# UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO

## Centro Tecnológico

**Programa de Pós-graduação em Engenharia Ambiental**


**Tese de Doutorado**

# *O uso e interpretação de análise de componentes principais, em séries temporais, com enfoque no gerenciamento da qualidade do ar*


**Orientador:**
***Prof. Dr. Valdério Anselmo Reisen***

**Co-orientador:**
***Prof. Dr. Neyval Costa Reis Junior***

**Aluno:**
***Bartolomeu Zamprogno***


**Ano 2013**

**Bartolomeu Zamprogno**

# *O USO E INTERPRETAÇÃO DE ANÁLISE DE COMPONENTES PRINCIPAIS, EM SÉRIES TEMPORAIS, COM ENFOQUE NO GERENCIAMENTO DA QUALIDADE DO AR*

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.
Co-orientador: Prof. Dr. Neyval C. Reis Junior.

**Ano 2013**

*A minha esposa que, praticamente sozinha, teve de suportar de tudo ao longo desses intermináveis anos. Só cheguei ao fim por causa dela e dos filhos que vieram, somente por isso e nada mais. Foram a minha energia para seguir adiante completamente destruído.*

## Agradecimentos

A Deus.
A minha esposa e filhos.
Ao meu orientador Valdério A. Reisen que muitíssimo colaborou como professor e, muitas vezes, como colega para o fim dessa tese.
Ao meu co-orientador Neyval Costa Reis Júnior.
Aos colegas Fabio Fajardo, Alessandro Sarnaglia e Nátaly Jiménez agradeço de coração toda a ajuda. A contribuição foi enorme.
Ao corpo docente do PPGEA que acreditou no término deste trabalho.
A todos que de alguma forma contribuíram para que este trabalho fosse concluído.

# Sumário

# Lista de Figuras

# Lista de Tabelas

# Lista de Abreviaturas e/ou siglas

| | |
|---|---|
| ACP | Análise de componentes principais |
| AC | Análise de clusters |
| ARFIMA | Fracionário autotorregressivo médias móveis |
| ARMA | Autorregressivo médias móveis |
| AQMN | Air quality manegement network |
| ARUMA | Fractionaly autoregressive unit circle moving average |
| $B$ | Operador atraso |
| BQM | Balanço químico de massa |
| CO | Carbono Orgânico |
| ES | Espírito Santo |
| FMP | Fatoração de matriz positiva |
| GARMA | Gegenbauer ARMA process |
| GPH | Estimador de memória longa de Geweke and Porter-Hudak |
| IEMA | Instituto Estadual do Meio Ambiente |
| MAD | Median absolute desviation |
| MAG | Modelo de regressão não linear aditivo generalizado |
| MAG-ACP | Modelo MAG com uso de ACP |
| $PM_{10}$ | Particulate Matter (Material Particulado) menor que $10\mu m$ de diâmetro aerodinâmico |
| PTS | Partículas Totais em Suspenção |
| RGV | Região da Grande Vitória |
| RR | Risco relativo |
| $SO_2$ | Dióxidos de enxofre |
| SARFIMA | Sazonal fracionário autotorregressivo médias móveis |
| $NO_2$ | Óxidos de nítrico |
| OLS | Ordinary least squares |
| $O_3$ | Ozônio |
| PC | Principal component |
| PCA | Principal componenents analysis |
| RPCA | Robust principal components analysis |
| tr | traço de uma matriz |
| $\mu m$ | micrômetro |
| $\Gamma_X(0)$ | Matriz de covariância do modelo $X$ na defasagem zero |
| $\Gamma_Y(l)$ | Matriz de covariância do modelo $Y$ na defasagem $l$ |
| $\Phi$ | Matriz com coeficientes do VARMA |
| $\Sigma$ | Matriz de covariância do processo ruído branco |
| $\Psi(B)$ | Operador atraso de um processo linear |
| VAR-MAG-ACP | Modelo MAG aplicada no resíduo de um processo VAR |
| VAR | Vetorial autorregressivo |
| VARMA | Vetorial autorregressivo médias móveis |
| VOD90 | 90% da variância original |

## Resumo

Este trabalho foi motivado pela aplicação da técnica análise de componentes principais em diferentes contextos da área poluição do ar, em especial no uso do gerenciamento de rede. Essa metodologia estatística, em termos práticos, produz informações com precisões na tomada de decisões importantes para qualidade do ar. Essa técnica é usualmente utilizada, assim como na análise de regressão, como ferramenta para análise e interpretação dos fenômenos dos dados. Entretanto, de acordo com a literatura estatística que fomenta base para o uso dessa ferramenta em qualquer área de aplicação, a técnica exige pressuposto, nesse caso o uso de variáveis independentes, característica que praticamente não é observada em situações práticas na área poluição do ar. Em geral, os dados disponíveis para resolução de problemas como gerenciamento de rede, identificação de fonte poluidora, estudos espaços-temporais e associação do número de internações por causas respiratórias por poluentes são séries temporais que apresentam estrutura curta e longa de dependência temporal, ou seja, autocorrelação. Os resultados dessa pesquisa mostram, no domínio do tempo, que a técnica de análise de componentes principais, dependendo da estrutura de autocorrelação das séries, podem ser baseadas em resultados espúrios. Quando a estrutura de autocorrelação é fraca o efeito da autocorrelação é praticamente nulo, dessa forma a técnica pode ser empregada sem maiores problemas. No contexto do uso da técnica de análise de séries temporais no domínio da frequência foi avaliado a extensão de métodos existentes para o caso de dados de séries temporais com memória longa. Os resultados evidenciam que o uso de métodos do domínio da frequência podem ser utilizados, mas algumas considerações devem ser observadas e alguns tipos de aplicações, da poluição do ar, merecem mais estudos devido a dificuldade de interpretação no domínio da frequência.



Palavras-chave: análise de componentes principais, poluição do ar, análise de séries temporais, domínio do tempo, domínio da frequência

## Abstract

This work was motivated by the application of principal component analysis technique in different contexts of area air pollution, especially in the use of network management. This statistical methodology in practical terms produces information with accuracies in making important decisions for quality air. This technique is commonly used, as well as in the regression analysis as a tool for analysis and interpretation of the phenomena of the data. However, according to the statistical literature that fosters basis for the use of this tool in any area of application, the technique requires the assumption in this case the use of independent variables, a characteristic which is hardly observed in practical situations in the field of air pollution. In general, the data available for troubleshooting management network, identification of pollutant source, studies spatio-temporal association and the number of hospitalizations for respiratory pollutants are by series displaying structure of short and long time dependence, that is, autocorrelation. The research results show, in the field of time that the technique of principal components analysis, depending on the structure autocorrelation of the series, can be based on spurious results. When the structure is weak, the autocorrelation effect of autocorrelation is practically zero, so that the method can be used without further problems. In the context of the use of the technique of time series analysis in the frequency domain was reported the extension of existing methods for the case of time series data memory long. The results show that the use of frequency domain methods can be used, but some considerations should be observed and some types of applications, the air pollution, deserve further study because of the difficulty of interpreting the frequency domain.

Keywords: principal component analysis, air pollution, time series analysis, time domain, frequency domain

# 1 Introdução

A realização de estudos ambientais é motivada por diversos fatores, por exemplo, o crescimento populacional mundial leva, em geral, a um anseio de diminuir a taxa de desemprego, de aumentar áreas de plantio, de construir novos lares, de aumentar o nível de produção das industrias, além de diversas outras questões. Dessa forma o progresso, em geral, tem como consequência a elevação dos níveis dos poluentes emitidos, o que acarreta a uma necessidade de investigações mais apuradas. Em particular, estudos da qualidade do ar são fundamentados por questões como, por exemplo, o desenvolvimento tecnológico, as condições sociais, de saúde, de transporte e as climáticas entre outras. Portanto, identificar fontes poluidoras e regiões que mais sofrem impactos da poluição, avaliar os níveis dos poluentes, seja na fonte poluidora ou no local receptor, e comparar com o disposto na legislação vigente são etapas primordiais para um bom controle das emissões de poluentes ou para provocar melhorias para o meio ambiente.

Diante do avanço das tecnologias computacionais e dos equipamentos de medições, a área da poluição, em especial, a atmosférica, tanto no Brasil como no exterior, produzem uma grande quantidade de dados. Analisá-los e interpretá-los depende de metodologias específicas de acordo com os fenômenos investigados. Em geral, as áreas de modelagem estocástica contribuem com vários métodos de análise. Em particular, diversos métodos estatísticos multivariados são empregados para avaliar os efeitos da poluição. Entre as técnicas multivariadas, análise de componentes principais (ACP) tem destaque em estudos ambientais. Como forma de quantificar a relevância dessa técnica, Richman (1986) mostrou que entre 1983 e 1985 mais de 60 aplicações de ACP, ou técnicas similares, apareceram em periódicos meteorológicos/climatológicos. Mais recentemente, entre os anos 1999 e 2000, 53 dos 215 artigos do *International Journal of Climatology* utilizaram alguma forma de ACP, que significa uma taxa de 25%, não alcançada por nenhuma outra técnica de estatística, Jollife (2002, pg. 71). O recente artigo de Belis et al. (2013) exemplifica e quantifica a importância do uso dessa técnica no problema de identificação de fontes. Os autores afirmam que, aproximadamente, um quarto (24%) dos estudos europeus de identificação de fonte poluidora para material particulado foram baseados em ACP e suas variantes. O uso de ACP como modelo receptor é devido a vantagem do não conhecimento a *priori* dos inventários de emissões.

A técnica de ACP é capaz de reduzir a dimensão de um conjunto de variáveis sem ocorrer perda considerável da variabilidade original. O emprego da técnica ACP não se restringe apenas à redução da dimensão de conjuntos de dados. Por exemplo, no domínio do tempo, Karar & Gupta (2007) utilizaram ACP como uma análise de cluster para identificar fontes poluidoras e Cohen (1983), White et al. (1991) e Romero et al. (1999) fizeram uso de ACP como uma análise de *cluster* para identificar subregiões homogêneas de estações de variáveis climáticas em grande área geográfica. Além do uso da ACP como uma análise de *cluster*, diversos trabalhos utilizam esse procedimento para eliminar a multicolinearidade em modelos de regressão e para detectar outliers veja, por exemplo, Liu (2009). A ACP também pode participar na execução de outras metodologias multivariadas como a análise fatorial, a análise de correlação canônica, a análise discriminante entre outras.

Em diversas áreas de aplicação, o método de ACP é útil para identificação de fontes, para a localização de regiões com comportamentos de poluição similares, para estudos espaço-temporais, entre outras finalidades. No caso de estudos ambientais relacionados à poluição do ar, ACP foi utilizada por Lowell et al. (1984) e Richman (1986) para detectar padrões espaciais da concentração de enxofre ($SO_2$) no oeste dos Estados Unidos. Por sua vez, Sanchez et al. (1996) utilizaram ACP para identificar diferentes fontes de emissões de $SO_2$. A mesma técnica foi empregada para associar a variabilidade espacial de poluentes por Yu & Yu (2004) e Pires et al. (2008*a*,*b*). No contexto de padrões meteorológicos e temporais, a ACP foi utilizada

por Oanh et al. (2005), Yu & Chang (2006), Ibarra-Berastegi et al. (2008) e Ezcurra et al. (2008). No contexto de identificação de fontes poluidoras, Guo et al. (2004) utilizaram uma combinação de ACP e regressão linear múltipla, denominada ACPS, para identificar fontes de poluição em uma área rural do leste da China. Karar & Gupta (2007) analisaram as componentes químicas do $PM_{10}$ em uma região urbana de Kolkata, India, utilizando ACP/ACPS. Shi et al. (2009) combinaram vários modelos receptores, entre eles ACP, em dados simulados e reais na China. Lehman et al. (2004) realizaram uma caracterização espaço-temporal do ozônio troposférico. Em muitos casos, ACP também tem sido utilizada para estes e outros similares propósitos combinada com outros grupos de técnicas, como análises de cluster (McGregor (1996), Pires et al. (2008*a*,*b*), Lau et al. (2009) ou *Self-Organizing Maps*, Ezcurra et al. (2008)). A ACP também pode ser utilizada como método de estimação da análise fatorial (AF), outro modelo receptor multivariado utilizado em estudos da poluição do ar (veja, e.g., Viana et al. 2008).

Todas essas formas de emprego da técnica ACP a colocam como método de enorme potencial na resolução de problemas, seja de forma direta ou indireta. Promover o conhecimento por meio de métodos precisos é uma necessidade atual de diversas áreas do saber. A independência dos dados é a base teórica da técnica usual de ACP, em geral, essa baseada na matriz de autocovariância, denominada assim ACP no domínio do tempo. Um fato que chama atenção é o uso indiscriminado dessa metodologia mesmo quando as variáveis são séries temporais. Para esse tipo de dados a matriz de covariância pode ser expressa, equivalentemente, em termos da função espectral da série, o que possibilita que a análise possa ser realizada também no domínio da frequência. Negligenciar a estrutura temporal das observações pode acarretar em análises e interpretações totalmente equivocadas ou, até mesmo, inviabilizar o cálculo das componentes quando o processo vetorial é não estacionário na média.

Os problemas discutidos nos parágrafos anteriores são bases para as pesquisas apresentadas nesta tese que investiga, sob vários ângulos, o uso de ACP, no domínio do tempo e da frequência, com diferentes estruturas de dependência da série, isto é, séries definidas como memórias curta e longa. Séries temporais de memórias curta e longa são caracterizadas, no domínio do tempo, pela propriedade de que a soma absoluta das autocovariâncias seja finita e não finita, respectivamente. No domínio da frequência, séries com memória longa possuem a função espectral ilimitada em pelo menos uma frequência no intervalo de $(0, \pi)$. Um modelo de séries temporais que representa processos com as propriedades acima é o modelo ARFIMA$(p, d, q)$, ver Reisen et al. (2010), onde $d$ é o parâmetro de regência de memória do processo. Se $d = 0$, o processo é dito memória curta (exemplo, modelos ARMA), se $0 < d < 0.5$ o processo é dito ter memória longa e é estacionário, se $d \geq 0.5$ o processo tem memória longa, mas não é estacionário. Definições do processo com memória longa de forma equivalente, sob certas condições matemáticas, são apresentadas no domínio da frequência [veja, por exemplo, Palma (2007)].

As principais contribuições científicas desta tese estão centradas no estudo de ACP em processos com diferentes estruturas de memórias e aplicações (solução de problemas) em dados da poluição do ar coletadas na Região da Grande Vitória (RGV). Os resultados são apresentados de forma empírica e teórica e contribuem de forma científica para análise, estimação, teste e previsão, mas com foco específico em problemas da poluição do ar. Com destaque, a técnica ACP pode ser utilizada mesmo com a violação da propriedade de independência. Em especial, para processos com fraca correlação temporal, entretanto, sob certa cautela. Caso contrário, inferências e interpretações espúrias são obtidas. Neste caso, na análise de identificação de fontes poluidoras e no gerenciamento da rede de monitoramento as conclusões mostram que sobre os autovetores sofrem o efeito da estrutura de autocorrelação. Uma outra linha de investigação é o uso da ACP em modelos de regressão. Os estudos desta tese indicam que as componentes principais resultantes de séries temporais também são séries temporais.

Por exemplo, no estudo da relação entre poluição e saúde, as variáveis explicativas que, em geral, são poluentes com forte estrutura de dependência, o uso de ACP nesse contexto gera componentes principais com forte autocorrelação e com correlação cruzada significativa entre componentes, em *lags* diferentes de zero. Como forma de contornar essas propriedades da correlação temporal na ACP, esta tese também contribui por propor um método de filtragem, por meio de processos lineares, nas variáveis multivariadas, que elimina os efeitos descritos nas aplicações mencionadas. Os contextos de ACP robusto e ACP no domínio da frequência são também motivos desta tese. Os resultados obtidos, de forma empírica e aplicada, mostram a importância dessas metodologias nas aplicações práticas.

Os problemas práticos investigados, identificação de fontes poluidoras, gerenciamento de rede de monitoramento e a associação do número de internações a poluentes atmosféricos foram analisados com dados obtidos na Região da Grande Vitória. Para o estudo de identificação de fonte poluidora foi utilizado os dados de partícula total em suspensão investigados por Soares (2011). No caso de gerenciamento da rede de monitoramento, as séries das concentrações de $PM_{10}$ e $SO_2$ foram consideradas pois apresentaram uma estruturas de autocorrelação mais atraente para modelagem. No contexto de regressão, foram examinadas as variáveis de internação hospitalar e poluentes atmosféricos pesquisadas por Souza (2013).

As contribuições acima são motivos dos artigos apresentados nesta tese. O artigo 1 é o coração central, pois explora a técnica usual de ACP, domínio do tempo, no sentido de estudar o efeito da autocorrelação, propõe teste para verificar se a correlação temporal pode causar qualquer tipo de interpretação espúria e, em base a essas discussões, o artigo também propõe a utilização de filtros para que a ACP possa ser aplicada sem causar os problemas estudados. São também fornecidas importantes aplicações da área poluição do ar como identificação de fontes, gerenciamento de rede e, também, o relevante uso da técnica em modelos de regressão, em particular, na análise de dados de internação associada a poluentes atmosféricos. Esses estudos contribuem de forma bastante significativa na utilização de ACP no domínio do tempo.

O artigo 2 avalia, por meio de simulação e aplicação, o uso de ACP no domínio da frequência, técnica desenvolvida para dados correlacionados no tempo. Entretanto, devido a sua complexidade metodológica, ainda não muito explorada em áreas correlacionadas com a estatística como a engenharia ambiental, o artigo aborda o uso do método ACP em séries com memória longa, que são comuns na área da poluição do ar, conforme já discutido no primeiro artigo. Além da contribuição empírica no uso da ACP no domínio da frequência, em problemas da poluição do ar (gerenciamento de rede entre outros), o artigo fomenta diferentes linhas de pesquisa que podem, de maneira significativa, contribuir no uso e na popularização para identificar fontes, construir intervalos de confiança e aprimorar as ferramentas de análise espectral na poluição do ar.

O artigo 3 também é uma contribuição não teórica da técnica ACP, no domínio do tempo, sob condições de observações atípicas, ou *outliers*, tipo observações bastante comum na poluição atmosférica. Embora algumas observações não ultrapassem os limites estabelecidos pelos órgãos ambientais, se comparadas com o restante dos dados, pode ser caracterizadas como *outliers*. Esse artigo contribui na aplicação e fomenta vertentes de pesquisa nessa linha de estudos de robustez que pode ser estendida para ACP no domínio da frequência.

Para finalizar, o artigo 4 apresenta estudos de séries temporais em processos sazonais com memória longa, fenômeno frequentemente observado na poluição do ar. Esse artigo tem como objetivo nesta tese de apresentar definições de estimação de processos sazonais ARFIMA, esses que são bases dos capítulos anteriores.

Como forma de ilustrar as potencialidades do reconhecimento científico dos resultados desta tese, na Tabela 1 são apresentados os periódicos onde os artigos 1, 2 e 3 serão submetidos. O artigo 4 já passou pelo segundo processo de revisão, de acordo com as sugestões dos *referees* do periódico "Mathematics and Computers in Simulation".

Tabela 1: Situação dos artigos.

| **Artigos** | **Submetido** | **Revista** |
|---|---|---|
| 1 | Não | Environmental Modelling & Software |
| 2 | Não | Applied Numerical Mathematics / EnvironMetrics |
| 3 | Não | Atmospheric Research |
| 4 | Sim | Mathematics and Computers in Simulation |

Esta tese está dividida da seguinte forma: a Seção 2 apresenta os objetivos, gerais e específicos, que motivaram esta pesquisa. A revisão das principais referências está descrita na Seção 3. Os Artigos estão anexados após a Seção 3. As Seções 4 e 5 apresentam, respectivamente, a discussão geral e as conclusões com as recomendações para pesquisas futuras.

# 2 Objetivos

## 2.1 Objetivo Geral

Estudar o modelo multivariado ACP na presença de dados correlacionados com diferentes estruturas de dependência tais como curta e longa, nos domínios do tempo e da frequência, e aplicar a metodologia em problemas da área da poluição do ar.

## 2.2 Objetivos Específicos

1. Estudar empírica e analiticamente as propriedades estatísticas da técnica ACP quando os vetores de dados amostrais são correlacionados no tempo;

2. Verificar a extensão do método de Brillinger no domínio da frequência em processos memória longa e aplicar a metodologia no problema gerenciamento da rede de monitoramento;

3. Avaliar e comparar o uso de ACP, domínio do tempo e frequência, como metodologia para gerenciar uma rede de monitoramento;

4. Comparar a identificação da contribuição das fontes de PTS pelo modelo receptor ACP, no domínio do tempo, com o método proposto nesta tese;

5. Analisar o uso de ACP, domínio do tempo, como variáveis explicativas em modelos de regressão para associar o número de internações por problemas respiratórios à poluentes do ar.

# 3 Revisão Bibliográfica

Statheropoulos et al. (1998) analisaram as concentrações de CO, NO, $NO_2$, $O_3$, fumaça e $SO_2$ usando ACP. Os registros desses dados são de cinco anos e foram obtidos em uma estação de monitoramento da qualidade do ar da cidade de Atenas. A técnica de ACP também foi aplicada a dados meteorológicos da umidade relativa, temperatura, radiação solar, e velocidade e direção do vento para o igual período de cinco anos. As análises foram separadas em período de verão e inverno. As componentes principais extraídas para a concentração de poluentes da poluição do ar foram relacionadas a combustão da gasolina e do óleo diesel, e interações do ozônio. A principal componente dos dados meteorológicos foi relacionada às condições de seca (período do verão) e à alta velocidade dos ventos sudoeste. Este estudo não considerou a correlação existente da série temporal.

Uma caracterização espaço-temporal do ozônio troposférico no leste dos Estados Unidos foi realizada por Lehman et al. (2004). Os autores caracterizaram a concentração máxima de ozônio por oito horas, em área não urbana, do leste dos Estados Unidos para um período de 1993-2002. A análise procedeu com a seleção de um conjunto de dados de $O_3$ sendo que os valores faltantes foram preenchidos usando um esquema de interpolação espacial, com a aplicação de uma rotação na ACP para delinear regiões espaciais de concentração homogênea, e investigar o padrão temporal exibido pela concentração em cada região. Espacialmente, a análise resulta na divisão do leste dos Estados Unidos em cinco regiões, sendo que entre cada região o padrão temporal (sazonalidade, tendência e persitência) da concentração do $O_3$ foram bem diferentes. Nesse estudo, de 194 lugares que foram avaliados em 2140 dias através da usual ACP, foi identificado as 5 regiões do estudo, e com os *scores* da ACP rotacionada eles realizaram a análise temporal.

Pires et al. (2008*a*) analisaram a rede de monitoramento da qualidade do ar da área metropolitana de Oporto. Os autores verificaram que o número de lugares coletores podem ser otimizados, com objetivo de reduzir significantemente os gastos associados. Idealmente, um único local de monitoramento deve operar em uma área caracterizada com igual comportamento de poluição do ar. O principal objetivo do estudo foi avaliar a performance dos métodos estatísticos para um gerenciamento mais eficiente da rede de monitoramento da qualidade do ar. Os objetivos específicos foram: (i) identificar áreas da cidade com similar comportamento da poluição do ar; e (ii) localizar as fontes de emissão. As técnicas estatísticas empregadas no estudo foram ACP e análise de cluster (AC) e foram aplicadas à concentração de massa do dióxido de enxofre e $PM_{10}$ obtidas para o período de janeiro de 2003 a dezembro de 2005. Os principais resultados mostram que dos 10 lugares de monitoramento da rede de qualidade do ar, somente são necessários 6 para o $SO_2$ e dois para o $PM_{10}$. Os resultados também indicam que vários locais cobertos pela rede de monitoramento são caracterizados pelo mesmo comportamento de poluição do ar, sugerindo um gerenciamento inefetivo do sistema da qualidade do ar. Além disso, o estudo mostrou que somente uma fonte poluidora foi detectada para o poluente $SO_2$, e para o $PM_{10}$ três principais fontes foram detectadas. Neste artigo ACP também foi conduzida sem considerar a correlação existente nos dados.

Pires et al. (2009) utilizaram ACP para identificar medidas redundantes na rede de monitoramento da qualidade do ar da área metropolitana de Oporto. A quantidade mínima de lugares para monitoramento da qualidade do ar na área metropolitana de Oporto foi avaliada usando ACP considerando os poluentes $NO_2$, $O_3$ e $PM_{10}$ e então comparado com o estabelecido na legislação. ACP foi aplicado aos dados que correspondem aos trimestres anuais de 2003 e 2004, totalizando oito trimestres, para verificar a persistência dos resultados de ACP. O ano de 2005 foi utilizado para validar os resultados da ACP. Dois critérios para escolha do número de componentes principais foram utilizados, o critério de Kaiser (componentes principais com autovalores maiores do que 1) e o critério que representa um percentual mínimo

desejado de variabilidade explicada, onde foi escolhido mínimo de 90% de variabilidade. Os resultados indicaram discordância para os dois critérios, pois o critério de Kaiser indicou uma menor quantidade de componentes, e então adotou-se o critério de manter componentes que expliquem no mínimo 90% da variância original dos dados (VOD90). Com o critério VOD90, do total de 9 lugares, cinco foram escolhidos para o $NO_2$, três para o $O_3$ e sete para o $PM_{10}$ com intuito de caracterizar a região. O número de lugares monitorados está de acordo com o estabelecido na legislação para os poluentes $NO_2$ e $O_3$. No entanto, a região em estudo precisa de mais dois lugares de monitoramento para o poluente $PM_{10}$. Os resultados da ACP foram validados com modelos de regressão múltipla para estimação da concentração dos poluentes do ar dos lugares testes que tiveram o monitoramento removido. A boa performance obtida com os modelos mostram que os locais selecionados para monitoramento foram suficientes para inferir a concentração da poluição do ar. Os equipamentos da concentração da poluição do ar que apresentaram medidas redudantes podem ser instalados em outros locais para um melhor gerenciamento da rede de monitoramento da qualidade do ar.

Liu (2009) realizou simulações para a concentração média do $PM_{10}$ em Ta-Liao, sul de Taiwan. O estudo fez uso de um modelo de regressão considerando os erros modelos de séries temporais, e incluiu uma variável resultante da análise de componentes principais para completar a simulação de $PM_{10}$. Diferentes resultados da ACP foram introduzidos no modelo de regressão, resultando em quatro tipos de modelos: um que não considera a análise de ACP com intuito de confrontar com os outros três modelos que levam em consideração a própria cidade e (ou) cidades vizinhas. Os resultados indicam que as previsões são melhores para os modelos que fizeram uso da ACP, mas são ainda melhores quando considera as cidades vizinhas. Essas conclusões foram obtidas através das análises de medidas estatísticas que avaliam a qualidade do ajuste.

Juneng et al. (2009) realizaram um estudo espacial e temporal da concentração de $PM_{10}$ na Malasia utilizando ACP rotacionada. Os resultados sugerem que a variabilidade da concentração de $PM_{10}$ pode ser decomposta em quatro modos dominante, cada uma caracterizando diferentes variações espaciais e temporais. Neste trabalho, os valores *missing* foram preenchidos via interpolação considerando as regiões mais próximas e a ACP foi aplicada na forma tradicional, domínio do tempo, mas quando analisaram os ciclos anuais as flutuações irregulares foram tratadas via análise espectral.

Ibarra-Berastegi et al. (2008) utilizaram ACP para confirmar, independentemente da análise de cluster e *Self-Organizing Maps*, se as quatro áreas avaliadas no estudo apresentam comportamento similar ou diferente para a variabilidade espacial do $SO_2$ medido. Os resultados mostram que as três técnicas produzem os mesmos resultados, mas a informação obtida via ACP pode ajudar não somente para esse propósito, como também lançar uma luz sobre os principais mecanismos envolvidos. Este poderá ser utilizado em um futuro estágio de otimização da rede.

Guo et al. (2004) analisaram dados de compostos orgânicos voláteis (em inglês, VOCs) e monóxido de carbono (em inglês, CO) obtidos em uma área rural no leste da China para investigar a natureza das fontes emissoras e suas contribuições relativas ao meio ambiente. A ACP mostrou que as emissões por veículos e combustão de biocombustível, combustão da biomassa e emissões industriais são as principais fontes de VOCs e CO na área rural. A identificação da fontes foram analisadas usando os escores da técnica APCS, uma vertente de ACP, com um modelo de regressão linear multipla. No geral, os resultados foram bem consistentes com os inventários de emissões.

Karar & Gupta (2007) analisaram $PM_{10}$ e as concontrações das espécies químicas de massa de umA região urbana de Kolkata. Neste estudo, o modelo PCA/APCS foi aplicado na concentração de massa do $PM_{10}$ e suas espécies químicas identificando cinco possíveis fontes.

Shi et al. (2009) combinaram os métodos FMP, BQM, APC/RLM para identificação de

fontes. Devido a alta similaridade entre os perfis de fontes, vários problemas foram relacionados quando somente um modelo receptor foi aplicado. A colinearidade gera contribuições negativas no modelo BQM; e certas fontes não podem ser separadas quando é aplicado FMP ou ACP. Os resultados indicados são plausíveis e indicam que o combustível de carvão é que mais representa o percentaul de massa.

Todos artigos acima citados fizeram uso da técnica da ACP no domínio do tempo, metodologia descrita em Johnson & Wichern (1998). Em uma linha paralela ao que iremos considerar, *outliers* é um dos problemas que devemos tomar cuidado em dados que devem ser analisados por ferramentas estatísticas. Os cálculos de autovalores e autovetores são extremamente sensíveis, pois o estimador usual da matriz de covariância é afetado por essas observações. Croux & Haesbroeck (2000) utilizaram análise de componentes principais robusta para obtenção dos autovalores e autovetores via estimador robusto da matriz de covariâncias. Eles utilizaram funções de influência e variância assintótica para os estimadores robustos dos autovalores e autovetores. Os comportamentos de vários estimadores foram estimados via simulação e resultados teóricos foram apresentados. Na área da poluição do ar, Ahn & James (1999) avaliaram a deposição atmosférica seca de fósforo em uma região no Sul da Flórida. O estudo buscou detectar e remover *outliers* das medidas das taxas de fluxo de fósforo de amostras de deposição seca e intervalos de confiança foram obtidos para esse elemento. Outras pesquisas da poluição do ar também consideraram a presença de dados atípicos no conjunto de observações (ver, e.g., Cosemans et al. 2008, Jorquera et al. 2000, entre outros).

Em termos teóricos, Furrer (2005) introduziu o estimador natural da matriz de covariâncias e mostrou que sob a estrutura de dependência ele é viesado. O autor avaliou o viés assintótico de dois diferentes modelos. Para o primeiro modelo ele mostrou a taxa de convergência do viés e da covariância entre as componentes da matriz estimada da covariância. O segundo modelo assintótico serviu para derivar uma rápida e exata correção de viés. O autor também propõe um teste para verificar se o viés é significante.

Stefaniak (2009) analisou dados autocorrelacionados com ACP no domínio do tempo com objetivo em controle de processo estatístico. Através de simulações ela mostra que existe um efeito dos dados autocorrelacionados na análise de componentes principais no domínio do tempo. Sua evidência empírica considerou curta dependência.

Keller (2000) analisa variáveis psicológicas de cuidados intensivos utilizando a usual técnica de ACP, domínio do tempo, e o método de componentes principais de Brillinger (1969), domínio da frequência. No estudo foi identificado que padrões clinicamente relevantes, como outliers e mudança de nível, são melhores capturados com o método de Brillinger, além de utilizar um menor número de componentes com essa metodologia. Dentre as técnicas de ACP no domínio da frequência, destacam-se os métodos propostos por Stoffer et al. (1993) e Brillinger (1969). O método de Stoffer et al. (1993) avalia as frequências mais expressivas no periodograma e aplica o método usual de componentes principais em cada frequência em destaque. Os valores das componentes principais resultantes deste método são números complexos, o que dificulta o uso. Além disso, não foi realizado estudo de correlação entre componentes. Brillinger (1969) considera as frequências em conjunto em um coeficiente temporal e as componentes geradas pelo método tem a vantagem de apresentar coerência zero em todas as frequências.

O método proposto por Brillinger é uma versão dinâmica de ACP que considera várias defasagens temporais. O método supõe que a matriz de autocovariância seja absolutamente somável. Dados da concentração de poluentes atmosféricos apresentam diversos problemas, entre eles estrutura de memória longa. Processos com memória longa podem ser estacionários, mas sua matriz de autocovariância não é absolutamente somável. No domínio do tempo, a usual definição de memória longa é a condição $\sum_{h=0}^{\infty} |\gamma(h)| = \infty$, onde $\gamma(h)$ é a função de autocovariância no *lag* $h$ do processo, e, no domínio da frequência, essa propriedade é definida

pelo fato da densidade espectral do processo torna-se ilimitada em algumas frequências entre $[0, \pi]$. Dessa forma, estudos de ACP para séries memória longa são necessários.

O uso indevido da técnica ACP pode causar problemas em metodologias multivariadas que fazem uso dessa técnica nos procedimentos de modelagem, como, por exemplo, o modelo receptor análise fatorial estimado via ACP e o modelo receptor ACP. Essas questões corroboram para a importância dos estudos na direção da proposta desta pesquisa.

Em geral, em situações práticas, os estudos que utilizam técnicas estatísticas para observações multivariadas não consideram o pressuposto básico de independência entre observações da amostra.

De forma não diferente, o uso da técnica de ACP, domínio do tempo, aplicada em dados de séries temporais viola o princípio básico do método que é a independência entre as observações da amostra. Os dados da concentração de poluentes do ar são dados múltiplos correlacionados e registrados ao longo do tempo. Esses apresentam correlação entre diversos poluentes (dentro da amostra), entre as estações de monitoramento (espacialmente) e ao longo do tempo (entre amostras).

A técnica de ACP vem sendo amplamente utilizada sem considerar tal pressuposto, como por exemplo em Pires et al. (2008*a*,*b*), Liu (2009), Juneng et al. (2009) e Pires et al. (2009). Esta pesquisa propõe estudar ACP na presença de dados correlacionados com diferentes estruturas de dependência tais como curta e longa, nos domínios do tempo e da frequência, e aplicar a metodologia em linhas de interesse da área poluição do ar.

# Análise de componentes principais no domínio do tempo e suas implicações em dados autocorrelacionados

Bartolomeu Zamprogno, Valdério Anselmo Reisen
*DEST-CCE, PPGEA-CT – Universidade Federal do Espírito Santo*
e Neyval C. Reis Jr.
*PPGEA-CT – Universidade Federal do Espírito Santo*

7 de outubro de 2013

**Resumo**

Este artigo contribui no estudo, na análise, na interpretação e na utilização das componentes principais quando obtidas de matriz de variância e autocovariância de processos correlacionados no tempo, isto é, de séries temporais. A análise de componente principais (ACP) é uma das técnicas multivariadas mais explorada nas diversas áreas do conhecimento, pois fornece novas variáveis ortogonais para análise e interpretação da variabilidade do vetor de observação. A técnica é usualmente empregada sob a condição de replicações independentes das variáveis, suposição não observada na maioria das situações práticas, especialmente nos casos de variáveis temporalmente correlacionadas. Nesse contexto, o objetivo deste artigo é avaliar o efeito da estrutura de autocorrelação do processo vetorial sob diferentes ângulos de aplicação da ACP. A redução espúria da dimensão do espaço vetorial é uma das questões investigadas. Como as ACPs são combinações lineares das variáveis originais, as propriedades de correlação temporal dessas variáveis são transladadas para as componentes principais, transformando-as em variáveis autocorrelacionadas e com correlação cruzada. Como forma de atenuar o efeito da correlação temporal nas aplicações das ACPs, o procedimento de filtragem, por meio de modelos vetoriais autorregressivo média móvel (VARMA), é sugerido para que a técnica seja aplicada no processo ruído branco. O estudo é justificado de forma teórica e empírica e dados reais são considerados como exemplo de aplicação da pesquisa ora proposta. Resultados assintóticos dos estimadores e testes são discutidos. Os estudos empíricos corroboram com os apresentados teoricamente e fundamentam em termos práticos três problemas de interesse na área da poluição do ar: a identificação de fontes de partículas totais em supensão (PTS); o gerenciamento de rede; e o uso das ACPs em modelos de regressão. Todos os exemplos considerados são observações de contaminantes atmosféricos medidos na Região da Grande Vitória, Brasil. Entretanto, as contribuições científicas propostas neste artigo podem ser aplicada em qualquer estudo que envolve a técnica ACP.

