, which uses the Lomb-Scargle periodogram, can be used to find periodic components in time series with missing data. The modified version of the Fisher test was applied to analyse the mean daily concentrations of monitored PM_{10} in RGV. It should be noted that it was possible to identify an expected periodic component of the series even under extreme conditions of missing data. The results presented in this work suggest that the methodology studied is an alternative for the spectral analysis of time series of concentrations of atmospheric pollutants with missing data and can be applied to data sets with different percentages of lost data.

Finally, the Lomb-Scargle periodogram and the modified version of Fisher test proved to be useful tools to identify significant periodicities in time series without the need to fill in missing observations or formulate previous hypotheses about anticipated patterns.

Acknowledgements

The authors gratefully acknowledge partial financial support from FAPES/ES, CAPES/Brazil and CNPq/Brazil.

References

[1] P. Bloomfield, *Fourier analysis of time series: an introduction*, John Wiley & Sons, 2004.

[2] D.R. Bowdalo, M.J. Evans, and E.D. Sofen, *Spectral analysis of atmospheric composition: application to surface ozone model–measurement comparisons*, Atmospheric Chemistry and Physics 16 (2016), pp. 8295–8308.

[3] G. Box, G. Jenkins, and G. Reinsel, *Time Series Analysis: Forecasting and Control*, 4th ed., Prentice Hall, 2008.

[4] P. Brockwell and R. Davis, *Introduction to Time Series and Forecasting*, 2nd ed., Springer Verlag, 2002.

[5] Z. Chen and Y. Yang, *Time series models for forecasting: Testing or combining*, Studies in Nonlinear Dynamics & Econometrics 11 (2007), pp. 56–90.

[6] S. Dilmaghani, I.C. Henry, P. Soonthornnonda, E.R. Christensen, and R.C. Henry, *Harmonic analysis of environmental time series with missing data or irregular sample spacing*, Environmental science & technology 41 (2007), pp. 7030–7038.

[7] C. Drake, O. Knapik, and J. Leśkow, *EM-based inference for cyclostationary time series with missing observations*, in *Cyclostationarity: Theory and Methods*, Springer, 2014, pp. 23–35.

[8] S.J. Dutton, B. Rajagopalan, S. Vedal, and M.P. Hannigan, *Temporal patterns in daily measurements of inorganic and organic speciated PM2.5 in Denver*, Atmospheric Environment 44 (2010), pp. 987–998.

[9] R.A. Fisher, *Tests of significance in harmonic analysis*, Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character 125 (1929), pp. 54–59.

[10] E.F. Glynn, J. Chen, and A.R. Mushegian, *Detecting periodic patterns in unevenly spaced gene expression time series using Lomb–Scargle periodograms*, Bioinformatics 22 (2006), pp. 310–316.

[11] C.W.J. Granger and M. Hatanaka, *Spectral Analysis of Economic Time Series.(PSME-1)*, Princeton university press, 2015.

[12] A.H. Hallett and C.R. Richter, *Spectral analysis as a tool for financial policy: An analysis of the short-end of the british term structure*, Computational Economics 23 (2004), pp. 271–288.

[13] T. Hies, R. Treffeisen, L. Sebald, and E. Reimer, *Spectral analysis of air pollutants. Part 1: elemental carbon time series*, Atmospheric Environment 34 (2000), pp. 3495–3502.

[14] K. Hocke and N. Kämpfer, *Gap filling and noise reduction of unevenly sampled data by means of the Lomb-Scargle periodogram*, Atmospheric chemistry and physics 9 (2009), pp. 4197–4206.

[15] P. Iglesias, H. Jorquera, and W. Palma, *Data analysis using regression models with missing observations and long memory: an application study*, Computational statistics and Data Analysis 50 (2005), pp. 2028–2043.

[16] Instituto Brasileiro de Geografia e Estatística, *Banco de dados. Cidades*, Rio de Janeiro (2014). Available at `http://www.cidades.ibge.gov.br/xtras/home.php`.

[17] Instituto Estadual de Meio Ambiente e Recursos Hídricos do Estado do Espírito Santo, *Relatório da qualidade do ar da Região da Grande Vitória*, Vitória (2014). Available at `http://www.meioambiente.es.gov.br/download/Relat%C3%B3rio_Anual_de_Qualidade_do_Ar_2013.pdf`.

[18] W.L. Junger, *Análise, imputação de dados e interfaces computacionais em estudos de séries temporais epidemiológicas. PhD Dissertation*, Programa de Pós-graduação em Saúde Coletiva, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, Brazil., 2008.

[19] W.L. Junger and A.P. Leon, *Imputation of missing data in time series for air pollutants*, Atmospheric Environment 102 (2015), pp. 96–104.

[20] H. Junninen, H. Niskaa, K. Tuppurainenc, J. Ruuskanena, and M. Koleh-Mainen, *Methods for imputation of missing values in air quality data sets*, Atmospheric Environment 38 (2004), pp. 2895–2907.

[21] U. Kumar and K. De Ridder, *Garch modelling in association with FFT–ARIMA to forecast ozone episodes*, Atmospheric Environment 44 (2010), pp. 4252–4265.

[22] P. Laguna, G.B. Moody, and R.G. Mark, *Power spectral density of unevenly sampled data by least-square analysis: performance and application to heart rate signals*, IEEE Transactions on Biomedical Engineering 45 (1998), pp. 698–715.

[23] B. Leroy, *Fast calculation of the Lomb-Scargle periodogram using nonequispaced fast fourier transforms*, Astronomy & Astrophysics 545 (2012), p. A50.

[24] N.R. Lomb, *Least-squares frequency analysis of unequally spaced data*, Astrophysics and space science 39 (1976), pp. 447–462.

[25] K. Metaxoglou and A. Smith, *Maximum Likelihood Estimation of VARMA Models Using a State-Space EM Algorithm*, Journal of Time Series Analysis 28 (2007), pp. 666–685.

[26] P. Morettin and C. Toloi, *Análise de Séries Temporais*, ABE - Projeto Fisher, 2006.

[27] M.N. Norazian, Y.A. Shukri, R.N. Azam, and A.M.M. Al Bakri, *Estimation of missing values in air pollution data using single imputation techniques*, ScienceAsia 34 (2008), pp. 341–345.

[28] A. Plaia and A.L. Bondì, *Single imputation method of missing values in environmental pollution data sets*, Atmospheric Environment 40 (2006), pp. 7316–7330.

[29] W.H. Press and G.B. Rybicki, *Fast algorithm for spectral analysis of unevenly sampled data*, The Astrophysical Journal 338 (1989), pp. 277–280.

[30] M.B. Priestley, *Spectral Analysis and Time Series*, Academic Press, 1981.

[31] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria (2017). Available at `https://www.R-project.org/`.

[32] J.D. Scargle, *Studies in astronomical time series analysis. ii-statistical aspects of spectral*

analysis of unevenly spaced data, The Astrophysical Journal 263 (1982), pp. 835–853.

[33] J.D. Scargle, *Studies in astronomical time series analysis. iii-fourier transforms, autocorrelation functions, and cross-correlation functions of unevenly spaced data*, The Astrophysical Journal 343 (1989), pp. 874–887.

[34] A. Schuster, *On the investigation of hidden periodicities with application to a supposed 26 day period of meteorological phenomena*, Journal of Geophysical Research 3 (1898), pp. 13–41.

[35] L. Sebald, R. Treffeisen, E. Reimer, and T. Hies, *Spectral analysis of air pollutants. Part 2: ozone time series*, Atmospheric Environment 34 (2000), pp. 3503–3509.

[36] H.Y. Toda and C. Makenzie, *LM tests for unit roots in the presence of missing observations: small sample evidence*, Mathematics and computers in simulation 48 (1999), pp. 457–468.

[37] R. Townsend, *Fast calculation of the lomb-scargle periodogram using graphics processing units*, The Astrophysical Journal Supplement Series 191 (2010), p. 247.

[38] W. Wei, *Time Series Analysis: Univariate and Multivariate Methods*, Addison Wesley, 2006.

[39] P. Whittle, *The simultaneous estimation of a time series harmonic components and covariance structure*, Trabajos de estadística 3 (1952), pp. 43–57.

[40] P. Wilson and J. Okunev, *Spectral analysis of real estate and financial assets markets*, Journal of Property Investment & Finance 17 (1999), pp. 61–74.

engenharia sanitária e ambiental

Análise estatística das concentrações de poluentes atmosféricos na Região da Grande Vitória, ES, Brasil, no período de 2008 a 2017

Journal:	*Engenharia Sanitária e Ambiental*
Manuscript ID	ESA-2019-0102
Manuscript Type:	Technical Article
Keyword:	Testes de comparação, Séries temporais, Poluição atmosférica, Região da Grande Vitória

Análise estatística das concentrações de poluentes atmosféricos na Região da Grande Vitória, ES, Brasil, no período de 2008 a 2017

Statistical analysis of atmospheric pollutants concentrations in the Region of Grande Vitória, ES, Brazil, from 2008 to 2017

RESUMO

Este estudo teve como objetivo avaliar, estatisticamente, os dados de séries temporais de PM_{10} e PTS na RGV, entre 2008 e 2017, verificando se as séries de cada poluente monitoradas em diferentes estações são geradas por um mesmo processo estocástico. Para tanto, utilizaram-se os testes propostos por Coates e Diggle (1986), por Quenouille (1958) e o procedimento de diferença de séries desenvolvido por Silva *et al*. (2000). Compararam-se, duas a duas, as séries das estações Laranjeiras (E1), Carapina (E2), Jardim Camburi (E3), Enseada do Suá (E4), Vitória (Centro) (E5), IBES (E6) e Cariacica (E8), para o poluente PM_{10}, e as séries das estações E3, E4, E5, E6 e E8, para o PTS. Os resultados indicam que, para um nível de significância de 5%, as estações E2, E3, E4, E5 e E6 para o PM_{10} e, E3, E4, E5 e E6, para o PTS, apresentam séries temporais geradas pelo mesmo processo estocástico. Considera-se assim, que os resultados obtidos apresentam-se como um indicativo da necessidade de reformulação do projeto inicial da RAMQAr que, se somados a um estudo de dispersão de contaminantes, podem garantir a ampliação da área de cobertura da rede, com destaque para a reespacialização das estações já existentes, visando melhorar sua representatividade de dados e instalação de novas estações em locais ainda desprovidos de monitoramento.

Palavras-chave: Testes de comparação. Séries temporais. Poluição atmosférica. Região da Grande Vitória.

ABSTRACT

This study aimed to evaluate, statistically, the data of PM_{10} and TSP time series in the RGV, between 2008 and 2017, in order to verify if the series of each pollutant monitored in different stations are generated by the same stochastic process. Therefore, was used the tests proposed by Coates and Diggle (1986), by Quenouille (1958) and the series difference

procedure developed by Silva et al. (2000). Were compared, two to two, the series of the Laranjeiras (E1), Carapina (E2), Jardim Camburi (E3), Enseada do Suá (E4), Vitória (E5), IBES (E6) and Cariacica (E8) stations, for the pollutant PM_{10}, and the series of E3, E4, E5, E6 and E8, stations for TSP. The results indicate that, for a significance level of 5%, stations E2, E3, E4, E5 and E6 for the PM_{10} and, E3, E4, E5 and E6, for the TSP, present time series generated by the same stochastic process. It is, therefore, considered that, the results obtained are an indicative of the need of reformulation for the initial project of RAMQAr, which if, added to a study of contaminants dispersion, can guarantee the expansion of the area of coverage of the network, with emphasis on the re-spatialisation of existing stations, aiming at improve their data representativeness and the installation of new stations in locations still lacking in monitoring.

Keywords: Comparison tests. Time series. Atmospheric pollution. Region of Grande Vitória.

INTRODUÇÃO

Até o final do século XIX, o ar necessário para a respiração de todos os seres vivos da Terra ainda não era abordado de forma tão evidente, pois julgava-se que este nunca deixaria de estar disponível na forma adequada à manutenção da vida (RUSSO, 2010). Entretanto, alguns episódios de elevadas concentrações de poluentes na atmosfera, ocorridos no século XX foram responsáveis por impulsionar a formulação de padrões de qualidade do ar e estudos na área de epidemiologia, relacionados aos efeitos da poluição atmosférica na saúde, tornando a qualidade do ar uma das maiores preocupações da humanidade.

O primeiro desses episódios ocorreu na Bélgica, em 1930, no vale do rio Meuse. Foi caracterizado por uma inversão térmica que resultou no acúmulo das concentrações de poluentes emitidos pelas indústrias locais e, ao longo de três dias, culminou em, aproximadamente, seis mil pessoas com problemas respiratórios e 60 mortes. (JUN, 2009). Anos mais tarde, em 1948, ocorreu um episódio semelhante ao da Bélgica na cidade de Donora, (Estados Unidos da América), que resultou em 20 óbitos e na metade da população local hospitalizada (JACOBS , BURGESS e ABBOTT, 2018).

O terceiro, e mais conhecido entre os grandes desastres da poluição atmosférica, ocorreu na capital britânica, Londres, no ano de 1952. O episódio foi marcado por uma intensa inversão térmica com duração de quatro dias, que acarretou a morte de quatro mil pessoas por doenças cardíacas e respiratórias, principalmente bronquite e pneumonia. O aumento do índice de

mortalidade afetou pessoas de todas as idades, entretanto, foi maior entre as crianças, com destaque para a mortalidade de recém-nascidos, que quase duplicou, e a de crianças de 1 a 12 meses, a qual mais do que dobrou e, entre os adultos com mais de 45 anos (LOGAN, 1952).

Entre os poluentes atmosféricos, o material particulado (PM) apresenta grande relevância devido à sua complexidade, em termos de composição química e propriedades físicas, visto que, este abrange uma grande classe de poluentes constituídos por partículas primárias e secundárias (ARAUJO e NEL, 2009) e devido ao tamanho da partícula, classificada em diferentes frações de tamanhos, conforme seu diâmetro aerodinâmico: PTS (partículas totais em suspensão, cujo diâmetro aerodinâmico é menor ou igual a 50 μm e maior que 10 μm), PM_{10} (partículas grossas/inaláveis, com diâmetro aerodinâmico menor que 10 μm e maior que 2,5 μm), $PM_{2,5}$ (partículas finas/respiráveis, com diâmetro aerodinâmico menor que 2,5 μm e maior que 0,1 μm) e $PM_{0,1}$ (partículas ultrafinas, com diâmetro aerodinâmico inferior a 0,1 μm).

Conforme Chaloulakou *et al.* (2003), desde o início dos anos 1950, um grande número de estudos epidemiológicos vem evidenciando as consequências do material particulado e a necessidade de monitorar suas frações, principalmente o PM_{10} e o $PM_{2.5}$, pois as partículas menores possuem a capacidade de penetrar mais fundo no sistema respiratório, chegando assim, até aos alvéolos pulmonares, onde ocorre a troca de oxigênio e dióxido de carbono dentro do sistema cardiovascular podendo, com exposição repetida, se acumular e recobrir a superfície de troca dos sacos alveolares, tornando a respiração cada vez mais difícil e, uma vez passada através dos alvéolos, as partículas podem ser incorporadas à corrente sanguínea, afetando assim, muitos órgãos.

Dentre os estudos realizados, as principais consequências observadas comprovam a relação entre concentração de PM no ambiente e o aumento dos casos diários de mortalidade e morbidade, tornando a poluição por PM um crescente problema de saúde pública (VARDOULAKIS e KASSOMENOS, 2008). Os estudos mais comuns referem-se a morbidades respiratórias e cardiovasculares associadas à exposição ao PM, que podem ser observadas a curto e médio prazo, como por exemplo, o estudo conduzido por Freitas *et al.* (2016), na cidade de Vitória, Espírito Santo, no qual, utilizando modelos de séries temporais via Regressão de Poisson, os autores analisaram o impacto da poluição atmosférica na morbidade respiratória e cardiovascular de crianças e adultos, no período de 2001 a 2006, chegando a conclusão de que, para cada incremento de 10 $\mu g/m^3$ do poluente PM_{10}, há um

aumento do risco relativo percentual para as hospitalizações por doenças respiratórias totais de 9,67 e de 6,60 para doenças respiratórias em menores de cinco anos.

Entretanto, estudos recentes realizados a longo prazo, tem comprovado a associação entre exposição à poluição por PM e maiores chances de uma criança ter Transtorno do Espectro do Autismo (RAZ *et al.*, 2015), impactos cognitivos, estruturais e metabólicos no desenvolvimento do cérebro de crianças e constituição de fator de risco para o desenvolvimento da doença de Alzheimer e esclerose múltipla ao decorrer da vida (CALDERÓN-GARCIDUEÑAS *et al.*, 2016), alterações na performance cognitiva humana (ZHANG, CHEN e ZHANG, 2018), constituição de fator de risco cerebrovascular e neurodegenerativo modificável visto que, o PM contribui para cerca de um terço da carga global de AVC e um quinto da carga global de demência (BÉJOT *et al.*, 2018) e o aumento do risco de incidência de diabetes melitus (BOWE *et al.*, 2018).

Diante do exposto, o monitoramento da qualidade do ar, através das chamadas redes de monitoramento, destaca-se como uma ferramenta de grande importância para o gerenciamento da qualidade do ar. No entanto, por motivos de caráter econômico e administrativo, o número de pontos de medida de uma rede é limitado e, acima de tudo, sua disposição espacial pode não ter sido estudada cuidadosamente, podendo estar posicionada em locais pouco representativos (MOREIRA, TIBASSI e MORAES, 2008).

Devido a isso, tem-se a possibilidade das séries temporais medidas pelas estações de uma rede de monitoramento da qualidade do ar serem geradas pelo mesmo processo estocástico, isto é, apresentarem padrões de comportamento semelhantes, do ponto de vista estatístico, para determinado poluente monitorado, conforme descrevem os estudos de Costa e Safádi (2010), analisando as séries temporais de PM_{10} da cidade de São Paulo, Zamprogno (2013), estudando as concentrações de PM_{10} e SO_2 na Região da Grande Vitória, ES, e Cotta (2016), avaliando as séries históricas de PM_{10}, também para a Região da Grande Vitória. Desta forma, caso seja comprovado que duas, ou mais estações, apresentam o mesmo padrão de comportamento para um poluente em comum, pode-se realocar os equipamentos de medições desse poluente para uma outra estação (COTTA, 2016), ampliando assim, a área de abrangência da Rede de Monitoramento, o que justifica o presente estudo.

Neste contexto, objetivou-se avaliar, estatisticamente, os dados de concentração de Partículas Inaláveis (PM_{10}) e de Partículas Totais em Suspensão (PTS), na Região da Grande Vitória (RGV), entre 2008 e 2017, visando detectar se as séries temporais de concentração dos poluentes são geradas por um mesmo processo estocástico.

METODOLOGIA

CARACTERIZAÇÃO DA ÁREA DE ESTUDO

A Região da Grande Vitória (RGV) situa-se no litoral do estado do Espírito Santo e é composta pelos municípios de Vitória, Vila Velha, Cariacica, Serra e Viana. Possui clima Aw (tropical quente) cujas temperaturas variam entre 24º e 30ºC, de acordo com a classificação climática de Köppen (KÖPPEN, 1900). Apresenta uma área de 1.456 km², com cerca de 1.565.393 habitantes que representam 44,5% da população total do estado do Espírito Santo, sendo que, 98,6% dessa população vive em área urbana (IBGE, 2010).

A RGV conta com uma Rede Automática de Monitoramento da Qualidade do Ar (RAMQAr), de propriedade do Instituto Estadual de Meio Ambiente e Recursos Hídricos (IEMA), composta por nove estações distribuídas por quatro municípios da RGV, cuja localização espacial está representada na Figura 1.

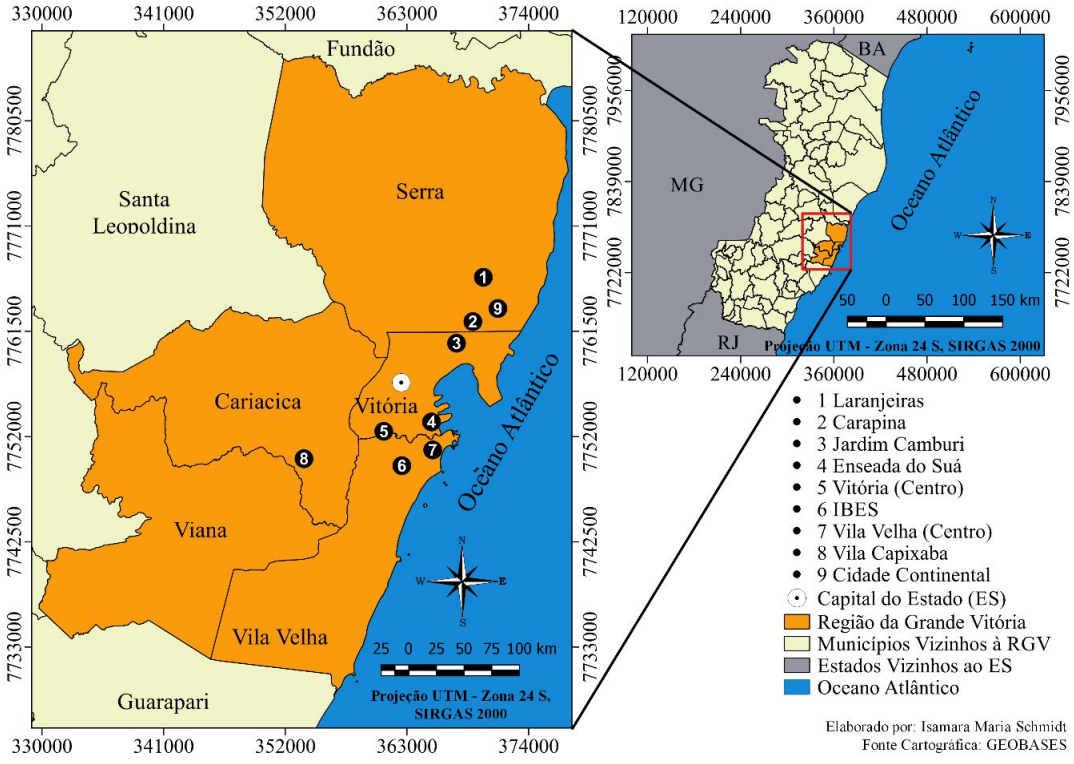


Figura 1 - Localização das estações componentes da RAMQAr na RGV, Espírito Santo.

Na Tabela 1, são apresentados os códigos de identificação das estações junto ao IEMA, seus respectivos códigos de identificação neste estudo (ID), a localização das estações de acordo com os bairros em que estão alocadas, o ano de início da operação de cada estação e suas coordenadas planas (UTM).

Tabela 1 - Caracterização das estações da RAMQAr.