*Palavras-chave:* análise de componentes principais, autocorrelação, correlação cruzada, autovalores.

## 1 Introdução

A técnica de análise de componentes principais (ACP) é amplamente empregada na redução da dimensão do conjunto de dados com objetivo de reduzir um conjunto de variáveis de dimensão $k$ para $m < k$ de tal maneira que o novo conjunto capte grande parte da variabilidade dos dados originais. O emprego dessa metodologia em diversos problemas se dá em todas as áreas do conhecimento. Como forma de quantificar a importância dessa técnica, Richman (1986) mostrou que entre 1983 e 1985 mais de 60 aplicações de ACP, ou técnicas similares, apareceram em periódicos meteorológicos/climatológicos. Mais recentemente, entre

*Endereço para correspondências*: Bartolomeu Zamprogno, Departamento de Estatística, CCE, UFES Av. Fernando Ferrari, s/n, Goiabeiras, Vitória ES, 29060-900 tel: 027-33352822/2481, Brasil.
E-mail: bzamprogno@yahoo.com.br (Bartolomeu Zamprogno)

os anos 1999 e 2000, 53 dos 215 artigos do *International Journal of Climatology* utilizaram alguma forma de ACP, que significa uma taxa de 25%, não alcançada por nenhuma outra técnica de estatística, Jollife (2002, pg. 71).

O emprego da técnica análise de componentes principais não se concentra apenas à redução da dimensão do conjunto de dados. Por exemplo, Karar & Gupta (2007) utilizaram ACP como uma análise de cluster para identificar fontes poluidoras e Cohen (1983), White et al. (1991) e Romero et al. (1999) fizeram uso de ACP como um cluster para identificar subregiões homogêneas de estações de variáveis climáticas em grande área geográfica. Além do uso da ACP como uma análise de cluster, diversos trabalhos utilizam a técnica para eliminar a multicolinearidade em regressão e para detectar outliers ver, por exemplo, Liu (2009). O emprego da técnica também ocorre como parte do processo de execução de outras técnicas multivariadas como a análise fatorial, a análise de correlação canônica, a análise discriminante entre outras.

Nas engenharias, em geral, ACP é um método que pode ser utilizado para descrever padrão de temperatura, de pressão e de várias outras variáveis de interesse, como a avaliação da qualidade (poluição) do corpo de água de um rio ver, por exemplo, Zimmermann et al. (2008) entre outros.

Em particular, na área da poluição do ar, as motivações para uso de ACP se dão em resoluções de diversas problemáticas. A identificação de fontes poluídoras através da ACP foi tratada por diversos estudos como, por exemplo, Statheropoulos et al. (1998), Borbon et al. (2002), Wang & Shooter (2004), Karar & Gupta (2007) e Shi et al. (2009). No contexto de gerenciamento de rede, os recentes trabalhos Pires et al. (2008*a*,*b*) utilizaram a técnica ACP nas concentrações de poluentes monitorados para gerenciar a rede de monitoramento na área Metropolitana de Oporto-MA em Portugual com o objetivo de reduzir gastos excessivos. O príncípio adotado pelos autores para atingir os objetivos propostos foi baseado na escolha de somente uma estação entre âquelas indicadas para o mesmo cluster (concentrações homogêneas). Eles concluiram que do total de 10 estações que monitoram a concentração de $SO_2$, 6 foram suficientes para medir o nível da concentração do $SO_2$. Em relação ao $PM_{10}$, não mais do que duas foram necessárias para registros da concentração desse poluente.

A técnica de ACP também é utilizada em estudos espaço-temporal, entre eles, por exemplo, Lehman et al. (2004), Ibarra-Berastegi et al. (2008) e Juneng et al. (2009). Nos trabalhos que usam modelo de regressão linear e não linear, as componentes principais são utilizadas como variáveis preditoras. Liu (2009) utilizou componentes principais como variáveis explicativas em um modelo de regressão para simular a concentração de $PM_{10}$. Em outro enfoque, a aplicação da técnica ACP para associar concentração de poluentes a números de internações de crianças e/ou idosos (estudos epidemiológicos) também é amplamente empregada ver, por exemplo, Souza (2013), Gonçalves et al. (2005), Arditsoglou & Samara (2005) e Namdeo & Bell (2005). Essas referências são algumas das diversas possibilidades de aplicação de ACP em uma única área de conhecimento, isto é, na poluição do ar.

Especificamente para identificação de fonte poluidora, ACP é utilizada para formar *clusters*. A formação do cluster, através dos maiores coeficientes de cada componente principal, é subjetiva, mas isso não diminui sua ampla utilização para esse fim. O recente artigo de Belis et al. (2013) exemplifica e quantifica a importância do uso dessa técnica no problema de identificação de fontes. Os autores afirmam que, aproximadamente, um quarto (24%) dos estudos europeus de identificação de fonte poluidora para material particulado foram baseados em ACP e suas variantes. Por exemplo, Rao (1964) mostrou que a representação bidimensional

pode dar um significado visual simples para detectar ou verificar a existência de cluster, desde que a maior parte da variação fique concentrada no subespaço bidimensional. Jollife (2002, pg. 212) aponta que se a variação entre os cluster for maior do que a variação dentro do cluster, as componentes principais muitas vezes refletem com sucesso a estrutura do cluster. Nessa direção, Ding & He (2004) provaram que a redução da dimensão obtida via ACP executa o agrupamento de dados de acordo com o método K-médias (K-means), uma técnica de agrupamentos não hierárquica.

Todas essas formas de emprego da técnica ACP no domínio do tempo a colocam como método de enorme potencial na resolução de problemas, seja de forma direta ou indireta. Uma das suposições usuais do uso da técnica ACP é baseada na independência dos dados. As componentes são combinações lineares das covariáveis. Portanto, as características dessas covariáveis serão transmitidas de forma linear para as componentes, por exemplo, a autocorrelação das covariáveis. Entretanto, um fato que chama atenção é o uso indiscriminado dessa técnica mesmo quando as variáveis são séries temporais. Por exemplo, os artigos acima citados entre tantos outros violam a suposição de independência sem apresentar justificativas para tal uso. Negligenciar a estrutura temporal das observações pode acarretar em análises e interpretações totalmente equivocadas ou, até mesmo, inviabilizar o cálculo das componentes quando o processo vetorial é não estacionário na média. Nesse contexto, a matriz de covariância não é definida.

Nessa direção, o objetivo deste artigo é avaliar o efeito de diferentes estruturas de correlação do processo vetorial estacionário $\mathbf{X}_t$ na análise, na interpretação e na inferência das componentes principais da matriz de covariância do processo $\mathbf{X}_t$. O estudo é justificado de forma teórica e empírica e dados reais são considerados como exemplo de aplicação da pesquisa ora proposta. Sob certas condições do processo $\mathbf{X}_t$, este artigo apresenta resultados teóricos que mostram que as componentes principais obtidas de dados autocorrelacionados são autocorrelacionadas e podem apresentar correlação cruzada entre elas, característica que depende da estrutura de autocorrelação presente no processo. O efeito temporal nas interpretações e inferências das ACPs pode ser atenuado por meio de filtros lineares vetoriais que é uma das contribuições deste artigo, isto é, utilizar o vetor autoregressivo média móveis (VARMA) para obter as ACPs da matriz de variância e covariancia do processo vetorial ruido branco.

Resultados assintóticos dos estimadores e testes são discutidos. Nos estudos empíricos, os resultados corroboram com os discutidos teoricamente. Esses estudos fundamentam em termos práticos três problemas de interesse na área da poluição do ar: a identificação de fontes de partículas totais em supensão (PTS); gerenciamento de rede baseada na proposta recente de Pires et al. (2008*a*,*b*); e o estudo de associar poluentes ao número de atendimentos hospitalares por causas de doenças respiratórias. Todos os exemplos considerados são observações medidas na Região da Grande Vitória, Brasil. Entretanto as contribuições científicas propostas neste artigo podem ser aplicadas em qualquer estudo que envolve a técnica ACP.

Este artigo está dividido da seguinte forma. A Seção 2 apresenta o processo e as propriedades teóricas das ACPs. Na Seção 3 são descritas propriedades dos estimadores da ACP em dados autocorrelacionados. Estudos empíricos são considerados na Seção 4. A Seção 5 discute 3 aplicações da pesquisa proposta neste artigo.

## 2 Metodologia de ACP em dados autocorrelacionados

Seja $\mathbf{X}_t = (X_{1t}, \ldots, X_{kt})^{'}$ um processo linear vetorial da forma

$$\mathbf{X}_t = \boldsymbol{\mu} + \sum_{j=0}^{\infty} \Psi(j)\boldsymbol{\xi}_{t-j}, \tag{1}$$

onde $\boldsymbol{\mu}^{'} = (\mu_1, \ldots, \mu_k)$ é o vetor de médias, $\Psi(0)$ é a matriz identidade, $\Psi(j)$ são matrizes $K \times K$ absolutamente somáveis, isto é, $(\sum_{j=0}^{\infty} |\Psi(j)| < \infty)$ e $\boldsymbol{\xi}_j^{'} = (\xi_{1j}, \ldots, \xi_{kt})$ são processos vetoriais ruído branco com média zero e matriz de covariância

$$\Gamma_{\boldsymbol{\xi}}(h) = E(\boldsymbol{\xi}_j \boldsymbol{\xi}_{j+h}^{'}) = \begin{cases} \Sigma, & h = 0 \\ 0, & h \neq 0 \end{cases} \tag{2}$$

onde $\Sigma$ é uma matriz não negativa definida. Observa-se pela Equação 2 que os elementos do vetor $\boldsymbol{\xi}_j$, em diferentes tempos, não apresentam correlação cruzada, isto é, $E(\boldsymbol{\xi}_j \boldsymbol{\xi}_{j+h}^{'}) = 0$, $\forall\ h \neq 0$. O processo $\mathbf{X}_t$ é estacionário de segunda ordem e tem matriz de covariância

$$\Gamma_{\mathbf{X}}(h) = \sum_{j=-\infty}^{\infty} \Psi(j+h)\Sigma\Psi(j)^{'},$$

onde o $ij$-ésimo elemento da matriz $\Gamma_{\mathbf{X}}(h)$ é dado por $\gamma_{ij}(h) = E[(X_{t+h,i} - \mu_i)(X_{t,j} - \mu_j)]$, $i, j = 1, \ldots, k$. Equivalentemente, $\Gamma_{\mathbf{X}}(h)$ é absolutamente somável desde que individualmente cada um dos seus elementos formam um sequência absolutamente somável, Hamilton (1994, pg. 262). O processo $\mathbf{X}_t$ tem matriz densidade espectral dada por

$$f(\omega) = \frac{1}{2\pi}\Psi(e^{i\omega})\Sigma\Psi(e^{-i\omega})^{'}, \quad -\pi \leq \omega \leq \pi.$$

Maiores detalhes ver, por exemplo, Reinsel (1997, pg. 85).

Uma classe paramétrica de modelos de séries temporais pertencente ao processo linear definido em 1 é o vetor autorregressivo média móveis (VARMA) que é a solução do sistema

$$\Phi(B)(\mathbf{X}_t - \mu) = \triangle^d(B)\Theta(B)\boldsymbol{\varepsilon}_t, \tag{3}$$

onde $B$ é o operador atraso, $\triangle^d(B) = \text{diag}\left\{(1-B)^{d_1}, (1-B)^{d_2}, \ldots, (1-B)^{d_k}\right\}$, $\boldsymbol{\mu}$ é o vetor de médias e $\boldsymbol{\varepsilon}_t$ é o ruído branco com $\text{E}(\boldsymbol{\varepsilon}_t) = 0$ e $Var(\boldsymbol{\varepsilon}_t) = \Sigma$. Os operadores $\Phi(B) = I - \sum_{i=1}^{p} \Phi_i B^i$ e $\Theta(B) = I + \sum_{i=1}^{q} \Theta_i B^i$ são matrizes polinomiais com ordem $p, q$ respectivamente, $d \in \mathbb{N}$ e $I$ é a matriz identidade de dimensão $k \times k$ e $\Phi_i$ e $\Theta_i$ são matrizes $k \times k$ de constantes.

No espaço de dados multivariados, análise de componentes principais busca as direções que capturam o maior percentual de variação dos dados mensurados. Essa técnica depende exclusivamente da matriz de covariâncias dos dados; ver, por exemplo, Anderson (2003) e Johnson & Wichern (1998). A técnica de ACP é baseada na teoria algébrica de vetores e, usualmente, aplicada a dados não autocorrelacionados, isto é, processos com $\Gamma(h) = 0$, $\forall\ h \neq 0$. A técnica consiste em obter da matriz de covariâncias $\Gamma(0)$, os autovalores $(\lambda_1, \ldots, \lambda_k)$ e os correspondentes autovetores $ß = (\boldsymbol{\beta}_1^{'}, \ldots, \boldsymbol{\beta}_k^{'})^{'}$. O vetor de componentes é dado por $\mathbf{Y} = ß\mathbf{X}$, onde $Y_i = \boldsymbol{\beta}_i^{'}\mathbf{X}$, $i = 1, \ldots, k$, $\boldsymbol{\beta}_i^{'} = (\beta_{i1}, \ldots, \beta_{ik})$ é um autovetor da matriz de covariância $\Gamma(0)$, e a variabilidade explicada de cada componente principal $Y_i$ é dada pelo

autovalor $\lambda_i$, associado a $\boldsymbol{\beta}_i$. Dessa maneira, o cálculo e estudo da ACP, baseado em matrizes de covariâncias, incorpora conceitos estatísticos na teoria usual de autovalores e autovetores.

A teoria inferencial de ACP foi estendida para processos autocorrelacionados que satisfazem as suposições do processo $\mathbf{X}_t$ ( Equação 1). Esses resultados foram primeiramente publicados por Taniguchi & Krishnaiah (1987). O Teorema 1 proposto pelos autores mostra que a distribuição assintótica dos autovalores e autovetores depende da caracterização do modelo, isto é, da estrutura da matriz de autocorrelação do processo. A Seção 3.2 deste artigo sumariza os resultados do Teorema 1 de Taniguchi & Krishnaiah (1987), pois é uma importante contribuição para a inferência estatística multivariada. No entanto, o cálculo da distribuição assintótica dos autovalores e autovetores não é simples, o que torna complicado o uso dessa teoria em situações práticas. Como as componentes $Y_i$ são combinações lineares dos autovetores $(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k)$ e de covariáveis que são dependentes do tempo, isto é, são autocorrelacionadas, essa propriedade também é transladada para $Y_i$, $i = 1, \ldots, k$. Portanto, a componente principal torna-se um processo indexado no tempo $t$, isto é, $Y_i = Y_{it}$, $t \in \mathbb{Z}$.

O Teorema 1 em Taniguchi & Krishnaiah (1987) não aborda a propriedade temporal de $Y_i$ e essa problemática é um dos resultados formalizados na Proposição 1 abaixo [itens (c) e (d)].

**Proposição 1.** *Seja* $\mathbf{X}_t^{'} = [X_{1t}, X_{2t}, \ldots, X_{kt}]$ *o processo definido em 1 e considere os pares de autovalores e autovetores* $(\lambda_1, \boldsymbol{\beta}_1), (\lambda_2, \boldsymbol{\beta}_2), \ldots, (\lambda_k, \boldsymbol{\beta}_k)$ *da matriz de covariância* $\Gamma_{\mathbf{X}}(0)$, *onde* $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k \geq 0$. *Então, a i-ésima componente principal é dada por*

$$Y_{it} = \boldsymbol{\beta}_i^{'}\mathbf{X} = \boldsymbol{\beta}_{i1}X_{1t} + \boldsymbol{\beta}_{i2}X_{2t} + \cdots + \boldsymbol{\beta}_{ik}X_{kt}, \quad i = 1, 2, \ldots, k, \quad t \in \mathbb{Z},$$

*onde:*

*a)* $Var(Y_{it}) = \boldsymbol{\beta}_i^{'}Cov(X_t, X_t^{'})\boldsymbol{\beta}_j = \boldsymbol{\beta}_i^{'}\Gamma_{\mathbf{X}}(0)\boldsymbol{\beta}_i = \lambda_i, \quad i = 1, 2, \ldots, k,$

*b)* $Cov(Y_{it}, Y_{jt}) = \boldsymbol{\beta}_i^{'}Cov(X_t, X_t^{'})\boldsymbol{\beta}_j = \boldsymbol{\beta}_i^{'}\Gamma_{\mathbf{X}}(0)\boldsymbol{\beta}_j = 0, \quad i \neq j,$

*e* $\forall\ h \neq 0,$

*c)* $Cov(Y_{it}, Y_{it+h}) = \boldsymbol{\beta}_i^{'}Cov(X_t, X_{t+h}^{'})\boldsymbol{\beta}_j = \boldsymbol{\beta}_i^{'}\Gamma_{\mathbf{X}}(h)\boldsymbol{\beta}_i, \quad i = 1, 2, \ldots, k,$

*d)* $Cov(Y_{t,i}, Y_{jt+h}) = \boldsymbol{\beta}_i^{'}Cov(X_t, X_{t+h}^{'})\boldsymbol{\beta}_j = \boldsymbol{\beta}_i^{'}\Gamma_{\mathbf{X}}(h)\boldsymbol{\beta}_j, \quad i \neq j.$

*Observação: No caso de multiplicidade de alguns* $\lambda_i$, *o vetor de coeficientes* $\boldsymbol{\beta}_i$, *e consequentemente* $Y_i$, *não são únicos.*

*Demonstração.* Suponha que $A$ seja uma matriz de dimensão $n \times n$ e não negativa definida. Então, $A$ pode ser escrita da forma $A = P\Lambda P^{'}$, onde $P$ é uma matriz ortogonal ($P^{'} = P^{-1}$) e $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$, isto é, $\Lambda$ é matriz diagonal, onde $\lambda_1, \ldots, \lambda_n$ são autovalores (todos não negativos) de $A$ (Teorema da decomposição espectral ver, por exemplo, proposição 1.6.3, Brockwell & Davis (1987, pg. 33)).

No processo $\mathbf{X}_t$ definido em 1, $\Gamma_{\mathbf{X}}(0)$ é matriz de covariância e possui a propriedade não negativa definida [mais detalhes em Lütkepohl (2005, pg. 30) e Reinsel (1997, pg. 3)]. Portanto, $\exists\ P$ tal que $\Gamma_{\mathbf{X}}(0) = P\Lambda P^{'}$. Seja ß $= (\boldsymbol{\beta}_1^{'}, \ldots, \boldsymbol{\beta}_k^{'})$ a matriz de autovetores de $\Gamma_{\mathbf{X}}(0)$ e $\mathbf{Y}_t$ o vetor das componentes principais obtida de $\Gamma_{\mathbf{X}}(0)$, isto é, $\mathbf{Y}_t =$ ß$\mathbf{X}$. Logo,

$$\Gamma_Y(h) = E(\mathbf{Y}_t \mathbf{Y}_{t+h}^{'}) = E(ß^{'} \mathbf{X}_t \mathbf{X}_{t+h}^{'} ß) = ß^{'} \Gamma_{\mathbf{X}}(h) ß. \quad (4)$$

Para $h = 0$, temos

$$\Gamma_Y(0) = E(\mathbf{Y}_t \mathbf{Y}_t^{'}) = E(ß^{'} \mathbf{X}_t \mathbf{X}_t^{'} ß) = ß^{'} \Gamma_{\mathbf{X}}(0) ß = \Lambda, \quad (5)$$

onde $\Lambda = \text{diag}\{\lambda_1, \ldots, \lambda_k\}$ é uma matriz diagonal com autovalores de $\Gamma_{\mathbf{X}}(0)$.

Portanto, os itens (a) e (b) são obtidos diretamente da equação 5 e os itens c) e d) da equação 4.

□

Os resultados (a) e (b) da Proposição 1 são equivalentes aos apresentados em Anderson (2003) e Johnson & Wichern (1998) para o caso $\mathbf{X}_t = \xi_t$ (dados multivariados não autocorrelacionados). Os itens (c) e (d) mostram, respectivamente, autocorrelação e correlação cruzada das componentes principais. Essas evidenciam o problema em usar ACP em séries temporais. Isto é, por (c) nota-se que $\mathbf{Y}_t$ é uma série temporal e, portanto, o uso de $\mathbf{Y}_t$ em qualquer contexto de modelagem não pode ignorar a estrutura de correlação temporal. No caso do item (d), esse torna-se o maior problema nos procedimentos descritivos e inferenciais da ACP, pois as componentes principais são correlacionadas para qualquer defasagem $h \neq 0$. Essas questões são estudadas empiricamente e os resultados e discussões encontram-se na Seção 4 deste artigo.

**Observação 1.** *Para o processo definido em 1, isto é, sejam* $(\lambda_1^*, \lambda_2^*, \ldots, \lambda_k^*)$ *autovalores de* $\Gamma_{\boldsymbol{\xi}}(0) = \Sigma$ *(processo ruído branco) e* $(\lambda_1, \lambda_2, \ldots, \lambda_k)$ *autovalores de* $\Gamma_{\mathbf{X}}(0)$*, onde* $\mathbf{X}_t = \sum_{j=0}^{\infty} \Psi_j \xi_{t-j}$*. Portanto,* $\sum_{j=0}^{k} \lambda_i \geq \sum_{j=0}^{k} \lambda_i^*$*, pois* $Var(\mathbf{X}_t) \geq Var(\boldsymbol{\xi}_t)$*, ou seja,* $Var(X_i) \geq Var(\xi_i)$ $\forall i = 1, \ldots, k$*. Isto é, a soma das variâncias das componentes obtidas de* $\Gamma_{\mathbf{X}}(0)$ *é maior do que a soma das variâncias das componentes obtidas de* $\Gamma_{\boldsymbol{\xi}}(0)$*. Essse resultado mostra que as componentes principais de* $\mathbf{X}_t$ *possuem maior variabilidade que as de* $\boldsymbol{\xi}_t$*.*

O modelo VAR(1) é um caso particular do Modelo 3, com $p = 1$ e $q = 0$, e é usualmente utilizado no contexto de modelagens multivariadas, devido a propriedade de parsimônia desse modelo. Isso motiva a apresentação da Proposição 2, que é um caso particular da Proposição 1.

**Proposição 2.** *Suponha que* $\mathbf{X}_t$ *seja um processo estacionário VAR(1) com coeficiente matricial* $\Phi$ *de dimensão* $(k \times k)$*, isto é,* $\mathbf{X}_t = \Phi \mathbf{X}_{t-1} + \xi_t = \sum_{j=0}^{\infty} \Phi^j \boldsymbol{\xi}_{t-j}$*, e* $\mathbf{Y}_t = ß^{'} \mathbf{X}_t$*, onde* $ß$ *é a matriz de autovetores de* $\Gamma_{\mathbf{X}}(0)$ *e* $\Lambda$ *a matriz diagonal com os autovalores de* $\Gamma_{\mathbf{X}}(0)$*. Considere a matriz de covariâncias de* $\mathbf{X}_t$*,* $\Gamma_{\mathbf{X}}(l) = \sum_{j=0}^{\infty} \Phi^j \Sigma \Phi^{'l+j}$*. Então,*

$$\Gamma_{\mathbf{X}}(0) = \sum_{j=0}^{\infty} \Phi^j \Sigma \Phi^{'j} \quad e \quad \Gamma_{\mathbf{X}}(l) = \Gamma_{\mathbf{X}}(0) \Phi^{'l}.$$

*Portanto,*

$$\Gamma_{\mathbf{Y}}(l) = \Lambda ß^{'} \Phi^l ß.$$

*Demonstração.* Com $l = 0$ em $\Gamma_{\mathbf{X}}(l)$, obtem-se $\Gamma_{\mathbf{X}}(0) = \sum_{j=0}^{\infty} \Phi^j \Sigma \Phi'^j$. De forma recusirva, pode-se facilmente mostrar que, para qualquer $l \geq 0$, $\Gamma_{\mathbf{X}}(l) = \Gamma_{\mathbf{X}}(0)\Phi'^l$.

Logo,

$$\Gamma_{\mathbf{Y}}(l) = ß'\Gamma_{\mathbf{X}}(l)ß = ß'\Gamma_{\mathbf{X}}(0)\Phi'^l ß = ß'ßΛß'\Phi^l ß = Λß'\Phi^l ß.$$

□

A Proposição 2 mostra de forma clara o efeito direto da estrutura de correlação temporal no cálculo da covariância de $\mathbf{Y}_t$, isto é, $\Gamma_{\mathbf{Y}}(l)$ depende de forma direta de $\Phi$.

As proposições acima acarretam nas seguintes observações.

**Observação 2.** *Considere o caso especial da proposição 2 com $\Phi = diag(\phi_1, \ldots, \phi_k)$, onde $\phi_i = \phi$, $\forall\ i = 1, \ldots, k$. Então, $\Gamma_{\mathbf{Y}}(l) = \Lambda\phi^l = diag(\lambda_1\phi^l, \ldots, \lambda_k\phi^l)$. Nota-se que a autocorrelação de $\mathbf{Y}_t$ decai exponencialmente. Em particular, se $\phi \to 1$, isto é, processo com raiz unitária, resultará em um processo não estacionário, o que inviabiliza o cálculo da ACP. Portanto, atenção deve ser dada no caso de usar ACP em séries temporais, mesmo para finalidade descritiva. Isso contesta a discussão em Jollife (2002, pg. 299) onde o autor relaxa essa questão.*

**Observação 3.** *O processo VAR(p) pode ser escrito na forma VAR(1) ver, por exemplo, Lütkepohl (2005, pg. 15) e Hamilton (1994, pg. 259). Portanto, a proposição 2 pode ser facilmente estendida para qualquer VAR(p).*

Como forma de ilustrar a propriedade da observação 1, o modelo VAR(1) com $k = 2$ é comparado ao processo ruído branco.

Seja $\boldsymbol{\varepsilon}_t = \{\varepsilon_{1t}, \varepsilon_{2t}\}'$ um processo ruído branco bi-dimensional com matriz de covariância

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}.$$

O polinômio característico dessa matriz é dado por

$$\lambda^2 - (\sigma_1^2 + \sigma_2^2)\lambda + \sigma_1^2\sigma_2^2 - \sigma_{12}^2. \tag{6}$$

Logo, pela relação de Girard, a soma das raízes do polinômio 6 é $S_\Sigma = \sigma_1^2 + \sigma_2^2$.

Seja $\mathbf{W}_t = \Phi\mathbf{W}_{t-1} + \boldsymbol{\varepsilon}_t$ um processo AR(1), bi-dimensional estacionário, onde

$$\Phi = \begin{bmatrix} \phi_1 & 0 \\ 0 & \phi_1^* \end{bmatrix}.$$

A matriz de covariância de $\mathbf{W}$, obtida pela resolução do sistema $\Gamma(0) = \Phi\Gamma(0)\Phi' + \Sigma$ [ver, por exemplo, Reisel (1997)], é

$$\Gamma_{\mathbf{W}}(0) = \begin{bmatrix} \frac{\sigma_1^2}{1-\phi_1^2} & \frac{\sigma_{12}}{1-\phi_1\phi_1^*} \\ \frac{\sigma_{12}}{1-\phi_1\phi_1^*} & \frac{\sigma_2^2}{1-\phi_1^{*2}} \end{bmatrix}. \tag{7}$$

O polinômio característico de $\Gamma_{\mathbf{W}}(0)$ é dado por

$$\lambda^2 - \Big[\frac{\sigma_1^2}{1-\phi_1^2} + \frac{\sigma_2^2}{1-\phi_1^{*2}}\Big]\lambda + \frac{\sigma_1^2\sigma_2^2}{(1-\phi_1^2)(1-\phi_1^{*2})} - \frac{\sigma_{12}^2}{(1-\phi_1\phi_1^*)^2}.$$

A soma das raízes do polinômio característico de 7 é $S_{\Gamma_{\mathbf{W}}(0)} = \frac{\sigma_1^2}{1-\phi_1^2} + \frac{\sigma_2^2}{1-\phi_1^{*2}}$. Portanto, nota-se que isso $S_{\Gamma_{\mathbf{W}}(0)} > S_\Sigma$, que exemplifica o aumento da variabilidade das componentes principais em função da introdução da correlação temporal no processo. Observe que quando $\phi_1$ e $\phi_1^* \to 1$ a variância torna-se infinita.

# 3 Propriedades inferenciais de ACP

## 3.1 Estimação de $\Gamma(h)$

É conhecido, no caso univariado, que os estimadores clássicos das funções de autocovariância e autocorrelação são viciados, e o vício depende da defasagem $h$, do tamanho amostral $n$, e também da estrutura de correlação do processo ver, por exemplo, Priestley (1983). Essa propriedade é também observada para o caso multivariado $\Gamma_{\mathbf{X}}(h)$. Estimar erroneamente a matriz de covariância afeta diretamente as propriedades estatísticas dos autovalores e autovetores de $\Gamma_{\mathbf{X}}(0)$.

Seja $\{\mathbf{X}_1, \dots \mathbf{X}_n\}$ uma amostra do processo $\mathbf{X}_t$ definido em 1. O estimador da matriz de covariância no *lag* $h$ é dado pela matriz de covariância amostral

$$\hat{\Gamma}_{\mathbf{X}}(h) = \frac{1}{n}\sum_{j=1}^{n-h}(\mathbf{X}_j - \bar{\mathbf{X}})(\mathbf{X}_{j+h} - \bar{\mathbf{X}})', \quad h = 0, 1, 2, \dots, \tag{8}$$

onde $\bar{\mathbf{X}} = (\bar{X}_1, \dots, \bar{X}_k)' = n^{-1}\sum_{t=1}^{n}\mathbf{X}_t$ é o vetor de médias amostrais, com $\hat{\Gamma}_{\mathbf{X}}(-h) = \hat{\Gamma}_{\mathbf{X}}(h)'$. Em particular, $\hat{\Gamma}_{\mathbf{X}}(0) = n^{-1}\sum_{j=1}^{n}(\mathbf{X}_j - \bar{\mathbf{X}})(\mathbf{X}_j - \bar{\mathbf{X}})'$ é a matriz de covariância amostral de $\mathbf{X}_t$.

A seguinte proposição mostra o vício de $\hat{\Gamma}_{\mathbf{X}}(h)$ e a demonstração pode ser derivada algebricamente do Teorema 2.1 de Furrer (2005).

**Proposição 3.** *Sejam $\{\mathbf{X}_1, \dots \mathbf{X}_n\}$ uma amostra do processo $\mathbf{X}_t$ definido em 1 com matriz de covariância $\Gamma_{\mathbf{X}}(h)$ e $\hat{\Gamma}_{\mathbf{X}}(h)$ definido em 8. $\hat{\Gamma}_{\mathbf{X}}(h)$ é estimador viciado de $\Gamma_{\mathbf{X}}(h)$.*

$$\begin{aligned} E(\hat{\Gamma}_{\mathbf{X}}(h)) = \frac{n-h}{n}\Gamma_{\mathbf{X}}(h) - \frac{1}{n^2}\sum_{i=1}^{n-h}\sum_{j=1}^{n}\Gamma_{\mathbf{X}}(j-i) \\ - \frac{1}{n^2}\sum_{i=1}^{n-h}\sum_{j=1}^{n}\Gamma_{\mathbf{X}}(i-j+h) + \frac{n-h}{n^3}\sum_{l=-(n-1)}^{n-1}(n- \mid l \mid)\Gamma_{\mathbf{X}}(l). \end{aligned}$$

Pode-se provar que, para fixo $h$, $\lim_{n\to\infty}E(\hat{\Gamma}(h) = \Gamma(h)$, isto é, $\hat{\Gamma}_{\mathbf{X}}(h)$ é um estimador assintoticamente não viciado de $\Gamma_{\mathbf{X}}(h)$ [ver, por exemplo, Hannan (1970, cap. 4) e Hannan & Deistler (1988, sec. 4.1)].

### 3.2 Distribuição assintótica dos autovalores e autovetores

Sejam $l_1 \geq \ldots \geq l_k$ autovalores de $\hat{\Gamma}_{\mathbf{X}}(0)$, isto é , $l_i$ é o estimador de $\lambda_i$. Como $\hat{\Gamma}_{\mathbf{X}}(0)$ é uma matriz simétrica, pode-se expressar $\hat{\Gamma}_{\mathbf{X}}(0) = BLB^{'}$, onde $L = \text{diag}(l_1, \ldots, l_k)$ e $B$ uma matriz ortogonal. Suponha que os autovalores de $\Gamma_{\mathbf{X}}(0)$ satisfazem $\lambda_1 > \ldots > \lambda_k$, e que $\Gamma_{\mathbf{X}}(0) = \boldsymbol{\beta}\Lambda\boldsymbol{\beta}^{'}$, onde $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_k)$ e ß $= (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k)$ é a matriz ortogonal. Defina $G = \sqrt{n}(B - ß)$ e a matriz diagonal $D = \sqrt{n}(L - \Lambda)$.

Teorema 1 de Taniguchi & Krishnaiah (1987) mostra que, sob certas suposições, os autovalores e autovetores do estimador de $\Gamma_{\mathbf{X}}(0)$ seguem distribuição assintótica normal. Em particular, se o processo 1 tem distribuição normal, então as distribuições limites de $D = (d_1, \ldots, d_k)$ e $G = (g_1, \ldots, g_k)$ são normais com $D$ e $G$ independentes e os elementos diagonal de $D$ são independentes. Os elementos da diagonal $d_i$ de $D$ têm distribuição limite

$$N\Big(0, 4\pi \int_{-\pi}^{\pi} g_{ii}(w)^2 d\omega\Big),$$

onde $g_{ij}(\omega) = \tilde{f}_{ii}(\omega)\delta(i,j)$, $\delta(i,j)$ é o delta de Kronecker, $\tilde{f}_{ii}(\omega) = (\Sigma_{ii}/2\pi)|\tilde{k}_{ii}(\omega)|^2$, $\tilde{k}_{ii}(\omega) = \sum_{i=0}^{\infty} \Psi_{ii}(l)e^{il\omega}$. A matriz de covariância do $i$-ésimo autovetor é

$$\text{Var}(g_i) = \sum_{j=1,\ j\neq i}^{k} \frac{2\pi}{(\lambda_i - \lambda_j)^2} \int_{-\pi}^{\pi} g_{ii}(\omega) g_{jj}(\omega) d\omega \boldsymbol{\beta}_j \boldsymbol{\beta}_j^{'},$$

e a matriz de covariância de $g_i$ e $g_j$ na distribuição limite é

$$\text{Cov}(g_i, g_j) = -\frac{2\pi}{(\lambda_i - \lambda_j)^2} \int_{-\pi}^{\pi} g_{ii}(\omega) g_{jj}(\omega) d\omega \boldsymbol{\beta}_j \boldsymbol{\beta}_i^{'}.$$

**Observação 4.** *O teorema 13.5.1 em Anderson (2003, cap. 13), dados independentes, é um caso particular do apresentado acima [ver Corolário 3.1 de Taniguchi & Krishnaiah (1987)]. Nesse contexto, o elemento diagonal $d_i$ de $D$ tem distribuição limite $N(0, 2\lambda_i^2)$. A matriz de covariância de $g_i$ da distribuição limite de $G$ é dada por*

$$Var(g_i) = \sum_{k=1,\ k\neq i}^{p} \frac{\lambda_i \lambda_k}{(\lambda_i - \lambda_k)^2} \boldsymbol{\beta}_k \boldsymbol{\beta}_k^{'},$$

*onde $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)$, e a matriz de covariância de $g_i$ e $g_j$ na distribuição limite é*

$$Cov(g_i, g_j) = -\frac{\lambda_i \lambda_j}{(\lambda_i - \lambda_j)^2} \boldsymbol{\beta}_j \boldsymbol{\beta}_i^{'}, \quad i \neq j.$$

Em base a teoria assintótica dos autovalores e autovetores descritos no caso da Observação 4, testes estatísticos são sugeridos na literatura para tomadas de decisões através da análise de componentes principais. Nesta pesquisa, serão considerados dois testes de hipóteses, um para testar autovalor individualmente (Anderson 2003, cap. 13) e outro que considera a proporção de variação total, Fujikoshi (1980).

Sob a hipótese nula $H_0 : \lambda_i = \lambda_i^0$, a distribuição assintótica dos autovalores $l_i$ para dados independentes segue distribuição normal com média $\lambda_i$ e variância $2\lambda_i^2/n$, veja Observação 4. Como $l_i$ é um estimador consistente para $\lambda_i$, a distribuição limite de

$$\sqrt{n}\frac{l_i - \lambda_i}{\sqrt{2}l_i}$$

é $N(0,1)$. O teste bilateral $\lambda_i = \lambda_i^0$ tem região de aceitação

$$-z_{\frac{\alpha}{2}} \leq \sqrt{\frac{n}{2}}\frac{l_i - \lambda_i^0}{\sqrt{2}l_i^0} \leq z_{\frac{\alpha}{2}},$$

onde $z_{\frac{\alpha}{2}}$ é obtido da distribuição $N(0,1)$.