Código da Estação	ID	Bairro	Início da Operação	Coordenada (m)	
				Leste	Norte
RAMQAr 1	E1	Laranjeiras	2000	369917	7766305
RAMQAr 2	E2	Carapina	2000	368945	7762315
RAMQAr 3	E3	Jardim Camburi	2000	367429	7760371
RAMQAr 4	E4	Enseada do Suá	2000	365266	7753279
RAMQAr 5	E5	Vitória (Centro)	2005	360857	7752450
RAMQAr 6	E6	IBES	2000	362532	7749346
RAMQAr 7	E7	Centro (Vila Velha)	2000	365354	7750721
RAMQAr 8	E8	Vila Capixaba	2000	353697	7749998
RAMQAr 9	E9	Cidade Continental	2011	371218	7763588

Fonte: adaptado de IEMA, 2018.

DADOS

As análises foram realizadas para o período de 01 de janeiro de 2008 a 31 de dezembro de 2017, sendo os dados referentes a concentração de PM_{10} e PTS, para a RGV, fornecidos em médias horárias de 24 horas e coletados através do banco de dados do IEMA (2018). Inicialmente, foi realizada a análise dos dados brutos para identificar dados faltantes e imputá-los via algoritimo EM (*expectation-maximisation*), através da plataforma *mstdi*, implementada por Junger (2008), no *software* R (R CORE TEAM, 2018). A porcentagem de dados faltantes aceita para que a série de determinada estação pudesse ser utilizada foi de 35%. Após a imputação dos dados faltantes foram calculadas as médias diárias de concentração dos poluentes, sendo consideradas, assim, para o período analisado, 3.653 observações de concentração de PM_{10} nas estações E1, E2, E3, E4, E5, E6 e E8, e 3653 observações de concentração de PTS nas estações E3, E4, E5, E6 e E8.

ANÁLISE ESTATÍSTICA

Para análise das concentrações dos poluentes estudados, foram utilizadas metodologias estatísticas na forma de métodos descritivos, para resumo e organização dos dados em análise, e inferenciais, para estimação e verificação de hipóteses. Toda análise estatística foi realizada a um nível de 5% de significância através do *software* livre R 3.4.4 (R CORE TEAM, 2018).

A característica de estacionariedade de uma série temporal relaciona-se ao desenvolvimento da série em torno de uma média (MORETTIN e TOLOI, 2006; WEI, 2006). A não-estacionariedade de uma série temporal, fato bastante comum, decorre da presença de fatores intrínsecos ao fenômeno estudado que são descritos nos modelos teóricos como componentes sazonais, tendências e heterogeneidade de variâncias (AMARAL, 2014). Desta forma, foram aplicados testes para verificação de tendência e sazonalidade para avaliar a necessidade de tratamento estatístico dos dados antes de sua análise. Para a detecção de tendências nas séries de concentração de PM_{10} e PTS, foram utilizados os testes de Mann-Kendall (MANN, 1945; KENDALL,1975) e de Cox-Stuart (WEI, 2006).

A avaliação da existência de periodicidade, também denominada sazonalidade, nas séries sob estudo foi realizada a partir da aplicação do teste G de Fisher (WEI, 2006). Tendo em vista que este teste baseia-se na avaliação da periodicidade da séries por meio da verificação de periodogramas, faz-se necessária a apresentação de alguns conceitos referentes à análise espectral para melhor compreensão.

Considere $\{Z(t), t \in T\}$ um processo estocástico que obedece uma condição de independência assintótica da forma

$$\sum_{\tau=-\infty}^{\infty} |\gamma(\tau)| < \infty. \quad (1)$$

A transformada de Fourier de $\gamma(\tau)$, também chamada de função de densidade espectral $f(\lambda)$, é dada por

$$f(\lambda) = \frac{1}{2\pi} \sum_{\tau=-\infty}^{\infty} \gamma(\tau) e^{-i\lambda\tau}, \ -\infty \leq \lambda \leq \infty, \quad (2)$$

em que $e^{-i\lambda} = \cos\lambda + i\sin\lambda$, com $i = \sqrt{-1}$ e λ é igual ao valor de frequência que será utilizado para verificação de um período a ser analisado.

A função $f(\lambda)$ é periódica com período 2π, e par, sendo sua representação dada no intervalo $[0, \pi]$. Visto que o interesse é verificar períodos em séries temporais discretas e finitas, a transformada de Fourier finita de um processo estacionário com média zero, $\{Z_t, t = 1, ..., N\}$, é definida pela Equação 3.

$$d^{(N)}(\lambda) = \frac{1}{\sqrt{2\pi N}}\sum_{t=1}^{N} Z_t e^{-i\lambda\tau}, \ -\infty < \lambda < \infty. \quad (3)$$

Se $\lambda_j = \frac{2\pi j}{N}$ e $-\left[\frac{N-1}{2}\right] \leq j \leq \left[\frac{N}{2}\right]$, em que j é o índice de cada observação de uma trajetória Z_t qualquer, obtem-se a transformada de Fourier discreta (Equação 4).

$$d_j^{(N)}(\lambda) = \frac{1}{\sqrt{2\pi N}}\sum_{t=1}^{N} Z_t e^{\left(\frac{-i2\pi jt}{N}\right)} \rightarrow$$

$$d_j^{(N)}(\lambda) = \frac{1}{\sqrt{2\pi N}}\sum_{t=1}^{N} Z_t \cos\left(\frac{2\pi jt}{N}\right) + i\frac{1}{\sqrt{2\pi N}}\sum_{t=1}^{N} Z_t \sin\left(\frac{2\pi jt}{N}\right), j = 0, 1, ..., \left[\frac{N}{2}\right]. \quad (4)$$

Para uma realização do processo estacionário $\{Z_t, t = 1, ..., N\}$, o objetivo é encontrar um estimador para $f(\lambda_j)$. Sob a suposição de independência assintótica estabelecida pela Equação 1, tem-se que

$$I_j^{(N)}(\lambda) = \left|d_j^{(N)}(\lambda)\right|^2 = \frac{1}{2\pi N}\left|\sum_{t=1}^{N} Z_t e^{-i\lambda_j t}\right|^2, \quad (5)$$

é denominado periodograma, um estimador não viciado de $f(\lambda_j)$ cuja distribuição assintótica é definida pelo Teorema 1, conforme Morettin e Toloi (2006).

Teorema 1. As ordenadas do periodograma $I_j^{(N)}(\lambda)$ são variáveis aleatórias assintóticamente independentes com distribuição assintótica múltipla de uma variável aleatória com distribuição qui-quadrado, isto é,

$$I_j^{(N)}(\lambda) \xrightarrow{D} \begin{cases} \frac{1}{2} f(\lambda_j)\chi_2^2, \ j \neq 0, \frac{N}{2}, \\ f(\lambda_j)\chi_1^2, \ j = 0, \frac{N}{2}. \end{cases}$$

Dado o periodograma, conforme Equação 5, utiliza-se o teste G de Fisher para verificar o período sazonal da série temporal em estudo, testando as hipóteses:

H_0: Não há sazonalidade, para todo $I_j^{(N)}(\lambda)$;

H_1: Há sazonalidade para algum $I_j^{(N)}(\lambda)$.

A estatística do teste G de Fisher é dada pela Equação 6:

$$G = \frac{\max\left[I_j^{(N)}(\lambda)\right]}{\sum_{j=1}^{\left(\frac{N}{2}\right)} I_j^{(N)}(\lambda)}, \quad (6)$$

em que: $I_j^{(N)}(\lambda)$ = valor do periodograma na ordenada j; e, N = número de observações da série.

A distribuição exata de G é dada por $Z_\alpha = 1 - \left(\frac{\alpha}{c}\right)^{\frac{1}{c-1}}$, sendo $c = \left(\frac{N}{2}\right)$ e α o nível de significância adotado. Se $G > Z$, a hipótese H_0 é rejeitada e a série apresenta periodicidade no período i (WEI, 2006). A determinação do período sazonal (s) consiste em verificar a qual frequência está associado o maior valor $I_j^{(N)}(\lambda)$ e, então, dividir 1 por esse valor de frequência, isto é,

$$s = \frac{1}{\lambda} \quad (7)$$

Para detalhes, consultar Morettin e Toloi (2006), Wei (2006) e Brockwell and Davis (2006).

Um problema recorrente em áreas práticas do conhecimento que fazem o uso de séries temporias é que a utilização da teoria sobre as mesmas nem sempre é suficiente para determinados estudos (AMARAL, 2014). Assim, a comparação dessas séries é de grande interesse em estudos sobre séries temporais coletadas em locais próximos. Para efetuar-se esta análise, aplicaram-se os testes propostos por Coates e Diggle (1986), por Quenouille (1958), sendo estes, testes univariados para comparação das funções de densidade espectral e de autocorrelação, respectivamente, e o procedimento de diferença de séries proposto por Silva *et al*. (2000).

Teste das Somas Acumuladas (COATES e DIGGLE, 1986)

Considere $I_j^{(N)}(\lambda)$ o periodograma da série $\{Z_i(t); t = 1, ..., N\}$, com $i = 1, 2$. Assintoticamente, $I_j^{(N)}(\lambda) \sim f_j(\lambda)\chi_2^2/2$, $\lambda \not\equiv 0, \pi$, conforme o Teorema 1. Assim, definem-se as razões espectrais

$$J(\lambda) = \frac{I_1^{(N)}(\lambda)}{I_2^{(N)}(\lambda)} \quad \text{e} \quad U(\lambda) = \frac{f_1(\lambda)}{f_2(\lambda)}, \; 0 < \lambda < \pi,$$

que, para $Z_1(t)$ e $Z_2(t)$ independentes, tem-se

$$J(\lambda) = \frac{I_1^{(N)}(\lambda)}{I_2^{(N)}(\lambda)} \sim U(\lambda) F_{2,2}, \quad (8)$$

na qual, F é distribuição de Fisher-Snedecor.

Conforme metodologia de Coates e Diggle (1986), é possível demonstrar que

$$z_i = \ln\left(1 + J^{-1}(\lambda_j)\right) \sim U(\lambda_j) \exp(1), \quad (9)$$

em que $\lambda_j = \frac{2\pi j}{N}$ $(j = 1, ..., m)$, no qual $m = \left[\frac{N-1}{2}\right]$ e $\exp(1)$ é a distribuição exponencial de média 1.

Sendo $f_1(\lambda)$ e $f_2(\lambda)$ as funções de densidade espectral das séries temporais $Z_1(t)$ e $Z_2(t)$, respectivamente, serão testadas as hipóteses:

$$H_0: f_1(\lambda) = f_2(\lambda) \text{ para todo } 0 < \lambda < \pi,$$

$$H_1: f_1(\lambda) \neq f_2(\lambda) \text{ para todo } 0 < \lambda < \pi.$$

Considerando a hipótese H_0 do teste de Coates e Diggle (1986), tem-se que $U(\lambda_j) = 1$ e $\exp(1)$ é a distribuição exponencial a qual pertencem, assintoticamente, as amostras aleatórias z_i. Desta forma, $c_j = \sum_{i=1}^{j} z_i$ compõem os pontos de um processo de Poisson e, por conseguinte, a razão $o_j = \frac{c_j}{c_m}$, $j = 1, ..., m$, é o vetor das estatísticas de ordem da distribuição uniforme [0,1].

Este teste proposto por Coates e Diggle (1986), consiste basicamente em construir as estatísticas o_j e usar o teste de Kolmogorov-Smirnov para avaliar afastamentos da distribuição $U(0,1)$. Se o p-valor for maior que α, não rejeita-se a hipótese H_0, ao nível de significância α. O teste apresentado requer que as duas séries a serem testadas apresentem o mesmo tamanho (N).