Um problema importante em ACP é saber quantas componentes principais devem ser escolhidas. Vários critérios são propostos na literatura. Entre eles destacam-se o gráfico de autovalores Jollife (2002), o método baseado no autovalor médio (Perez-Neto et al. 2005), o teste de hipóteses de igualdade dos últimos autovalores (Ferreira 2008) e o teste de Fujikoshi (1980), esse último sumarizado a seguir.

O teste de Fujikoshi (1980) não exige que a distribuição dos dados independentes seja normal multivariada. Seja $\tau_m = \frac{\sum_{i=1}^m \lambda_i}{\sum_{i=1}^k \lambda_i}$, $1 \leq m \leq k$, a proporção da variabilidade explicada pelas $m$ primeiras componentes principais populacionais. Sob $H_0 : \tau_m \geq \tau_{m0}$, a estatística de teste

$$Z = \frac{R_m - \tau_{m0}}{\frac{\vartheta}{\sqrt{n}}},$$

tem distribuição assintótica $N(0,1)$, onde $R_m = \frac{\sum_{i=1}^m l_i}{\sum_{i=1}^k l_i}$. O teste é uma derivação do resultado $\sqrt{n}(R_m - \tau_m) \sim N(0, \vartheta^2)$, onde

$$\vartheta^2 = \frac{2\text{tr}(\Sigma^2)}{\text{tr}(\Sigma)^2}[(\rho_m^2)^2 - 2\eta\rho_m^2 + \eta] \quad \text{e} \quad \eta = \frac{\sum_{i=1}^m \lambda_i^2}{\sum_{i=1}^k \lambda_i^2},$$

sendo tr o traço da matriz. O estimador para $\vartheta$ é dado por

$$\hat{\vartheta} = \sqrt{\frac{2\text{tr}(\hat{\Gamma}(0)^2)}{\text{tr}(\hat{\Gamma}(0))^2}[(\rho_{m0}^2)^2 - 2\hat{\eta}\rho_{m0}^2 + \hat{\eta}]}, \text{ onde } \quad \hat{\eta} = \frac{\sum_{i=1}^m l_i^2}{\sum_{i=1}^k l_i^2}.$$

# 4 Simulações

Com o objetivo de quantificar e exemplificar o efeito das autocorrelações nas estimativas, nas interpretações e nos testes dos autovalores da ACP, esta seção divide os estudos em dois casos. Primeiramente, propriedades discutidas anteriormente são checadas somente por meio das análises matriciais, isto é, sem invocar processos empíricos. O segundo estudo elucida as propriedades estatísticas (inferências) através de simulações de Monte Carlo.

Para ambos casos, seja $\mathbf{X}_t$ processo gerado de acordo com o modelo VAR(1), com $\mathbf{X}_t = (X_{1t}, \ldots, X_{4t})'$. A matriz de covariância de $\boldsymbol{\xi}_t$ é dada por

$$E(\boldsymbol{\xi}_t\boldsymbol{\xi}_t') = \begin{bmatrix} 127.4089 & 30.5878 & 47.4390 & 62.4214 \\ 30.5878 & 58.7881 & 33.8929 & 70.6551 \\ 47.4390 & 33.8929 & 64.1786 & 58.4933 \\ 62.4214 & 70.6551 & 58.4933 & 172.2163 \end{bmatrix}.$$

Os valores dos coeficientes do modelo VAR(1) são dados na Tabela 1. Para efeito de comparação, os resultados de ACP aplicados a $\mathbf{X}_t$ são comparados com os do vetor ruído branco $\boldsymbol{\xi}_t$, denotado por Modelo 1.

Como já mencionado, o estudo desta seção tem como objetivo elucidar o efeito das autocorrelações em ACP. Nessa direção, os Modelos 2, 3, 4 e 5 (Tabela 1) usados no experimento apresentam matrizes de coeficientes com diferentes graus de correlação positiva, isto é, pequenas, médias e fortes forças de correlação temporal. O Modelo 2 é um processo que não carrega efeito temporal fora da diagonal principal, ou seja, não há efeito de autocorrelação cruzada temporal presente nos dados. Em contrapartida ao Modelo 2, o Modelo 3 apresenta estrutura fraca de autocorrelação e também considera uma fraca correlação cruzada entre as séries, pois o valor máximo de $\phi_{ij}$ é 0.12, $\forall\ i, j$. Os Modelos 4 e 5 consideram uma estrutura de autocorrelação mais complexa, com destaque para o Modelo 5 que tem forte efeito de correlação entre variáveis, por exemplo, $\phi_{32} = 0.8$.

Tabela 1: Matrizes de $\Phi$ para os processos VAR(1).

| $\Phi_1$ (**Modelo 2**) | | | | $\Phi_1$ (**Modelo 3**) | | | |
|---|---|---|---|---|---|---|---|
| 0.3 | 0.0 | 0.0 | 0.0 | 0.12 | 0.00 | 0.03 | 0.00 |
| 0.0 | 0.3 | 0.0 | 0.0 | 0.05 | 0.08 | 0.00 | 0.01 |
| 0.0 | 0.0 | 0.3 | 0.0 | 0.00 | 0.00 | 0.10 | 0.00 |
| 0.0 | 0.0 | 0.0 | 0.3 | 0.01 | 0.02 | 0.00 | 0.05 |
| $\Phi_1$ ( **Modelo 4**) | | | | $\Phi_1$ ( **Modelo 5**) | | | |
| 0.2 | 0.0 | 0.6 | 0.1 | 0.6 | 0.3 | 0.0 | 0.3 |
| 0.0 | 0.3 | 0.0 | 0.0 | 0.1 | 0.2 | 0.0 | 0.1 |
| 0.2 | 0.0 | 0.5 | 0.0 | 0.1 | 0.8 | 0.4 | 0.2 |
| 0.0 | 0.0 | 0.0 | 0.4 | 0.2 | 0.0 | 0.2 | 0.5 |

As matrizes de covariâncias de cada modelo VAR(1) estão apresentadas na tabela 2. Como esperado, o aumento da correlação no VAR(1) causa o aumento das autocovariâncias, o que confirma o resultado da Observação 1.

Tabela 2: $\Gamma(0)$ dos modeles VAR(1) considerados no estudo.

| $\Gamma(0)$ (**Modelo 2**) | | | | $\Gamma(0)$ (**Modelo 3**) | | | |
|---|---|---|---|---|---|---|---|
| 140.01 | 33.61 | 52.13 | 68.59 | 129.68 | 31.92 | 48.21 | 63.15 |
| 33.61 | 64.60 | 37.24 | 77.64 | 31.92 | 59.95 | 34.47 | 71.42 |
| 52.13 | 37.24 | 70.52 | 64.27 | 48.21 | 34.47 | 64.83 | 58.90 |
| 68.59 | 77.64 | 64.27 | 189.25 | 63.15 | 71.42 | 58.90 | 172.90 |
| $\Gamma(0)$ ( **Modelo 4**) | | | | $\Gamma(0)$ ( **Modelo 5**) | | | |
| 232.17 | 43.32 | 131.75 | 98.40 | 1029.30 | 264.14 | 858.22 | 833.75 |
| 43.32 | 64.60 | 42.93 | 80.29 | 264.14 | 120.50 | 249.43 | 272.21 |
| 131.75 | 42.93 | 133.09 | 82.95 | 858.22 | 249.43 | 832.23 | 766.16 |
| 98.40 | 80.29 | 82.95 | 205.02 | 833.75 | 272.21 | 766.16 | 847.09 |

A Tabela 3 apresenta os autovalores dos Modelos 1, 2, 3, 4 e 5 com os respectivos percentuais de variabilidade $\frac{\lambda_i}{\sum_{j=1}^{k}\lambda_j}$, onde $\lambda_i$ é autovalor de $\Gamma_{\mathbf{X}}(0)$. Observa-se que os autovalores do

Modelo 1 e do Modelo 3 são bem similares, o que é justificado pelo fato do Modelo 3 apresentar pequena correlação fora da diagonal. Consequentemente, os percentuais de variabilidade dos autovalores são também bastante próximos. O Modelo 2 tem coeficientes somente na diagonal, mas de valores superiores ao Modelo 3 e, portanto, de maior variabilidade. Isso acarreta no aumento no valor dos autovalores do modelo, mas os percentuais são equivalentes. A similaridade entre os percentuais de variabilidade dos modelos 1 e 2 é um resultado esperado. Por exemplo, com $\phi_1 = \phi_1^*$ na matriz $\Gamma_{\mathbf{W}}(0)$ (matriz 7), resulta $\Gamma_{\mathbf{W}}(0) = \frac{1}{1-\phi_1^2}\Gamma_{\boldsymbol{\xi}}(0)$. Logo, os autovalores são equivalentes. Note que se a matriz de autocorrelação é usada ao invés da matriz de covariância, os valores dos $\lambda_i$, $i = 1, \ldots, k$, nos modelos 1 e 2 serão os mesmos. Os Modelos 4 e 5 mostram claramente o efeito da estrutura fortemente positiva da correlação do VAR(1) nos autovalores. Outro fato observado resultante desses modelos é o percentual da variabilidade da 1ª componente, que aumenta de forma significativa (ver Tabela 3, Modelo 5). Observa-se que o percentual de explicação da primeira componente do Modelo 5 aumentou 27.13% em relação a primeira componente do Modelo 1. Portanto, a autocorrelação pode direcionar quase toda variabilidade para a 1ª componente, o que reduz a dimensão do espaço gerado pelos vetores de $\mathbf{X}_t$. Essa é uma questão de relevância prática e, será discutida no que segue.

Tabela 3: Autovalores de $\Gamma(0)$ dos modelos VAR(1) considerados no estudo e percentual de variabilidade.

| Modelos | $\lambda_1$ | $\lambda_2$ | $\lambda_3$ | $\lambda_4$ | % $\lambda_1$ | % $\lambda_2$ | % $\lambda_3$ | % $\lambda_4$ |
|---|---|---|---|---|---|---|---|---|
| **1** | **278.20** | **87.65** | **34.24** | **22.49** | **65.83** | **20.74** | **8.10** | **5.32** |
| 2 | 305.71 | 96.32 | 37.63 | 24.71 | 65.83 | 20.74 | 8.10 | 5.32 |
| 3 | 281.29 | 88.58 | 34.50 | 22.97 | 65.82 | 20.73 | 8.07 | 5.38 |
| 4 | 429.96 | 136.74 | 41.67 | 26.50 | 67.72 | 21.54 | 6.56 | 4.17 |
| 5 | 2630.00 | 109.37 | 60.91 | 28.76 | 92.96 | 3.87 | 2.15 | 1.02 |

As propriedades empíricas inferênciais (caso de estudo 2) relacionadas a estimação dos autovalores são apresentadas e discutidas a seguir. Amostras aleatórias de tamanhos $n = 100, 500$ e 1000 foram geradas de uma distribuição normal multivariada com variância $E(\boldsymbol{\xi}_j \boldsymbol{\xi}_{j+h}^{'})$ (Modelo 1) replicando 1000 vezes os Modelos 1, 2, 3, 4 e 5. O estudo avalia o comportamento empírico dos autovalores dos Modelos 2, 3, 4 e 5 em relação aos autovalores do Modelo 1 através dos testes descritos no final da seção 3. A correlação cruzada entre as componentes principais também foi avaliada e disponibilizada graficamente.

A Tabela 4 mostra resultados de rejeição de $H_0$: $\lambda_i = \lambda_i^0$ ao nível de significância $\alpha = 5\%$. O valor $\lambda_i^0$ corresponde aos percentuais de variabilidade do Modelo 1, isto é, do ruído branco. Observa-se que o tamanho dos testes estão próximos de 5% para o caso do Modelo 1, que é um resultado esperado. As taxas de rejeição no caso do Modelo 3 são similares ao Modelo 1. O que é justificado pelos resultados discutidos da Tabela 3. Taxas significativas de rejeição de $H_0$ são observadas para os casos dos Modelos 4 e 5, resultado também em acordo com os discutidos da Tabela 3. Isto é, a autocorrelação do processo influencia fortemente nas estimativas das componentes principais.

A Tabela 5 mostra resultados simulados do comportamento do teste Fujikoshi (1980). Nesse estudo considera-se o percentual de explicação da primeira componente ($m = 1$), ou seja, $H_0$: $\tau_1 \geq 0.6583$, e o percentual de explicação das duas primeiras componentes ($m = 2$),

Tabela 4: Tamanho do teste $H_0 : \lambda_i = \lambda_i^0$ sobre $E(\boldsymbol{\xi}_j \boldsymbol{\xi}_{j+h}^{'})$.

| Modelos | $n$ | $\hat{\lambda}_1$ | $\hat{\lambda}_2$ | $\hat{\lambda}_3$ | $\hat{\lambda}_4$ |
|---|---|---|---|---|---|
| 1 | 100 | 5.4% | 5.6% | 4.7% | 5.1% |
| | 500 | 4.4% | 4.6% | 4.7% | 4.8% |
| | 1000 | 5.0% | 5.0% | 4.8% | 5.5% |
| 2 | 100 | 13.0% | 14.0% | 9.0% | 6.0% |
| | 500 | 36.0% | 33.0% | 34.0% | 29.0% |
| | 1000 | 53.0% | 51.7% | 54.0% | 49.0% |
| 3 | 100 | 6.0% | 7.0% | 4.0% | 6.0% |
| | 500 | 6.0% | 5.0% | 5.0% | 5.0% |
| | 1000 | 4.0% | 7.7% | 6.0% | 5.7% |
| 4 | 100 | 71.0% | 80.0% | 29.0% | 13.0% |
| | 500 | 100.0% | 100.0% | 75.0% | 69.0% |
| | 1000 | 100.0% | 100.0% | 100.0% | 93.0% |
| 5 | 100 | 100.0% | 35.66% | 93.4% | 39.0% |
| | 500 | 100.0% | 95.4% | 100.0% | 95.4% |
| | 1000 | 100.0% | 100.0% | 100.0% | 100.0% |

$H_0$: $\tau_2 \geq 0.8657$. Esses percentuais são apresentados na Tabela 3 para o Modelo 1.

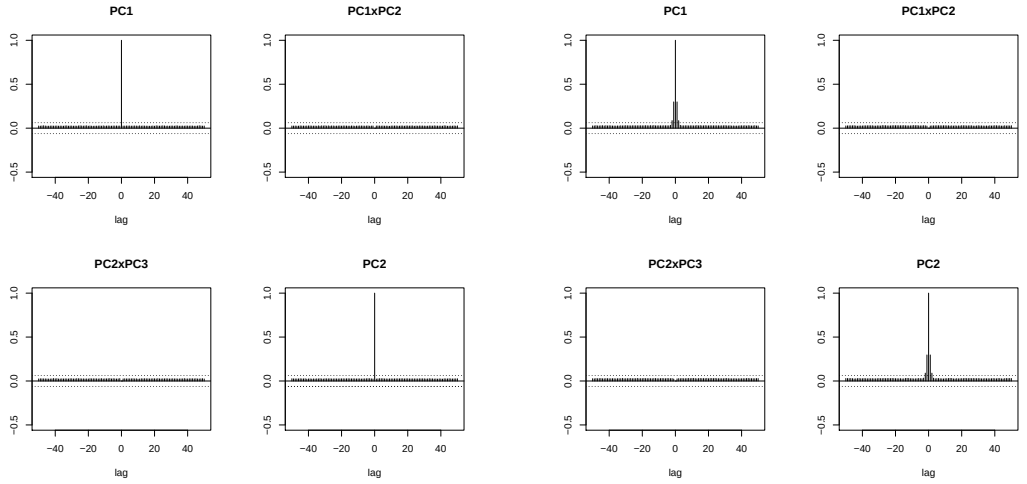
Para o nível de significância de $\alpha = 5\%$, percebe-se que a fraca autocorrelação dos Modelos 2 e 3 gera resultados equivalentes ao Modelo 1 para o tamanho amostral $n = 1000$. Entretanto, a estrutura de autocorrelação dos Modelos 4 e 5 evidenciam claramente o efeito dessas na inferência e teste estatísticos por meio dos autovalores. Nota-se que o aumento da estrutura de dependência temporal do processo leva uma redução da rejeição de $H_0$. Fenômeno discutido teoricamente nas seções anteriores. Percebe-se também que essa propriedade torna-se mais evidente ao aumento do tamanho da amostra. Nessa situação, as estimativas são mais precisas e, consequentemente, o teste acarreta maior não rejeição de $H_0$. Por exemplo, no Modelo 4 para $m = 1$ e $n = 100$, a taxa de rejeição é de $\hat{\alpha} = 5\%$. Esse valor é reduzido significativamente para $\hat{\alpha} = 0.66\%$ quando $n = 1000$. Similar interpretação é para caso $m = 2$. No Modelo 5, todos casos apresentaram $\hat{\alpha} = 0\%$. Isto é, correlação temporal de certa significância pode acarretar na redução espúria do espaço de variabilidade. Em termos de aplicação prática na área da poluição do ar esse fenômeno pode levar a uma má intrerpretação ou redução do espaço de variabilidade devido a força da autocorrelação presente nos dados. Em questões de gerenciamento de uma rede de monitoramento e na escolha do números de componentes principais que entram como variável explicativa em um modelo de regressão é possível que a autocorrelação reduza toda a variabilidade a uma única componentes principal. Na Seção 5 discutiremos algumas aplicações de ACP com essa problemática.

A Figura 1 mostra resultado da estrutura de autocorrelação e correlação cruzada entre as componentes principais para o processo ruído branco e o Modelo 2, com $n = 1000$. Nota-se que ambos casos evidenciam os resultados discutidos na Proposição 2 e Observação 2. No caso do ruído branco (Figura 1a), os gráficos mostram que as componentes principais não são autocorrelacionadas e a correlação cruzada é nula. O cenário muda quando existe a correlação temporal como no caso da Figura 1b, Modelo 2. Observa-se autocorrelação nas componentes principais e a correlação entre componentes é nula, pois o Modelo contempla autocorrelação

Tabela 5: Taxa de rejeição do teste $H_0$: $\tau_m \geq \tau_{m0}$ sobre $E(\boldsymbol{\xi}_j \boldsymbol{\xi}'_{j+h})$.

| Modelos | $m$ | $n$ | | |
|---|---|---|---|---|
| | | 100 | 500 | 1000 |
| 1 | 1 | 5.0% | 6.0% | 6.0% |
| | 2 | 3.0% | 3.3% | 5.3% |
| 2 | 1 | 7.0% | 8.6% | 9.6% |
| | 2 | 4.3% | 4.6% | 6.3% |
| 3 | 1 | 4.0% | 8.0% | 4.0% |
| | 2 | 4.0% | 5.3% | 8.3% |
| 4 | 1 | 5.0% | 3.3% | 0.6% |
| | 2 | 0.3% | 0.0% | 0.0% |
| 5 | 1 | 0.0% | 0.0% | 0.0% |
| | 2 | 0.0% | 0.0% | 0.0% |

na diagonal da matriz $\Phi$.

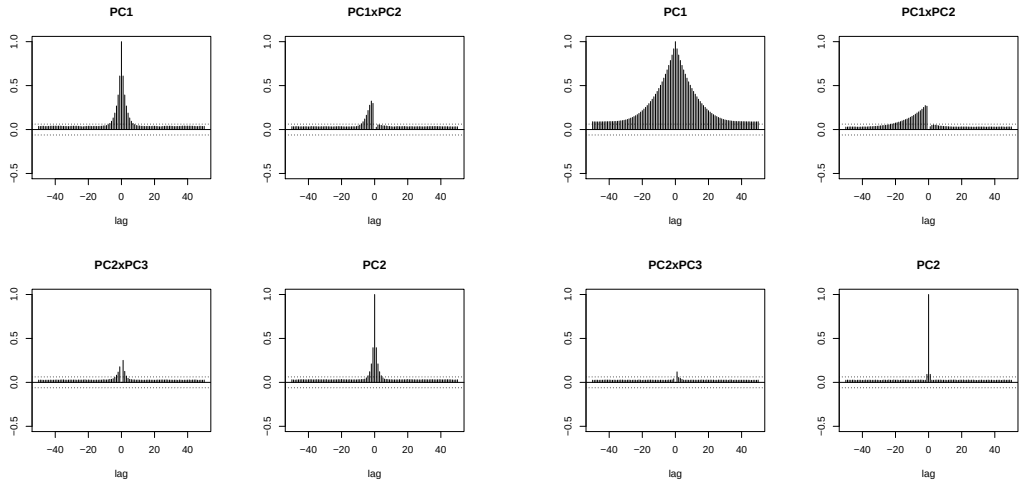



(a) Ruído branco (b) Modelo 2

Figura 1: Autocorrelação e correlação cruzada das componentes principais geradas do processo.

Diferentemente dos casos apresentados na Figura 1, os modelos explorados na Figura 2 mostram claramente que a estrutura de correlação é transladada para as componentes, pois as funções de autocorrelação e correlação cruzadas são significativas mesmo para grandes "lags". Esse é mais um exemplo que confirma empiricamente a Proposição 1. Os exemplos das Figuras 1 e 2 indicam que o uso, por exemplo, das ACPs em modelos de regressão deve ser feito sob certa cautela, isto é, propriedades de correlação temporal das ACPs não devem ser negligenciadas.

Em face as problemáticas do uso da técnica ACP em séries temporais discutidas anteriormente, torna-se necessário o uso de procedimentos ou métodos alternativos que permitam o uso dessa técnica, no domínio do tempo, em um contexto onde a suposição de independência




(a) Modelo 4 (b) Modelo 5

Figura 2: Autocorrelação e correlação cruzada das componentes principais geradas do processo.

do vetor $\mathbf{X}$ possa a ser relaxada, como o processo considerado neste artigo.

Nessa direção, este artigo propõe como umas das etapas do uso de ACP em modelos do tipo $\mathbf{X}_t$, definido em 1, o procedimento de transformação dos dados para atenuar a correlação temporal por meio do filtro linear multivariado ARMA (VARMA). O método pode ser classificado como paramétrico no sentido que a ordem do modelo seja corretamente especificada por meio dos procedimentos usuais de ajuste do modelo VARMA.

O procedimento consiste em filtrar a série $\mathbf{X}_t$ pelo filtro $\Psi(B) = \Phi(B) \triangle^d \Theta^{-1}(B)$ para resultar no ruído branco. Estudos empíricos foram conduzidos para verificar o desempenho desse procedimento nos modelos anteriores. Como esperado, o método de filtragem, sob ordem correta do modelo, mostrou-se eficaz no sentido de pequeno erro quadrático médio empírico na estimação das componentes principais do ruído. Os resultados estão disponíveis com os autores. Esse estudo empírico motivou a construção do teste de hipótese para verificar se a estrutura temporal dos dados pode ser negligenciada sem causar distorções nas interpretações e inferências das componentes principais.

Na sequência da proposta da transformação VARMA no vetor $\mathbf{X}_t$ para obter as componentes principais dos resíduos, o seguinte exercício empírico corresponde ao teste de hipóteses para verificar, estatísticamente, o efeito da correlação temporal no cálculo da proporção de variabilidade das componentes para a estrutura temporal considerada neste estudo.

Os resultados apresentados na Tabela 6 correspondem ao teste de Fujikoshi (1980), onde $\tau_{m0}$ refere-se ao percentual de explicação dos dados originais, isto é, dos Modelos 1, 2, 3, 4 e 5 discutidos anteriormente. Os dados originais são transformados através do filtro VAR(1), onde os parâmetros são estimados por procedimentos usuais. O percentual de variabilidade da componente principal dos resíduos é testado. Por exemplo, gera-se os dados originais de acordo com o Modelo 4, adota-se para $m = 1$ $H_0$: $\tau_1 \geq 0.6772$, filtra dos dados gerados e obtém o percentual de estimação $R_1$ dos resíduos.

Em acordo com os resultados apresentados nas Tabelas 3 e 5, a Tabela 6 mostra que

a estrutura de correlação temporal dos modelos 2 e 3 não levam a rejeição da hipótese de proporção de variabilidade equivalente ao ruído branco. Embora o percentual do Modelo 4 mostra-se próximo numericamente dos modelos 1, 2 e 3 (ver Tabela 3), o teste evidencia diferença significativa de pelo menos 12% ($n = 100$). No caso do Modelo 5, a rejeição do teste é de 100% para qualquer tamanho amostral. Ao analisar as Tabelas 5 e 6, percebe-se que os Modelos 4 e 5 apresentam estrutura de correlação que causam mudanças significativas nas interpretações da técnica ACP. Esse procedimento de filtragem e execução do teste permite que a técnica ACP, no domínio do tempo, seja utilizada em processos mesmo com forte estrutura de correlação temporal.

Note que a proposta de transformar os processos, por meio de filtros, para atenuar a estrutura temporal em técnicas multivariadas foi motivo também dos recentes trabalhos, Jaimungal & Ng (2007) e Greenaway-McGrevy et al. (2012) onde utilizam as técnicas funcional ACP e análise fatorial, respectivamente, sob um olhar que não negligencia a característica temporal das variáveis.

Tabela 6: Taxa de rejeição do teste $H_0$: $\tau_m \geq \tau_{m0}$ sobre $\Gamma_{\mathbf{X}}(0)$.

| Modelos | $m$ | $n$ | | |
|---|---|---|---|---|
| | | 100 | 500 | 1000 |
| 1 | 1 | 6.6% | 5.0% | 6.0% |
| | 2 | 4.0% | 4.0% | 4.6% |
| 2 | 1 | 7.3% | 6.6% | 6.0% |
| | 2 | 4.3% | 4.3% | 4.0% |
| 3 | 1 | 4.6% | 5.3% | 5.3% |
| | 2 | 4.3% | 4.0% | 3.3% |
| 4 | 1 | 12.0% | 32.3% | 48.6% |
| | 2 | 41.0% | 97.0% | 100.0% |
| 5 | 1 | 100.0% | 100.0% | 100.0% |
| | 2 | 100.0% | 100.0% | 100.0% |

Como já mencionado no início desta seção, os exemplos anteriores consideram uma particular estrutura de correlação temporal, isto é, processos com autocorrelação e correlação cruzada positiva. Esses são exemplos onde a proporção de variabilidade aumenta para as primeiras componentes, fenômeno de principal interesse deste artigo e de maior problemática na utilização das ACPs. Entretanto, as interpretações e inferências das ACPs depende intrisincamente da estrutura matricial de correlação do processo. Nessa direção, outros modelos, com estruturas de autocorrelação positivas e negativas, diferentes dos anteriores são discutidas a seguir. As matrizes estão dispostas na Tabela 7. As análises empíricas, com respeito a proporção de variabilidade das componentes, mostram que, nesse contexto, a proporção de variabilidade da primeira componente dos modelos (Tabela 8) é reduzida em relação ao ruído branco (Tabela 3), o que acarreta em uma distribuição da proporção da variabilidade para as demais componentes. Essa evidência empírica contrária as obtidas pelos Modelos 2, 3, 4 e 5 mostra que as estruturas de correlação temporal dos Modelos 6, 7, 8 e 9 não provocam efeitos drásticos nas interpretações e inferências das ACP comparado com os casos anteriores.

Para finalizar, a proposta de filtrar a série com objetivo de atenuar a correlação temporal torna-se um método alternativo para utilizar a técnica ACP, no domínio do tempo,

independente da estrutura de correlação temporal.

Tabela 7: Matrizes de $\Phi$ para os processos VAR(1).

| $\Phi_1$ (**Modelo 6**) | | | | $\Phi_1$ (**Modelo 7**) | | | |
|---|---|---|---|---|---|---|---|
| 0.7 | 0.0 | 0.0 | 0.0 | -0.7 | 0.0 | 0.0 | 0.0 |
| 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 0.0 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | -0.1 | 0.0 |
| 0.0 | 0.0 | 0.0 | 0.3 | 0.0 | 0.0 | 0.0 | 0.3 |
| $\Phi_1$ ( **Modelo 8**) | | | | $\Phi_1$ ( **Modelo 9**) | | | |
| 0.7 | 0.0 | 0.3 | 0.1 | 0.6 | 0.3 | 0.0 | 0.3 |
| 0.0 | 0.0 | 0.0 | 0.0 | 0.1 | 0.2 | 0.0 | 0.1 |
| 0.2 | 0.0 | 0.0 | 0.0 | 0.1 | -0.8 | 0.4 | 0.2 |
| 0.0 | 0.0 | 0.0 | -0.4 | 0.2 | 0.0 | 0.2 | -0.5 |

Tabela 8: Autovalores de $\Gamma(0)$ dos modelos VAR(1) considerados no estudo e percentual de variabilidade.

| Modelos | % $\lambda_1$ | % $\lambda_2$ | % $\lambda_3$ | % $\lambda_4$ |
|---|---|---|---|---|
| 2' | 61.46 | 27.05 | 7.20 | 4.27 |
| 3' | 57.24 | 31.31 | 7.17 | 4.26 |
| 4' | 59.99 | 31.29 | 5.24 | 3.46 |
| 5' | 55.90 | 28.38 | 10.78 | 4.92 |

# 5 Estudo de Casos

Nesta seção alguns exemplos reais são analisados no contexto dos estudos propostos neste artigo. Os dados (séries temporais) são observações obtidas na Região da Grande Vitória (RGV). A RGV é composta de 7 cidades e está situada na Costa Sul do Atlantico (latitude 20°19S, longitude 40°20W). O clima é tropical úmido, com média de temperaturas no intervalo de 24°C a 30°C. A preciptação é bem distribuída ao longo do ano, sendo o total anual de $1300mm$.

A seção está dividida nos seguintes três estudos: identificação de fonte poluidora (Subseção 5.1), gerenciamento da rede (Subseção 5.2) e uso de ACP no modelo de regressão para associar poluentes no ar à taxa de internação (Subseção 5.3). Para a classificação da componentes principais em *clusters* foi considerado um ponto de corte e o critério de *clusters* de componentes principais

$$r_m = \lambda_j^{1/2}(a_j^{k'} S_k^{-1} a_j^k)^{1/2} \tag{9}$$

discutido em Cadima & Jolliffe (1995), onde $\lambda_j$ é o autovalor da $j-$ésima componente, $a_j^k$ é subvetor de $\beta_j$ que considera $k$ coeficientes retidos e $S_k$ é a sub-matriz de covariância que envolve somente linhas e colunas correspondente aos coeficientes retidos. A medida $r_m$ é um coeficiente de correlação e quanto mais próximo de 1 melhor será a representação entre a componente e sua interpretação.

## 5.1 Identificação de fonte poluidora

Soares (2011) avaliou a composição química das partículas totais em suspensão (PTS) no ar em Vitória - ES, por meio dos modelos balanço químico de massas (BQM) e de fatoração de matriz positiva (FMP). No total foram realizadas 100 coletas, durante um período de 50 dias (de 14/09/2010 a 04/11/2010). Dessas amostras, 6 foram descartadas devido a imperfeições nos filtros. Das 94 restantes, somente 79 foram possíveis de validar, pois algumas apresentaram valores faltantes de elementos majoritários ou minoritários.

O ponto de coleta foi a estação de monitoramento da qualidade do ar localizada na Enseada do Suá devido às características da região. Esse ponto de monitoramento sofre influência de intenso tráfego de veículos, de inúmeras obras de construção cívil e de diversos pontos comerciais e habitações.

O filtro de quartzo utilizado para obtenção das amostras é composto, basicamente, de silício e por isso não foi possível determinar a concentração desse elemento químico nas amostras de PTS. Esse fato dificulta as análises de identificação de fonte porque o silício é majoritário em alguns perfis de fontes como, por exemplo, solos.

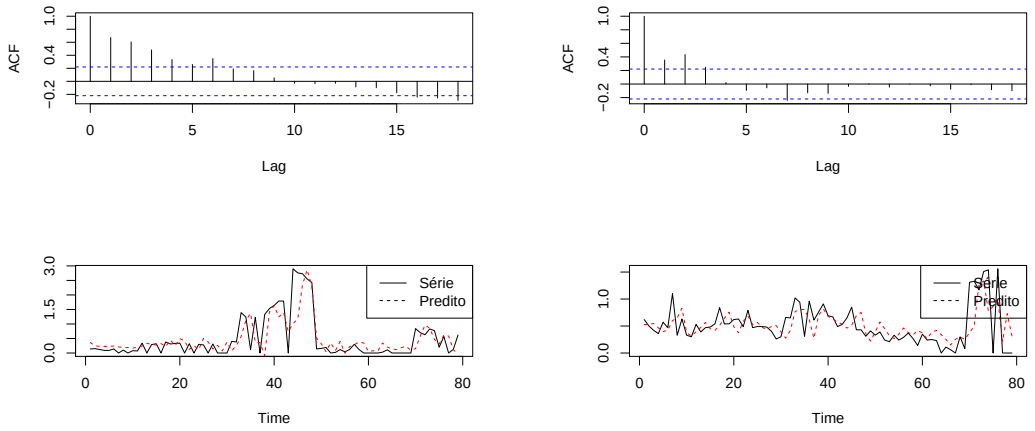
As análises das amostras foram realizadas pelo laboratório francês *Centre Commun de Mesure - Université du Littoral Côte d'Opale* pelo método *Inductively Coupled Plasma/Mass Spectrometry* através da metodologia da U.S.EPA 68-D-00-264. Nas 79 amostras de PTS foram encontrados 29 elementos químicos (detalhes com o autor). Do total dos elementos químicos, somente os elementos arsênio, selénio, rubídio e prata apresentam autocorrelação significativa. Os modelos de séries temporais ajustados para esses elementos estão apresentados na Tabela 9. Esses ajustes foram feitos com base na metodologia de modelos de Box-Jenkins, isto é, identificação, testes e análise de resídual. A Figura 3 apresenta as séries temporais, a função de autocorrelação e o ajustes do modelo dos elementos químicos arsênio e rubídio. Note que o ajuste indicado na Tabela 9 encaixa perfeitamente aos dados.

Tabela 9: Elementos químicos e ajustes de modelos.

| Elementos químicos | Modelos |
|---|---|
| Arsênico | ARMA (1,4) |
| Selênio | ARMA (1,0) |
| Rubídio | ARMA (1,2) |
| Prata | ARMA (2,0) |

As estimativas dos parâmetros dos modelos na Tabela 9 (disponíveis com o autor) indicam processo com fraca autocorrelação. Em base as discussões teóricas e empíricas apresentadas nas seções anteriores, as estruturas de autocorrelação dos quatros elementos possivelmente não influenciará nas análises por ACP.

A Tabela 10 apresenta os resultados de ACP em dois contextos; 1ª parte mostra as estimativas dos autovalores e autovetores para os dados originais. A 2ª parte da tabela mostra as estimativas dessas quantidades obtidas após aplicar nos dados originais os filtros lineares dos modelos da Tabela 9. Essa sugestão de filtrar os dados tem amparo no trabalho de Greenaway-McGrevy et al. (2012), onde os autores filtram a série e trabalham um modelo de análise fatorial. Na Tabela, verifica-se em ambos casos que os valores estimados da ACP são muitos próximos. Isso corrobora com a conclusão obtida no parágrafo anterior obtida através da visualização gráfica das estruturas de correlação temporal das séries e das estimativas dos




(a) Arsênio (b) rubídio

Figura 3: Autocorrelação, série temporal e ajuste dos elementos químicos.

parâmetros dos modelos. Esse exemplo prático apresenta similares conclusões discutidas para o Modelo 3 na Seção 4. Na Tabela observa-se também que as cinco primeiras componentes principais de cada situação mencionada captam aproximadamente 85% da variabilidade total, ou seja, a autocorrelação presente não gera efeito drástico de mudança de variabilidade. Como última análise, o teste de Fujikoshi aplicado nos resíduos indica não haver mudança do percentual de variabilidade (valor-p = 0.4999) entre os dados originais e os dados filtrados.

Com intuito de identificar a fonte poluidora, tomou-se os coeficientes dos autovetores com valor absoluto superior a 0.2 e comparou-se com os perfis das fontes próximas ao ponto de monitoramento. Em geral, os estudos com ACP para identificação de fonte considera uma rotação dos autovetores para melhor entendimento do que representa a componente. Nesse estudo não foi utilizado nenhum método de rotação para que este não influencie os resultados obtidos da aplicação da técnica ACP nos dados originais que são comparados com os resultados obtidos para os dados filtrados.