Teste de Igualdade das Funções de Autocorrelação (QUENOUILLE,1958)

O teste de Igualdade das Funções de Autocorrelação, proposto por Quenouille (1958), tem a finalidade de verificar se duas séries temporais distintas apresentam a mesma estrutura de correlação. Para tanto, considere $\rho_1(j)$ e $\rho_2(j)$ as funções de autocorrelação das séries $Z_1(t)$ e $Z_2(t)$, respectivamente. Serão testadas as hipóteses:

$$H_0: \rho_1(j) = \rho_2(j) \text{ para todo } j = \pm 1, \pm 2, ...,$$

$$H_1: \rho_1(j) \neq \rho_2(j) \text{ para algum } j = \pm 1, \pm 2, ...$$

O procedimento para aplicação do método proposto por Quenouille (1958) consiste na seguinte metodologia, conforme descrito por Toloi e Echeverry (2000):

1. Obter as funções de autocorrelação $\hat{\rho}_1(j)$ e $\hat{\rho}_2(j)$, com $j = 0,1,..., J$, das séries $Z_1(t)$ e $Z_2(t)$, respectivamente;

2. Calcular a função de autocorrelação comum às duas séries, utilizando a Equação 10, em que n_1 e n_2 são o número de observações das séries $Z_1(t)$ e $Z_2(t)$, respectivamente;

$$\hat{\rho}(j) = \frac{n_1\hat{\rho}_1(j) + n_2\hat{\rho}_2(j)}{n_1 + n_2}. \quad (10)$$

3. Calcular a função de autocorrelação parcial comum estimada $(\hat{\Phi}(k))$, a partir da função de autocorrelação comum $(\hat{\rho}(j))$;

4. Identificar a ordem autorregressiva, p, utilizando $\hat{\Phi}(k)$;

5. Estimar os p coeficientes do modelo autorregressivo resolvendo as equações de Yule-Walker. Para detalhes, consultar Costa (2010);

6. Ajustar às séries $Z_1(t)$ e $Z_2(t)$ o modelo autorregressivo com os coeficientes obtidos em 5, encontrando assim as séries residuais $\hat{a}_1$ e $\hat{a}_2$;

7. Calcular as funções de autocorrelação parcial υ_j e υ_j' para as séries residuais $\hat{a}_1$ e $\hat{a}_2$, respectivamente;

8. Testar se $\frac{\upsilon_j - \upsilon_j'}{\sqrt{\frac{1}{n_{1-j}} + \frac{1}{n_{2-j}}}}$ apresenta distribuição, aproximadamente, N(0,1), ou, de forma equivalente, testar se

$$SQ = \sum_{j=1}^{J} \frac{(\upsilon_j - \upsilon_j')^2}{\frac{1}{n_1 - j} + \frac{1}{n_2 - j}} \sim \chi_j^2. \quad (11)$$

9. Se $SQ > C_\alpha$, em que C_α é tal que $P(\chi_j^2 > C_\alpha) = \alpha$, rejeita-se H_0 ao nível de significância α.

Método para comparação de séries temporais (SILVA, FERREIRA e SÁFADI, 2000)

O procedimento para comparação das séries temporais $Z_1(t)$ e $Z_2(t)$, como proposto por Silva, Ferreira e Sáfadi (2000), consiste em efetuar a diferença entre as duas séries sob estudo, como demonstrado na Equação 12,

$$Z_d = Z_1(t) - Z_2(t),\ t = 1, 2, \ldots N, \quad (12)$$

resultando na série residual Z_d, a qual aplica-se o teste de Cox-Stuart para verificação de tendência, o teste G de Fisher para verificação de sazonalidade e o teste de Box & Pierce (BOX e PIERCE, 1970) para verificar se os resíduos são independentes e identicamente distribuídos, com média zero e variância constante.

Se Z_d for estacionária, isto é, não apresentar tendência nem sazonalidade, e os resíduos comportarem-se como ruído branco, conclui-se que as duas séries provém do mesmo processo estocástico, ou seja, são iguais no período analisado.

RESULTADOS E DISCUSSÃO

A Tabela 2 apresenta algumas estatísticas básicas dos poluentes em estudo e, de acordo com esta, para o PM_{10}, a maior média foi registrada na E8 (40,31 $\mu g.m^{-3}$), seguida pela E1 (32,73 $\mu g.m^{-3}$) e, para o PTS, o maior valor médio também foi encontrado na E8 (72,14 $\mu g.m^{-3}$). Todas as estações apresentaram alto desvio padrão e coeficiente de variação, sugerindo que a média dos dados é pouco representativa. Entre os valores máximos observados para o PM_{10}, todos ultrapassaram as diretrizes estabelecidas pela Organização Mundial de Saúde (WHO, 2005) para esse poluente e, o valor máximo registrado na E8 para o PM_{10} (120,83 $\mu g.m^{-3}$) e para o PTS (275,21 $\mu g.m^{-3}$) ultrapassaram o Padrão Intermediário 1 (PI1) determinado pela Resolução CONAMA 491/2018 (CONAMA, 2018) e a Meta Intermediária 1 (MI1) estabelecida pelo Decreto Estadual nº 3463 – R/2013 (ESPÍRITO SANTO, 2013), o que é alarmante, visto que, mesmo estes poluentes sendo encontrados em concentrações abaixo dos padrões permitidos já são observados danos à saúde humana. Os coeficientes de assimetria e curtose encontrados, inferem que as séries em estudo não pertencem a uma distribuição normal de probabilidades.

Tabela 2 – Medidas descritivas dos poluentes sob estudo.

Estações		Medidas Descritivas							
		μ	M	σ^2	CV	Máx.	Mín.	Assimetria	Curtose
PM_{10}	E1	32,73	31,04	13,17	40,23	118,79	5,97	1,00	1,93
	E2	20,35	19,12	7,28	35,81	88,25	4,42	1,83	6,87
	E3	23,45	22,61	7,43	31,67	66,88	3,54	0,95	2,00
	E4	28,15	27,21	8,74	31,06	83,58	8,83	0,95	2,14
	E5	25,48	21,42	8,08	31,70	83,12	6,79	1,03	2,21
	E6	26,93	25,96	9,54	35,43	88,13	5,00	0,95	2,08
	E8	40,31	38,34	15,22	37,75	120,83	5,50	1,19	2,63
PTS	E3	45,22	43,96	15,55	34,38	157,79	9,12	1,11	3,09
	E4	38,38	37,08	11,99	31,25	128,79	10,08	1,32	4,36
	E5	40,15	39,08	14,23	35,43	139,92	8,75	1,14	3,45

Análise estatística das concentrações de poluentes atmosféricos

	E6	43,32	41,21	16,58	38,27	146,04	6,45	0,99	1,83
	E8	72,14	67,32	31,31	43,41	275,21	3,63	1,30	3,26

Unidade de Medida: μ, M, σ^2, Máx., Mín. = ($\mu g.m^{-3}$); CV = %.

As Figuras 2 (a) e 2 (b) delineiam a variação espacial entre as concentrações de PM_{10} e PTS, respectivamente. Pode ser verificado que a E8 se destaca por apresentar maiores valores para ambos poluentes. Para o $PM_{10,}$ além da E8, a E1 apresenta expressiva diferença das demais estações. Em relação ao monitoramento do PTS, a E3 apresenta as menores concentrações monitoradas do poluente, e as estações E4, E5 e E6 apresentam valores similares entre si.

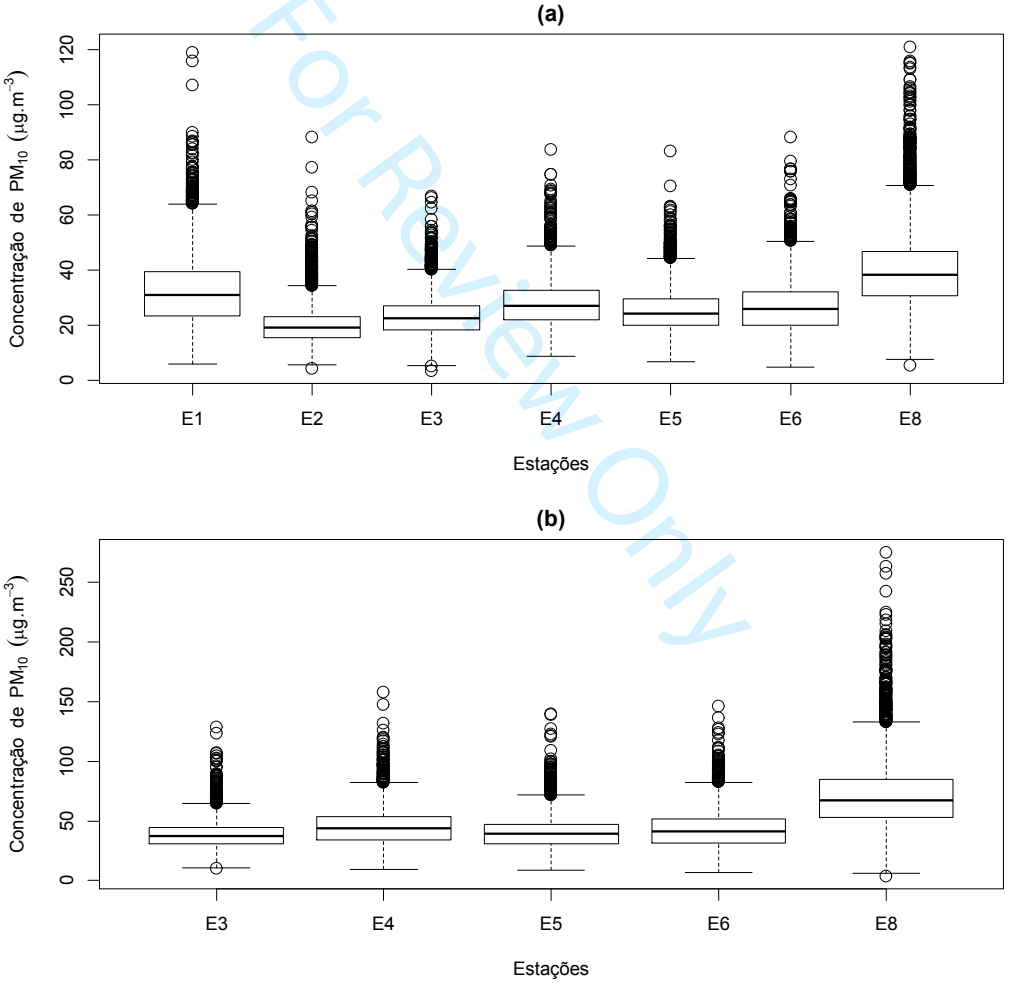
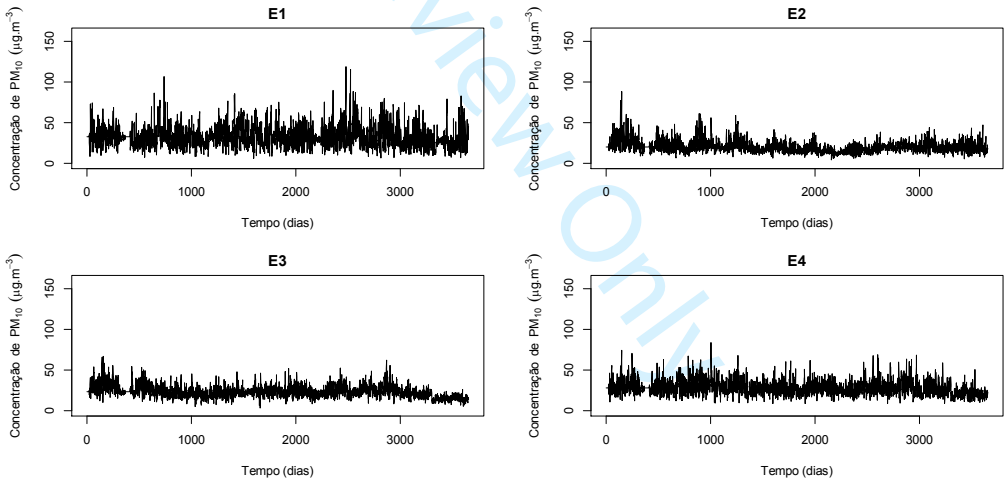


Figura 2 - Boxplot das séries temporais para cada estação observada em relação a distribuição espacial para os poluentes (a) PM_{10} e (b) PTS.

É possível visualizar que todas estações apresentaram *outliers*, indicativos da existência de níveis de concentração atípicos, acima do padrão médio do conjunto de dados, que ocorrem devido emissões excessivas de fontes de poluição, em curtos intervalos de tempo, ou por condições meteorológicas adversas. Os altos valores da média e, também, de máxima concentração de PM_{10} e PTS para a E8 estão associadas a localização desta na RGV, pois a mesma encontra-se próxima a Centrais de Abastecimento do Espírito Santo (CEASA/ES), onde há um intenso fluxo de veículos e, consequentemente, um maior nível de poluição devido a emissão oriunda do processo de combustão veicular e, também, a ressuspensão de poluentes proveniente do movimento destes.

As Figuras 3 e 4, apresentam a evolução temporal das séries de PM_{10} e PTS para as estações em estudo. Realizando uma análise visual, é possível verificar que o comportamento das séries de poluentes analisadas foi de estabilidade, apresentando, entretanto, uma leve tendência de decaímento, principalmente, nos últimos anos, hipótese confirmada pelos testes de Mann Kendell e Cox-Stuart.



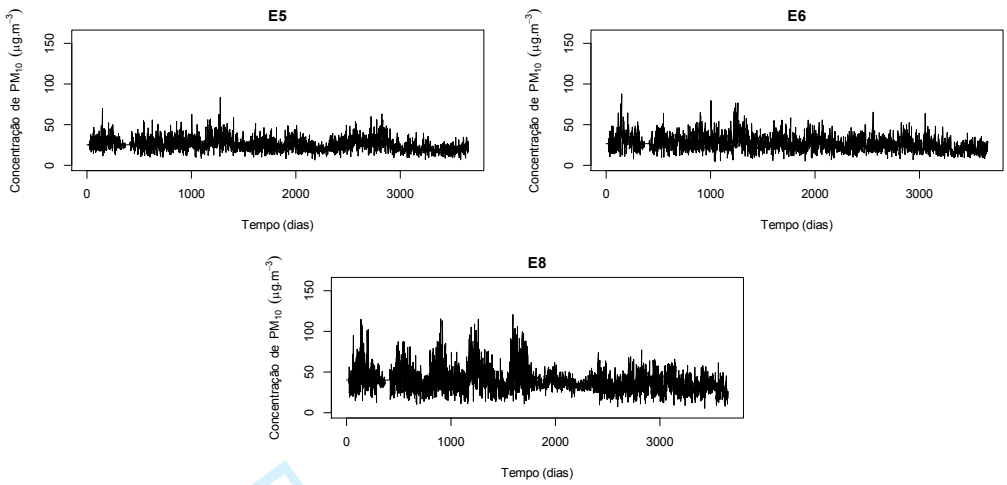


Figura 3 – Evolução temporal das séries de PM_{10} em cada estação da RAMQAr.

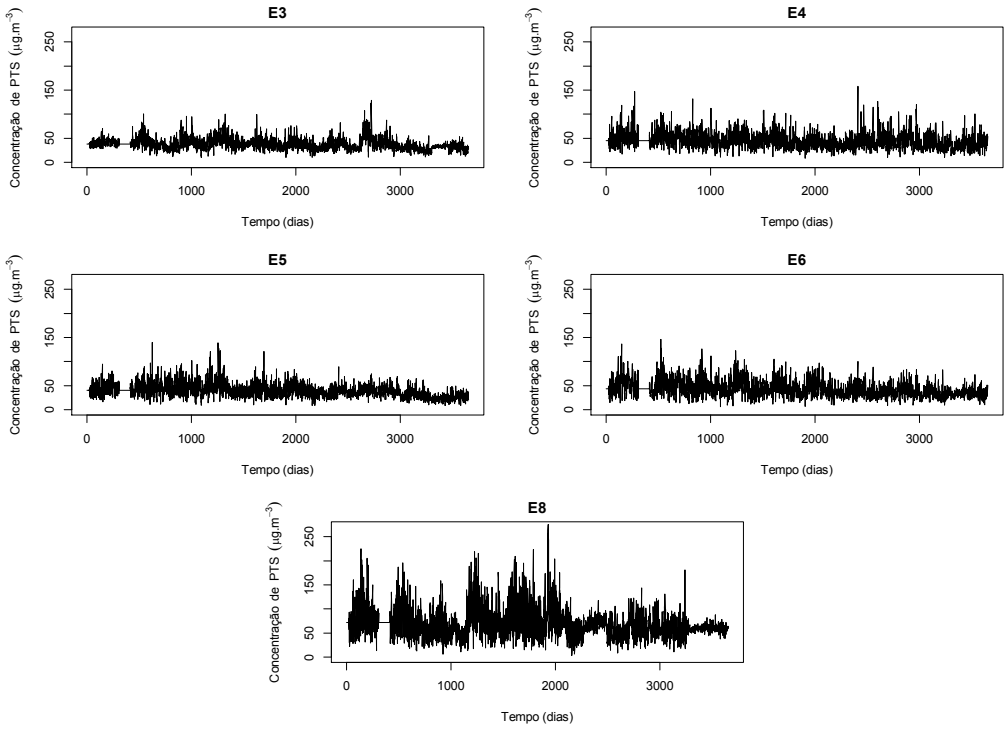


Figura 4 – Evolução temporal das séries de PTS em cada estação da RAMQAr.

Conforme os resultados dos testes de tendência de Mann-Kendal e Cox-Stuart, em todas as estações e, para ambos poluentes, existe tendência negativa ao nível de significância de 5% de probabilidade, ou seja, houve um decrescimento das concentrações ao longo do tempo. Tal tendência de redução nas concentrações de PM_{10} e PTS sugerem que, as medidas mais

restritivas impostas pelo Decreto 3463 – R/2013 (ESPÍRITO SANTO, 2013), influenciaram na redução das concentrações dos poluentes. Resultados semelhantes são encontrados nos estudos de Ramachandra e Shwetmala (2009), Nesamani (2010), Petro e Konečnýa (2017), Oliveira (2017) e Abe e Miraglia (2018), que evidenciam a relação entre a adoção de programas para controle da poluição do ar e a diminuição dos níveis de concentração dos principais poluentes atmosféricos.

Se faz importante também a análise da componente sazonalidade, afim de compreender o comportamento das concentrações de PM_{10} e PTS durante o tempo e possíveis fatores responsáveis pelas suas emissões. Normalmente, os dados de fenômenos naturais apresentam oscilações que se repetem em determinado período de tempo idêntico. Desse modo, para verificar a propriedade de sazonalidade nos dados foi realizada a decomposição espectral das séries de PM_{10} e de PTS, conforme mostram as Figuras 5 e 6.

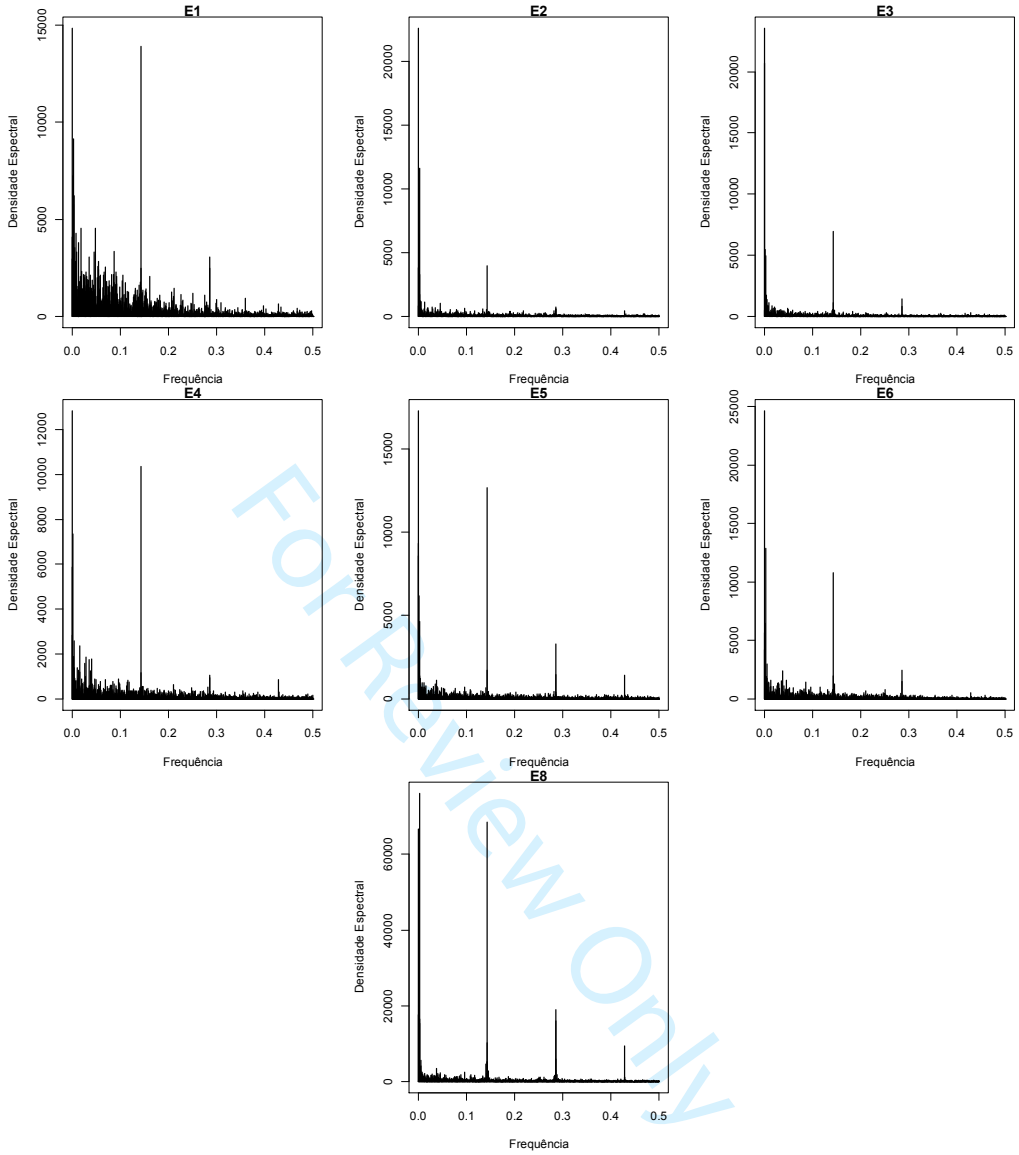


Figura 5 – Periodogramas das séries de PM_{10} para cada estação da RAMQAr.