As análises de ACP nos dados originais indicam que a primeira componente principal representa 55.79% de variabilidade. Os valores dos coeficientes cério, bismuto, urânio, estrôncio molibidênio, titânio, cromo, manganês associados às fontes próximas ao local do monitoramento relaciona esta componente a processos siderúrgicos (e exaustores de veículos a diesel). Magnésio, alumínio, rubídio e mercúrio são elementos químicos com coeficientes mais representativo na segunda componente principal. Esses elementos associam a segunda componente à fonte construção civil. A terceira componente principal tem maior correlação com os elementos químicos arsênio, selênio, prata, magnésio e alumínio. A combinação desses elementos indica como fonte poluídora ressuspensão de vias. Na quarta componente principal destacam-se, com maior correlação, os elementos químicos cobalto, arsênico, selênio, mercurío, magnésio e vanádio marcando essa componente como representativa de exaustores de veículos à gasolina (e desgaste de pneus). E a quinta componente principal evidencia o presença de elementos químicos de fornos de sinterização, tais como potássio, ferro, berílio, prata e chumbo. Esses resultados coincidem com os obtidos por Soares (2011), que fez uso de modelos receptores

BQM e FMP, que são modelos de uso mais sofisticados.

Tabela 10: Coeficientes da ACP para dados das PTS: dados normais e filtrados

| **Elementos** | ACP dados originais | | | | | ACP sobre dados filtrados | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 |
| Magnésio | -0.03 | **-0.42** | **-0.28** | **0.25** | 0.10 | 0.02 | **-0.47** | **0.26** | 0.04 | 0.10 |
| Alumínio | -0.02 | **-0.39** | **-0.30** | 0.19 | **0.20** | 0.02 | **-0.44** | **0.27** | -0.01 | 0.11 |
| Potassio | 0.00 | 0.06 | -0.10 | -0.03 | **-0.40** | 0.00 | 0.13 | 0.14 | 0.09 | **0.63** |
| Ferro | 0.00 | -0.09 | **0.32** | **-0.24** | **0.57** | 0.00 | -0.01 | **-0.44** | -0.17 | -0.15 |
| Vanadio | -0.02 | -0.15 | **0.30** | **-0.24** | 0.09 | 0.01 | -0.06 | **-0.40** | -0.13 | 0.14 |
| Cromo | **0.24** | 0.03 | -0.07 | 0.10 | 0.04 | **-0.24** | -0.01 | 0.10 | 0.07 | -0.03 |
| Manganês | **0.24** | 0.03 | 0.03 | -0.01 | 0.11 | **-0.24** | 0.04 | -0.03 | 0.02 | -0.05 |
| Cobalto | 0.15 | -0.03 | -0.07 | **-0.45** | -0.15 | -0.15 | 0.06 | -0.09 | **-0.43** | **0.22** |
| Nikel | **0.24** | -0.08 | -0.01 | 0.03 | -0.03 | **-0.24** | -0.06 | 0.02 | 0.02 | 0.00 |
| Cobre | **0.24** | 0.02 | -0.07 | 0.11 | 0.05 | **-0.24** | 0.00 | 0.09 | 0.09 | -0.09 |
| Arsênico | 0.01 | **-0.23** | **-0.38** | **-0.44** | -0.02 | -0.05 | **-0.29** | 0.07 | **-0.53** | **-0.21** |
| Selênio | -0.01 | -0.09 | **-0.36** | **-0.46** | 0.00 | 0.01 | -0.13 | 0.19 | **-0.52** | **0.26** |
| Lítio | **0.24** | -0.03 | 0.02 | -0.03 | -0.11 | **-0.24** | -0.01 | -0.02 | -0.02 | 0.12 |
| Berílio | 0.12 | -0.09 | 0.16 | -0.02 | **-0.53** | -0.12 | -0.04 | -0.14 | 0.04 | **0.47** |
| Titânio | **0.24** | 0.03 | -0.07 | 0.04 | 0.01 | **-0.24** | 0.02 | 0.09 | 0.03 | -0.01 |
| Zinco | **0.21** | -0.03 | -0.17 | 0.18 | 0.05 | **-0.21** | -0.08 | **0.20** | 0.09 | -0.18 |
| Rubídio | -0.02 | **-0.44** | 0.14 | -0.15 | -0.13 | 0.02 | **-0.35** | **-0.30** | -0.05 | -0.11 |
| Estrôncio | **0.25** | 0.02 | -0.04 | 0.06 | 0.02 | **-0.25** | 0.00 | 0.06 | 0.06 | -0.05 |
| Molibidênio | **0.24** | 0.01 | -0.08 | 0.09 | 0.01 | **-0.24** | -0.02 | 0.10 | 0.06 | -0.05 |
| Prata | -0.02 | **-0.36** | **0.41** | 0.05 | **-0.21** | 0.02 | **-0.31** | **-0.34** | **0.30** | **0.22** |
| Cádmio | **0.22** | -0.05 | 0.19 | -0.10 | -0.02 | **-0.22** | 0.03 | **-0.21** | -0.02 | 0.08 |
| Estanho | **0.25** | -0.03 | 0.00 | 0.01 | -0.02 | **-0.25** | -0.01 | 0.01 | 0.02 | 0.02 |
| Bário | **0.24** | 0.01 | -0.06 | 0.05 | 0.03 | **-0.24** | -0.01 | 0.07 | 0.02 | -0.07 |
| Cério | **0.25** | -0.01 | -0.02 | 0.02 | 0.00 | **-0.25** | 0.00 | 0.03 | 0.02 | 0.00 |
| Mercurío | -0.06 | **-0.46** | 0.08 | **0.22** | -0.03 | 0.03 | **-0.46** | -0.12 | **0.25** | -0.03 |
| Tálio | **0.23** | 0.04 | 0.14 | -0.05 | 0.08 | **-0.23** | 0.07 | -0.14 | 0.03 | 0.01 |
| Chumbo | **0.21** | -0.07 | 0.12 | -0.16 | **0.21** | **-0.21** | -0.02 | **-0.20** | -0.12 | -0.12 |
| Bismuto | **0.25** | -0.03 | 0.04 | 0.02 | 0.02 | **-0.25** | -0.01 | -0.04 | 0.04 | 0.01 |
| Urânio | **0.24** | -0.10 | 0.05 | -0.03 | -0.08 | **-0.24** | -0.07 | -0.06 | -0.03 | 0.09 |
| % $\hat{\lambda}$ acumulado | 55.79 | 65.60 | 74.42 | 81.51 | 85.42 | 55.73 | 66.09 | 74.10 | 79.63 | 83.72 |

Na comparação entre os coeficientes dos dados originais com os coeficientes dos dados filtrados, verifica-se praticamente o mesmo valor, o que leva em ambas situações às mesmas fontes poluidoras. No que tange as componentes principais, nenhuma correlação cruzada foi encontrada entre as componentes, como é esperado em ACP. Esses resultados obtidos assemelham-se ao estudo de simulação com fraca autocorrelação. Assim, o uso da técnica ACP pode ser usada sem nenhuma preocupação neste estudo, pois os resultados não são afetados pela autocorrelação presente.

Para finalizar, Soares (2011) utilizou os modelos receptores BQM e FMP para identificar as principais fontes poluidoras que impactam no local amostrado sem considerar o efeito da autocorrelação presente.

Um outro ponto que deve ser abordado é a exigência de normalidade do conjunto de dados, tanto para aplicação nas técnicas BQM e FMP quanto em ACP. A análise da normalidade da concentração dos elementos químicos presentes mostram que nenhum dos elementos apresentam ter distribuição normal. Dessa forma, os testes estatísticos não foram aplicados neste estudo.

## 5.2 Gerenciamento de rede

Seguindo a idéia adotada por Pires et al. (2008*a*,*b*) de usar ACP para reduzir o número de locais de monitoramento para determinado poluente, esta seção apresenta análise do conjunto de dados de poluentes obtidos da rede automática de monitoramento da qualidade do ar (RAMQAR) da região da Grande Vitória, ES, Brasil, mas sem negligenciar a estrutura de correlação temporal, contexto não considerado pelos autores supracitados. A RAMQAR é composta de oito estações e todas estão localizadas em região urbana. Das setes cidades que compõem a RGV, 4 são monitoradas pela RAMQAR. A área total das quatro cidades monitoradas é de 1.139,88 km$^2$, que representa 48,90% da área da RGV.

A RAMQAR registra concentrações de poluentes atmosféricos e parâmetros meteorológicos. Os poluentes monitorados são PTS, $PM_{10}$, ozônio ($O_3$), óxidos de nitrogênio ($NO_x$), monóxido de carbono (CO) e hidrocarbonetos (HC), e os parâmetros meteorológicos são direção e velocidade do vento, precipitação pluviométrica, umidade relativa do ar, temperatura, pressão atmosférica e radiação solar. De todas estações, a de Sua e a estação de Ibes são as únicas que registram as concentrações de todos os poluentes. A estação de Carapina monitora todos os parâmetros meteorológicos.

Nas subseções abaixo serão analisados registros dos poluentes $PM_{10}$ e $SO_2$, obtidos na RAMQAR, compreendidos entre o período de janeiro de 2005 a dezembro de 2009.

### 5.2.1 Análise da concentração do $PM_{10}$

Séries das concentrações médias diárias de $PM_{10}$ nas 8 estações de RAMQAR estão apresentadas na Figura 4. As respectivas autocorrelações são dadas na Figura 5. Com base na Figura 5, observa-se estrutura sazonal e de significativas autocorrelações, mesmo para distantes lags. O decaimento lento das ACF sugerem que as séries possivelmente possuem estrutura de correlação de processos de memória longa [Reisen et al. (2010)].

Embora este artigo esteja centralizado no Modelo 1, isto é, estrutura de correlação que satisfaz $\sum_{j=0}^{\infty} |\Gamma(j)| < \infty$, a aplicação desta subseção é estendida para o modelo sazonal ARFIMA, isto é, processo que não satifaz a propriedade de covariâncias absolutamente somáveis. Como forma de estimar os parâmetros fracionários, denotado pelos autores $d$ e $D$, do modelo SARFIMA nas séries, utilizou-se o método de estimação de Reisen et al. (2010) para séries univariadas que apresentam característica de memória longa de longo prazo e nos períodos sazonais, associadas respectivamente aos parâmetros $d$ e $D$. Os autores propõem métodos para estimar séries sazonais com diferentes parâmetros de memória longa em diferentes períodos sazonais. Esta aplicação torna-se mais uma motivação para estender os estudos discutidos neste artigo para processos memória longa, isto é, processos que não possuem a propriedade de autocorrelações absolutamente somável. Como já investigado em Reisen et al. (2010), a série da concentração de $PM_{10}$ da RGV apresentam característica memória longa e essa propriedade pode ser observada através da Figura 5. A função de autocorrelação tem decaimento lento (hiperbólico) e picos sazonais para concentrações de $PM_{10}$ das localidades Carapina, VVCentro, Cariacica, Ibes e Camburi. O estimador de Reisen et al. (2010) fornecem as estimativas dos parâmetros $d$ e $D$ apresentadas na Tabela 11. Como $\hat{d}+\hat{D} < 0.5$, esse resultado indica que as séries são estacionários. As estimativas dos parâmetros fracionários $d$ e $D$ foram obtidas considerando diferentes valores de bandwidth $m = n^{\alpha}$, onde $0 < \alpha < 1$ e $n$ é o tamanho da série. O valor escolhido para $\alpha$ foi 0.5 [ maiores detalhes da escolha de $\alpha$, ver Reisen et al. (2010)].

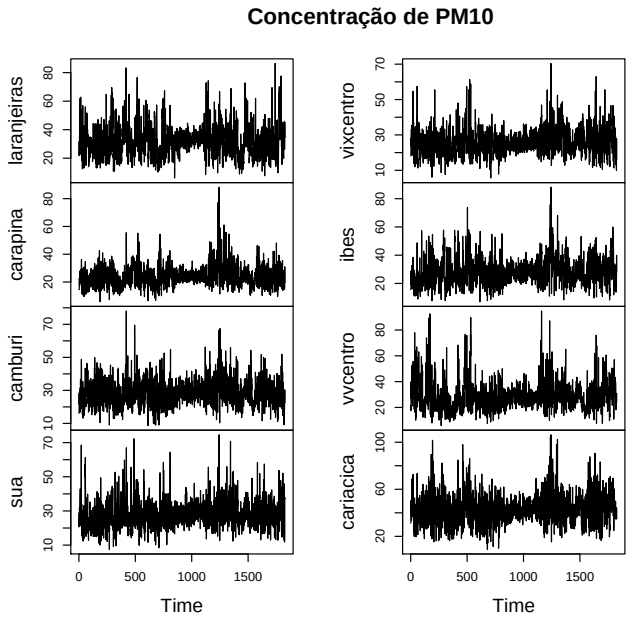


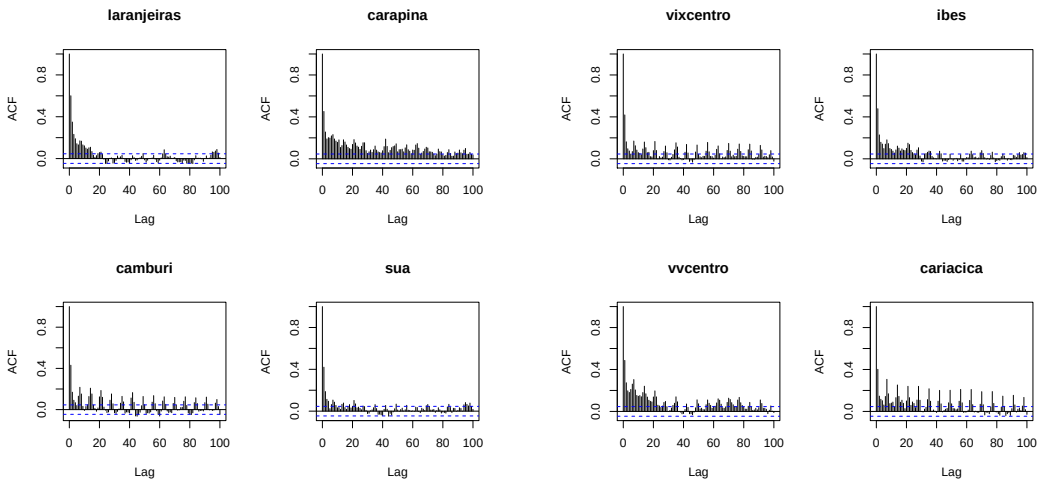

Figura 4: Séries da concentração de $PM_{10}$ da RAMQAR.

Tabela 11: Estimativas dos parâmetros fracionários para poluente $PM_{10}$ (* $H_0$: $d = 0$, $H_0$: $D = 0$ rejeitadas).

| Estação | $\hat{d}$ | d. p. $\hat{d}$ | $\hat{D}$ | d. p. $\hat{D}$ |
|---|---|---|---|---|
| Carapina | 0.2622 | 0.0141 | 0.1572 | 0.0021 |
| Camburi | * | — | 0.1059 | 0.0022 |
| Ibes | 0.2813 | 0.0146 | * | — |
| VVCentro | 0.2664 | 0.0161 | 0.1198 | 0.0024 |
| Cariacica | 0.3043 | 0.0140 | 0.1679 | 0.0021 |

A Tabela 12 apresenta resultados da técnica ACP sobre a concentração de $PM_{10}$ em dois contextos. O primeiro considera os dados originais para obtenção dos autovalores e autovetores. O segundo mostra estimativas dessas quantidades após aplicar o filtro VARFIMA$(1, d, 0)(0, D, 0)$, onde as estimativas dos parâmetros fracionários, $\hat{d}$ e $\hat{D}$, estão apresentados na Tabela 11. Nota-se, para ambos os casos, que grande parte da variabilidade fica concentrada na primeira componente principal e houve redução de 4% na primeira componente após filtrar os dados originais. As outras componentes também sofreram redução após usar o filtro. A metodologia proposta para o teste de Fujikoshi nos resíduos indica mudança do percentual de variabilidade (valor-p = 0) entre os dados originais e os dados filtrados.

Um ponto que chama a atenção é a proximidade dos coeficientes da primeira componente, pois leva a suspeitar de igual correlação entre as concentrações de $PM_{10}$, ver, por exemplo, Johnson & Wichern (1998, pg. 470). Nessa direção, a execução do teste de igual correlação entre variáveis, descrito em Johnson & Wichern (1998, pg. 457 e 458), foi aplicado e resultou na rejeição de igual estrutura de correlação para os dois contextos.




(a) ACF $PM_{10}$ (b) FACP $PM_{10}$

Figura 5: ACF das séries $PM_{10}$.

Tabela 12: Resultado de ACP para concentração de $PM_{10}$.

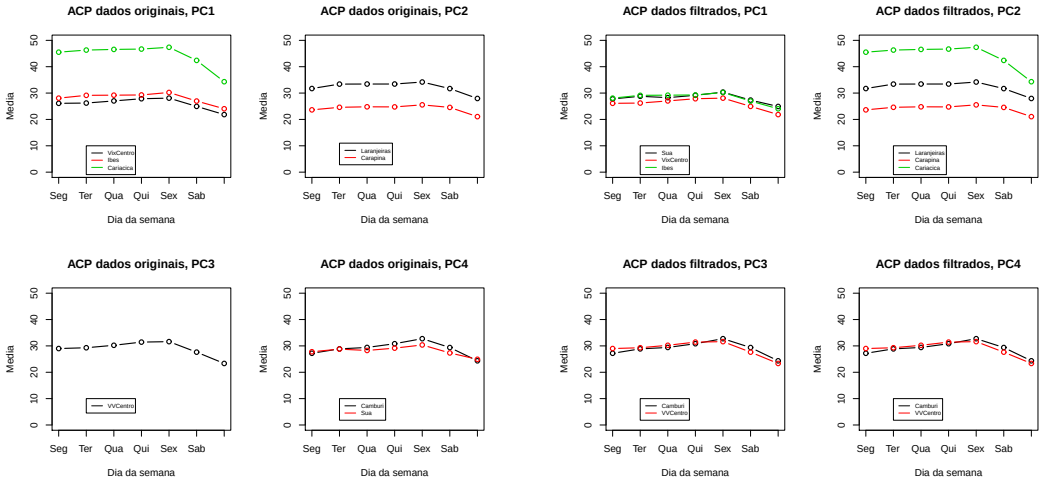
| **Estações** | ACP dados originais | | | | ACP sobre dados filtrados | | | |
|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| Laranjeiras | -0.3002 | **0.7193** | -0.1756 | 0.1460 | -0.3067 | **0.7090** | -0.0529 | 0.1606 |
| Carapina | -0.3554 | **-0.4004** | 0.2628 | 0.1750 | -0.3536 | **-0.5233** | 0.0368 | 0.0669 |
| Camburi | -0.3472 | 0.1700 | 0.0502 | **0.7019** | -0.3166 | 0.0560 | **0.7079** | **0.5055** |
| Sua | -0.3632 | 0.2163 | 0.0406 | **-0.6118** | **-0.3722** | 0.2283 | -0.3546 | -0.1360 |
| VixCentro | **-0.3864** | -0.2265 | -0.1026 | -0.1629 | **-0.3856** | -0.0222 | -0.2168 | -0.2125 |
| Ibes | **-0.3869** | 0.1787 | 0.2359 | -0.2271 | **-0.3935** | 0.0625 | -0.1563 | 0.1426 |
| VVCentro | -0.3055 | -0.2942 | **-0.8391** | 0.0141 | -0.3222 | -0.0087 | **-0.4764** | **-0.7571** |
| Cariacica | **-0.3721** | -0.2766 | 0.3542 | 0.0507 | -0.3669 | **-0.4044** | -0.2652 | 0.2383 |
| Eigenvalue | 4.8971 | 0.7744 | 0.6282 | 0.4973 | 4.5586 | 0.7462 | 0.6412 | 0.6050 |
| Proportion | 61.22 | 9.68 | 7.85 | 6.22 | 56.98 | 9.32 | 8.01 | 7.56 |
| Cumulative | 61.22 | 70.90 | 78.75 | 84.97 | 56.98 | 66.30 | 74.31 | 81.87 |

A fórmula 9 resultou no valor 0.96 e juntamente com análises gráficas desta seção foi decidido considerar os coeficientes dos autovetores com valores superiores a 0.37, em módulo. Assim, primeira componente principal dos dados originais indica semelhança entre as concentrações das estações VixCentro, Ibes e Cariacica. A segunda componente principal aponta similaridade entre Laranjeiras e Carapina. A concentração de $PM_{10}$ da estação VVCentro não apresenta semelhança com nenhuma outra estação de acordo com a terceira componente e a quarta componente principal aponta mesmo perfil para as estações Camburi e Sua. Observe que as quatro primeiras componentes principais engloba as 8 estações da rede de monitoramento.

O cenário obtido com as componentes principais dos dados filtrados é mais sucinto. A primeira componente principal indica semelhança de padrão entre as estações Sua, VixCentro e Ibes. A segunda componente principal aponta mesmo padrão para as concentrações das estações de Laranjeiras, Carapina, e até aqui iguala-se aos resultados dos dados originais, mas acrescenta a estação de Cariacica. A terceira componente principal indica semelhança

entre as concentrações de Camburi e VVCentro e esse mesmo padrão é indicado na quarta componente. Ou seja, três componentes são suficientes suficientes para englobar as 8 estações.

Para confirmar que os resultados de ACP indicam semelhança entre as concentrações das estações destacadas por cada componente, a Figura 6 mostra resultados obtidos, nos dois contextos avaliados, por dia da semana. Observa-se claramente que os resultados obtidos com os dados filtrados são bem superiores aos apresentados com os dados originais, onde a técnica discrimina os dados com mais clareza, pois fica mais evidente o igual comportamento após passar filtro nos dados.




(a) Dados originais

(b) Dados filtrados

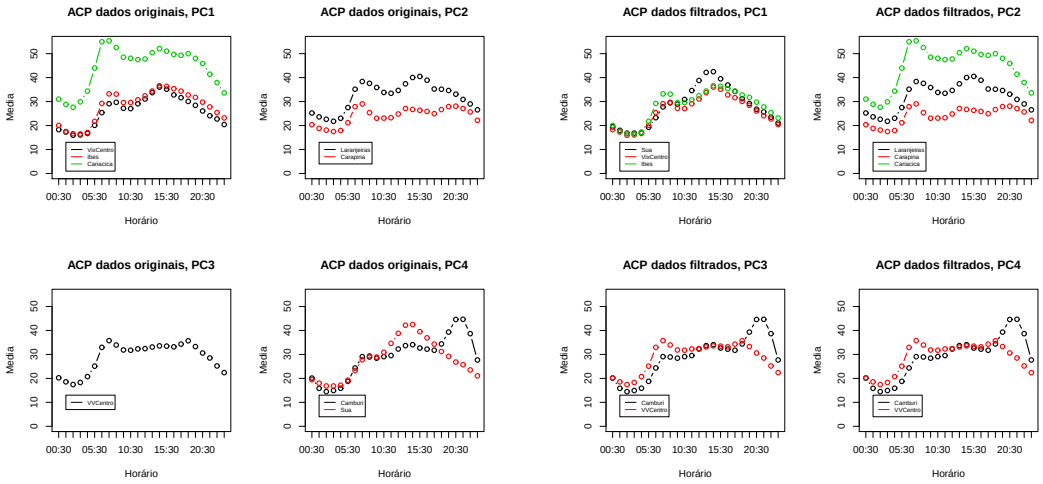
Figura 6: Concentração semanal de $PM_{10}$ com mesmo padrão de acordo com ACP.

A Figura 7 apresenta análise semelhante â Figura 6, mas nesse caso as concentrações são disponibilizadas por média horária. Por exemplo, ao comparar a quarta componente dos dados originais com a terceira componente dos dados filtrados é verificado que a concentração horária na estação VVCentro somente não acompanha a concentração em Camburi em um pico noturno que ocorre nessa estação. Outro ponto é que na comparação da segunda componente em ambos contextos a indicação de três componentes principais com semelhante padrão mostra que eliminar a presença da autocorrelação torna ACP mais discriminativa, ou seja, obtém-se resultados mais expressivos e assertivos.

Para complementar o estudo, na Figura 8 são apresentados os comportamentos temporais das componentes principais obtidas das séries de concentrações do $PM_{10}$ e das componentes principais dos resíduos, obtidos com as estimativas dadas na Tabela 11 juntamente com um VAR(1).

Verifica-se na Figura 8a que a componente é autocorrelacionada e apresenta correlação cruzada entre componentes, o que está de acordo com a Proposição 1. A Figura 8b mostra que o filtro elimina quase toda correlação cruzada entre componentes principais. Os valores de correlação cruzada fora do intervalo de confiança pode ser causado pela estrutura de heterogeneidade dos dados.

Desse estudo concluímos que das oito estações de monitoramento de concentração do poluente $PM_{10}$, três são suficientes para o monitoramento do poluente. Os outros 5 equipamentos




(a) Dados originais (b) Dados filtrados

Figura 7: Concentração horária de $PM_{10}$ com mesmo padrão de acordo com ACP.

podem ser deslocados para outras regiões de interesse a fim de abranger uma maior área da RGV. O estudo leva em consideração o igual padrão entre as estações de monitoramento independente da fonte poluídora. Análise semelhante pode ser elaborada para os demais poluentes e dados meteorológicos da rede de monitoramento da qualidade do ar. Na subseção abaixo apresentamos resultados para o poluente $SO_2$.

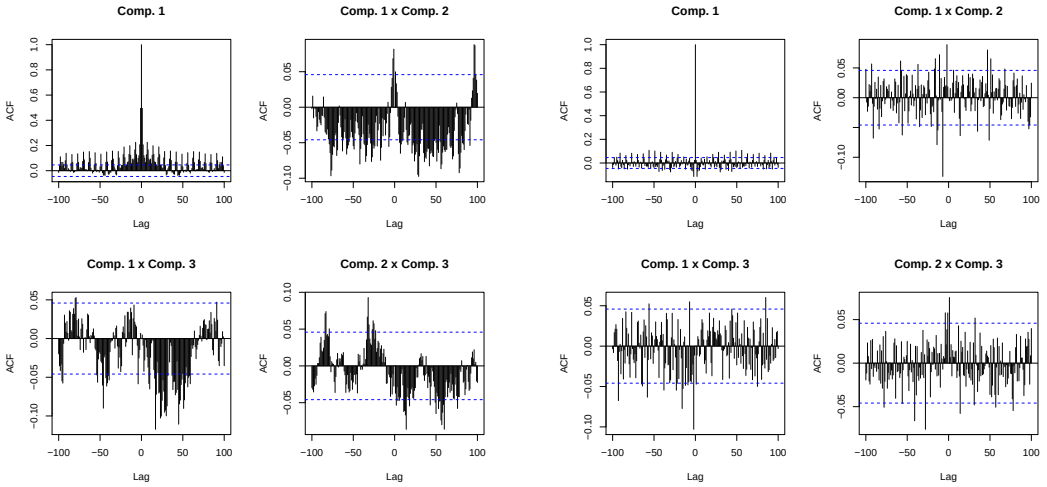
### 5.2.2 Análise da concentração do $SO_2$

De maneira semelhante a análise do $PM_{10}$ as séries das concentrações de $SO_2$ apresentam forte estrutura de autocorrelação. O estimador de Reisen et al. (2010) fornecem as estimativas do parâmetro $d$ apresentadas na Tabela 13. As estimativas para o parâmetro $D$ não foram significativas para rejeitar $H_0 : D = 0$, algo que está de acordo com a volatilidade desse poluente, pois não permanece no ar muito tempo sem reagir com outros elementos. Com exceção da concentração do $SO_2$ para estação de Laranjeiras, todas as demais concentrações do $SO_2$ resultam em séries estacionárias, pois $d < 0.5$.

Tabela 13: Estimativas dos parâmetros fracionários para poluente $SO_2$.

| Estação | $\hat{d}$ | desvio padrão $\hat{d}$ | $\alpha$ |
|---|---|---|---|
| Laranjeiras | 0.4741 | 0.0537 | 0.35 |
| Camburi | 0.3554 | 0.0073 | 0.55 |
| Sua | 0.4299 | 0.0300 | 0.40 |
| VixCentro | 0.3715 | 0.0186 | 0.45 |
| Ibes | 0.3631 | 0.0060 | 0.60 |
| VVCentro | 0.4069 | 0.0063 | 0.60 |
| Cariacica | 0.3841 | 0.0312 | 0.38 |

A Tabela 14 apresenta resultados da técnica ACP sobre a concentração de $SO_2$ em dois con-




(a) Dados originais (b) Dados filtrados

Figura 8: CCF das componentes das séries $PM_{10}$.

textos. O primeiro considera os dados originais para obtenção dos autovalores e autovetores. O segundo mostra estimativas dessas quantidades após aplicar um processo VARFIMA$(2, d, 0)$, onde as estimativas para $d$ estão apresentados na Tabela 13. Nota-se que as quatros primeiras componentes explicam aproximadamente 85% da variabilidade total. No entanto, os resultados mostram que o processo sem efeito da autocorrelação capta maior parte da variabilidade presente. Nesse caso a autocorrelação impacta para reduzir a variabilidade dos dados originais, similar interpretação referente aos Modelos 6, 7, 8 e 9. Resultado do teste de Fujikoshi nos resíduos indica não haver mudança do percentual de variabilidade (valor-p = 0.9838) entre os dados originais e os dados filtrados.

Tabela 14: Resultado de ACP para concentração de $SO_2$.

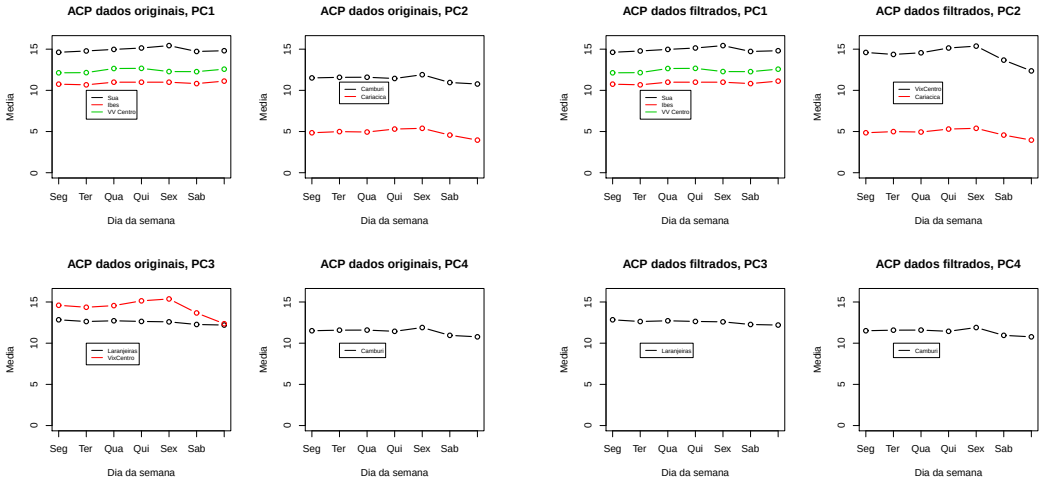
| **Estações** | ACP dados originais | | | | ACP sobre dados filtrados | | | |
|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| Laranjeiras | -0.0162 | -0.3442 | **0.7343** | 0.3687 | 0.0583 | 0.0094 | **0.9927** | 0.0543 |
| Camburi | 0.1986 | **-0.5212** | 0.1713 | **-0.7209** | 0.2076 | -0.4675 | -0.0523 | **0.7793** |
| Sua | **-0.4931** | -0.3505 | 0.0656 | 0.0240 | **-0.5309** | -0.2823 | 0.0724 | -0.1749 |
| VixCentro | 0.2129 | -0.4232 | **-0.5366** | 0.4189 | 0.1448 | **-0.5816** | -0.0372 | -0.4505 |
| Ibes | **-0.5438** | -0.2789 | -0.0817 | 0.2176 | **-0.5837** | -0.1842 | 0.0445 | -0.0439 |
| VVCentro | **-0.4986** | -0.0662 | -0.3039 | -0.3435 | **-0.4942** | -0.1338 | -0.0188 | 0.3564 |
| Cariacica | 0.3568 | **-0.4749** | -0.2001 | 0.0530 | 0.2558 | **-0.5580** | 0.0525 | -0.1647 |
| Eigenvalue | 2.2883 | 1.7169 | 1.1199 | 0.6840 | 2.4164 | 1.8715 | 1.0008 | 0.6973 |
| Proportion | 32.69 | 24.52 | 15.99 | 9.77 | 34.52 | 26.73 | 14.29 | 9.96 |
| Cumulative | 32.69 | 57.21 | 73.20 | 82.97 | 34.52 | 61.25 | 75.54 | 85.50 |

Na Tabela 14, os coeficientes dos autovetores com valores em módulo superior a 0.47 foram considerados para detectar estações com igual padrão de concentração. Os resultados de ACP para os dados originais indicam, para a primeira componente, que as concentrações de $SO_2$ das estações de Sua, Ibes e VVCentro têm igual padrão de comportamento. A segunda componente principal indica mesmo padrão entre as estações de Camburi e Cariacica. As

estações Laranjeiras e VixCentro têm igual padrão de concentração de acordo com a terceira componente principal.

Em contrapartida, os resultados de ACP sobre os dados filtrados evidenciam comportamento diferentes dos resultados anteriores a partir da interpretação da segunda componente principal. Os coeficientes dessa componente indicam que as concentrações de VixCentro e Cariacica são semelhantes. A terceira e quarta componentes principais colocam, respectivamente, a estação de Laranjeiras e a estação de Camburi, com padrão de concentração diferentes das demais estações.

Para ilustrar os resultados obtidos com o uso dos dados originais e filtrados, as Figuras 9 e 10 mostram, respectivamente, o nível semanal e horário das séries da concentração do $SO_2$ de acordo com os resultados obtidos das componentes. Em ambas as Figuras percebe-se, para a primeira componente, que o padrão semanal e horário são os mesmos nas estações Sua, Ibes e VVCentro, pois é observado igual comportamento no decorrer da semana e das horas de uma dia. Resultado semelhante é dado pela segunda componente, embora a concentração da estação de Cariacica seja indicada como semelhante a Camburi, dados originais, e a VixCentro nos dados filtrados. Chama atenção que a quarta componente, dados originais, aponta a estação de Camburi com um padrão distinto das demais, o que gera uma contratição com a segunda componente. A terceira componente principal evidência a influência temporal na técnica ACP. Percebe-se que a concentração de $SO_2$ da estação de Laranjeiras é distinta das demais, praticamente constante, como pode ser verificado nas Figuras 9 e 10.

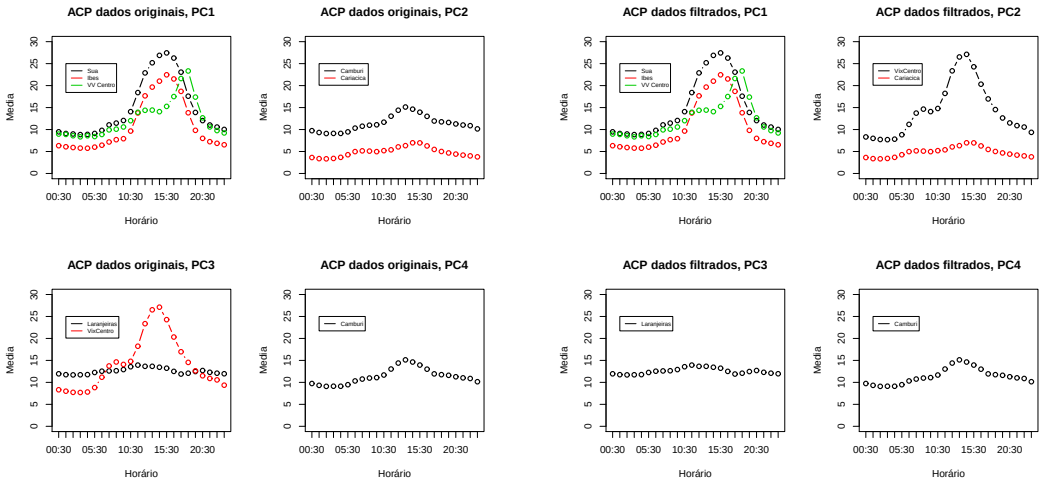



(a) Dados originais (b) Dados filtrados

Figura 9: Concentração semanal de $SO_2$ com mesmo padrão de acordo com ACP.

Esses resultados confirmam que a influência da autocorrelação modifica as análises a ponto de gerar erros cruciais, fato não considerado por Pires et al. (2008*a*,*b*). O filtro nos dados originais melhora as análises tornando a técnica mais discriminativa.

Dos resultados acima, das 7 estações, é suficiente manter 4 estações de monitoramento, sendo duas provenientes das indicadas pela primeira e segunda componentes principais. A escolha da estação nessas duas primeiras componentes pode ser por aquela que capta o maior nível de concentração ou o tipo de fonte poluidora para do poluente. As outras duas são

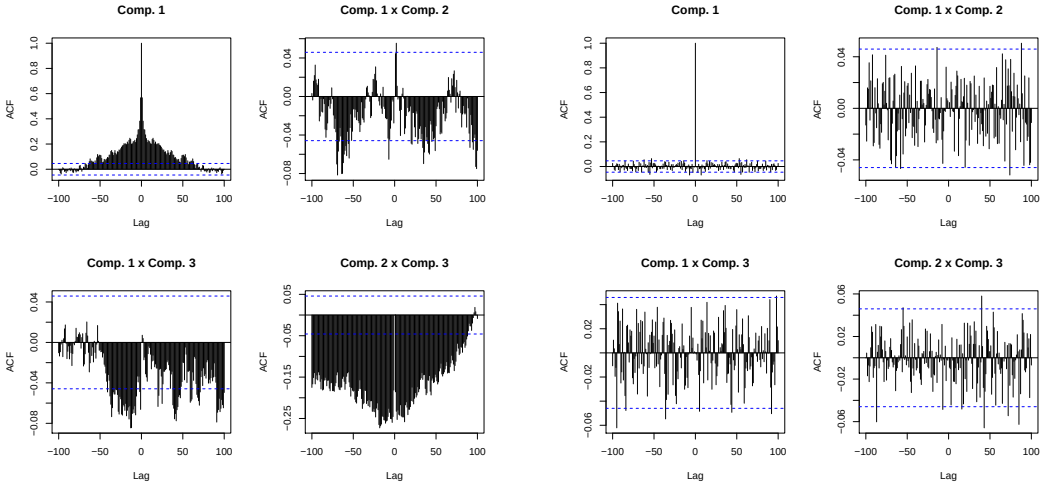



(a) Dados originais (b) Dados filtrados

Figura 10: Concentração horária de $SO_2$ com mesmo padrão de acordo com ACP.

indicadas pela terceira (Laranjeiras) e quarta (Camburi) componentes principais.

Para finalizar esse estudo, na Figura 11 são apresentados os comportamento temporais das componentes principais obtidas das séries de concentrações do $SO_2$ e das componentes principais dos resíduos, obtidos com as estimativas dadas na Tabela 13 juntamente com um VAR(2). Percebe-se claramente que uma componente influencia os resultados das outras componentes.




(a) Dados originais (b) Dados filtrados

Figura 11: CCF das componentes das séries $SO_2$.

## 5.3 Associação do poluente com a taxa de internação

A poluição atmosférica afeta a saúde da população mesmo quando seus íveis encontram-se abaixo do que determina a legislação vigente. Vários estudos têm mostrado associações significativas entre os níveis diários de concentração de poluentes e o número de atendimentos por causas respiratórias ou cardiovasculares ( para uma recente revisão bibliográfica no tema, ver Souza (2013), Arditsoglou & Samara (2005) entre outros).

Pelas características da variável de desfecho em saúde (número de atendimentos), o modelo de regressão não linear aditivo generalizado (MAG), com distribuição marginal de Poisson, tem sido, em geral, a ferramenta estatística para medir e quantificar a associação entre os poluentes atmosféricos e os efeitos adversos â saúde. Nessa metodologia, a inclusão das covariáveis (por exemplo, os poluentes) no modelo de regressão ocorre de forma individual, pois os poluentes são correlacionados. Uma forma de contornar essa questão é utilizar a análise de componentes principais da matriz de covariância dos poluentes ( Roberts & Martin (2006), Wang & Pham (2011)).

Como discutido nas secções anteriores (ver, por exemplo, Proposição 1), as componentes principais obtidas por meio de séries temporais carregam a estrutura temporal das covariáveis e, portanto, o seu uso em regressão deve ser feito de forma cautelosa. Uma maneira de contornar o problema da correlação temporal das ACPs na aplicação no modelo de regressão, o uso do procedimento de filtros de modelos multivariados, sugerido na seção de simulação, nas covariáveis torna-se uma metodologia alternativa para estimar as relações lineares e não lineares entre as variáveis dependendente e explicativas (covariáveis).

A metodologia hibrida proposta acima, isto é, o casamento entre a técnica de ACP e dos modelos de regressão, é baseada nos estudos teóricos e empíricos deste artigo e justificada por meio de aplicação a dados reais por Souza (2013). Nesse trabalho, os autores consideraram essa proposta para estimar o efeito da associação entre a exposição atmosférica dos poluentes $PM_{10}$, $SO_2$, $NO_2$, $O_3$ e CO e o número de atendimentos por doenças respiratórias em crianças menores de 6 anos na região da Grande Vitória, ES, Brasil. A seguir são apresentados de forma sumarizada os resultados referentes ao estudo supracitado. Os modelos de regressão propostos são definidos como segue. MAG-ACP denota o modelo onde as componentes principais são utilizadas no MAG como variáveis explicativas. Entretanto, devido as ACPs também apresentam as propriedades de correlação temporal, no ajuste final do modelo MAG-ACP foi necessário a inclusão de modelos do tipo SARMA$(p,q)(P,Q)$ nos resíduos, com o objetivo de eliminar as estruturas de autocorrelação presente nas componentes, como é esperado pelo resultado na Proposição 1. Para atenuar a correlação temporal das componentes, o método de filtragem, por meio do modelo Vetorial Autoregressivo (VAR(7)), é utilizado como procedimento alternativo para transformar os dados atmosféricos num processo Ruído Branco. A matriz de resíduos é utilizada para obter as componentes e essas aplicadas ao modelo MAG. Esse método é denominado de VAR-MAG-ACP.

Os resultados empirícos mostraram que o modelo VAR-MAG-ACP removeu as autocorrelações das componentes principais (ver Figura 12) e indicou estimativas mais significantes do Risco Relativo (RR) para cada poluente, além de gerar melhores ajustes residuais, resultados estão na Tabela 15. Comparadas â modelagem MAG usual, em geral, as duas vertentes propostas (MAG-ACP e VAR-MAG-ACP) apresentaram melhores resultados, tanto na estimativa do RR quanto na qualidade do ajuste. Por exemplo, um aumento de 10.49 $\mu g/m^3$

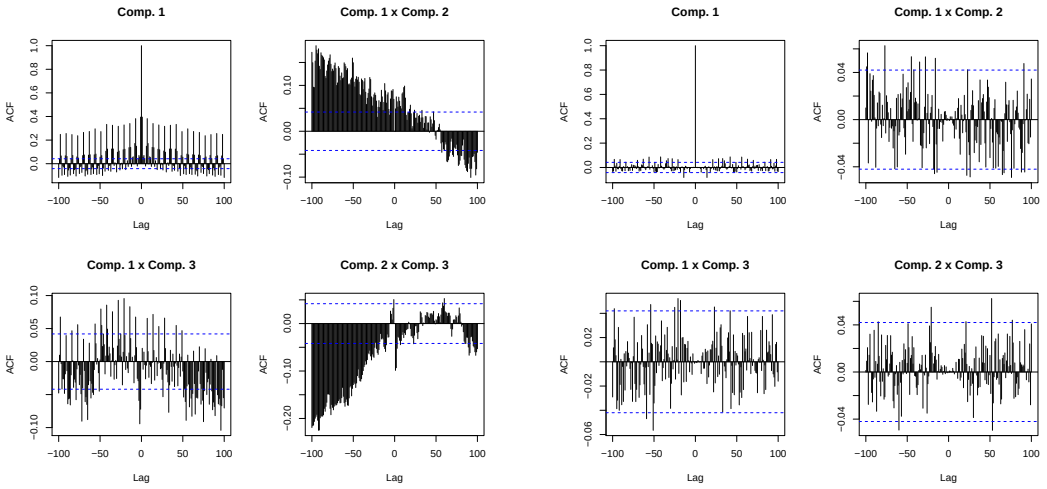
(Intervalo interquatílico) nos níveis de $PM_{10}$ resultou num aumento de pelo menos 5% do valor do RR estimado, por meio do modelo VAR-MAG-ACP, comparado ao valor obtido no modelo MAG usual.

Os riscos relativos estimados $\widehat{RR}$, $\widehat{RR}^{*}$ e $\widehat{RR}^{**}$, apresentados na Tabela 15, correspondem aos RR estimados calculados por meio dos modelos MAG, MAG-ACP e VAR-MAG-ACP, respectivamente. Note que, para um nível de significância $\alpha$, a hipótese a ser testada é definida com $H_0 : RR(x) = 1$ contra $H_1 : RR(x) > 1$. A não rejeição de $H_0$ implica estatisticamente que o poluente estudado não causa efeito adverso à saúde.

A Figura 12 apresenta gráficos das componentes principais para os dados originais e filtrados. Percebe-se na Figura 12b que a correlação temporal foi removida das componentes principais. Assim, este procedimento é uma sugestão para utilizar as componentes principais em modelos de regressão.

Tabela 15: Risco Relativo(RR) e intervalo de confiança de 95% nos atendimentos por doenças respiratórias em crianças menores de 6 anos para uma variação interquatílica dos poluentes $PM_{10}$, $SO_2$, $NO_2$ e $O_3$ e $CO$ na RGV, jan-2005 a dez-2010.

| RR | $\widehat{RR}$ | $\widehat{RR}^{*}$ | $\widehat{RR}^{**}$ |
|---|---|---|---|
| $PM_{10}$ | 1.020(1.010,1.039) | **1.029(1.001,1.090)** | 1.075(1.001,1.092) |
| $SO_2$ | 1.040(1.010,1.080) | 0.982(0.972,1.001) | 1.027(1.010,1.040) |
| CO | 1.020(1.010,1.030) | **1.048(1.002,1.071)** | 1.077(1.020,1.100) |
| $NO_2$ | 1.000(0.990,1.020) | 1.028(1.010,1.040) | 1.012(1.010,1.030) |
| $O_3$ | 0.980(0.972,1.001) | **1.081(1.003,1.093)** | 0.992(0.992,1.020) |




(a) Dados originais

(b) Resíduos do VAR

Figura 12: Autocorrelação e correlação cruzada das componentes principais das séries $PM_{10}$, $SO_2$, $NO_2$, $O_3$ e CO.

# 6 Conclusão

Este artigo considera efeito de diferentes estruturas de autocorrelação na técnica ACP. O principal feito deste artigo foi mostrar que as componentes principais apresentam correlação cruzada dependendo da estrutura de autocorrelação presente nos dados originais. Foi verificado que a técnica pode ser empregada quando uma fraca autocorrelação está presente, pois seja de forma descritiva ou inferencial a autocorrelação não influencia nos resultados finais. Esses resultados foram confirmados de forma teórica e empírica e aplicada a problemas reais da área de poluição do ar.

# Referências


Anderson, T. W. (2003), *An Introduction to Multivariate Statistical Analysis*, 3rd edn, John Wiley & Sons.

Arditsoglou, A. & Samara, C. (2005), 'Levels of total suspended particulate matter and major trace elements in Kosovo: a source identification and apportionment study', *Chemosphere* **59**, 669–678.

Belis, C. A., Karagulian, F., Larsen, B. & Hopke, P. K. (2013), 'Critical review and meta-analysis of ambient particulate matter source apportionment using receptor models in Europe', *Atmospheric Environment* **69**, 94–108.

Borbon, A., Locoge, N., Veillerot, M., Galloo, J. C. & Guillermo, R. (2002), 'Characterisation of NMHCs in a french urban atmosphere: overview of the main sources', *the Science of the Total Environment* **292**, 177–191.

Brockwell, P. J. & Davis, R. A. (1987), *Time Series: Theory and Methods*, Springer-Verlag.

Cadima, J. & Jolliffe, I. T. (1995), 'Loading and correlations in the interpretation of principle components', *Journal of Applied Statistics* **22**, 203–214.

Cohen, S. J. (1983), 'Classification of 500 mb height anomalies using obliquely rotated principal components', *J. Climate Appl. Meteorol.* **22**, 1975–1988.

Ding, C. & He, X. (2004), K-means clustering via principal components analysis, pp. 225–232.

Ferreira, D. F. (2008), *Estatística Multivariada*, Lavras: UFLA.

Fujikoshi, Y. (1980), 'Asymptotic expansions for the distributions of the sample roots under nonnormality', *Biometrika* **67**, 45–51.

Furrer, R. (2005), 'Covariance estimation under spacial dependence', *Journal of Multivariate Analysis* **94**, 366–381.

Gonçalves, F. L. T., Carvalho, L. M. V., Conde, F. C., Latorre, M. R. D. O., Saldiva, P. H. N. & Braga, A. L. F. (2005), 'The effects of air pollution and meteorological parameters on respiratory morbidity during the summer in São Paulo city', *Environment International* **31**, 343–349.

Greenaway-McGrevy, R., Han, C. & Sul, D. (2012), 'Estimating the number of common factors in serially dependent appoximate factor models', *Economics Letters* **116**, 531–534.

Hamilton, J. D. (1994), *Time Series Analysis*, Princeton University Press.

Hannan, E. J. (1970), *Multiple Time Series*, John Wiley.

Hannan, E. J. & Deistler, M. (1988), *The Statistical Theory of Linear Systems*, John Wiley.

Ibarra-Berastegi, G., Elias, A., Barona, A., Saenz, J., Ezcurra, A. & Diaz de Argandonã, J. (2008), 'From diagnosis to prognosis for forecasting air pollution using neural networks: air pollution monitoring in Bilbao', *Environmental Modelling and Software* **23**, 622–637.

Jaimungal, S. & Ng, E. K. H. (2007), Consistent functional pca for financial time-series, pp. 103–108.

Johnson, R. A. & Wichern, D. W. (1998), *Applied Multivariate Statistical Analysis*, 6th edn, Prentice Hall.

Jollife, I. T. (2002), *Principal component analysis*, 2th edn, Prentice Hall.

Juneng, L., Latif, M. T., Tangang, F. T. & Mansor, H. (2009), 'Spatio-temporal characteristics of $PM_{10}$ concentration across Malaysia', *Atmospheric Research* **43**, 4584–4594.

Karar, K. & Gupta, A. (2007), 'Source apportionment of $PM_{10}$ at residential and industrial sites of an urban region of Kolkata, India', *Atmospheric Research* **84**, 30–41.

Lehman, J., Swinton, K., Bortnick, S., Hamilton, C., Baldridge, E., Eder, B. & Cox, B. (2004), 'Spatio-temporal characterization of tropospheric ozone across the eastern United States', *Atmospheric Environment* **38**, 4357–4369.

Liu, P.-W. G. (2009), 'Simulation of the daily average $PM_{10}$ concentrations at Ta-Liao with Box-Jenkins time series models and multivariate analysis', *Atmospheric Environment* **43**, 2101–2113.

Lütkepohl, H. (2005), *New Introduction to Multiple Time Series Analysis*, Springer-Verlag.

Namdeo, A. & Bell, M. (2005), 'Characteristics and health implications of fine and coarse particulates at roadside, urban background and rural sites in UK', *Environment International* **31**, 565–573.

Perez-Neto, P. R., Jackson, D. A. & Somers, K. M. (2005), 'How many principal components? Stopping rules for determining the number of non-trivial axes revisited', *Computacional Statistics & Data Analysis* **49**(4), 974–997.

Pires, J. C. M., Souza, S. I. V., Pereira, M. C., Alvim-Ferraz, M. C. M. & Martins, F. G. (2008*a*), 'Management of air quality monitoring using principal component and cluster analysis — part I: $SO_2$ and $PM_{10}$', *Atmospheric Environment* **42**, 1249–1260.

Pires, J. C. M., Souza, S. I. V., Pereira, M. C., Alvim-Ferraz, M. C. M. & Martins, F. G. (2008*b*), 'Management of air quality monitoring using principal component and cluster analysis — part II: CO, $NO_2$ and $O_3$', *Atmospheric Environment* **42**, 1261–1274.

Priestley, M. B. (1983), *Spectral Analysis and Time Series, Vol. I e II*, Academic Press.

Rao, C. R. (1964), 'The use and interpretation of principal component analysis in applied research', *Sankhya A* **26**, 329–358.

Reinsel, G. C. (1997), *Elements of Multivariate Time Series Analysis*, 2th edn, Springer-Verlag.

Reisen, V. A., Zamprogno, B., Palma, W. & Arteche, J. (2010), 'Seasonal fractional long-memory processes. A semiparametric estimation approach', *arXiv:1011.5631* .

Richman, M. B. (1986), 'A principal component analysis of sulphur concentrations in the western United States', *Atmospheric Environment* **20**, 606–607.

Roberts, S. & Martin, M. (2006), 'Using supervised principal components analysis to assess multiple pollutant effects', *Environmental Health Perspectives* **116**(12).

Romero, R., Ramis, C., Guijarro, J. A. & Sumner, G. (1999), 'Daily rainfall affinity areas in Mediterranean Spain', *Int. J. Climatol* **19**, 557–578.

Shi, G.-L., Li, X., Yin-Chang, F., Wang, Y.-Q., Wu, J.-H., Jun, L. & Tan, Z. (2009), 'Combined source apportionment, using positive matrix factorization-chemical mass balance and principal component analysis/multiple linear regression-chemical mass balance models', *Atmospheric Environment* **43**, 2929–2937.

Soares, I. P. (2011), Avaliação do uso de diferentes modelos receptores para determinação da contribuição das fontes de partículas totais em suspensão, Master's thesis, Departamento de Engenharia Ambiental, Universidade Federal do Espírito Santo, Vitória.

Souza, J. B. (2013), Análise de componentes principais e a modelagem linear generalizada: uma associação entre o número de atendimentos hospitalares por causas respiratórias e a qualidade do ar, na região da grande vitória, es, Master's thesis, Departamento de Engenharia Ambiental, Universidade Federal do Espírito Santo, Vitória.

Statheropoulos, M., Vassiliadis, N. & Pappa, A. (1998), 'Principal component and canonical correlation analysis for examining air pollution and meteorological data', *Atmospheric Environment* **32**(6), 1087–1095.

Taniguchi, M. & Krishnaiah, P. R. (1987), 'Asymptotic distributions of functions of the eigenvalues of sample covariance matrix and canonical correlation matrix in multivariate time series', *Journal of Multivariate Analysis* **22**, 156–176.

Wang, H. & Shooter, D. (2004), 'Source apportionment of fine and coarse atmospheric particles in Auckland, New Zealand', *Science of the Total Environment* **340**, 189–198.

Wang, Y. & Pham, H. (2011), 'Analyzing the effects of air pollution and mortality by generalized additive models with robust principal components', *Int J. Syst. Assur. Eng. Manag* **2**, 253–259.

White, D., Richman, M. & Yarnal, B. (1991), 'Climate regionalization and rotation of principal components', *Int. J. Climatol.* **11**, 1–25.

Zimmermann, C. M., Guimarães, O. M. & Peralta-Zamora, P. G. (2008), 'Avaliação da qualidade do corpo híbrido do rio Tibagi na região de Ponta Grossa', *Química Nova* **31**(7), 1727–1732.

# Sobre o gerenciamento de redes de monitoramento da qualidade do ar: uma aplicação da análise de componentes principais com dados correlacionados

*Bartolomeu Zamprogno*[*]
Programa de Pós-Graduação em Engenharia Ambiental - UFES, Vitória, ES.
Departamento de Estatística, UFES, Vitória, ES.

**Resumo**

Este artigo explora a técnica de análise de componentes principais proposta por Brillinger (1969) em dados de séries temporais com característica de longa dependência. O método original considera que a matriz de covariância do processo seja absolutamente somável, situação que não ocorre em processos de memória longa. No estudo, diferentes estruturas de correlação são consideradas e os resultados simulados e aplicados mostram que a técnica pode ser empregada em processos com essa característica. A técnica é aplicada a dados de concentração do poluente $PM_{10}$, com enfoque de gerenciar rede de monitoramento do poluente, e comparada com a usual metodologia de análise de componentes principais no domínio do tempo.

*Palavras-chave:* Componentes principais, rede de monitoramento, poluição do ar.

## 1 Introdução

A qualidade do ar urbano tem-se deteriorado principalmente como consequência da urbanização acelerada, do crescimento da população, da industrialização e da ausência de adequada política ambiental. Para controlar os níveis de poluição, normas específicas têm sido elaboradas de tal maneira que a concentração de determinado poluente seja considerada aceitável. Inúmeros estudos têm encontrado associações significativas entre os níveis diários de concentração dos poluentes e o número de atendimentos por causas respiratórias ou cardiovasculares. Para detalhes ver, e.g., Ostro et al. (1999), Wong et al. (2002), Villeneuve et al. (2012), entre outras.

Redes de monitoramento da qualidade do ar são importantes para avaliar se os valores das concentrações dos poluentes estão de acordo com os padrões estabelecidos pelas agências reguladoras de poluição de determinada localidade. Os equipamentos que monitoram os níveis dos poluentes podem ser adequadamente manejados em qualquer rede de monitoramento da qualidade do ar. Assim, o número de estações de monitoramento que constituem uma rede pode ser otimizada com intuito de reduzir despesas e ao mesmo tempo garantir a caracterização adequada da qualidade do ar regional (ver, e.g., Pires et al. 2008*a*,*b*).

Entre os métodos estatísticos utilizados para avaliar medidas redundantes destacam-se a análise de componentes principais (ACP) e análise de cluster (AC). Estas técnicas estatísticas

[*]Email: bzamprogno@yahoo.com.br

têm sido aplicadas em vários estudos que tratam a gestão de redes de monitoramento de água (ver, e.g., Kannel et al. 2007, Mendiguchía et al. 2004, Shrestha & Kazama 2007, Singh et al. 2004, 2005, entre outros). No entanto, a aplicação da ACP e AC tem sido pouco explorada para a avaliação da gestão da qualidade do ar. Destacam-se os estudos conduzidos por Gramsch et al. (2006) e Pires et al. (2008*a*,*b*). No primeiro estudo, a AC permitiu identificar as tendências sazonais e espaciais das concentrações de $PM_{10}$ e $O_3$ na cidade de Santiago de Chile (Chile). Os autores concluíram que a cidade tem quatro grandes setores com comportamentos semelhantes nos níveis de poluição do ar. Os resultados do estudo indicam que tal comportamento pode ser causado pelas características topográficas e meteorológicas da região.

Pires et al. (2008*a*,*b*) aplicaram a ACP e a AC na identificação dos locais de monitoramento com semelhante comportamento nas concentrações de $SO_2$, $PM_{10}$, CO, $NO_2$ e $O_3$ na rede de monitoramento da região metropolitana da cidade do Porto (Portugal). A metodologia permitiu identificar problemas no gerenciamento da rede de monitoramento de poluição do ar, deslumbrando a necessidade de realocação de equipamentos para outros locais.

ACP é uma metodologia estatística que consiste na criação de variáveis não-correlacionadas, resultado da combinação linear das variáveis originais e, tal que, a combinação linear entre as variáveis explique a maior fração de variabilidade do conjunto de observações. Para que os resultados obtidos a partir dessa metodologia sejam válidos é necessário que exista independência estatística entre as variáveis sob estudo (para detalhes ver, e.g., Jolliffe 2008). Na maioria dos casos, a suposição de independência entre as variáveis não é estatisticamente testada e, muitas vezes, adota-se naturalmente nas observações.

No cenário com variáveis autocorrelacionados, Zamprogno et al. (2013) evidenciaram a influência causada pela estrutura de correlação, dos dados originais, na construção das combinações lineares entre as variáveis sob estudo. Os autores mostraram que a metodologia usual de ACP fornece componentes correlacionadas, consequência da estrutura de dependência das variâveis originais. Quando a dependência entre as variáveis é fraca, o uso da metodologia usual não causa efeitos significativos nas análises.

Baseado na representação espectral das variáveis, Brillinger (1969) propõe uma metodologia alternativa para a ACP em variáveis com dependência fraca. Entre as vantagens da metodologia, em relação a técnica usual, destacam-se: a detecção de ciclos específicos entre as variáveis sob estudo e a contemplação das associações entre as variáveis em diferentes frequências. Utilizando um filtro linear, o autor obtêm componentes ortogonais, até mesmo em defasagens diferentes de zero, ou seja, as componentes resultantes têm valor zero para a coerência em todas as frequências.

Keller (2000) explora uma outra vantagem da metodologia no domínio da frequência, proposta por Brillinger (1969). O autor analisa variáveis psicológicas de cuidados intensivos comparando a metodologias de ACP usual e no domínio da frequência. No estudo identificou-se que padrões clinicamente relevantes, como outliers e mudança de nível, são melhor capturados com a metodologia no domínio da frequência, além de utilizar um menor número de componentes com essa metodologia.

No exemplo dado em Brillinger (1981, cap. 9), seção 9.6, a segunda componente principal resulta em amplitude e fase de difícil interpretação. Uma possibilidade de eliminar o efeito da autocorrelação nas componentes oriundas da usual metodologia de ACP e as dificuldades de interpretação das componentes resultantes dos métodos do domínio da frequência é considerar a proposta de Zamprogno et al. (2013). Os autores propõem utilizar o método usual de ACP

com filtro adequado que elimine os efeitos da autocorrelação presente nos dados.

Zamprogno et al. (2013) observou que as séries de registros da concentração de poluentes apresentam memória longa. No domínio do tempo, a usual definição de memória longa é a condição $\sum_{h=0}^{\infty} |\gamma(h)| = \infty$, onde $\gamma(h)$ é a função de autocovariância no *lag* $h$ do processo, e, no domínio da frequência, essa propriedade é definida pelo fato da densidade espectral do processo tornar-se ilimitada em algumas frequências entre $[0, \pi]$. O método de Brillinger aplica-se em processos com função de autocovariância absolutamente somável. A proposta principal deste artigo é avaliar, com simulações, o método de Brillinger em processos com memória longa e aplicar a metodologia em rede de monitoramento da qualidade do ar com foco no gerenciamento da rede. Este estudo também compara a extensão de Brillinger com a proposta de Zamprogno et al. (2013).

Este artigo está dividido como se segue. A Seção 2 descreve os conceitos básicos da análise de componentes principais para dados correlacionados e não-correlacionados. A Seção 3 apresenta as análises dos resultados empíricos obtidos dos experimentos de Monte Carlo. Na Seção 4 são comparadas as metodologias de ACP usual e a extensão do método de Brillinger (1969). Finalmente, a Seção 5 apresenta as conclusões e alguns comentários finais sobre a análise dos dados.

# 2 Metodologia

## 2.1 Processo de séries temporais

Seja $\mathbf{X}_t = (X_{1t}, \ldots, X_{kt})'$ um vetor de observações de dimensão $k \times 1$ no tempo $t$ que safisfaz a equação de um processo vetorial autorregressivo média móveis fracionário (VARFIMA) definido como

$$\mathbf{\Phi}(B)\mathbf{X}_t = \mathcal{D}(B)\mathbf{\Theta}(B)\boldsymbol{\xi}_t, \quad t = 1, 2, \ldots, n, \tag{1}$$

onde $B$ é o operador de defasagem, $\boldsymbol{\mu}$ é o vetor de médias e $\boldsymbol{\xi}_t$ é o ruído branco com $\mathrm{E}(\boldsymbol{\xi}_t) = 0$ e $\mathrm{Var}(\boldsymbol{\xi}_t) = \Sigma$. Os operadores $\mathbf{\Phi}(B) = I - \sum_{l=1}^{p} \Phi_l B^l$ e $\mathbf{\Theta}(B) = I + \sum_{l=1}^{q} \Theta_l B^l$ são matrizes polinomiais com ordem $p, q$ respectivamente, $I$ é a matriz identidade de dimensão $k \times k$ e $\Phi_l$ e $\Theta_l$ são matrizes $k \times k$ de constantes. O operador $\mathcal{D}(B)$ é a matriz do operador diferença fracionário tal que $\mathcal{D}(B) = \mathrm{diag}\left\{(1-B)^{d_1}, (1-B)^{d_2}, \ldots, (1-B)^{d_k}\right\}$, onde

$$(1-B)^{\text{-}d_l} = \sum_{k=0}^{\infty} \frac{\Gamma(d_l + k)}{\Gamma(d_l)\Gamma(k+1)} B^k, \quad l = 1, 2, \ldots, k,$$

com $d_l \in$ (-0.5, 0.5) e $\Gamma(\cdot)$ é a função Gamma. Assume-se que os polinômios $\Phi(B)$, $\Theta(B)$ e $d_l$ satisfazem as condições de estacionariedade e invertibilidade, ver Robinson (1995) para mais detalhes.

A função densidade espectral do processo $\mathbf{X}_t$, na frequência $\omega$, é dada por

$$f(\omega) = \mathcal{D}\left(e^{i\omega}\right)^{\text{-}1} f_{ST}(\omega) \left[\mathcal{D}\left(e^{i\omega}\right)^{\text{-}1}\right]^*,$$

$f_{ST}(\omega) = \frac{1}{2\pi} \mathbf{\Phi}_p\left(e^{i\omega}\right)^{\text{-}1} \mathbf{\Theta}_q\left(e^{i\omega}\right) \Sigma\, \mathbf{\Theta}_q\left(e^{i\omega}\right)^* \left[\mathbf{\Phi}_p\left(e^{i\omega}\right)^{\text{-}1}\right]^*$ e $A^*$ representa o transposto conjugado da matriz complexa $A$.

O parâmetro $d_l$ representa a memória do $l$-ésimo elemento de $\mathbf{X}_t$, isto é, para $d_l = 0, d_l \in$ (-0.5, 0) ou $d_l \in (0, 0.5)$, o processo é dito ter memória curta, memória intermediária ou memória longa, respectivamente.

Um processo é dito ter matriz de autocovariância, $\Gamma_{\mathbf{X}}(h)$, absolutamente somável quando, individualmente, cada um dos elementos $\gamma_{ij}(h) = E[(X_{t+h,i} - \mu_i)(X_{t,j} - \mu_j)]$, $i, j = 1, \ldots, k$, da matriz formam uma sequência absolutamente somável, Hamilton (1994, p. 262). A propriedade de memória longa está relacionada com o comportamento da função de autocovariância, que não é absolutamente somável ou, alternativamente, a função densidade espectral torna-se ilimitada nas frequências zero e cíclicas.

Como $\omega \to 0^+$, nós temos

$$f_{ST}(\omega) = \frac{1}{2\pi} \mathbf{\Phi}_p\,(1)^{\text{-}1}\, \mathbf{\Theta}_q\,(1)\, \Sigma\, \mathbf{\Theta}_q\,(1)^*\, \left[\mathbf{\Phi}_p\,(1)^{\text{-}1}\right]^* \sim G,$$

onde $G$ é uma matriz positiva definida real e simétrica. Daí,

$$f(\omega) \sim \mathcal{D}\left(1 - \omega e^{i\frac{(\omega-\pi)}{2}}\right)^{\text{-}1} G \left[\mathcal{D}\left(1 - \omega e^{i\frac{(\omega-\pi)}{2}}\right)^{\text{-}1}\right]^*,$$

ou $f(\omega) \sim \Lambda(\omega; d) G \Lambda^*(\omega; d)$, onde $\Lambda(\omega; d) = \mathcal{D}\left(1 - \omega e^{i\frac{(\omega-\pi)}{2}}\right)^{\text{-}1}$ e o símbolo "$\sim$"significa que a razão entre os lados esquerdo e direito tende a 1.

Seja $I(\omega_j)$ a função periodograma de $\mathbf{X}_t$ avaliada nas frequências de Fourier $\omega_j = \frac{2\pi j}{n}$ dada por

$$I(\omega_j) = \frac{1}{2\pi n} \left(\sum_{t=1}^{n} \mathbf{X}_t e^{it\omega_j}\right) \left(\sum_{t=1}^{n} \mathbf{X}_t e^{it\omega_j}\right)^*,$$

onde $j = 1, 2, \ldots, \lfloor \frac{n}{2} \rfloor$ e $\lfloor \cdot \rfloor$ denota a parte inteira. A função periodograma é um estimador da função densidade espectral do processo $\mathbf{X}_t$ e pode ser calculada rapidamente pela transformada rápida de Fourier, mesmo quando $n$ é muito grande.

Do Teorema 7.2.4 em Brillinger (1983, p. 238), sendo $s(n)$ um inteiro com $2\pi s(n)/n$ perto $\omega \not\equiv 0 \pmod{\pi}$, a distribuição de $I(2\pi[s(n) + s]/n)$, $s = 0, \pm 1, \ldots, \pm m$, pode ser aproximada por $2m + 1$ distribuições independentes Wishart complexas, $W_r^c(1, f(\omega))$. Essas observações sugerem considerar como estimador da função espectral

$$f^{(n)}(\omega) = (2m+1)^{-1} \sum_{s=-m}^{m} I\left(\frac{2\pi[s(n) + s]}{n}\right) \quad \text{se} \quad \omega \not\equiv 0 (\text{mod } \pi). \tag{2}$$

A coerência ($C_{X_l X_j}$) entre duas séries $X_{lt}$ e $X_{jt}$ está definida em Shumway e Stoffer (2000) como

$$C_{X_l X_j} = \frac{|\, f_{lj}(\omega)\,|^2}{f_l(\omega) f_j(\omega)}.$$

A coerência é uma medida análoga ao coeficiente de correlação e satisfaz $0 \leq C_{X_l X_j} \leq 1$. Quando $C_{X_l X_j} = 0$ não há coerência e quando $C_{X_l X_j} = 1$ há perfeita coerência. Para mais detalhes, veja Priestley (1983).

## 2.2 ACP para dados independentes

A usual técnica de análise de componentes principais busca as direções que captura o maior percentual de variação entre os dados mensurados. Essa técnica depende exclusivamente da matriz de covariâncias dos dados ver, por exemplo, Anderson (2003) e Johnson e Wichern (1998). A técnica de ACP é baseada na teoria algébrica de vetores e, usualmente, aplicada a dados não autocorrelacionados, isto é, processos com $\Gamma(h) = 0,\ \forall\ h \neq 0$, por exemplo, a matriz $\Sigma$. A técnica consiste em obter da matriz de covariância $\Gamma(0)$, os autovalores $(\lambda_1, \ldots, \lambda_k)$ e os correspondentes autovetores $ß = (\boldsymbol{\beta}_1^{'}, \ldots, \boldsymbol{\beta}_k^{'})^{'}$. O vetor de componentes é dado por $\mathbf{Y} = ß\mathbf{X}$, onde $Y_i = \boldsymbol{\beta}_i^{'}\mathbf{X}$, $i = 1, \ldots, k$, e $\boldsymbol{\beta}_i^{'} = (\beta_{i1}, \ldots, \beta_{ik})$ é o autovetor da matriz de covariância $\Gamma(0)$ e a variabilidade explicada de cada componente principal $Y_i$ é dada pelo autovalor $\lambda_i$, associado a $\boldsymbol{\beta}_i$.

## 2.3 ACP no domínio da frequência

ACP descrita na subseção 2.2 captura relações somente entre medidas simultâneas. Brillinger (1969) sugere uma versão dinâmica de ACP para séries temporais que leva em consideração a correlação entre as séries em várias defasagens temporais assumindo que a série é estacionária com função de autocovariância absolutamente somável. O método proposto considera em conjunto todas as frequências, sendo que as componentes resultantes são séries temporais, com coerência zero em todas as frequências, ver Teorema 9.3.2 de Brillinger (1981, p. 345). Seja o número complexo $a + bi$, as componentes resultantes do método são analisadas através das funções amplitude $(\sqrt{a^2 + b^2})$ e fase $(\tan^{-1}(b/a)$, ou seja, é o argumento do número complexo). O método é descrito a seguir.