O periodograma referente ao PM_{10} (Figura 5) demonstra que o maior pico está associado a frequência 0,14293 em todas as sete estações estudadas, o que implica em $s = \frac{1}{0{,}14293} = 6{,}99$, ou seja, uma componente sazonal com periodicidade de 7 dias. Conforme os resultados do teste G de Fisher, em todas as estações o valor de G é superior ao de Z, portanto, rejeita-se a hipótese H_0, confirmando a existência de sazonalidade para os períodos de 7 dias ao nível nominal de significância de 5%.

A análise da presença de sazonalidade nos dados do PTS pode ser feita através da Figura 6 e, assim como ocorre no periodograma do PM_{10}, as séries temporais apresentaram maior pico na frequência 0,14293, implicando num padrão de comportamento repetindo-se num período de 7 dias. A aplicação do teste G de Fisher comprovou a existência de sazonalidade para esse período em todas as sete estações sob estudo para o PTS.

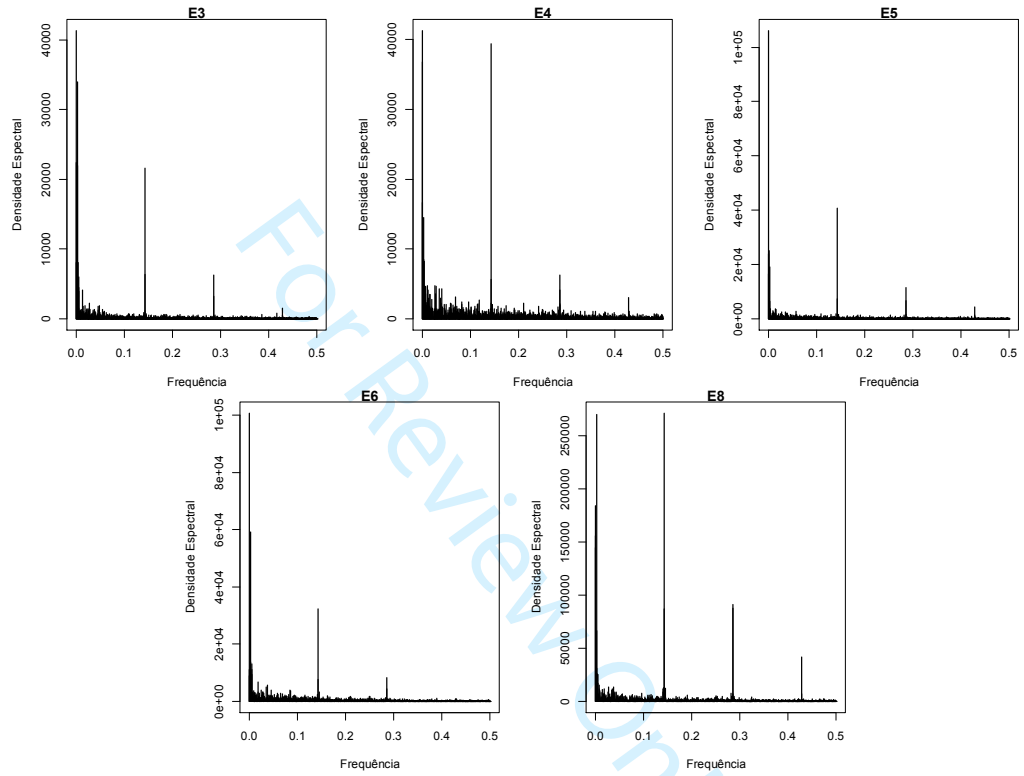


Figura 6 – Periodogramas das séries de PTS para cada estação da RAMQAr.

Esse comportamento apresentado por ambos poluentes é esperado pois, conforme o inventário de emissões atmosféricas da RGV, a principal fonte emissora de material particulado na região são os veículos automotores, representando aproximadamente 70% das emissões, decorrentes da ressuspensão de partículas em vias (ECOSOFT, 2011). Dessa forma, a variação sazonal apresentada pode ser explicada pelo ritmo de movimentação dos transportes, ou seja, as maiores concentrações de PM_{10} e PTS medidas, ocorrem durante os dias úteis, quando o fluxo de veículos é maior, e o oposto ocorre nos fins de semana, devido à redução da circulação veicular pelas ruas.

Os resultados encontrados também seguem as observações encontradas na região e em diferentes localidades do Brasil e do mundo, obtidos pelos estudos de Celis *et al.* (2007);

Stadlober, Horman e Pfeiler (2008), Leite *et al.* (2011), Reina e Olaya (2012), Wang *et al.* (2013), Liu *et al.* (2014) e Monte, Alburquerque e Reisen (2016). Os autores argumentam que o os dias úteis da semana proporcionam maiores concentrações de PM_{10} na atmosfera e, ainda, apontam que este fato está relacionado ao tráfego de veículos automotores, indicado pelos mesmos como a principal fonte de emissão do poluente.

As Tabelas 4 e 5 apresentam, respectivamente, os resultados dos testes de comparação de médias aplicados para todas as combinações entre as estações estudadas referente ao poluente PM_{10} e os resultados encontrados para o PTS. Para que as séries comparadas entre duas estações fossem consideradas provenientes de processos estocásticos diferentes, adotou-se como critério a condição de que, no mínimo, dois testes deviam apresentar o mesmo resultado, isto é, dentre os três testes aplicados, dois, obrigatoriamente, deviam concordar, para a condição de rejeição de H_0.

Assim, analisando-se os resultados apresentados, tem-se que, a série medida na E1 é, estatisticamente diferente, das séries medidas nas demais estações com as quais foi comparada, isto é, os resultados dos testes de comparação de séries temporais apresentaram p-valor menor que o nível de significância adotado, indicando que a série da E1 é gerada por um processo estocástico diferente das demais estações da RAMQAR. Isso pode ser explicado pelo fato de que a E1, além de receber influências diretas das indústrias da Ponta de Tubarão quando da ocorrência de ventos Sul, é a única pertencente a RAMQAr da RGV que cobre áreas sob a influência das indústrias do CIVIT (Centro Industrial de Vitória), quando há concorrência de ventos Nordeste, direção principal tomada pelos ventos na Região.

É possível observar, também, que, a E8 apresentou série temporal estatisticamente diferente das demais estações as quais foi comparada, exceto a E5. Isso se deve basicamente ao fato de que a E8 abrange áreas diretamente influenciadas pelas emissões veiculares, pois esta localiza-se próximo à CEASA, local caracterizado por um intenso tráfego veicular.

Verificou-se, após a aplicação dos testes aos dados de PM_{10} e PTS que, o monitoramento realizado na E2, na E3, na E4, na E5 e na E6, quando comparadas entre si, são geradas pelo mesmo processo estocástico, o que sugere que as referidas estações, apesar de operarem em diferentes locais da RGV, estão medindo poluentes emitidos pelas mesmas fontes, e para isso são gerados custos demasiados com operação, manutenção, coleta e análise dos dados que, estatisticamente, são iguais.

Um estudo semelhante para comparação das séries temporais de concentração PM medidas por diferentes estações de monitoramento, foi realizado por Singh e Perwez (2018) para a região mineira de Goa, na Índia para três localidades diferentes: em torno das minas, na zona de amortecimento e ao longo das rotas de transporte de minério. Entretanto, a metodologia utilizada plos autores foi a aplicação da ANOVA (Análise de Variância) e do teste t aos dados. Conforme os resultados encontrados por eles, os níveis PM não mudaram consideravelmente entre as minas. Mas na zona de amortecimento e ao longo das rotas de transporte de minério, os valores calculados de F foram maiores do que o F crítico, o que implica em uma variação significativa nos níveis de PM entre as estações de monitoramento. O resultado do teste 't' realizado usando concentrações sazonais médias de PM, comprovou a diferença significativa entre todos os três grupos de estações de monitoramento (minas, zonas tampão e rotas de transporte de minério), evidenciando que a diferença entre as séries temporais medidas nas diferentes estações advém das fontes geradoras das séries.

Além disso, desde a década de 1970, grandes projetos industriais foram implantados na RGV e, desta forma, a indústria era apontada como a principal fonte de poluição na área (FREITAS *et al.,* 2016). Assim, em 2000, quando a RAMQAr foi dimensionada, o objetivo era medir a exposição da população aos principais poluentes emitidos pela atividade industrial. Entretanto, nos últimos anos, o crescimento populacional, o aumento da frota de veículos e o desenvolvimento de grandes empreendimentos imobiliários tem alterado o perfil das fontes de poluição atmosférica na região. Esses fatores, conforme descrito a seguir, acabam por, corroborar com a hipótese de necessidade de redimensionamento da RAMQAr apresentada pelo resultado estatístico do presente estudo.

Tabela 4 - Resultados dos testes de comparação de séries temporais para o PM_{10}.

ESTAÇÕES	COATES & DIGGLE	QUENOUILLE	SANTOS *et al.*
E1 x E2	$3{,}23 \times 10^{-8}$	$7{,}428796 \times 10^{-5}$	$2{,}220446 \times 10^{-16}$
E1 X E3	$3{,}863 \times 10^{-6}$	0,009101863	$2{,}220446 \times 10^{-16}$
E1 X E4	$2{,}738 \times 10^{-5}$	0,5936195	$2{,}220446 \times 10^{-16}$
E1 X E5	$1{,}182 \times 10^{-8}$	$1{,}677112 \times 10^{-7}$	$2{,}220446 \times 10^{-16}$
E1 X E6	0,0001912	0,8858912	$2{,}220446 \times 10^{-16}$
E1 X E8	$3{,}737 \times 10^{-5}$	$2{,}220446 \times 10^{-16}$	$2{,}220446 \times 10^{-16}$
E2 X E1	$4{,}359 \times 10^{-5}$	$7{,}428796 \times 10^{-5}$	$2{,}220446 \times 10^{-16}$
E2 X E3	0,2778	0,9999978	$2{,}220446 \times 10^{-16}$
E2 X E4	0,2778	0,9999859	$2{,}220446 \times 10^{-16}$
E2 X E5	0,4252	0,9885811	$2{,}220446 \times 10^{-16}$
E2 X E6	0,1366	0,9855875	$2{,}220446 \times 10^{-16}$
E2 X E8	0,9754	$1{,}410633 \times 10^{-5}$	$2{,}220446 \times 10^{-16}$

Análise estatística das concentrações de poluentes atmosféricos

E3 X E1	0,004774	0,009101863	2,220446 x 10^{-16}
E3 X E2	0,04598	0,9999978	2,220446 x 10^{-16}
E3 X E4	0,357	0,9999981	2,220446 x 10^{-16}
E3 X E5	0,05502	0,998044	2,220446 x 10^{-16}
E3 X E6	0,5266	0,9944052	2,220446 x 10^{-16}
E3 X E8	0,3789	0,0001928085	2,220446 x 10^{-16}
E4 X E1	0,01039	0,5936195	2,220446 x 10^{-16}
E4 X E2	0,02613	0,9999859	2,220446 x 10^{-16}
E4 X E3	0,05502	0,9999981	2,220446 x 10^{-16}
E4 X E5	0,01286	0,9058954	2,220446 x 10^{-16}
E4 X E6	0,8476	1,0	2,220446 x 10^{-16}
E4 X E8	0,7195	7,164516 x 10^{-5}	2,220446 x 10^{-16}
E5 X E1	8,905 x 10^{-6}	1,647278 x 10^{-7}	2,220446 x 10^{-16}
E5 X E2	0,8699	0,9885811	2,220446 x 10^{-16}
E5 X E3	0,1366	0,998044	2,220446 x 10^{-16}
E5 X E4	0,09186	0,9058954	2,220446 x 10^{-16}
E5 X E6	0,03168	0,6189378	2,220446 x 10^{-16}
E5 X E8	0,7467	0,5627391	2,220446 x 10^{-16}
E6 X E1	0,05502	0,8858912	2,220446 x 10^{-16}
E6 X E2	0,0005044	0,9855875	2,220446 x 10^{-16}
E6 X E3	0,001837	0,9944052	2,220446 x 10^{-16}
E6 X E4	0,5266	1,0	2,220446 x 10^{-16}
E6 X E5	0,00111	0,6189378	2,220446 x 10^{-16}
E6 X E8	0,1169	0,0002398618	2,220446 x 10^{-16}
E8 X E1	2,34 x 10^{-5}	2,220446 x 10^{-16}	2,220446 x 10^{-16}
E8 X E2	0,0005044	1,410633 x 10^{-5}	2,220446 x 10^{-16}
E8 X E3	0,002348	0,0002231832	2,220446 x 10^{-16}
E8 X E4	0,357	7,164516 x 10^{-05}	2,220446 x 10^{-16}
E8 X E5	0,0002917	0,5627391	2,220446 x 10^{-16}
E8 X E6	0,2601	0,0002398618	2,220446 x 10^{-16}

Tabela 5 - Resultados dos testes de comparação de séries temporais para o PTS.