Seja $\mathbf{X}_t = (X_{1t}, X_{2t}, \ldots, X_{kt}),\ t = 0, \pm 1, \ldots$ um vetor estacionário de segunda ordem de dimensão $k$ com média $c_{\mathbf{x}}$, matriz de autocovariância absolutamente somável $\Gamma(h),\ h = 0, \pm 1, \pm 2, \ldots$ e matriz densidade espectral $f(\omega)$. Seja $\zeta_t = [\zeta_{1t}, \zeta_{2t}, \ldots, \zeta_{kt}],\ q \leq k$ a série de tempo multivariada que representa as componentes principais propostas por Brillinger definida por

$$\zeta_t = \sum_h b_{t-h}\mathbf{X}_h,$$

onde $b_t$ é um filtro matricial de ordem $q \times k$ que projeta a série original $\mathbf{X}_t$ em $\zeta_t$, dado por

$$b_t = \frac{1}{2\pi}\int_0^{2\pi} B(\omega)e^{it\omega}d\omega,$$

onde $B(\omega) = [\overline{V_1(\omega)^{'}}\ \ldots\ \overline{V_q(\omega)^{'}}]^{'}$, $\overline{V_j(\omega)}$ é o conjugado de $V_j(\omega)$ e $V_j(\omega)$ é o $j$-ésimo vetor próprio, $j = 1, \ldots, k$, associado ao $j$-ésimo maior autovalor $\lambda_j(\omega)$ da matriz espectral. Observe que $b_t$ é um coeficiente que considera a matriz espectral em todas as frequências do intervalo 0 a $2\pi$.

A escolha do número de componentes pode ser obtida pelo gráfico da máxima variabilidade (MV), definido por

$$MV = \int_0^{2\pi} \lambda_j(\alpha)d\alpha, \tag{3}$$

onde $\lambda_j(\alpha)$ é o $j-$ésimo autovalor da matriz da densidade espectral. Para mais detalhes do método veja Brillinger (1981, cap. 9).

# 3 Simulações

O estudo desta seção tem como objetivo elucidar se a proposta de Brillinger também é plausível para processos com característica de longa dependência, ou seja, processos onde a matriz de covariância não é absolutamente somável.

Seja $\mathbf{X}_t$ o processo (1) com $p = 1,\ q = 0$, onde $\mathbf{X}_t = (X_{1t}, \ldots, X_{4t})'$ tem matriz de covariância $\boldsymbol{\xi}_t$ dada por

$$E(\boldsymbol{\xi}_t \boldsymbol{\xi}_t') = \begin{bmatrix} 127.40 & 30.58 & 47.43 & 62.42 \\ 30.58 & 58.78 & 33.89 & 70.65 \\ 47.43 & 33.89 & 64.17 & 58.49 \\ 62.42 & 70.65 & 58.49 & 172.21 \end{bmatrix}.$$

A análise está dividida em processos com matriz de covariância absolutamente somável e processos com memória longa. Considere para ordem $p = 1$ as matrizes $\Phi_1$ apresentadas na Tabela 1. Primeiramente, seja $\mathcal{D}(B)$ com $d_1 = d_2 = d_3 = d_4 = 0$. Dessa forma, as matrizes $\Phi_1$ dispostas na Tabela 1 definem os Modelos 1, 2, 3 e 4. Tome agora $\mathcal{D}(B) = \text{diag}\left\{(1-B)^{0.35}, (1-B)^{0.25}, (1-B)^{0.3}, (1-B)^{0.4}\right\}$, ou seja, $d_1 = 0.35$, $d_2 = 0.25$, $d_3 = 0.3$ e $d_4 = 0.4$. Defina Modelo 5 o Modelo 1 multiplicado por $\mathcal{D}(B)$. Defina Modelo 6 o Modelo 2 combinado com $\mathcal{D}(B)$. Defina Modelo 7 o Modelo 3 combinado com $\mathcal{D}(B)$ e defina Modelo 8 o Modelo 4 combinado com $\mathcal{D}(B)$. Dessa forma, os Modelos 5, 6, 7 e 8 apresentam memória longa.

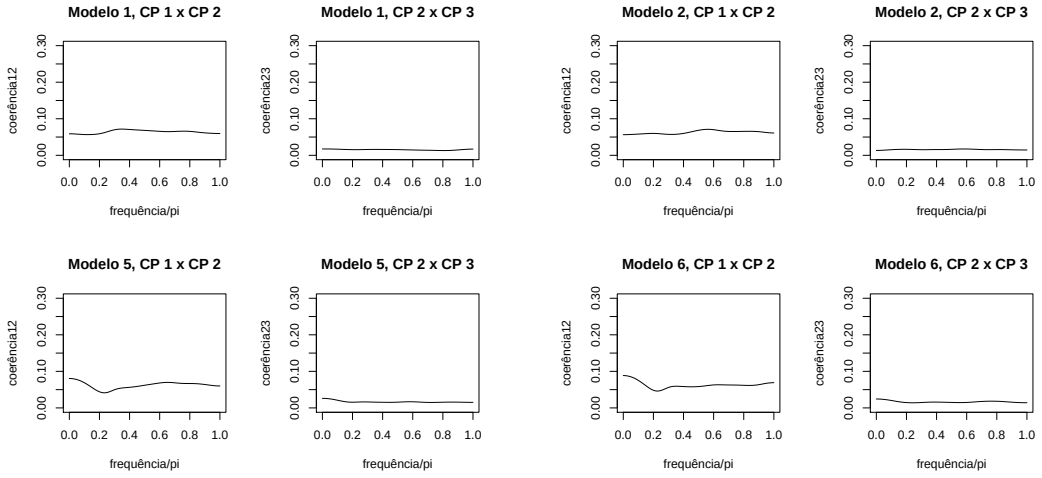
Dentre esses modelos, os autovalores da matriz $\Phi_1$ do Modelo 4 são 0.9914, 0.5053, 0.1760 e 0.0272. Observe que um dos autovalores está próximo de 1, ou seja, da região de não estacionariedade. Esse modelo deve portanto ser avaliado com maior atenção.

Tabela 1: Matrizes de $\Phi$ para os processos VAR(1).

| $\Phi_1$ (**Modelo 1**) | | | | $\Phi_1$ (**Modelo 2**) | | | |
|---|---|---|---|---|---|---|---|
| 0.0 | 0.0 | 0.0 | 0.0 | 0.12 | 0.00 | 0.03 | 0.00 |
| 0.0 | 0.0 | 0.0 | 0.0 | 0.05 | 0.08 | 0.00 | 0.01 |
| 0.0 | 0.0 | 0.0 | 0.0 | 0.00 | 0.00 | 0.10 | 0.00 |
| 0.0 | 0.0 | 0.0 | 0.0 | 0.01 | 0.02 | 0.00 | 0.05 |
| $\Phi_1$ ( **Modelo 3**) | | | | $\Phi_1$ ( **Modelo 4**) | | | |
| 0.4 | 0.0 | 0.2 | 0.1 | 0.6 | 0.3 | 0.0 | 0.3 |
| 0.0 | 0.1 | 0.0 | 0.0 | 0.1 | 0.2 | 0.0 | 0.1 |
| 0.3 | 0.0 | 0.2 | 0.0 | 0.1 | 0.8 | 0.4 | 0.2 |
| 0.0 | 0.1 | 0.0 | 0.6 | 0.2 | 0.0 | 0.2 | 0.5 |

Os resultados apresentados adiante consideram amostras aleatórias de tamanhos $n = 500$ geradas de uma distribuição normal multivariada com variância $E(\boldsymbol{\xi}_t \boldsymbol{\xi}_t')$ (Modelo 1) replicando 1000 vezes os Modelos 1, 2, 3, 4, 5, 6, 7 e 8. Os valores dos coeficientes $b_t$ foram obtidos por soma de Riemann.

As Figuras 1 e 2 apresentam os valores médios das coerências entre a primeira e segunda componentes (denominado CP1 × CP2) e entre a segunda e terceira componente (denominado CP2 × CP3) obtidas pela da metodologia de Brillinger para os 8 modelos considerados. Cada subfigura (a) e (b) compara o processo absolutamente somável com o respectivo processo memória longa. Por exemplo, a Figura 1a compara o processo ruído branco, Modelo 1, com o Modelo 5. Percebe-se nesse caso não haver implicação no uso da metodologia em processos longa dependência. Interpretação semelhante pode ser obtida da Figura 1b, apesar de existir uma suave modificação entre as frequências entre 0 e 0.1.




(a) Modelos 1 e 5 (b) Modelos 2 e 6

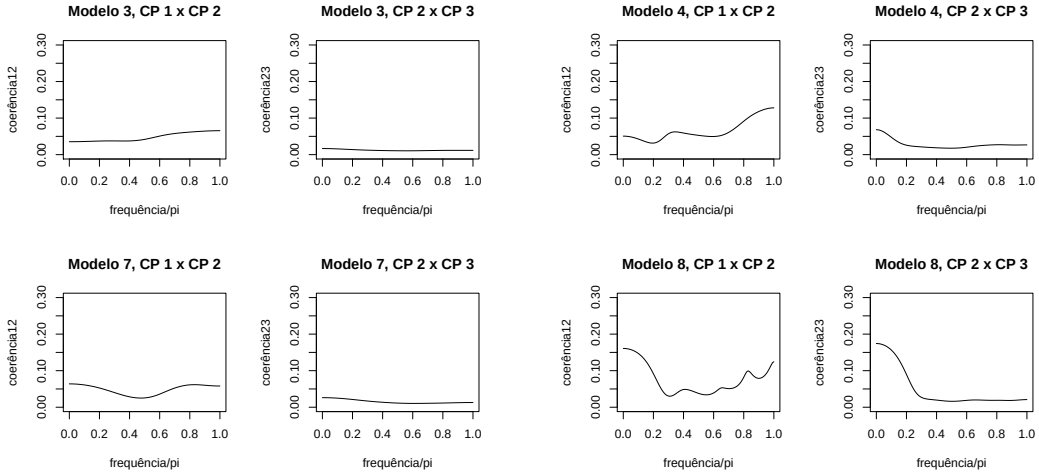
Figura 1: Coerência dos (a) Modelos 1 e 5 e (b) Modelos 2 e 6.

Resultados médios de coerência para os Modelos 3 e 4, Figura 2, também indicam uma leve alteração da coerência, principalmente entre a primeira e segunda componentes, como apresentada no Modelo 6. A média das coerências nesse caso fica próximo a 0.1 em algumas frequências. Deve ser lembrado que o Modelo 4 apresenta autovalores de $\Phi_1$ próximo à região de não estacionariedade.

O Modelo 7 comparado ao Modelo 3, Figura 2a, também não apresenta mudanças no comportamento da coerência. O Modelo 4 apresenta média de coerência próximo a 0.2 em algumas frequências. Esse fato pode ser explicado pelo fato dos autovalores de $\Phi_1$ estarem próximos a região de não estacionariedade e isso, somado ao efeito da matriz $d$, explica os resultados obtidos para o Modelo 8.

Na Figura 3 apresentamos a taxa de rejeição do teste de coerência, ver Priestley (1983, p. 706), entre as duas primeiras componentes principais em todas as frequências para os Modelos 3, 4, 7 e 8. O nível de significância adotado foi de $\alpha = 5\%$ e está representado pela linha tracejada. Percebe-se que há uma oscilação do percentual de rejeição ao redor de 5%, para os Modelos 3, 4 e 7 ao longo de todas as frequências como é esperado. Para o Modelo 8 o percentual de rejeição chega próximo de 20% nas frequências mais baixas. Recorda-se que o Modelo 8 apresenta um autovalor próximo a 1, ou seja, perto da região de não estacionariedade.

Os resultados dessa seção mostram a validade do método de Brillinger para processos com




(a) Modelos 3 e 7 (b) Modelos 4 e 8

Figura 2: Coerência dos (a) Modelos 3 e 7 e (b) Modelos 4 e 8.

característica de memória longa, pois as médias das coerências, em todas as frequências, estão próximas a zero e a taxa de rejeição do teste de coerência está ao redor do nível $\alpha = 5\%$, também em praticamente todas as frequências. Isso corrobora, pelo menos empiricamente, que vale a extensão do método de Brillinger para processos memória longa.

# 4 Aplicação a dados de rede de monitoramento

Essa seção compara proposta de gerenciamento da rede de monitoramento da qualidade do ar através da técnica análise de componentes principais em duas vertentes: domínio do tempo e domínio da frequência. A usual técnica de ACP é considerada no domínio do tempo e confrontada no domínio da frequência com a proposta de Brillinger.

Os dados analisados são registros de médias diárias das concentrações dos poluentes $PM_{10}$, obtidos na rede automática de monitoramento da qualidade do ar (RAMQAR) da Região da Grande Vitória (RGV), compreendidos entre o período de janeiro de 2005 a dezembro de 2009, totalizando 1826 observações.

## 4.1 Análise da concentração do $PM_{10}$

A análise da função de autocorrelação dos dados da concentração do $PM_{10}$ mostra decaimento lento (hiperbólico) e picos sazonais, $s = 7$, para as concentrações do poluente nas localidades Carapina, VVCentro, Cariacica entre outras. Considerando os indícios de sazonalidade e da característica de memória longa nos dados, será adotado a metodologia sazonal VARFIMA para análise dos modelos. O método proposto por Reisen, Zamprogno et al. (2010) foi considerado para estimação dos parâmetros fracionários $d$ e $D$. As estimatimas de $d$ e $D$, e seus desvios padrões (d.p.), estão apresentadas na Tabela 2 para os casos que $H_0$: $d = 0$ e $H_0$: $D = 0$ foram rejeitadas. Essas estimativas dos parâmetros fracionários foram obtidas considerando diferentes valores de bandwidth $m = n^{\alpha}$, onde $0 < \alpha < 1$ e $n$ é o tamanho da

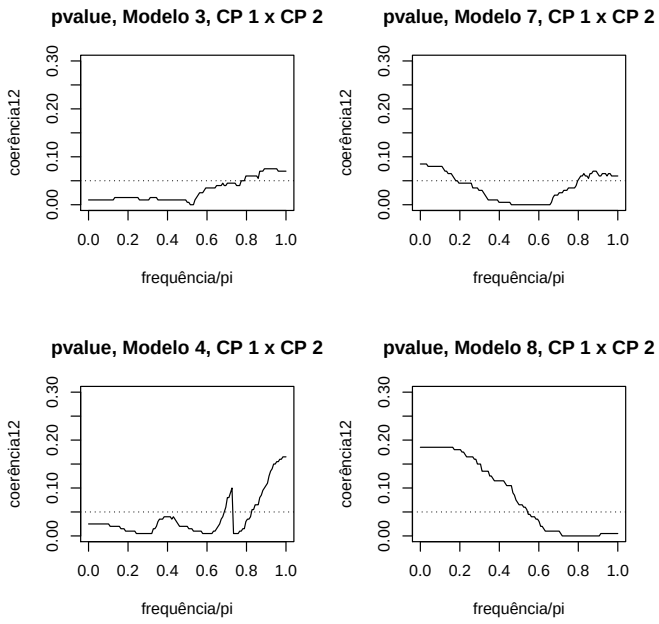



Figura 3: Taxa de rejeição obtida com p-valor para os Modelos 3, 4, 7 e 8.

série. O valor escolhido para $\alpha$ foi 0.5 [ maiores detalhes da escolha de $\alpha$, ver Reisen et al. (2010)]. Como $\hat{d} + \hat{D} < 0.5$, as séries são estacionárias. Análise dos ajustes dos modelos VARMA confirmam a estacionariedade das séries.

Tabela 2: Estimativas dos parâmetros fracionários para poluente $PM_{10}$ (* $H_0$: $d = 0$, $H_0$: $D = 0$ rejeitadas).

| Estação | $\hat{d}$ | d. p. $\hat{d}$ | $\hat{D}$ | d. p. $\hat{D}$ |
|---|---|---|---|---|
| Carapina | 0.2622 | 0.0141 | 0.1572 | 0.0021 |
| Camburi | * | — | 0.1059 | 0.0022 |
| Ibes | 0.2813 | 0.0146 | * | — |
| VVCentro | 0.2664 | 0.0161 | 0.1198 | 0.0024 |
| Cariacica | 0.3043 | 0.0140 | 0.1679 | 0.0021 |

A Tabela 3 apresenta resultados da usual técnica ACP, domínio do tempo, sobre a concentração de $PM_{10}$ em dois contextos. O primeiro considera os dados originais para obtenção dos autovalores e autovetores. O segundo mostra estimativas dessas quantidades após aplicar o filtro VARFIMA$(1, d, 0)(0, D, 0)$, onde as estimativas dos parâmetros fracionários, $\hat{d}$ e $\hat{D}$, estão apresentados na Tabela 2. Aplicar a usual técnica de ACP sobre os dados filtrados é uma proposta de Zamprogno et al. (2013). Nota-se que grande parte da variabilidade fica concentrada na primeira componente principal e que 4 componentes captam aproximadamente 85% da variabilidade total. Com base nessas 4 componentes faz-se as análise em busca de semelhantes padrões de concentração entre as estações de monitoramento.

Para análise da usual técnica de ACP, Tabela 3, será adotado que as estações têm igual padrão de concentração quando o coeficiente do autovetor for superior a 0.37 em módulo. A

Tabela 3: Resultado de ACP para concentração de $PM_{10}$.

| **Estações** | ACP dados originais | | | | ACP sobre dados filtrados | | | |
|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| Laranjeiras | -0.3002 | **0.7193** | -0.1756 | 0.1460 | -0.3067 | **0.7090** | -0.0529 | 0.1606 |
| Carapina | -0.3554 | **-0.4004** | 0.2628 | 0.1750 | -0.3536 | **-0.5233** | 0.0368 | 0.0669 |
| Camburi | -0.3472 | 0.1700 | 0.0502 | **0.7019** | -0.3166 | 0.0560 | **0.7079** | **0.5055** |
| Sua | -0.3632 | 0.2163 | 0.0406 | **-0.6118** | **-0.3722** | 0.2283 | -0.3546 | -0.1360 |
| VixCentro | **-0.3864** | -0.2265 | -0.1026 | -0.1629 | **-0.3856** | -0.0222 | -0.2168 | -0.2125 |
| Ibes | **-0.3869** | 0.1787 | 0.2359 | -0.2271 | **-0.3935** | 0.0625 | -0.1563 | 0.1426 |
| VVCentro | -0.3055 | -0.2942 | **-0.8391** | 0.0141 | -0.3222 | -0.0087 | **-0.4764** | **-0.7571** |
| Cariacica | **-0.3721** | -0.2766 | 0.3542 | 0.0507 | -0.3669 | **-0.4044** | -0.2652 | 0.2383 |
| Eigenvalue | 4.8971 | 0.7744 | 0.6282 | 0.4973 | 4.5586 | 0.7462 | 0.6412 | 0.6050 |
| Proportion | 61.22 | 9.68 | 7.85 | 6.22 | 56.98 | 9.32 | 8.01 | 7.56 |
| Cumulative | 61.22 | 70.90 | 78.75 | 84.97 | 56.98 | 66.30 | 74.31 | 81.87 |

seguir são descritas as análises da usual técnica de ACP aplicada sobre os dados originais. A primeira componente principal dos dados originais indica semelhança entre as concentrações das estações VixCentro, Ibes e Cariacica. Nesse caso a técnica indica que uma das três estações é suficiente para monitorar a concentração nas três localidades. A segunda componente principal aponta similaridade entre as estações localizadas em Laranjeiras e Carapina. No caso da terceira componente, a concentração de $PM_{10}$ da estação VVCentro não apresenta semelhança com nenhuma outra estação. Já a quarta componente principal aponta mesmo perfil para as estações Camburi e Sua. Nesse domínio, observa-se que as quatro primeiras componentes principais englobam as 8 estações da rede de monitoramento, indicando a redução de oito pontos de monitoramento para 4 ou o deslocamento de 4 equipamentos para outras regiões de interesse.

Os resultados obtidos com a usual técnica de ACP sobre os dados filtrados, Tabela 3, indicam resultados diferentes dos obtidos com os dados originais. A primeira componente principal indica semelhança de padrão entre as estações Sua, VixCentro e Ibes. A segunda componente principal aponta mesmo padrão entre as concentrações das estações de Laranjeiras e Carapina, igual aos resultados de ACP sobre os dados originais, mas inclui a estação de Cariacica como semelhante padrão a duas primeiras. A terceira componente principal indica semelhança entre as concentrações de Camburi e VVCentro e esse mesmo padrão é indicado na quarta componente. Ou seja, três componentes são suficientes suficientes para englobar as 8 estações.

Para análise no domínio da frequência serão disponibilizados apenas os gráficos das amplitudes de cada componente principal, pois as fases, em todas as frequências, apresentam valores distintos, ou seja, não há um padrão. Nas Figuras 4 e 5 são apresentadas as amplitudes para cada estação da primeira e da segunda componentes principais. Observa-se das figuras que a variação das amplitudes da segunda componente é superior em relação a primeira componente.

Para facilitar as análises, na Tabela 4 estão apresentadas as médias das amplitudes das três primeiras componentes principais, além da máxima variabilidade (3), em percentual, de cada componente principal. Observa-se que com três componentes o método de Brillinger reduz a dimensão dos dados em menor número de componentes principais em relação a metodologia usual. Três componentes captam 87,2% da variabilidade total pelo método de Brillinger enquanto que com 4 componentes a usual técnica de ACP concentra entre 81% e 85% de

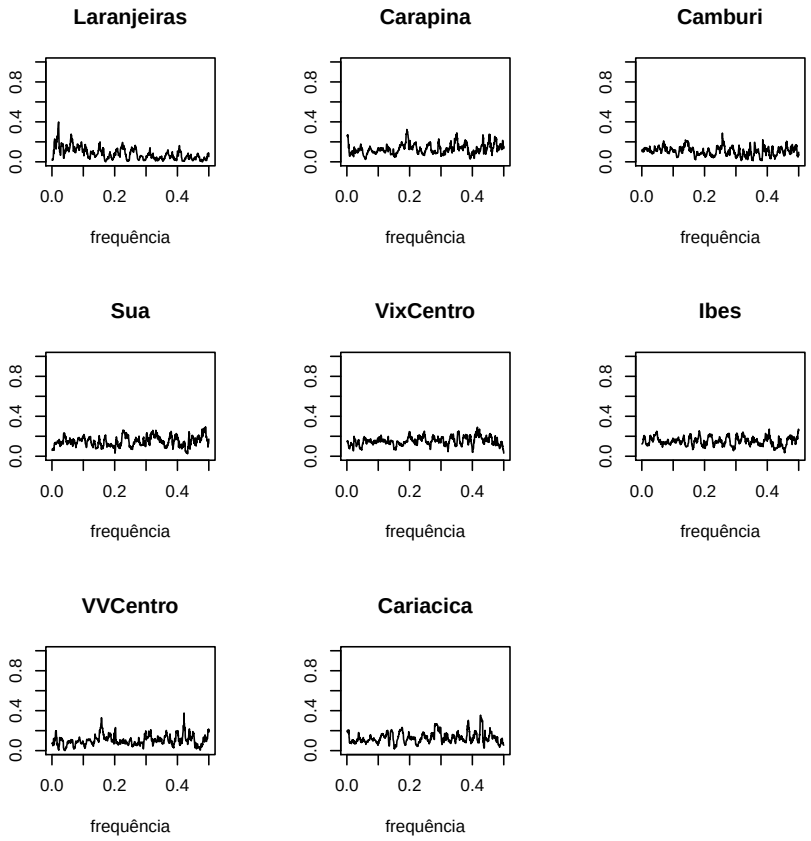



Figura 4: Amplitude da primeira componente principal.

variação. Esse fato mostra uma vantagem dessa metodologia.

De igual maneira à usual técnica de ACP de considerar um valor mínimo para os coeficientes das componentes, a fim de classificar estações em igual padrão de comportamento, adota-se para a ACP no domínio da frequência que o valor médio das amplitudes seja maior ou igual a 0.13. O cenário obtido com as componentes principais obtidas via método de Brillinger são os seguintes descritos. A primeira componente principal indica semelhança de padrão entre as estações Carapina, Sua, VixCentro, Ibes e Cariacica. Destaca-se que as médias das amplitudes das estações de Sua, VixCentro e Ibes (aproximadamente 0.15) estão bem próximas e distante da média de Carapina e Cariacica (igual a 0.13). Isso indica semelhança com a primeira componente principal obtida com a usual técnica da ACP sobre os dados filtrados. A segunda componente principal aponta mesmo padrão de concentração para as estações Laranjeiras, Camburi e VVCentro. E a terceira componente principal apresenta resultados semelhantes a segunda componente principal. Nesse caso, o método de Brillinger reduz a duas componentes a concentração de 8 estações. Logo, dois pontos de monitoramento são suficientes para monitorar o padrão da região nesse período avaliado.

Para verificar se os resultados dos três métodos de ACP empregados indicam semelhança de concentrações entre as estações destacadas por cada componente, alocamos as informações de concentração por dia da semana e por hora. A Figura 6 apresenta resultados da usual

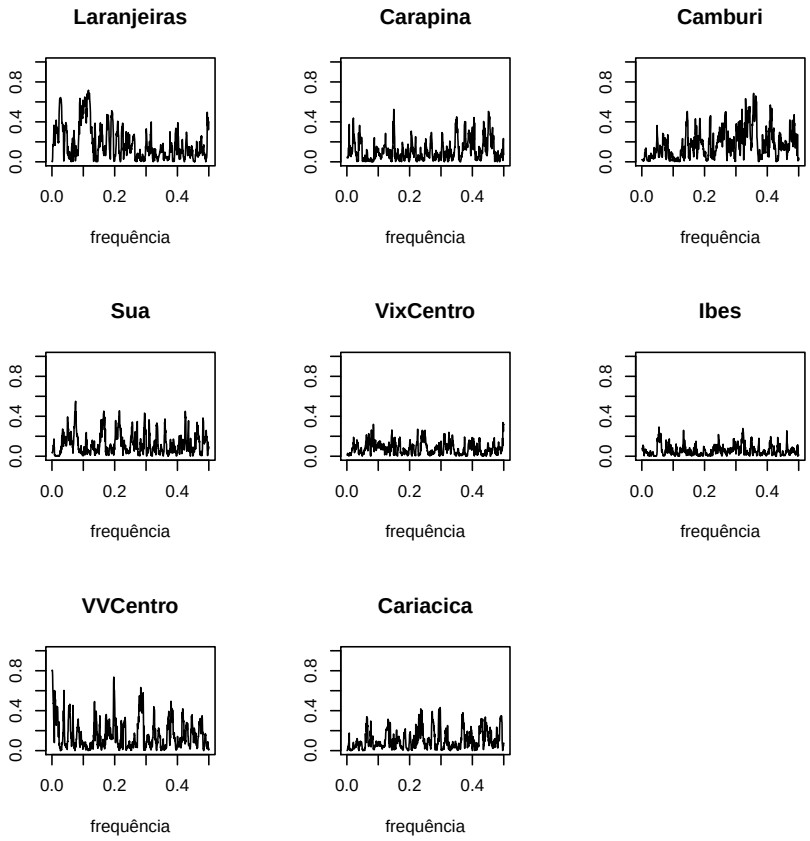



Figura 5: Amplitude da segunda componente principal.

técnica de ACP, sobre dados originais e filtrados, considerando os dados obtidos por dia da semana. Observa-se claramente que os resultados obtidos com os dados filtrados são bem superiores aos apresentados com os dados originais, onde a técnica discrimina os dados com mais clareza, pois fica mais evidente o igual comportamento após passar filtro nos dados.

A Figura 7 apresenta análise semelhante à Figura 6, mas nesse caso as concentrações são disponibilizadas por média horária. Por exemplo, ao comparar a quarta componente dos dados originais com a terceira componente dos dados filtrados é verificado que a concentração horária na estação VVCentro somente não acompanha a concentração em Camburi em um pico noturno que ocorre nesta estação, Figura 7b. E, fica incoerênte dizer que a estação de Camburi tem padrão de concentração semelhante à estação do Sua, Figura 7a, pois picos de concentração ocorrem em período distintos.

Outro ponto a ser destacado é que tanto semanalmente quanto por hora a concentração na estação de Cariacica tem sua trajetória seguida mais fielmente pela concentração de Laranjeiras e Carapina, Figura 6b e Figura 7b, do que de VixCentro e Ibes, Figura 6a e Figura 7a. Dessa forma ACP sobre dados filtrados torna ACP mais discriminativa, ou seja, obtém-se resultados mais expressivos e assertivos.

Na Figura 8 é apresentado comportamento semanal e horário de séries classificadas como semelhantes de acordo com metodologia de Brillinger. Pode-se observar na primeira com-

Tabela 4: MV e médias das amplitudes.

| Estações | Média das amplitudes | | |
|---|---|---|---|
| | CP 1 | CP 2 | CP 3 |
| Laranjeiras | 0.081 | **0.186** | 0.139 |
| Carapina | **0.130** | 0.112 | 0.132 |
| Camburi | 0.106 | **0.179** | 0.156 |
| Sua | **0.149** | 0.124 | 0.121 |
| VixCentro | **0.152** | 0.078 | 0.087 |
| Ibes | **0.150** | 0.056 | 0.088 |
| VVCentro | 0.103 | **0.159** | 0.155 |
| Cariacica | **0.130** | 0.106 | 0.121 |
| Proporção da MV | 0.652 | 0.137 | 0.083 |
| Acumulado da MV | 0.652 | 0.789 | 0.872 |

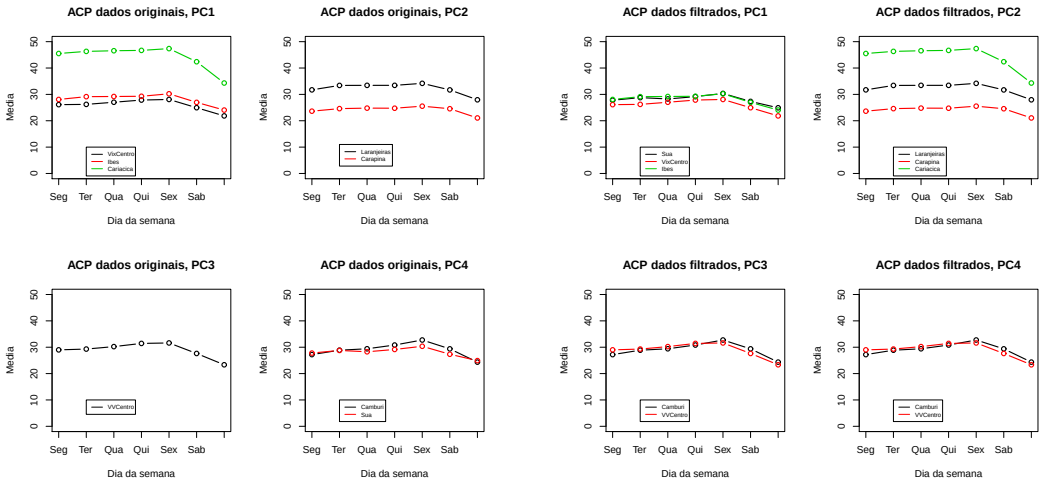
ponente que a concentração de Cariacica destoa das demais, mas cobre toda a concentração das demais estações mesmo sem ter igual comportamento. Já para a segunda componente principal fica incoerente dizer que são semelhantes as concentrações das estações de Laranjeiras e Camburi devido a picos em horários distintos. Dessa maneira, a redução em duas componentes principais não permite distinguir perfeitamente os locais com igual padrão de comportamento.

Na Figura 9 é apresentado a coerência entre a primeira e segunda componentes principais obtidas pelo método de Brillinger. Pode-se observar que apesar das séries de concentração $PM_{10}$ da RAMQAR apresentarem longa dependência sazonal o método resulta em coerência zero em todas as frequências. Isso reforça os resultaodos simulados obtidos.

Mesmo que a proposta de Zamprogno et al. (2013), nessa aplicação, seja levemente superior a técnica proposta por Brillinger, um ponto interessante ao adotar ACP no domínio da frequência é que podemos obsevar os picos das frequências e avaliar a informação sobre uma ótica mais apurada. Em geral, os gráficos da amplitude e fase apresentam frequências que se destacam em relação às demais, o que indica alguma característica do dados originais. Nessa aplicação nenhuma frequência sobresaiu em relação às demais.

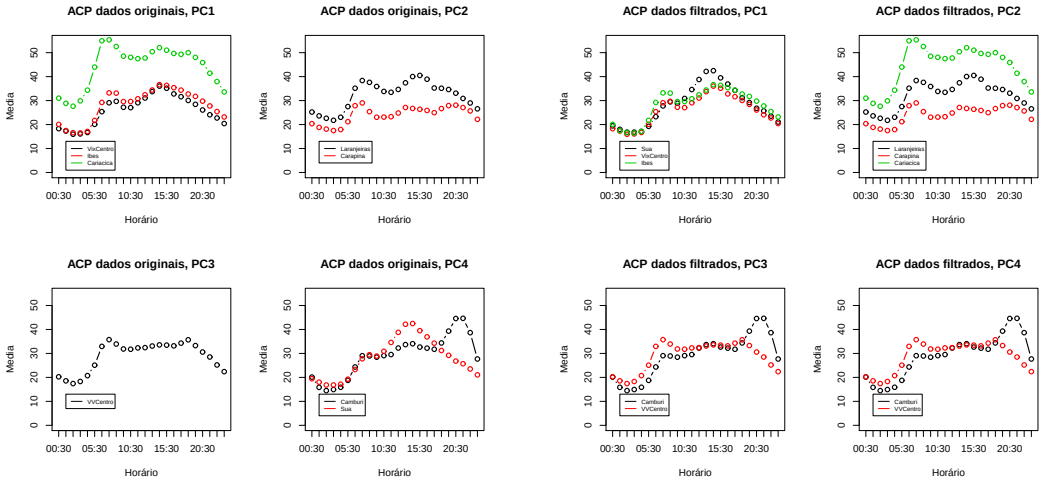
# 5 Conclusões

Estudos de simulação foram realizados para avaliar a extensão da metodologia de ACP proposta por Brillinger (1969) para séries com memória longa. Os resultados mostram a viabilidade da técnica nesse tipo de processo. No estudo aplicado de gerenciamento da rede de monitoramento, a técnica forneceu resultados satisfatórios, mas com resultado levemente inferior aos fornecidos pelo método proposto por Zamprogno et al. (2013). A vantagem do menor número de componentes principais, obtida com o método de Brillinger, não teve como contrapartida o melhor resultado final.




(a) Modelos 3 e 7 (b) Modelos 4 e 8

Figura 6: Concentração semanal de $PM_{10}$ com mesmo padrão de acordo com ACP de (a) Dados originais e (b) Dados filtrados.




(a) Modelos 3 e 7 (b) Modelos 4 e 8

Figura 7: Concentração horária de $PM_{10}$ com mesmo padrão de acordo com ACP de (a) Dados originais e (b) Dados filtrados.

# Referências

Brillinger, D. (1969), The canonical analysis of stationary time series, *in* P. R. Krishnaiah, ed., 'Multivariate analysis', Vol. II, New York: Academic Press, pp. 311–350.

Brillinger, D. (1981), *Time Series: Data Analysis and Theory*, Siam.

Gramsch, E., Cereceda-Balic, F., Oyola, P. & Baer, D. (2006), 'Examination of pollution

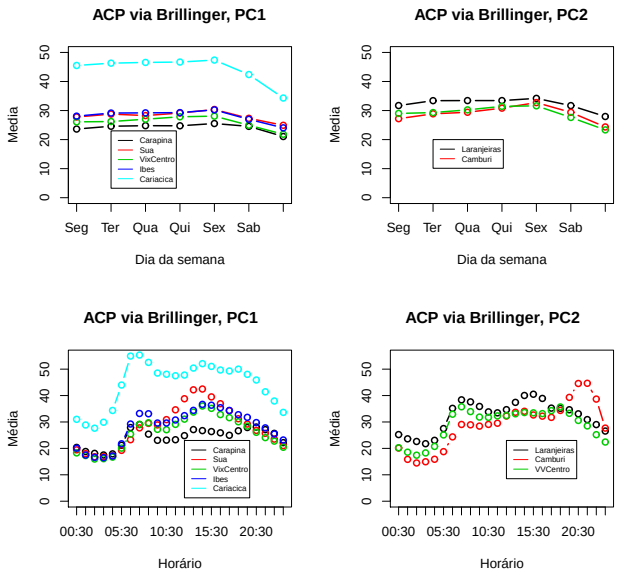



Figura 8: Concentração semanal e horária de $PM_{10}$ com mesmo padrão de acordo com método ACP de Brillinger.

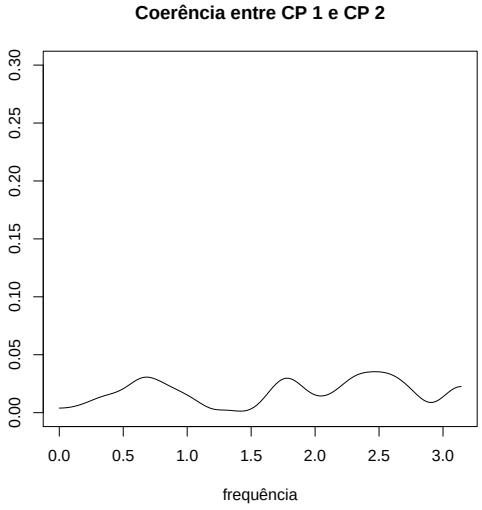



Figura 9: Coerência entre a primeira e segunda componentes principais obtidas com método ACP de Brillinger.

trends in Santiago de Chile with cluster analysis of $PM_10$ and ozone data', *Atmospheric Environment* **40**, 5464–5475.

Jolliffe, I. T. (2008), *Principal component analysis*, 2nd edn, New York: Springer-Verlag.

Kannel, P., Lee, S., Kanel, S. & Khan, S. (2007), 'Chemometric application in classification

and assessment of monitoring locations of an urban river system', *Analytica Chimica Acta* **582**, 390–399.

Keller, M. (2000), Hauptkomponentenanalyse für intensivmedizinische Zeitreihen (in German), PhD thesis, Department of Statistics, University of Dortmund, Germany.

Mendiguchía, C., Moreno, C., Galindo-Rian, M. D. & García-Vargas, M. (2004), 'Using chemometric tools to assess anthropogenic effects in river water: a case study: Guadalquivir river (Spain)', *Analytica Chimica Acta* **515**, 143–149.

Ostro, B. D., Eskekand, G. S., Sánchez, J. M. & Feyzioglu, T. (1999), 'Air pollution and health effects: A study of medical visits among children in Santiago, Chile', *Environmental Health Perspect* **107**, 69–73.

Pires, J. C. M., Souza, S. I. V., Pereira, M. C., Alvim-Ferraz, M. C. M. & Martins, F. G. (2008*a*), 'Management of air quality monitoring using principal component and cluster analysis — part I: $SO_2$ and $PM_{10}$', *Atmospheric Environment* **42**, 1249–1260.

Pires, J. C. M., Souza, S. I. V., Pereira, M. C., Alvim-Ferraz, M. C. M. & Martins, F. G. (2008*b*), 'Management of air quality monitoring using principal component and cluster analysis — part II: CO, $NO_2$ and $O_3$', *Atmospheric Environment* **42**, 1261–1274.

Reisen, V. A., Zamprogno, B., Palma, W. & Arteche, J. (2010), 'Seasonal fractional long-memory processes. a semiparametric estimation approach', *arXiv:1011.5631* .

Shrestha, S. & Kazama, F. (2007), 'Assessment of surface water quality using multivariate statistical techniques: a case study of the Fuji river basin, Japan', *Environmental Modelling and Software* **22**, 464–475.

Singh, K., Malik, A., Mohan, D. & Sinha, S. (2004), 'Multivariate statistical techniques for the evaluation of spatial and temporal variations in water quality of Gomti river (India): a case study', *Water Research* **18**, 3980–3992.

Singh, K., Malik, A. & Sinha, S. (2005), 'Water quality assessment and apportionment of pollution sources of Gomti river (India) using multivariate statistical techniques: a case study', *Analytica Chimica Acta* **538**, 355–374.

Villeneuve, P., Johnson, J., Pasichnyk, D., Lowes, J., Kirkland, S. & Rowe, B. (2012), 'Short-term effects of ambient air pollution on stroke: Who is most vulnerable?', *Science of the Total Environment* **430**, 193–201.

Wong, T. W., Tam, W., Yu, T. & Wong, A. H. (2002), 'Associations between daily mortalities from respiratory and cardiovascular diseases and air pollution in hong kong, china', *Occupational and Environmental Medicine* **59**, 30–35.

Zamprogno, B., Reisen, V. A. & Reis, N. C. (2013), Análise de componentes principais em séries temporais, nos domínios do tempo e da frequência: abordagem em problemas da poluição do ar. In press.

# $PM_{10}$ and $SO_2$ mass concentrations analysis: an application of robust principal component analysis

*Bartolomeu Zamprogno*[*a,b], *Nátaly A. Jiménez*[a,b], *Fabio A. Fajardo*[a]
*Neyval C. Reis Jr.*[b], *Valdério A. Reisen*[a,b]
[a]PPGEA, CT – Universidade Federal do Espírito Santo, Vitória, ES, Brasil.
[b]DEST, CCE – Universidade Federal do Espírito Santo, Vitória, ES, Brasil.

**Resumo**

This paper explores the performance of the Robust Principal Component Analysis (RPCA) technique, suggested by Huber et al. (2005), using $PM_{10}$ and $SO_2$ concentrations for the time period 2005-2009. The data sets used in this study were obtained from the air quality monitoring network of the Greater Vitoria Region, Brazil.

The Principal Component Analysis (PCA) technique can be used to detect spatial patterns, to establish different sources of pollution and to identify redundant air pollution measurements, among other purposes. Our empirical study compares the performance of the usual and the robust PCA for optimizing an air quality network, through the identification of monitoring stations with similar behavior. The results obtained show the RPCA approach as the most suitable technique due to its superior performance for obtaining more coherent results.

*Palavras-chave:* PCA, monitoring network, outliers, robust methods, air pollutants.

## 1 Introduction

In several regions around the world, mainly in developing countries, urban air quality has deteriorated gradually as a consequence of the accelerated urbanization, population growth, absence of adequate environmental policies and industrialization. The impacts of air pollution have local, regional and global effects.

Atmospheric pollutants release causes many serious problems to environment and human health. For example, effects associated with longer-term exposures to high concentrations of sulphur dioxide ($SO_2$), in conjunction with high levels of particulate matter, include respiratory illness, alterations in the lungs, defenses and aggravation of existing heart or lung disease. The most susceptible populations under these conditions include individuals with cardiovascular disease or chronic lung disease, children and older adults (U.S. EPA 1982).

On the other hand, one of the most studied pollutants is the particulate matter with aerodynamic diameter smaller than 10 $\mu$m ($PM_{10}$). Several epidemiological studies have indicated a strong association between elevated concentrations of $PM_{10}$ and increased mortality and morbidity (see Arditsoglou & Samara 2005, Namdeo & Bell 2005). It also influences many atmospheric processes including cloud formation, visibility, solar radiation and precipitation,

*Email: bzamprogno@yahoo.com.br

and plays a major role in acidification of clouds, rain and fog (Celis et al. 2004, Hong et al. 2002, Khoder 2002, among others).

Pedrero et al. (2009) stated that the study of air quality degraded by emissions of particulate matter (PM), $SO_2$, Nitrogen oxides (NOx) and volatile organic compounds (VOCs) requires the knowledge of several mathematical techniques. The knowledge of statistical methods is also very important. Particularly, one of the multivariate statistical techniques more frequently used in air pollution studies is the Principal Component Analysis (PCA). This technique is based on the linear combinations of the original variables, which allows for interpretation and better understanding of the different sources of variation.

PCA is frequently used in environmental studies to reduce the dimension of the data sets and to identify the main characteristics of the investigated phenomenon. The PCA method is also useful in order to identify pollution sources, to locate regions with similar pollution behavior and for space-time studies, among other purposes.

In environmental studies related to $SO_2$ air pollution, Lowell et al. (1984) and Malm et al. (1990) used PCA to detect sulphur spatial patterns in the western United States. Sanchez et al. (1996) applied the technique to identify $SO_2$ urban sources. Additionally, Oanh et al. (2005), Yu & Chang (2006) and Ibarra-Berastegi et al. (2008) used PCA to detect meteorological and temporal patterns.

With regard to $PM_{10}$, Karar & Gupta (2007) used PCA for source apportionment and Liu (2009) adopted the PCA in a model for simulating daily concentrations of the pollutant. In many cases, the technique has been used along with other techniques as cluster analysis (Pires et al. 2008*a*,*b*, Lau et al. 2009) and Self-Organizing Maps (Ezcurra et al. 2008). PCA can also be used as a estimation method in factor analysis, as showed by Viana et al. (2008). Shi et al. (2009) used PCA for source apportionment in the receptor model PCA/MLR.

Concentration levels of the pollutants are regulated by specific standards for each region. When the concentration of the pollutant exceeds the maximum levels established, there are increased health problems observed in people with respiratory, coronary and many other diseases. Clearly, the situation of low concentrations is the desirable one. When atmospheric pollutants are measured, it is common to find discrepant concentrations that can be quite large when compared with the values usually obtained. Even if those values are considered under control according to the standards, they can be characterized as atypical (or outlier) data in statistical analysis.

Cosemans et al. (2008) defined outliers as a large value that is not typical for the cumulative frequency distribution of the data in a sample, while it could be a correct value in the population from which the sample is taken. Rousseeuw & Van Zomeren (1990) defined outliers as observations that deviate from the estimates by a statistical model suggested by the majority of a data set. There exists a variety of statistical methods for detecting outliers, the most popular is to set an outlier bound at either two or three standard deviations from the mean (Barnett & Lewis 1984). Nowadays, there are sophisticated techniques to detect atypical data depending on the data structure and dimensionality (see, e.g. Filzmoser et al. 2005). The treatment of outliers linked to the analysis techniques must be considered in all studies, since the presence of atypical data drastically affect any statistical methodology, leading to unreliable results and wrong inferences.

In this context, to minimize the effects of possible atypical observations in the environmental data sets, we suggest the use of robust techniques when reducing the dimensionality of the set. We develop an empirical application to compare the performance of usual and robust

PCA methodologies in optimization of an air quality network through the identification of similar pollution patterns. The observations correspond to $PM_{10}$ and $SO_2$ concentrations, obtained from the air quality network of the Metropolitan Region of Greater Vitoria, Brazil. The robust PCA technique here analyzed was proposed by Huber et al. (2005), this is based on the robust estimation of the covariance matrix of the data set for selecting the number of principal components. A graphical comparison between the results from the robust and usual PCA techniques was conducted to observe the daily and hourly behavior. The results indicate the better performance of the RPCA technique as an alternative tool for air quality network management, reinforcing our suggestion of using robust techniques in the context of air quality studies.

# 2 Data and methodology

## 2.1 Data and monitoring network

The daily average $SO_2$ and $PM_{10}$ concentration (expressed in $\mu g/m^3$) are the data sets here analyzed. These data were obtained from the air quality monitoring network at the Metropolitan Region of Greater Vitoria (RGV) in Espírito Santo State, Brazil. The analyzed period was from January 2005 to December 2009.

The RGV is constituted by seven main cities: Vitória, Serra, Vila Velha, Cariacica, Viana, Guarapari and Fundão. The population is about 1.73 million inhabitants in an area of $2,331$ $km^2$. The population growth rate in Espírito Santo (ES) State is 3.2% per year and almost half of total population (46%) is covered by the RGV, which produces 58% of the wealth and consumes 55% of the total electric power produced.

The region is located in the Brazilian South Atlantic coast (latitude 20°19S, longitude 40°20W). The climate is tropical humid with average temperatures from 24°C to 30°C. The rainfall is fairly distributed over the entire year with drier periods from June to August (average precipitation of 60.8 mm per month) and more intense precipitation periods from October to January (average precipitation of 158.3 mm per month).

RGV topography is characterized by mountain range at Northeast and West, plain and steppes are mainly located at North and South. The entire region is interspersed by small and medium size rocky massif. Those conditions favors the wind flowing and pollutants spreading over a large extent of the region.

The main atmospherical flowing systems at RGV are the South Atlantic subtropical anticyclone, which produces the dominant eastern and northeastern winds; and the moving polar anticyclone, which generates the cold flows from the southern region of the continent, characterized by low temperatures, mist and strong winds.

The air quality monitoring network (AQMN) in Greater Vitoria Region was inaugurated in June 2000. It is managed by the Instituto Estadual do Meio Ambiente e Recursos Hídricos (IEMA). The network consists of eight monitoring stations, all located in urban areas. Within the seven cities of RGV, only four of them are part of the AQMN. Their total area represents only the 48.9% of the RGV area (see Figure 1).

The main industrial activities of RGV are related to iron and steel industry, stone quarry, cement and food industry, among others. Table 1 shows the main characteristics of each AQMN monitoring station, the names used in this paper for each one are also given.


Figura 1: Location of the AQMN monitoring stations in Greater Vitoria Region.

Tabela 1: General characteristics of the air quality monitoring network at RGV.

| Site | Main pollution sources | Station | City of RGV |
|---|---|---|---|
| Laranjeiras | Industrial and background | Laranjeiras | Serra |
| Carapina | Industrial, background and traffic | Carapina | Serra |
| Jardim Camburi | Industrial and traffic | Camburi | Vitória |
| Enseada do Sua | Port of Tubarão and traffic | Sua | Vitória |
| Vitoria-Downtown | Traffic, ports, Industrial | VixCentro | Vitória |
| Vila Velha-Ibes | Traffic and industrial | Ibes | Vila Velha |
| Vila Velha-Downtown | Traffic and industrial | VVCentro | Vila Velha |
| Cariacica | Traffic and industrial | Cariacica | Cariacica |

*Source:* IEMA.

## 2.2 The Principal Component Analysis

Many scientific researches are based on the study of the influence of specific variables on a data set and its simultaneous behavior. This influence may be studied using statistic multivariate methods, which give valuable conclusions about the relationship between variables. Sometimes, the results are distorted and difficult to interpret due to the large number of variables. There are some useful techniques to reduce the data set dimension, which summarize and group the most similar variables. Principal Component Analysis is one of them and is the technique here studied.

### 2.2.1 The usual PCA

The PCA is one of the most popular techniques of the multivariate statistics. The objective is to explain the covariance structure of a data set by means of a small number of components, i.e. the PCA is a dimension reduction technique which transforms the data to a smaller set of variables called *Principal Components* (PC). These components are linear combinations of the original variables, and often allow for interpretation and better understanding of the different sources of variation. In this paper, let $\{X_1, X_2, \ldots, X_r\}$ be the pollutant concentration data matrix of $r$ monitoring stations in the RGV. The PCA obtains an orthogonal

set $\{Z_1, Z_2, \ldots, Z_q\}$ such that $Z_i = \sum_{k=1}^{q} l_{ik} X_k$, where $l_{ik}$ is the $i-$th element of the $k-$th eigenvector of the covariance matrix of $\{X_1, X_2, \ldots, X_r\}$. The elements of $\{Z_1, Z_2, \ldots, Z_q\}$ are called *principal component*. The principal components are ordered so that the first PC explains the most part of the variance in the data, and each subsequent PC accounts for the largest proportion of variability that has not been accounted by its predecessors, see, e.g. Jolliffe (2002) and Statheropoulos et al. (1998) for detailed mathematical treatment of PCA.

The usual PCA is sensitive to the presence of atypical observations and the statistical analysis from this methodology could be highly affected. In view of this problem, we recommend the use of a robust technique to obtain reliable outcomes when there are doubts about the contamination of the data set by outliers.

### 2.2.2 Robust PCA

Huber et al. (2005) suggested the RPCA method as a combination of both projection pursuit and robust covariance estimation. The RPCA method proceeds in three steps: first, the data are preprocessed so that the transformed data belong to a subspace whose dimension is at most $n - 1$, where $n$ is sample size. Second, a preliminary covariance matrix $\Gamma_0$ is constructed. This covariance is used for selecting the number of components $q$ that will be retained in the sequel, yielding a $q$-dimensional subspace that fits the data well. Finally, the data points are projected on this subspace where their location and their scatter matrix are robustly estimated. The $q$ non-zero eigenvalues $l_1, \ldots, l_k$ are computed from the robust scatter matrix. The corresponding eigenvectors are the $q$ robust principal components. As the usual PCA, the RPCA method is location and orthogonal equivariant, i.e. when a shift and/or a rotation (or a reflection) is applied to the data, the robust center is also shifted and the loadings are rotated accordingly. Hence, the scores do not change under this kind of transformations.

## 2.3 Multiple outliers detection

In practice, a very influential factor that must be taken into account is the presence of dissonant values, or *outliers*, in the data set. The outliers must be considered in the analysis because they could disturb the results of the statistical analysis. We use the *PCOut* procedure based on PC decomposition for identifying multivariate outliers. The procedure was proposed by Filzmoser et al. (2005) and consists of two basic parts: a part that is especially good at detecting location outliers, and a step that yields particularly good results among scatter outliers. The algorithm proceeds in the following steps: first, the data set is normalized by subtracting the median and dividing by the median absolute deviation (MAD) in each variable, then the sample covariance matrix of the transformed data is calculated. Second, a principal component decomposition of the semirobust covariance matrix from the first step is computed and only those $p^*(< n)$ eigenvectors/values that contribute to at least 99% of the total variance are retained. The robust kurtosis weights are computed for each component as

$$w_j = \left| \frac{1}{n} \sum_{i=1}^{n} \frac{(y_{ij} - median\{y_{ij}\})^4}{MAD(y_{1j}, \ldots, y_{nj})^4} - 3 \right|,$$

where $y_{ij}$ indicates data that have been transformed via Principal Components. Third, the weights for each robust distance are determined according to the translated biweight, such

that $w_{1i} = \left[1 - \left(\frac{d_i - M}{c - M}\right)^2\right]^2$ if $M < d_i < c$, $w_{1i} = 0$ if $d_i \geq c$ and $w_{1i} = 1$ if $d_i \leq M$, where $M$ equal to the 0.33 quantile of the distances $\{d_1, \ldots, d_n\}$, $c = median\{d_1, \ldots, d_n\} + 2.5MAD\{d_1, \ldots, d_n\}$ and $d_i = \frac{MD_i}{median(MD_i)}\sqrt{\chi^2_{p^*,0.5}}$ with $\chi^2_{p^*,0.5}$ the 0.5 quantile of the chi-square distribution with $p^*$ degrees of freedom. $MD_i$ represents a robust version of the Mahalanobis distance. Finally, weights $w_{2i}$ for each robust distance are determined according to the translated biweight as in the third step with $c^2$ and $M^2$ equals to the 0.99 and 0.25 quantiles of the chi-square distribution with $p^*$ degrees of freedom, respectively. The weights $w_{1i}$ and $w_{2i}$ are used to determine final weights for all observations such that $w_i = \frac{(w_{1i}+0.25)(w_{2i}+0.25)}{1.5625}$. Finally, the outliers are classified as points that have weight $w_i < 0.25$.

# 3 Results and discussion

The PCA analysis was developed under the approaches described before. As our purpose is to compare these methodologies, all the analyses were conducted using standardized data. Aiming at refining the results and their interpretation, the principal components were rotated using the varimax method. The selection criteria for the number of components is to obtain a minimum of 80% explanation for the total variability.

The missing values for each one of the stations monitoring $SO_2$ and $PM_{10}$ over the 1826 analyzed days were filled using Gibbs sampling for multiple imputations of the incomplete multivariate data suggested by Aerts et al. (2002). This algorithm imputes an incomplete column (in this case, each column corresponds to a monitoring station) by generating plausible synthetic values given the other columns in the data. Each incomplete column must act as a target column, and has its own specific set of predictors. The default set of predictors for a given target consists of all other columns in the data. All computations were made using the language and environment for statistical computing `R` (http://www.r-project.org/) and the high-level language and interactive environment `MATLAB` (http://www.mathworks.com/products/matlab/).

## 3.1 Results for $SO_2$

For $SO_2$ levels, the primary standard demands that average concentration on a 24 hour period do not exceed $365\mu g/m^3$ more than once a year. Besides, the annual arithmetic average must not exceed $80\mu g/m^3$. In accordance with the secondary standard established by the Brazilian law, the average concentration on a 24 hours period could not exceed $100\mu g/m^3$ more than once a year and the annual arithmetic average must not exceed $40\mu g/m^3$. Table 2 presents the annual averages for each station and the maximum observed on the entire analyzed period.

Figure 2 shows the boxplots of daily average $SO_2$ concentration for each monitoring station of AQMN on the period 01/01/2005 to 12/31/2009. The comparison of the notches in the boxplots indicates that the median concentration differs for all stations. The concentrations are higher at Sua and VixCentro stations. Those high concentrations are caused by the industrial emissions from Port of Tubarão as well as the mobile sources that converge to the Sua station area, as observed from Table 1. The VixCentro station is influenced by traffic emissions, besides the contribution of port activities and industrial sources from Vila Velha and Cariacica cities. The Cariacica station is the one which presents the lowest concentration

Tabela 2: Annual average $SO_2$ concentration by monitoring station.

| **Station** | **Year** | | | | | **Maximum** |
|---|---|---|---|---|---|---|
| | 2005 | 2006 | 2007 | 2008 | 2009 | |
| Laranjeiras | 10.50 | 14.31 | 11.80 | 9.52 | 16.65 | 194.01 |
| Camburi | 10.82 | 13.85 | 10.98 | 10.59 | 10.72 | 223.90 |
| Sua | 15.92 | 17.65 | 12.91 | 11.26 | 16.88 | 257.41 |
| VixCentro | 11.31 | 13.60 | 16.04 | 16.89 | 13.62 | 218.41 |
| Ibes | 8.02 | 10.07 | 10.18 | 12.01 | 14.23 | 162.68 |
| VVCentro | 12.38 | 11.56 | 14.48 | 13.24 | 10.28 | 250.90 |
| Cariacica | 4.92 | 4.86 | 4.44 | 5.29 | 4.77 | 73.16 |

level. This fact suggests that the two main roads and the industrial region nearby do not contribute to a significant increase of the pollutants.

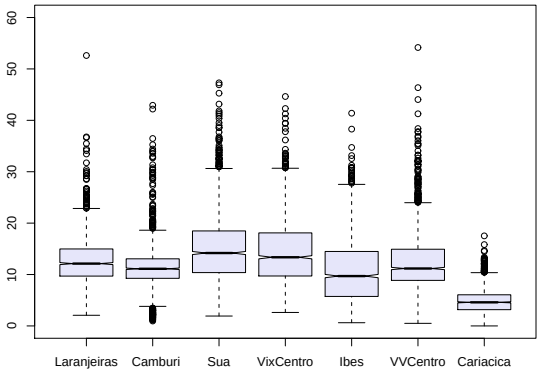



Figura 2: Daily average $SO_2$ concentration by monitoring station.

The presence of outliers was tested and they were detected in all monitoring stations. Figure 3 shows the different steps of the *PCOut* algorithm (distances and weights) for each robust distance. The different symbols represent the eight stations and the intermediate graphs provide more insight about the philosophy of the procedure. The kurtosis weights of the *PCOut* algorithm second step are shown in the upper left/right panel, together with the weight boundaries. Similarly, the distances and the weights from the third step are respectively shown in the left and right panel of the second row. The lower left panel shows the combined weights together with the outlier boundary 0.25, which results in 0/1 weights (lower right panel).

The results show potential multivariate outliers for daily average $SO_2$ concentration. The individual analysis showed that Camburi, Laranjeiras and VVCentro presented the most extreme records. Even with presence of the atypical values, the primary and secondary air quality standards established by the Brazilian law are met, but every statistical analysis is affected for the influence of atypical data. Therefore, the careful treatment of the data set under contamination plays important role to make decisions. Since the concentration levels

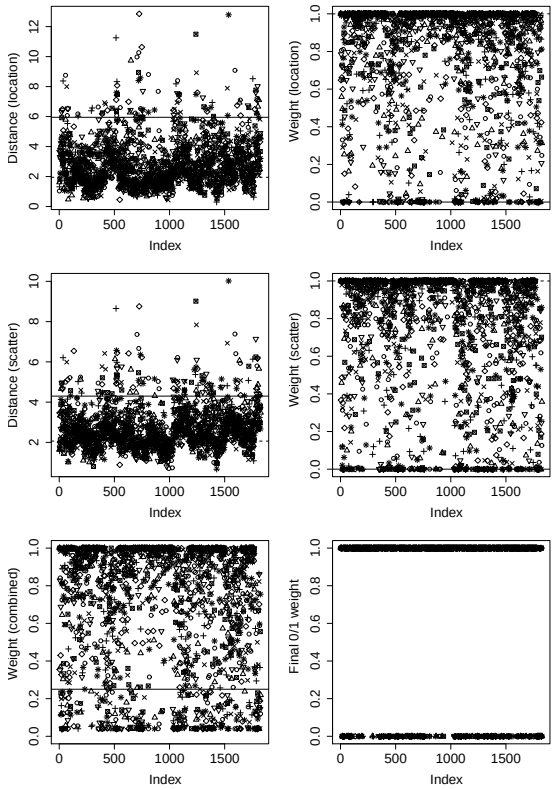



Figura 3: Potential multivariate outliers of the daily average $SO_2$ concentration by monitoring station.

are well controlled, there are incentives for the reallocation or even for the reduction of the number of monitoring stations for this pollutant at RGV.

Hourly average $SO_2$ concentration is showed in Figure 4a for each station. The highest pollution peaks are presented between noon and 5:00 p.m. at Sua and VixCentro stations. The highest concentration happens at 3:00 p.m. at Sua station, with average of $27.44\mu g/m^3$. It can also be observed that Laranjeiras station has the highest concentration values during a period of eleven hours, running from 8:00 p.m. to 7:00 a.m.. Nevertheless, the Sua station presents the highest concentrations along all the days, as shown in Figure 4b. Therefore, Sua and Laranjeiras could be considered as the principal monitoring stations for $SO_2$ control.

Table 3 presents the rotated principal components according to the usual technique. The first two components explain about 57.2% of total variability. Considering the first four components, the total variability explained rises to 83%.

The results from the first principal component (PC1) show that Sua and Ibes stations can be grouped. The second principal component (PC2) aggregates VixCentro and Cariacica stations. The third principal component (PC3) indicates Laranjeiras station as having a different behavior from the others. Similarly, the fourth principal component (PC4) shows the Camburi station as different.

Table 4 presents the principal components according to the RPCA technique. The first two principal components explain 58.1% of total variability. When the first four components

Tabela 3: Rotated usual PCA results for average $SO_2$ concentration.

| **Station** | **PC** | | | |
|---|---|---|---|---|
| | PC1 | PC2 | PC3 | PC4 |
| Laranjeiras | 0.196 | −0.057 | **0.865** | 0.057 |
| Camburi | −0.003 | 0.043 | −0.015 | **0.926** |
| Sua | **0.598** | −0.018 | 0.072 | 0.089 |
| VixCentro | 0.090 | **0.802** | −0.068 | −0.179 |
| Ibes | **0.634** | 0.069 | 0.018 | −0.145 |
| VVCentro | 0.436 | −0.132 | −0.486 | 0.141 |
| Cariacica | −0.062 | **0.574** | 0.071 | 0.240 |
| Eigenvalue | 2.288 | 1.716 | 1.119 | 0.684 |
| Proportion | 0.327 | 0.245 | 0.160 | 0.098 |
| Cumulative | 0.327 | 0.572 | 0.732 | 0.830 |

are taken, the variability explained reaches 83.1%. Comparing with the results obtained from the usual technique, the explaining percentage stays nearly the same but the influence exerted by the outliers on the first principal component is handled.

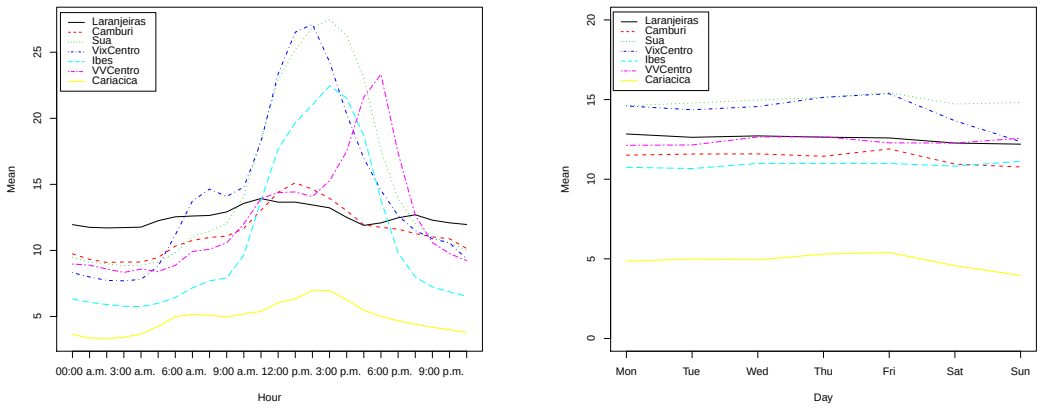
Tabela 4: Rotated RPCA results for average $SO_2$ concentration.

| **Station** | **PC** | | | |
|---|---|---|---|---|
| | PC1 | PC2 | PC3 | PC4 |
| Laranjeiras | −0.051 | 0.011 | **-0.934** | −0.002 |
| Camburi | −0.066 | **-0.708** | 0.012 | 0.163 |
| Sua | **-0.602** | −0.254 | −0.031 | 0.200 |
| VixCentro | −0.037 | −0.037 | 0.012 | **-0.917** |
| Ibes | **-0.637** | 0.170 | −0.139 | −0.160 |
| VVCentro | −0.462 | 0.063 | 0.327 | −0.044 |
| Cariacica | 0.099 | **-0.632** | −0.023 | −0.257 |
| Eigenvalue | 1.835 | 1.279 | 0.877 | 0.460 |
| Proportion | 0.343 | 0.239 | 0.164 | 0.086 |
| Cumulative | 0.342 | 0.581 | 0.744 | 0.831 |

The first principal component aggregates Ibes and Sua stations, as concluded using usual PCA. The PC2 indicates the similar behavior at Camburi and Cariacica stations. The third principal component agrees with the third one obtained by the usual PCA, where Laranjeiras station is detected as markedly different from the other stations. Finally, the PC4 shows the VixCentro station as dissimilar from any other station.

The results of usual and robust PCA were compared with the graphical behavior of the daily average $SO_2$ mass concentration. The first principal component of both techniques aggregates the same two monitoring stations, the similarity of their hourly and daily dynamics can be observed in Figure 4b.

The strong relationship between these stations become evident when observed their correlation of 0.703 on Table 5 and the similar behavior showed in Figure 4. Since Sua station has the highest concentration levels, this could be the selected station to continue collecting




(a) Hourly concentration.

(b) Daily concentration.

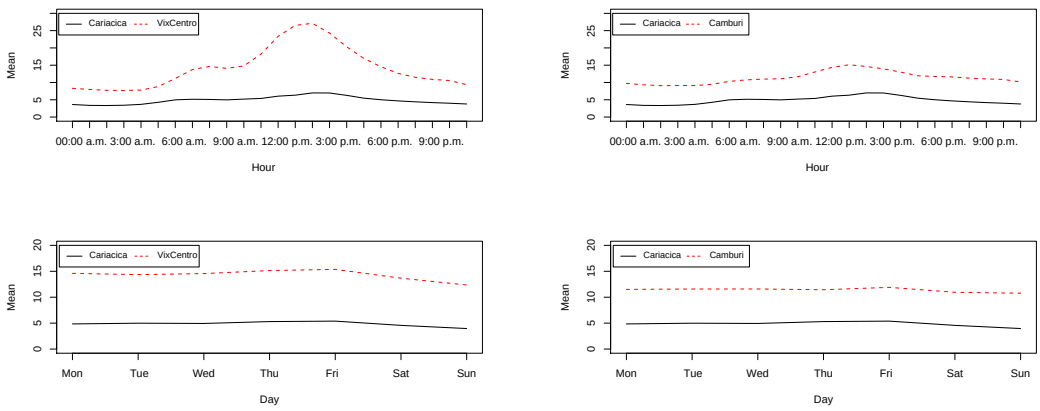
Figura 4: $SO_2$ average concentration by monitoring station.

$SO_2$ data, and the monitoring equipments of the remaining station could be reallocated in one of the other three cities in RGV that are not part of the AQMN, or any other region of interest. The results suggest that, if problems are presented in the monitoring network, the priority should be given to the equipments of the stations with highest concentration levels comparing with the stations having similar daily and hourly patterns.

Tabela 5: Correlation between monitoring stations for $SO_2$.

| **Station** | **Laranjeiras** | **Camburi** | **Sua** | **VixCentro** | **Ibes** | **VVCentro** | **Cariacica** |
|---|---|---|---|---|---|---|---|
| **Laranjeiras** | 1.000 | 0.252 | 0.182 | −0.038 | 0.126 | −0.130 | 0.086 |
| **Camburi** | 0.252 | 1.000 | 0.069 | 0.230 | −0.088 | −0.098 | 0.413 |
| **Sua** | 0.182 | 0.069 | 1.000 | −0.057 | 0.703 | 0.432 | −0.100 |
| **VixCentro** | −0.038 | 0.230 | −0.057 | 1.000 | 0.002 | −0.070 | 0.464 |
| **Ibes** | 0.126 | −0.088 | 0.703 | 0.002 | 1.000 | 0.548 | −0.171 |
| **VVCentro** | −0.130 | −0.098 | 0.432 | −0.070 | 0.548 | 1.000 | −0.275 |
| **Cariacica** | 0.086 | 0.413 | −0.100 | 0.464 | −0.171 | −0.275 | 1.000 |

The analysis of the PC2 shows clearly that there is a gain treating outliers. Both techniques indicate Cariacica station as similar to another station. In usual PCA case, the behavior is showed as similar to VixCentro station while in RPCA the results suggest more similarity with Camburi station. Figure 5 shows that the line slopes in daily dynamics are quite similar between Cariacica and the other stations. However, the hourly behavior showed in Figure 5b indicates a higher agreement between Camburi and Cariacica stations, as concluded by the RPCA technique. The distinct behavior of $SO_2$ in Laranjeiras station is verified observing Figure 4, where the concentrations are almost stable along all the time. The results obtained by both methods excluded VVCentro station, indicating that its measurements are not giving relevant information and this station should be reallocated to any other place of interest.




(a) Results for usual PCA.

(b) Results for RPCA.

Figura 5: PC2 usual and robust PCA comparison for $SO_2$.

## 3.2 Results for $PM_{10}$

Primary and secondary Brazilian air quality standards for $PM_{10}$ require that the 24 hour average concentration do not exceed $150\mu g/m^3$ more than once a year and the annual average do not exceed $50\mu g/m^3$. Table 6 shows the annual average for each station and the maximum hourly value observed along the analyzed period. The annual average allowed by law is not exceeded in any observed year and, in regard to the maximum value in 24 hour running periods, the standard was never exceeded despite the maximum value $745\mu g/m^3$ obtained at Laranjeiras station. The highest value obtained between the maximum values in 24 hours was $113.33\mu g/m^3$ at VVCentro station. The fact of the maximum level has not been exceeded encourages the reduction of the number of monitoring stations for $PM_{10}$ at RGV.

Tabela 6: Annual average $PM_{10}$ concentration by monitoring station.

| **Station** | **Year** | | | | | **Maximum** |
|---|---|---|---|---|---|---|
| | 2005 | 2006 | 2007 | 2008 | 2009 | |
| Laranjeiras | 31.28 | 33.03 | 32.82 | 32.75 | 31.39 | 745.00 |
| Carapina | 22.22 | 23.60 | 24.03 | 27.85 | 22.93 | 401.00 |
| Camburi | 28.40 | 28.65 | 28.54 | 31.07 | 28.17 | 359.00 |
| Sua | 25.98 | 28.32 | 28.07 | 29.77 | 28.28 | 452.00 |
| VixCentro | 24.51 | 25.65 | 24.64 | 28.17 | 27.06 | 287.00 |
| Ibes | 26.13 | 28.15 | 26.96 | 30.18 | 29.23 | 263.00 |
| VVCentro | 29.69 | 27.16 | 26.73 | 31.40 | 29.69 | 307.00 |
| Cariacica | 42.50 | 42.25 | 41.86 | 47.71 | 46.47 | 508.00 |

All the data for $PM_{10}$ levels were obtained from the eight monitoring stations in the Greater Vitoria AQMN in the period 01/01/2005 to 12/31/2009. For descriptive purposes, the daily average $PM_{10}$ concentration was summarized using boxplots for each monitoring station, as showed in Figure 6. The notches indicate that the median of the concentrations are

quite different for all stations. The Cariacica and Laranjeiras stations have the highest $PM_{10}$ average concentration. The Laranjeiras station is mainly influenced by industrial sources. On the other hand, traffic and industrial emissions constitute the main pollution sources to Cariacica station (Table 1). Again, the presence of multivariate outliers was verified using the *PCOut* algorithm and the results confirmed the presence for all stations.

The hourly average $PM_{10}$ concentration is showed in Figure 7a for each station. The records of Cariacica station have the maximum average during almost all the day, except for the period between 9:00 p.m. and 11:00 p.m., when Camburi station have the maximum values. On the other hand, Carapina concentration levels are completely dominated by all the other stations and presents the minimum average concentrations along a period of 13 hours. Considering the daily average, Figure 7b, the records of Cariacica and Laranjeiras stations dominate all the other ones along all the week. This fact suggests that these stations should be considered as the main monitoring stations for $PM_{10}$ controlling.