ESTAÇÕES	**COATES & DIGGLE**	**QUENOUILLE**	**SANTOS *et al.***
E3 X E4	1,234 x 10^{-5}	9,568 x 10^{-6}	<2,22 x 10^{-16}
E3 X E5	2,22 x 10^{-16}	1,688 x 10^{-14}	<2,22 x 10^{-16}
E3 X E6	5,91 x 10^{-5}	<2,22 x 10^{-16}	<2,22 x 10^{-16}
E3 X E8	2,22 x 10^{-16}	2,22 x 10^{-16}	<2,22 x 10^{-16}
E4 X E3	2,22 x 10^{-16}	9,568 x 10^{-6}	<2,22 x 10^{-16}
E4 X E5	2,22 x 10^{-16}	0,0003404	<2,22 x 10^{-16}
E4 X E6	2,22 x 10^{-16}	3,462 x 10^{-9}	<2,22 x 10^{-16}
E4 X E8	7,979 x x 10^{-5}	8,329 x 10^{-11}	<2,22 x 10^{-16}
E5 X E3	2,22 x 10^{-16}	1,684 x 10^{-14}	<2,22 x 10^{-16}
E5 X E4	2,22 x 10^{-16}	0,0003404	<2,22 x 10^{-16}
E5 X E6	2,22 x 10^{-16}	0,2327	<2,22 x 10^{-16}
E5 X E8	2,22 x 10^{-16}	3,696 x 10^{-7}	<2,22 x 10^{-16}
E6 X E3	2,22 x 10^{-16}	<2,22 x 10^{-16}	<2,22 x 10^{-16}
E6 X E4	2,22 x 10^{-16}	3,462 x 10^{-9}	<2,22 x 10^{-16}
E6 X E5	2,22 x 10^{-16}	0,2327	<2,22 x 10^{-16}

E6 X E8	0,05033	0,0001968	$<2,22 \times 10^{-16}$
E8 X E3	$2,22 \times 10^{-16}$	$<2,22 \times 10^{-16}$	$<2,22 \times 10^{-16}$
E8 X E4	$2,22 \times 10^{-16}$	$8,329 \times 10^{-11}$	$<2,22 \times 10^{-16}$
E8 X E5	$2,22 \times 10^{-16}$	$3,696 \times 10^{-7}$	$<2,22 \times 10^{-16}$
E8 X E6	$2,22 \times 10^{-16}$	0,0001968	$<2,22 \times 10^{-16}$

No tocante ao crescimento populacional, a necessidade de um redimensionamento advém do fato de uma rede de monitoramento ser dimensionada com base no tamanho da população que habita os locais escolhidos para sediarem as estações (USEPA, 2013). Assim, quando a RAMQAr foi instalada, a RGV abrigava uma população de 1.337.167 habitantes (IBGE, 2010). Atualmente, a região conta com, aproximadamente, 1.807.630 habitantes, cuja distribuição pelos municípios que compõe a RGV é apresentada na Tabela 6, a qual também traz o número mínimo de estações necessárias para uma boa representatividade dos dados fornecidos pela RAMQAr em cada município, conforme metodologia da USEPA.

Tabela 6 - Número mínimo de estações da RAMQAR da RGV baseado na população atual.

Município	População (habitantes)	Concentração de MP (média anual, $\mu g/m^3$)	Número mínimo de estações	Quantidade atual de estações
Serra	507.598	35,7	4	3
Vitória	358.267	29,0	2	3
Vila Velha	486.208	25,6	2	2
Cariacica	378.603	35,6	2	1
Viana	76.954	-	2	0

Infere-se da Tabela 6 que, a atual quantidade de estações que formam a RAMQAr da RGV não é suficiente para a população da área nos municípios de Serra, Cariacica e Viana, com destaque para este último que integra a RGV e não possui nenhuma estação para monitormanto da qualidade do ar. Ainda, soma-se a esta situação o fato de que, além das estações da Serra serem insuficientes para representar o impacto a sua população, os dados da E2 são formados pelo mesmo processo estocástico de outras estações da rede, como já discorrido.

Vitória, apesar de apresentar um número de estações superior ao mínimo requerido, apresenta outro problema: as suas três estações estão medindo os poluentes da mesma fonte, isto é, apesar de satisfazer o critério de dimensionamento no quesito quantidade, falha no quesito qualidade. Desta forma, sugere-se que as estações sejam realocadas para localidades em que suas medições sejam mais representativas dentro do município.

A alteração do perfil das fontes da RGV também se deve ao aumento da frota de veículos. Conforme disposto no Inventário de Fontes da RGV (ECOSOFT, 2011), com relação às

emissões de material particulado, identificou-se que a fonte preponderante desse poluente na região encontra-se nas emissões veiculares, decorrentes da ressuspensão de partículas (ECOSOFT, 2011), o que torna ainda mais enfática a necessidade de um redimensionamento da RAMQAr, pois a atual espacialização das estações encontra-se obsoleta observando-se esta fonte em potencial.

Outro fator que influi diretamente na alteração dos perfis de fontes da RGV é o desenvolvimento de grandes empreendimentos imobiliários que acabam se tornando barreiras à circulação de ventos para dispersão de poluentes. Além disso, em algumas vezes, ocorre a construção de edificações próximas às estações de monitoramento, o que que prejudica, não apenas o monitoramento da qualidade do ar, mas, principalmente, o monitoramento das variáveis meteorológicas que também é realizado pela RAMQAr (IEMA, 2018).

Diante dos resultados encontrados, sugerem-se revisões no projeto original da RAMQAr, com destaque para a reespacialização das estações, isto é, realocar estações que estão monitorando as mesmas fontes, visando melhorar sua representatividade de dados e abrangência espacial, tendo em vista a alteração dos perfis das fontes, e ampliação da rede, principalmente, devido ao aumento populacional da região.

CONCLUSÕES

Este trabalho avaliou as concentrações médias diárias de PM_{10} e PTS medidas em diferentes estações de monitoramento da RGV, Espírito Santo, Brasil, por meio da aplicação de métodos estatísticos descritivos e inferenciais, no período de 01/01/2008 a 31/12/2017.

Em relação a análise da série temporal das partículas, foi possível observar que as concentrações de ambos poluentes por vezes ultrapassaram os limites estabelecidos tanto pelos padrões estaduais e nacionais, quanto pelas diretrizes internacionais, fato este deve ser tratado com atenção, tendo em vista que até mesmo em baixas concentrações, principalmente, o $PM_{10,}$ tem potencial de provocar efeitos adversos à saúde. Dentre as fontes, o tráfego de veículos automotores é apresentado como o principal contribuinte de emissão de partículas respiráveis e partículas totais em suspensão na RGV.

No entanto, também foi observado para todas as estações tendência de decrescimento nas concentrações dos poluentes nos últimos anos, o que sugere que para melhoria da qualidade

do ar devem ser tomadas medidas mais restritivas em relação aos limites de emissão de contaminantes atmosféricos, a exemplo do Decreto Estadual 3463 – R/2013.

Conforme os resultados encontrados pelos testes de comparação de séries temporais aplicados neste estudo, as estações E2 , apenas para o poluente PM_{10}, e E3, E4, E5 e E6, para os poluentes PM_{10} e PTS, quando comparadas entre si, foram geradas pelo mesmo processo estocástico, ou seja, estas estações medem dados de concentrações de poluentes emitidos pela mesma fonte. Portanto, devido as transformações ocorridas nas cidades desde a implantação da RAMQAr, no ano 2000, o perfil estabelecido das estações não é mais suficiente para cobertura espacial de todas as fontes potencialmente poluidoras na região.

Deste modo, o presente trabalho pode ser usado como um indicativo da necessidade de reformulação do projeto inicial da RAMQAr, realocando os medidores de PM_{10} e PTS das estações cuja séries provém de mesmo processo estocástico. Entretanto, para a escolha dos novos locais a receber as estações, de forma a ampliar a área de cobertura da RAMQAr, deve ser realizado um estudo de dispersão de contaminantes de modo a orientar o monitoramento dos poluentes, visto que um estudo de dispersão possibilita verificar o comportamento destes na atmosfera e a abrangência do impacto ambiental. Além disso, também sugere-se que seja estudada a necessidade de ampliação da rede, para cobrir toda a RGV e, assim, fornecer uma base de dados ainda mais confiável e adequada a ser utilizada no planejamento e gerenciamento de estudos e estratégias relativas ao controle da poluição do ar, afim de minimizar e/ou prevenir possíveis impactos nocivos à população e ao meio ambiente em geral.

REFERÊNCIAS

ABE, Karina Camasmie; MIRAGLIA, Simone Georges El Khouri. Avaliação de impacto à saúde do programa de controle de poluição do ar por veículos automotores no município de São Paulo, Brasil. **RBCIAMB**, n.47, p.61-73, 2018.

AMARAL, Marcus Vinicius Silva Gurgel do. **Ajuste de modelos e comparação de séries temporais para dados de vazão específica em microbacias pareadas**. 2014. Dissertação (Mestrado em Ciências: Estatística e Experimentação Agronômica) – Universidade de São Paulo, Piracicaba, 2014.

ARAUJO, J. A.;NEL, A. E. Particulate matter and atherosclerosis: role of particle size, composition and oxidative stress. **Particle and Fibre Toxicology**, v. 6, n.24, 2009.

BÉJOT, Yannick; REIS, Jacques; GIROUD, Maurice; FEIGIN, Valery. A review of epidemiological research on stroke and dementia and exposure to air pollution. **International Journal of Stroke**, p. 1-9, 2018.

BOWE, Benjamin; XIE, Yan; LI, Tingting, YAN, Yan, XIAN, Hong; ALY-AL, Ziyade. The 2016 global and national burden of diabetes mellitus attributable to PM 2· 5 air pollution. **The Lancet Planetary Health**, v. 2, n. 7, p. e301-e312, 2018.

BOX, George EP; PIERCE, David A. Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. **Journal of the American statistical Association**, v. 65, n. 332, p. 1509-1526, 1970.

BROCKWELL, P. J.; DAVIS, R. A. **Time Series**: Theory and Methods. 2ª ed. Springer, 2006.

CALDERÓN-GARCIDUEÑAS, L.; LERAY, E.; HEYDARPOUR, P.; JARDÓN-TORRES, R.; REIS, J. Air pollution, a rising environmental risk factor for cognition, neuroinflammation and neurodegeneration: the clinical impact on children and beyond. **Revue Neurologique**, v. 172, p. 69-80, 2016.

CELIS, José E.; MORALES, José R.; ZAROR, Claudio A.; CARVACHIO, Omar F. Contaminación del aire atmosférico por material particulado em uma Ciudad intermedia: el caso de Chillán (Chile). **Información Tecnologica**, v.18, p.49-58, 2007.

CHALOULAKOU, A.; KASSOMENOS, P.; SPYRELLIS, N.; DEMOKRITOU, Philip; KOUTRAKIS, P. Measurements of PM10 and PM2.5 particle concentrations in Athens, Greece. **Atmospheric Environment**, v. 37, p. 649–660, 2003.

COATES, D. S.; DIGGLE, P. J. Tests for comparing two estimated spectral densities. **Journal of Time Series Analysis**, v. 7, n. 1, p. 7-20, 1986.

CONAMA – Conselho Nacional de Meio Ambiente. Resolução nº 491, de 19 de novembro de 2018. Dispõe sobre padrões de qualidade do ar**. Diário Oficial da União**: seção 1, Brasília, DF, p. 155. Disponível em: http://www2.mma.gov.br/port/conama/legiabre.cfm?codlegi=740. Acesso em: 16 dez. 2018.

COSTA, Franciella Marques da. **Comparação Estatística de duas séries de material particulado (MP_{10}) na cidade de São Paulo**. 2010. Dissertação (Mestrado em Estatística e Experimentação Agropecuária) – Universidade Federal de Lavras, Lavras, 2010.

COSTA, Franciella Marques da; SÁFADI, Thelma. Comparação estatística de duas séries de material particulado (MP_{10}) na cidade de São Paulo. **Rev. Bras. Biom**, v. 28, n. 3, p. 23-38, 2010.

COTTA, H. H. A. **Análise de componentes principais robusta em dados de poluição do ar: aplicação à otimização de uma rede de monitoramento**. 2016. Dissertação de mestrado. Universidade Federal do Espírito Santo, Vitória, ES, 2016.

ECOSOFT CONSULTORIA E SOFTWARES AMBIENTAIS. **Inventário de emissões atmosféricas da Região da Grande Vitória.** Vitória, 2011. Disponível em:

https://iema.es.gov.br/Media/iema/.../Inventário%20de%20fontes%20de%202010.pdf. Acesso em: 24 out. 2018.

ESPÍRITO SANTO (Estado). Decreto nº 3463-R, de 16 de dezembro de 2013. Estabelece novos padrões de qualidade do ar e dá providências correlatas. **Diário Oficial do Estado do Espírito Santo**, Vitória, ES, p. 9-11, 17 dez. 2013. Disponível em: https://iema.es.gov.br/Media/iema/CQAI/Documentos/DECRETO_N%C2%BA_3463_2013.pdf. Acesso em: 15 set. 2018.

FREITAS, Clarice Umbelino de; LEON, Antonio Ponce de; JUGER, Washington; GOUVEIA, Nelson. Air pollution and its impacts on health in Vitoria, Espirito Santo, Brazil. **Revista de Saúde Pública**, v. 50, p. 4, 2016.

IBGE – Instituto Brasileiro de Geografia e Estatística. **Sinopse do Censo Demográfico 2010 - Espírito Santo.** 2010. Disponível em: https://censo2010.ibge.gov.br/sinopse/index.php?dados=21&uf=32. Acesso em: 15 out. 2018.

IEMA – Instituto Estadual de Meio Ambiente e Recursos Hídricos do Estado do Espírito Santo. **Relatório da qualidade do ar da Região da Grande Vitória - 2016**. Vitória: IEMA. 2018. Disponível em: https://iema.es.gov.br/Media/iema/CQAI/Relatorios_anuais/Relat%C3%B3rio_Anual_de_Qualidade_do_Ar_2016.pdf. Acesso em: 04 jan. 2019.

IEMA – Instituto Estadual de Meio Ambiente e Recursos Hídricos. **Qualidade do Ar**. 2018. Disponível em: https://iema.es.gov.br/qualidadedoar/dadosdemonitoramento/automatica. Acesso em: 01 ago. 2018.

JACOBS, E. T.; BURGESS, J. L.; ABBOTT, M. B. The Donora Smog Revisited: 70 Years After the Event That Inspired the Clean Air Act. **American Journal of Public Health**, v. 108, n. S2, p. S85-S88, 2018.

JUN, K. Case Study of Air Pollution Episodes in Meuse Valley of Belgium, Donora of Pennsylvania, and London, UK. **Environmental Toxicology and Human Health**, v. I, p.78, 2009.

JUNGER, Washington Leite. **Análise, imputação de dados e interfaces computacionais em estudos de séries temporais epidemiológicas**. 2008. Tese (Doutorado em Epidemiologia) – Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2008.

KENDALL, M. G. **Rank correlation methods**. London: Charles Griffin, 1975. 120p.

KÖPPEN, Wladimir. Versuch einer Klassifikation der Klimate, vorzugsweise nach ihren Beziehungen zur Pflanzenwelt. **Geographische Zeitschrift**, v. 6, n. 11. H, p. 593-611, 1900.

LEITE, Renata Carvalho Macedo; GUIMARÃES, Ednaldo Carvalho; LIMA, Euclides Antonio Pereira de; BARROZO, Marcos Antonio de Souza; TAVARES, Marcelo. Utilização de regressão logística simples na verificação da qualidade do ar atmosférico de Urbelândia. **Engenharia Sanitária e Ambiental**, v. 16, n. 1, p. 175-180, 2011.

LIU, Ziruli; HU, Bo; Wang, Lili; WU, Fangkun; GAO, Wenkang; WANG, Yuesi. Seasonal and diurnal variation in particulate matter (PM_{10} and $PM_{2,5}$) at na urban site of Beijing: analyses from a 9-year study. **Environmental Science and Polluition Research**, 2014.

LOGAN, W. P. D. et al. Mortality in the London fog incident. **Lancet**, p. 336-338, 1952.

MANN, H. B. Nonparametric tests against trend. **Econometrica**, v.13, p.245-259, 1945.

MONTE, Edson Zambon; ALBURQUERQUE, Taciana Toledo de Almeida; REISEN, Valdério Anselmo. Impactos das variáveis meteorológicas n qualidade do ar da Região da Grande Vitória, Espírito Santo, Brasil. **Revista Brasileira de Meteorologia**, v. 31, n.4, p. 546-554, 2016.

MOREIRA, Davidson Martins; TIRABASSI, Tiziano; MORAES, Marcelo Romero de. Meteorologia e poluição atmosférica. **Ambiente & Sociedade**, v. 11, n. 1, p. 1-13, 2008.

MORETTIN, P. A.; TOLOI, C. M. C. **Análise de Séries Temporais**. 2. ed. São Paulo: Edgard. Blücher, 2006. 538p.

NESAMANI, K. S. Estimation of automobile emissions and control strategies in India. **Science of the Total Environment**, v. 408, n. 8, p. 1800-1811, 2010.

OLIVEIRA, L. M. D. **Contribuição do Programa Despoluir para a redução das emissões atmosféricas pela frota de ônibus da Região Metropolitana do Natal-RN**. 2017. Dissertação de mestrado. Instituto Federal do Rio Grande do Norte, Natal, RN, 2017.

PETRO, F.; KONEČNÝ, V. Calculation of Emissions from Transport Services and their use for the Internalisation of External Costs in Road Transport. **Procedia Engineering**, v. 192, p. 677-682, 2017.

QUENOUILLE, M. H. The comparison of correlations in time-series. **Journal of the Royal Statistical Society**. Series B (Methodological), p. 158-164, 1958.

R CORE TEAM (2018). *R:* **A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

RAMACHANDRA, T. V. Emissions from India's transport sector: statewise synthesis. **Atmospheric Environment**, v. 43, n. 34, pp. 5510-5517, 2009.

RAZ, R.; ROBERTS, A.L.; LYALL, K.; HART, J.E.; JUST, A.C.; LADEN, F.; WEISSKOPF, M.G. Autism spectrum disorder and particulate matter air pollution before, during, and after pregnancy: a nested case-control analysis within the Nurses' Health Study II cohort. **Environmental Health Perspectives**, v.123, p. 264-270, 2015.

REINA, Jhovana; OLAYA, Javier. Curve fitting nonparametric methods for studying behavior from air pollution PM_{10}. **Revista EIA**, n. 18, p. 19-31, 2012.

RUSSO, Paulo Roberto. A qualidade do ar no município do Rio de Janeiro: análise espaço--temporal de partículas em suspensão na atmosfera. **Revista de Ciências Humanas**, v. 10, n. 1, p. 78-93, 2010.

SILVA, Roberta Bessa Veloso; FERREIRA, Daniel Furtado; SÁFADI, Thelma. Modelos de séries temporais aplicados à série dos índices de preços ao consumidor na região de Lavras, MG, no período de 1992 a 1999. **Organizações Rurais & Agroindustriais**, v. 2, n. 2, p. 44-55, 2000.

SINGH, Gurdeep; PERWEZ, A. Assessment of ambient air quality around mines, in buffer zone and along ore transportation routes in iron ore mining region of Goa: emphasis on spatial distributions and seasonal variations. **Int. J. Environment and Pollution**, v. 63, p. 47-68, 2018.

STADLOBER, Ernst; HÖRMANN, Siegfried; PFEILER, Brigitte. Quality and performance of a PM10 daily forecasting model. **Atmospheric Environment**, v. 42, p.1098-1109, 2008.

TOLOI, C.M.C; ECHEVERRY, G.E.S. Testes para comparação de séries temporais: uma aplicação a séries de temperatura e salinidade da água, medidas em profundidades diferentes. **Rev. Bras. Estat**, p. 51-80, 2000.

USEPA – Unites States Environmental Protection Agency. **40 CFR Appendix D to Part 58: Network Design Criteria for Ambient Air Quality Monitoring**. 2013. Disponível em: https://www.govinfo.gov/content/pkg/CFR-2014-title40-vol6/pdf/CFR-2014-title40-vol6-part58-appD.pdf. Acesso em: 17 dez. 2018.

VARDOULAKIS, Sotiris; KASSOMENOS Pavlos. Sources and factors affecting PM10 levels in two European cities: implications for local air quality management. **Atmospheric Environment**, v. 42, p. 3949–3963, 2008.

WEI, W. W. **Time series analysis**: univariate and multivariate methods. 2ª. ed. Addison Wesley, 2006.

WANG, Jun; HU, Zimei; CHEN, Yuanyuan; CHEN, Zhenlou; XU, Shiyuan. Contamination characteristics and possible sources of PM10 and PM2,5 in diferente functional áreas of Shanghai, China. **Atmospheric Environment**, v. 68, p. 221 e 229, 2013.

WHO – World Health Organization. **WHO air quality guidelines global update 2005**. 2005. Disponível em: http://www.euro.who.int/__data/assets/pdf_file/0008/147851/E87950.pdf. Acesso em: 29 set. 2018.

ZAMPROGNO, B. **O uso e interpretação da análise de componentes principais, em séries temporais, com enfoque no gerenciamento da qualidade do ar**. 2013 Tese de Doutorado, Universidade Federal do Espírito Santo (UFES), Vitória, ES, 2013.

ZHANG, Xin; CHEN, Xi; ZHANG, Xiaobo. The impact of exposure to air pollution on cognitive performance. **Proceedings of the National Academy of Sciences**, v. 115, n. 37, p. 9193-9197, 2018.