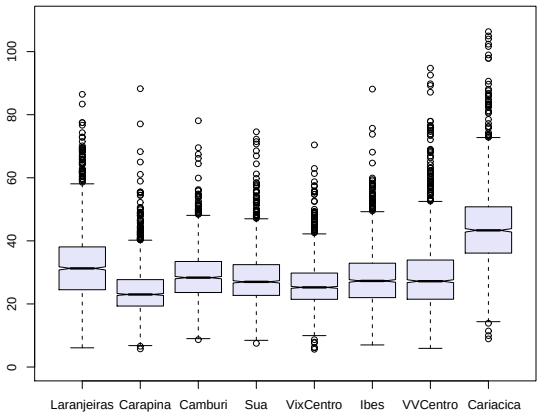


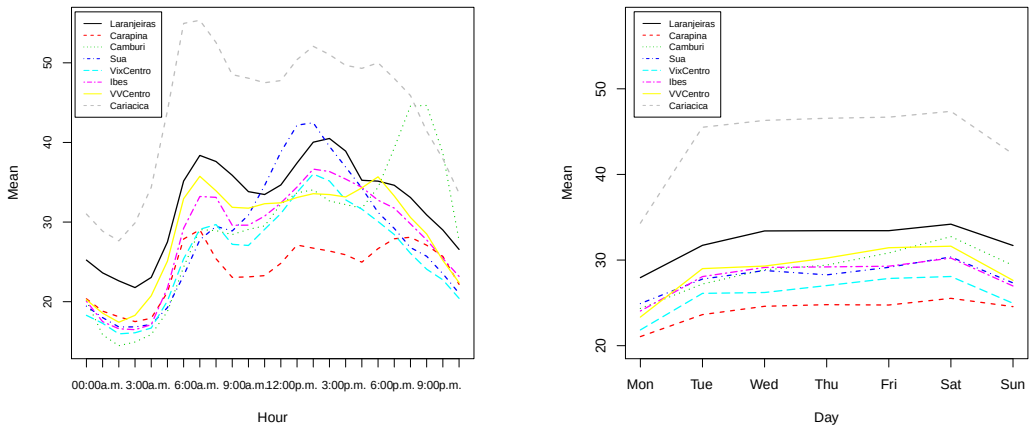

Figura 6: Daily average $PM_{10}$ concentration by monitoring station.

Table 7 shows the rotated principal components using the usual technique. The first two principal components explain 70.9% of the total variability, if the first four principal components are taken, the variability explained raises to 85%. The results of the first principal component (PC1) sets Carapina and Cariacica stations. The second principal component (PC2) aggregates Laranjeiras and Camburi stations. The third principal (PC3) component suggest that VVCentro station has a different behavior from the other ones, as well as Sua station enhanced on the fourth principal component (PC4).

Table 8 shows the results obtained for the RPCA technique. The percentage of explanation stays stable, if compared with the results from the usual PCA. The four principal components explain 85.2% of total variability, where only the first two can explain 71.8%. The PC1 indicates Laranjeiras station as having a different behavior from the other stations. The PC2 aggregates the same stations of the first principal component from the usual PCA, e.g., Carapina and Cariacica stations. Also, the results for the PC3 agrees with the third one obtained from usual PCA, indicating VVCentro station as different from the other ones. Finally, the PC4 shows that Camburi station has different behavior from the remaining

stations.

A comparison of the results obtained for PC2, according to the daily average behavior, is showed in Figure 8. It becomes evident from Figure 8b, that Carapina and Cariacica stations present similar daily and hourly behavior. The strength of the correlation between them is high (0.70), as can be observed in Table 9 in spite of the strong correlation between Cariacica and VixCentro stations (0.71) as well as Cariacica and Ibes stations (0.69). This shows that a strong correlation do not necessarily indicates similar pollution patterns.




(a) Hourly concentration.

(b) Daily concentration.

Figura 7: $PM_{10}$ average concentration by monitoring station.

A graphical analysis shows that grouping Laranjeiras and Camburi (correlation 0.52), result obtained from the usual PCA analysis, could not be feasible because the hourly and daily patterns are quite different (Figure 8a). The VVCentro station, indicated as different for both techniques, has a very different pattern from the other stations, as showed in Figure 7. The analysis of the fourth component obtained from usual and robust PCA, shows that Camburi station has higher merit when compared with Sua station, due to the nighttime pollution peak presented by the former station. Those results encourage the use of the robust technique to get more consistent principal components.

# 4 Final remarks

In this paper we compare two PCA techniques using $SO_2$ and $PM_{10}$ hourly concentration data. The data were obtained from the AQMN at Region of Greater Vitoria, Brazil. The results indicated that the RPCA technique has better performance when there are atypical observations in the data set, this situation is frequently presented by atmospherical pollutants measurement.

Regarding to the identification of groups with similar air pollution behavior, the results for the usual and robust PCA methods suggest that the eight monitoring stations system of the AQMN can be classified in four groups for both analyzed pollutants. This verification shows

Tabela 7: Rotated usual PCA results for average $PM_{10}$ concentration.

| **Station** | **PC** | | | |
|---|---|---|---|---|
| | PC1 | PC2 | PC3 | PC4 |
| Laranjeiras | −0.338 | **0.668** | −0.009 | 0.315 |
| Carapina | **0.618** | 0.023 | −0.061 | −0.030 |
| Camburi | 0.273 | **0.722** | 0.005 | −0.221 |
| Sua | −0.057 | −0.121 | −0.088 | **0.728** |
| VixCentro | 0.259 | −0.059 | −0.341 | 0.226 |
| Ibes | 0.170 | 0.110 | 0.079 | 0.492 |
| VVCentro | −0.085 | 0.040 | **-0.930** | −0.106 |
| Cariacica | **0.570** | 0.013 | 0.043 | 0.132 |
| Eigenvalue | 4.897 | 0.774 | 0.628 | 0.497 |
| Proportion | 0.612 | 0.097 | 0.079 | 0.062 |
| Cumulative | 0.612 | 0.709 | 0.787 | 0.850 |

Tabela 8: Rotated RPCA results for average $PM_{10}$ concentration.

| **Station** | **PC** | | | |
|---|---|---|---|---|
| | PC1 | PC2 | PC3 | PC4 |
| Laranjeiras | **-0.660** | −0.333 | 0.043 | 0.253 |
| Carapina | 0.024 | **0.538** | 0.025 | 0.059 |
| Camburi | 0.054 | 0.082 | 0.002 | **0.946** |
| Sua | −0.560 | 0.102 | 0.066 | −0.191 |
| VixCentro | −0.114 | 0.401 | 0.218 | −0.032 |
| Ibes | −0.477 | 0.186 | −0.096 | −0.011 |
| VVCentro | 0.035 | 0.011 | **0.958** | 0.007 |
| Cariacica | −0.074 | **0.622** | −0.139 | 0.014 |
| Eigenvalue | 3.548 | 0.615 | 0.399 | 0.376 |
| Proportion | 0.612 | 0.106 | 0.068 | 0.065 |
| Cumulative | 0.612 | 0.718 | 0.787 | 0.852 |

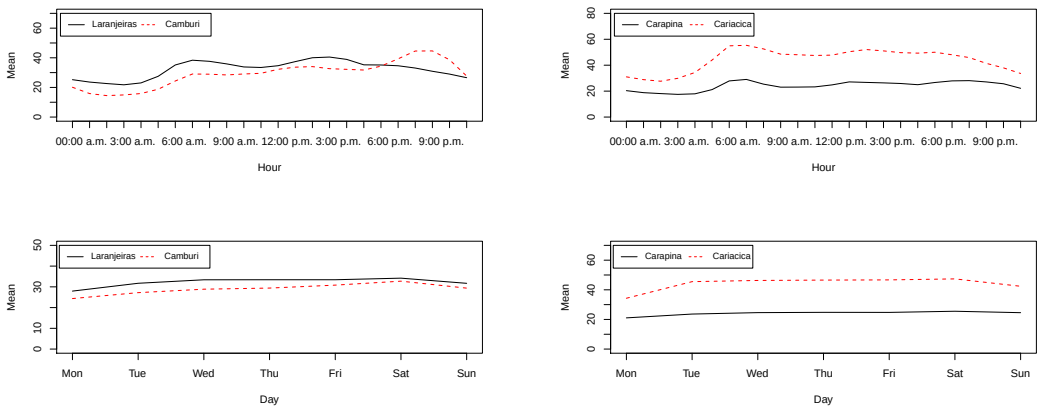
that some monitoring stations are being inefficiently used. We suggest that those monitoring stations should be relocated to the cities of RGV which are not covered by the network yet.

## Acknowledgements

The financial support from CAPES is gratefully acknowledged. We also thank Instituto Estadual de Meio Ambiente e Recursos Hídricos - IEMA - (Vitória/ES, Brazil) for providing the data set.

## Referências

Aerts, M., Claeskens, G., Hens, N. & Molenberghs, G. . (2002), 'Local multiple imputation', *Biometrika* **89**, 375–388.




(a) Results for usual PCA.

(b) Results for RPCA.

Figura 8: PC2 usual and PCA methods comparison for $PM_{10}$.

Tabela 9: Correlation between monitoring stations for $PM_{10}$.

| Station | Laranjeiras | Carapina | Camburi | Sua | VixCentro | Ibes | VVCentro | Cariacica |
|---|---|---|---|---|---|---|---|---|
| **Laranjeiras** | 1.000 | 0.353 | 0.524 | 0.532 | 0.447 | 0.579 | 0.376 | 0.421 |
| **Carapina** | 0.353 | 1.000 | 0.551 | 0.535 | 0.634 | 0.611 | 0.492 | 0.702 |
| **Camburi** | 0.524 | 0.551 | 1.000 | 0.533 | 0.591 | 0.605 | 0.444 | 0.556 |
| **Sua** | 0.532 | 0.535 | 0.533 | 1.000 | 0.670 | 0.720 | 0.456 | 0.544 |
| **VixCentro** | 0.447 | 0.634 | 0.591 | 0.670 | 1.000 | 0.636 | 0.607 | 0.713 |
| **Ibes** | 0.579 | 0.611 | 0.605 | 0.720 | 0.636 | 1.000 | 0.460 | 0.692 |
| **VVCentro** | 0.376 | 0.492 | 0.444 | 0.456 | 0.607 | 0.460 | 1.000 | 0.455 |
| **Cariacica** | 0.421 | 0.702 | 0.556 | 0.544 | 0.713 | 0.692 | 0.455 | 1.000 |

Arditsoglou, A. & Samara, C. (2005), 'Levels of total suspended particulate matter and major trace elements in Kosovo: a source identification and apportionment study', *Chemosphere* **59**, 669–678.

Barnett, V. & Lewis, T. (1984), *Outliers in Statistical Data*, second edn, John Wiley, New York.

Celis, J., Morales, J., Zaror, C., Inzunza, J. & Areitio, J. (2004), 'A study of the the particulate matter $PM_{10}$ composition in the atmosphere of Chillan, Chile', *Chemosphere* **54**, 541–550.

Cosemans, G., Kretzschmar, J. & Mensink, C. (2008), 'Pollutant roses for daily averaged ambient air pollutant concentrations', *Atmospheric Environment* **42**, 6982–6991.

Ezcurra, A., S., Ibarra-Berastegi, G. & Areitio, J. (2008), 'Electrical storm rainfall yield characteristics observed in the Spanish Basque Country area during the period 1992-1996', *Atmospheric Research* **89**, 233–242.

Filzmoser, P., Maronna, R. & M., W. (2005), 'Outlier identification in high dimensions', *Computational Statistics and Data Analysis* **52**, 1694–1711.

Hong, Y., Lee, B., Park, K., Kang, M., Jung, Y., Lee, D. & Kim, M. (2002), 'Atmospheric nitrogen and sulfur containing compounds for three sites of South Korea', *Atmospheric Environment* **36**, 3485–3494.

Huber, M., Rousseeuw, P. J. & Branden, K. V. (2005), 'ROBPCA: A new approach to robust principal component analysis', *Technometrics* **47**, 64–79.

Ibarra-Berastegi, G., Elias, A., Barona, A., Saenz, J., Ezcurra, A. & de Argandoña, J. D. (2008), 'From diagnosis to prognosis for forecasting air pollution using neural networks: Air pollution monitoring in Bilbao', *Environmental Modelling & Software* **23**(5), 622 – 637.
**URL:** *http://www.sciencedirect.com/science/article/B6VHC-4R1733C-1/2/66a384b6b6286545206406c5d1bd0814*

Jolliffe, I. T. (2002), *Principal Component Analysis*, second edn, Springer.

Karar, K. & Gupta, A. (2007), 'Source apportionment of $PM_{10}$ at residential and industrial sites of an urban region of Kolkata, India', *Atmospheric Research* **84**, 30–41.

Khoder, M. (2002), 'Atmospheric conversion of sulfur dioxide to particulate sulfate and nitrogen dioxide to particulate nitrate and gaseous nitric acid in an urban area', *Chemosphere* **49**, 675–684.

Lau, J., Hung, H. G. & Cheung, C. S. (2009), 'Interpretation of air quality in relation to monitoring station's surroundings', *Atmospheric Environment* **43**, 769–777.

Liu, P.-W. G. (2009), 'Simulation of the daily average $PM_{10}$ concentrations at Ta-Liao with Box-Jenkins time series models and multivariate analysis', *Atmospheric Environment* **43**(13), 2104–2113.

Lowell, L. A., Leonard, O. M. & Flocchini, R. G. (1984), 'A Principal Component Analysis of sulphur concentrations in the western United States', *Atmospheric Environment (1967)* **18**(4), 783–791.
**URL:** *http://www.sciencedirect.com/science/article/B757C-48CFYCR-2N4/2/e1aae49ed3414f6add20e2ab61bfedf0*

Malm, W. C., Gebhart, K. A. & Henry, R. C. (1990), 'An investigation of the dominant source regions of fine sulfur in the western United States and their areas of influence', *Atmospheric Environment. Part A. General Topics* **24**(12), 3047 – 3060.
**URL:** *http://www.sciencedirect.com/science/article/B757D-4893PHV-7C/2/efc7f38363b74d3663c7142192ac696d*

Namdeo, A. & Bell, M. (2005), 'Characteristics and health implications of fine and coarse particulates at roadside, urban background and rural sites in UK', *Environment International* **31**, 565–573.

Oanh, N. T. K., Chutimon, P., Ekbordin, W. & Supat, W. (2005), 'Meteorological pattern classification and application for forecasting air pollution episode potential in a mountain-valley area', *Atmospheric Environment* **39**, 1211–1225.

Pedrero, P., Tardón, C. & López, E. (2009), 'Descriptive mathematical techniques to study historical data: An application to sulfur dioxide pollution in the city of Talcahuano - Chile', *Atmospheric Environment* **43**, 6279–6286.

Pires, J. C. M., Souza, S. I. V., Pereira, M. C., Alvim-Ferraz, M. C. M. & Martins, F. G. (2008*a*), 'Management of air quality monitoring using Principal Component and Cluster Analysis – Part I: $SO_2$ and $PM_{10}$', *Atmospheric Environment* **42**(6), 1249–1260.
**URL:** *http://www.sciencedirect.com/science/article/B6VH3-4R17V4K-C/2/5548b0653d7d70612ef4ae7b729766ba*

Pires, J. C. M., Souza, S. I. V., Pereira, M. C., Alvim-Ferraz, M. C. M. & Martins, F. G. (2008*b*), 'Management of air quality monitoring using Principal Component and Cluster Analysis – Part II: CO, $NO_2$ and $O_3$', *Atmospheric Environment* **42**(6), 1261–1274.
**URL:** *http://www.sciencedirect.com/science/article/B6VH3-4R17V4K-7/2/e6194569f68a3a0cf0d8eee7f4f990d7*

Rousseeuw, P. & Van Zomeren, B. (1990), 'Unmasking multivariate outliers and leverage points (with discussion)', *Journal of the American Statistical Association* **85**, 633–651.

Sanchez, M. L., Casanova, J. L., Ramos, M. C. & Sanchez, J. L. (1996), 'Studying the spatial and temporal distribution of $SO_2$ in an urban area by Principal Component Factor Analysis', *Atmospheric Research* **20**, 53–65.

Shi, G.-L., Xiang Li, Feng Yin-Chang, Wang, Y.-Q., Wu, J.-H., Li, J. & Zhu, T. (2009), 'Combined source apportionment, using positive matrix factorization-chemical mass balance and principal component analysis/multiple linear regression-chemical mass balance models', *Atmospheric Environment* **43**, 2929–2937.

Statheropoulos, M., Vassiliadis, N. & Pappa, A. (1998), 'Principal Component and Canonical Correlation Analysis for examining air pollution and meteorological data', *Atmospheric Environment* **32**, 1087–1095.

U.S. EPA (1982), 'Air quality criteria for particulate matter and sulfur oxides'. EPA/600/P-82/020a-c. Research Triangle Park, NC.

Viana, M., Pandolfi, M., Minguillón, M. C., Querol, X., Alastuey, A., Monfort, E. & Celades, I. (2008), 'Inter-comparison of receptor models for PM source apportionment: Case study in an industrial area', *Atmospheric Environment* **42**, 3820–3832.

Yu, T.-Y. & Chang, I.-C. (2006), 'Spatiotemporal features of severe air pollution in northern Taiwan', *Environmental Science and Pollution Research* **13**, 268–275.

Originally submitted to the *Mathematics and Computers in Simulation*, 2013

# A semiparametric approach to estimate two seasonal fractional parameters in the SARFIMA model

*VALDERIO A. REISEN*[a*], *BARTOLOMEU ZAMPROGNO*[a], *WILFREDO PALMA*[b] *and JOSU ARTECHE*[c]

[a]Department of Statistics, CCE and PPGEA-CT-UFES, Vitoria - ES, Brazil.

[b]Department of Statistics, PUC-Chile

[c]Department of Econometrics and Statistics, Universidad del Pais Vasco-Spain

**Resumo**

This paper explores seasonal and long-memory time series properties by using the fractional ARIMA model when the data have one and two seasonal periods and short-memory components. The stationarity and invertibility parameter conditions are established for the model studied. To estimate the seasonal fractional long-memory parameters, a semiparametric estimation method is proposed. The asymptotic properties of the estimator are established and the accuracy of the method is investigated through Monte Carlo experiments. The good performance of the estimator indicates that it can be an alternative procedure to estimate long-memory time series data with two seasonal periods. Series of $PM_{10}$ concentrations and electricity hourly demand are considered as examples of applications of the proposed estimation method.

## 1 Introduction

Time series exhibiting seasonal or cyclical characteristics are very common in economics, hydrology, and many other disciplines. As a consequence, several methodologies have been developed to deal with these features (see, for example, Gould et al. (2008)). One of the most well known of these tools is the class of seasonal autoregressive integrated moving average (SARIMA) process. This model can describe many time series containing a mixture of seasonal phenomena of different periods (see, for example, Taylor (2010)). It is well known that many series may contain a persistent seasonal structure along with a long run trend (see, for example, Ferrara & Guegan (2006) and Ray (1993)).

Let $X_t \equiv \{X_t\}_{t\in\mathbb{Z}}$ be a time series with a zero mean and a constant variance. A multiple

*Corresponding author. E-mail: valderioanselmoreisen@gmail.com

seasonal ARIMA model can be written as follows:

$$\prod_{j=1}^{M} \phi_j(B)\nabla^{\mathbf{d}}(B)X_t = \prod_{\ell=1}^{N} \theta_\ell(B)\varepsilon_t, \tag{1}$$

where $\nabla^{\mathbf{d}}(B) \equiv \prod_{\iota=1}^{k}(1-B^{s_\iota})^{d_\iota}$, $M$ and $N$ are, respectively, the number of factors of the AR and MA components, $d_\iota \in \mathbb{N}$ , $\iota = 1, ..., k$, is the differencing parameter and $k$ is the number of differencing factors, $s_\iota$ is the $\iota$-seasonal period, $BX_t = X_{t-1}$, and $\varepsilon_t$ is a white noise process with zero-mean and variance $\sigma_\varepsilon^2$. $\phi_j(B), j = 1, ..., M$, and $\theta_\ell(B), \ell = 1, ..., N$, can also be polynomials with seasonal effects. The stationarity and invertibility properties of model (1) are established based on certain parameter conditions. The multiple seasonal ARIMA model belongs to a class of models with a general difference operator given by

$$\nabla^{\mathbf{d}}(B) \equiv \prod_{\ell=1}^{L}[(1 - Be^{i\lambda_\ell})(1 - Be^{-i\lambda_\ell})]^{d_\ell}, \tag{2}$$

where now $d_\ell \in \mathbb{R}$ ($d_\ell > -1$) and it is defined as the *fractionally differencing parameter* and $\lambda_\ell$, $\ell = 1, ..., L$, are fixed frequencies in the range $[-\pi, \pi]$. For suitable choices of the fractional parameters, time series models with filter given in (2) may have a finite number of zeros or singularities of order $d_1$,...,$d_L$ on the unit circle which allows the modeling of long and short memory data containing seasonal periodicities. In the time domain, the usual definition of long memory is the non-summability condition

$$\sum_{h=0}^{\infty} |\gamma(h)| = \infty,$$

where $\gamma(h)$ is the autocovariance at lag $h$ of the process, whereas, in the frequency domain, this property is defined by the fact that the spectral density of the process becomes unbounded at some frequencies in [0,$\pi$]. See, for example, Reisen et al. (2010) for a recent survey on the estimation methodology of time series with long-memory.

A time series with both seasonal and non-seasonal fractional differencing parameters has a spectral density specified by

$$f(\lambda) = f^*(\lambda)|\lambda|^{-2d} \prod_{\iota=1}^{L} \prod_{j=1}^{\xi_\iota} |\lambda - \lambda_{\iota j}|^{-2d_\iota}, \tag{3}$$

where $d_\iota \in \mathbb{R}$ ($d_\iota > -1$), $\lambda \in (-\pi, \pi]$, $f^*(\lambda)$ is a continuous function, bounded above and away from zero and $\lambda_{\iota j} \neq 0$ are poles for $j = 1, ..., \xi_\iota$, $\iota = 1, ..., L$. Processes with a spectral density given by (3) have been discussed by Arteche & Robinson (1999), Giraitis & Leipus (1995), Leipus & Viano (2000), Palma & Chan (2005) and Palma (2007), among others, to

model time series with seasonal and cyclical long-memory behavior.

The main interest in models which have filter (2) and spectral density of the form (3) is related to the estimation of fractional memory parameters $d_1$,...,$d_L$. Ray (1993) used IBM product revenues to illustrate the usefulness of modeling seasonal fractionally differenced ARMA models by allowing two seasonal fractional differencing parameters in the model, one at lag 3 and the other at lag 12. Other papers related to this topic are, for example, Hassler (1994), Gray et al. (1989), Gray et al. (1994), Giraitis & Leipus (1995), Woodward et al. (1998), Ooms & Franses (2001), Arteche & Robinson (2000), Palma & Chan (2005), Reisen et al. (2006*b*,*a*) , Gil-Alana (2001), among others. Hassler (1994) introduced the rigid and flexible seasonal filters and an application of this methodology to the economic activities in the Euro area is discussed in Ferrara & Guegan (2006). Arteche & Robinson (2000), Arteche (2002) and Arteche & Velasco (2005) dealt with robust semiparametric estimators and testing procedures for the seasonal fractional memory parameter. Woodward et al. (1998) extended the Gegenbauer ARMA process (GARMA). Independently of these works, a time series model for fitting long or short-memory data containing seasonal periodicities was introduced by Giraitis & Leipus (1995) which is called the Fractionally Autoregressive Unit Circle Moving Average model (ARUMA). These authors discussed the asymptotic properties of the ARUMA model and the estimation of its parameters.

Another equally relevant publication related to the asymptotic properties of seasonal and periodic time series is the work by Viano et al. (1995) in which the authors provided theoretical results of the extended fractional ARMA processes. Reisen et al. (2006*b*,*a*) dealt with the estimation of the seasonal ARFIMA model with long-memory innovations (SARFIMA $(p, d, q) \times (P, D, Q)_s$) by using different estimation procedures for the seasonal and non-seasonal memory parameters, that is, for $D$ and $d$ respectively. The estimators are based on the multilinear regression equation of $\log f(\cdot)$, where $f(\cdot)$ is the spectral density of the process.

Necessary conditions that guarantee the stationary and invertibility of the model were also established. Through Monte Carlo experiments, they compared their proposed methodology with other well-known parametric estimation procedures such as the Whittle and the maximum likelihood methods. The empirical evidence showed that the multilinear regression estimators are very promising. The SARFIMA $(p, d, q) \times (P, D, Q)_s$ model was also considered by Marques (2011) where the parameters were estimated by the maximum likelihood method.

Most of the works referred to above deal with the estimation of only one seasonal long-memory parameter in the context of finite sample size investigation and applied situation. However, in many practical problems the time series exhibits more than one seasonal component. In order to explore these more complex situations, this paper focuses on the estimation of models containing one and two seasonal periods which encompass long and short-memory dependence structures. Specifically, a consistent semiparametric ordinary least squares (OLS)

procedure, based on a log-periodogram regression, is proposed to estimate all fractional parameters simultaneously. The seasonal ARFIMA model considered here is a particular case of the ARUMA process studied by Giraitis & Leipus (1995). In the paper, the authors discussed the parameter estimation based on the parametric Whittle´s approach. However, the semiparametric OLS estimator proposed here has three main advantages over the one discussed in Giraitis & Leipus (1995). The calculation of the estimates is simple and fast; the semiparametric setup, by definition, is less restrictive than a parametric approach and the parametric method gives estimates that are highly biased under order misspecification.

Let now $X_t \equiv \{X_t\}_{t\in\mathbb{Z}}$ be a zero-mean time series defined by

$$X_t = (\nabla^{\mathbf{d}}(B))^{-1}\nu_t = (1-B^{s_1})^{-d_1}(1-B^{s_2})^{-d_2}\nu_t, \tag{4}$$

where $\mathbf{d} = (d_1, d_2)^{'}$, $s_1$ and $s_2$ are seasonal periods and $d_1, d_2 \in \mathbb{R}$ ($d_\iota > -1$) are their seasonal memory parameters, respectively, and $\nu_t$ has a spectral density that satisfies the following assumption.

**Assumption 1:** The spectral density of $\nu_t$ satisfies as $\lambda \to 0$

$$f_\nu\left(\frac{2\pi k}{s'} + \lambda\right) = f_{\nu,k} + c_k|\lambda|^{\alpha_k} + O(|\lambda|^{\alpha_k+\varsigma})$$

for some $\varsigma > 0$, $f_{\nu,k} \equiv f_\nu\left(\frac{2\pi k}{s'}\right)$, $k = 0, 1, ..., [s'/2]$, $s' = \max(s_1, s_2) = s_1$ (without loss of generality) and $\alpha_k = \alpha_1$ for $k = 0, s'/2$ (if $s'$ even) and $\alpha_k = \alpha_2$ otherwise. If $\nu_t$ is a stationary and invertible ARMA process then $\alpha_1 = 2$, $\alpha_2 = 1$ and $\varsigma = 1$. In this case, the process $X_t$ is usually defined as Seasonal ARFIMA (SARFIMA) model.

In the next section some properties of the model given by (4) are discussed. In particular, the stationarity and invertibility conditions of model (4) are established in Proposition 1. The estimation of these models is discussed in Section 3, where the proposed ordinary least squares (OLS) estimator is introduced. Some asymptotic properties of these estimators are established in Theorems 1 and 2. For example, the proposed OLS estimator is shown to be asymptotically unbiased and normally distributed. The finite sample performance of the proposed estimator is investigated in Section 4 while Section 5 discusses some applications. Final remarks are presented in Section 6.

## 2 Model properties

Let $X_t$ be a time series process defined by (4). For simplicity, it is assumed that $s_1$ and $s_2$ are even numbers. The fractional $d_\iota$ difference is a generalization of the binomial expression

$(1-B)^d$ and it can be written as

$$(1-B^{s_\iota})^{d_\iota} = \sum_{k=0}^{\infty} \pi_{k,\iota} B^{k s_\iota},$$

where

$$\pi_{0,\iota} = 1, \qquad \pi_{k,\iota} = \frac{\Gamma(k-d_\iota)}{\Gamma(k+1)\Gamma(-d_\iota)}, \quad \iota = 1, 2,$$

and $\Gamma(\cdot)$ is the Gamma function.

In the literature of seasonal long-memory processes, there are some specific time series models of interest obtained from the solution of the general fractional operator (2) and the spectral density of form (3). The specific filters and their models are: (a) $(1-B)^d$ is the filter of the fractional integrated I($d$) process see, for example, Hosking (1981), among others; (b) $(1-B)^{d_1}(1-B^s)^{d_2}$ is the filter in the SARFIMA process that has been explored in the literature of seasonal fractional ARMA model, see for example, Porter-Hudak (1990), Hassler (1993), Arteche & Robinson (2000), Arteche (2002) and Reisen et al. (2006*b*,*a*) ; (c) $(1-\upsilon_1 B - \cdots - \upsilon_v B^d)$ is the filter that belongs to the ARUMA and $k$-GARMA processes, proposed by Giraitis & Leipus (1995) and independently by Woodward et al. (1998) (1998), respectively; (d) $(1-B^3)^{d_1}(1-B^{12})^{d_2}$ is the filter used by Ray (1993) to model and forecast a monthly IBM revenue data under the restriction $d_1 + d_2 = 1$.

Returning to our specific model of interest for $X_t$ given in (4), the filter may be written as follows:

$$(1-B^{s_1})^{d_1}(1-B^{s_2})^{d_2} = \prod_{\iota=1}^{2}\prod_{j=0}^{\xi_\iota}[(1-Be^{i\lambda_{\iota j}})(1-Be^{-i\lambda_{\iota j}})]^{d_{\iota j}}, \tag{5}$$

where $\lambda_{\iota j} = \frac{2\pi j}{s_\iota}$ $(j = 0, 1, \ldots \xi_\iota)$ are the frequencies of the period $\iota$, $\xi_\iota = [\frac{s_\iota}{2}]$ $(\iota = 1, 2)$. For example, when $s_1$ and $s_2$ are even numbers and $\iota = 1$, the memory parameters in (5) are given by

$$\begin{array}{ll} d_{1j} = d_1 & \text{when } \lambda_{1j} \neq \lambda_{2j}, 0, \pi; \\ d_{1j} = d_1/2 & \text{when } \lambda_{1j} \neq \lambda_{2j} \text{ and } \lambda_{1j} = 0, \pi; \\ d_{1j} + d_{2j} = \frac{d_1+d_2}{2} & \text{when } \lambda_{1j} = \lambda_{2j} = 0, \pi; \\ d_{1j} + d_{2j} = d_1 + d_2 & \text{when } \lambda_{1j} = \lambda_{2j} \neq 0, \pi; \end{array} \tag{6}$$

and similarly when $\iota = 2$.

It is easy to show that the filter (5) is a particular case of the operator (2) by using the

equality

$$1 - z^s = (1 - z)(1 + z) \prod_{k=1}^{\frac{s}{2}-1} (1 - ze^{2\pi ik/s})(1 - ze^{-2\pi ik/s}), \quad \text{for} \quad s \quad \text{even.}$$

When $s$ is an odd number, the term $(1 + z)$ does not appear in the above equation and the product goes up to $\frac{s-1}{2}$. From the expression of $1 - z^s$, the following proposition is reached:

**Proposition 1.** *Let the process $X_t$ be a solution of equation*

$$X_t = (\nabla^{\boldsymbol{d}}(B))^{-1}\nu_t = (1 - B^{s_1})^{-d_1}(1 - B^{s_2})^{-d_2}\nu_t, \tag{7}$$

*where $\boldsymbol{d} = (d_1, d_2)^{'}$, $\nu_t$ is a covariance stationary ARMA process ( $\nu_t = \frac{\Theta(B)}{\Phi(B)}\epsilon_t$), $\epsilon_t$ is an i.i.d Gaussian sequence with zero mean and variance $\sigma_\epsilon^2$ , and $d_\iota \in \mathbb{R}$ is the fractional parameter at seasonal period $s_\iota$ for $\iota = 1, 2$. Then,*

*(a) The process $X_t$ is stationary and invertible if $|d_1 + d_2| < 1/2$ and $|d_\iota| < 1/2$, $\iota = 1, 2$.*

*(b) The spectral density of $X_t$ is given by*

$$\begin{aligned} f(\lambda) &= f_\nu(\lambda) \prod_{\iota=1}^{2} \prod_{j=0}^{\xi_\iota} |2\sin(\tfrac{\lambda-\lambda_{\iota j}}{2}) 2\sin(\tfrac{\lambda+\lambda_{\iota j}}{2})|^{-2d_{\iota j}} \\ &= f_\nu(\lambda)\Big(2\sin\tfrac{\lambda s_1}{2}\Big)^{-2d_1}\Big(2\sin\tfrac{\lambda s_2}{2}\Big)^{-2d_2}, \end{aligned}$$

*where $f_\nu(\lambda)$ ( $0 \le \lambda \le \pi$) is the spectral density of $\nu_t$, $\lambda_{\iota j} = \frac{2\pi j}{s_\iota}$, $\iota = 1, 2$ and $j = 0, 1, ..., \frac{s_\iota}{2}$, and $d_{\iota j}$ are given by (5).*

*(c) Assuming that $\max\{d_{\iota j}\} > 0$, the asymptotic autocovariance of $X_t$, $\gamma(h) = \mathbb{E}(X_h X_0)$, is given by*

$$\gamma(h) = \sum_{\iota=1}^{2} \sum_{j=1}^{\xi_\iota} a_{\iota j} \mid h \mid^{2d_{\iota j}-1} (\cos h\lambda_{\iota j} + o(1)) \quad as \quad k \to \infty,$$

*where*

$$a_{\iota j} = \begin{cases} a_{\iota j}^{'} & \lambda_{\iota j} = 0, \pi \\ 2a_{\iota j}^{'} & 0 < \lambda_{\iota j} < \pi, \end{cases}$$

*$d_{\iota j}$ is specified as from (5) and*

$$a_{\iota j}^{'} = \mid \frac{\Theta(e^{-2\pi\lambda_{\iota j}})}{\Phi(e^{-2\pi\lambda_{\iota j}})} \mid^2 \frac{\sigma_\epsilon^2}{\pi}\Gamma(1 - 2d_{\iota j})\sin(d_{\iota j}\pi)D_{\iota j}^2,$$

*where*

$$D_{\iota j} = \begin{cases} \mid 2\sin\lambda_{\iota j} \mid^{-d_{\iota j}} \prod_{\ell \neq j} \mid 2(\cos\lambda_{\iota j} - \cos\lambda_{\iota \ell}) \mid^{-d_{\iota \ell}}, & 0 < \lambda_{\iota j} < \pi, \\ \prod_{\ell \neq j} \mid 2(\cos\lambda_{\iota j} - \cos\lambda_{\iota \ell}) \mid^{-d_{\iota j \ell}}, & \lambda_{\iota j} = 0, \pi. \end{cases}$$

*Demonstração.* (a) As previously noted, filter (5) is a particular case of (2) which is the operator of the ARUMA$(p, d_1, ..., d_L, q)$ model where $p$ and $q$ are the polynomial orders of a stationary and invertible ARMA$(p, q)$ process. From Theorem 1 of Giraitis & Leipus (1995), the ARUMA $(0, d_1, ..., d_L, 0)$ process is stationary and invertible if the fractional parameters $d_\ell$, $\ell = 1, ..., L$, in (2) satisfy $|d_\ell| < 1/2$ when $\lambda_\ell \neq 0, \pi$ and $|d_\ell| < 1/4$ otherwise. From this fact and by means of equation (5), it is straightforward to establish the stationary and invertibility properties. The proof of (b) is immediately obtained from (3) and Theorem 2 in Giraitis & Leipus (1995). This theorem is also used to prove the asymptotic covariance given in $(c)$ where $d_{\iota j}$ is obtained by (5). □

# 3 Seasonal fractional parameter estimators

This section deals with the estimation method based on the regression equation of $\log f(\lambda)$ to obtain the estimates of model (7). Since the procedure proposed here provides simultaneous estimates for multiple seasonal memory parameters, the method is a more general approach than those discussed in Reisen et al. (2006$b$,$a$) and related references. Let $n$ be the sample size and let $X_1, \ldots, X_n$ be a realization of the process defined by (7), where $\nu_t$ is a Gaussian ARMA process. The well-known periodogram function $I(\lambda) = (2\pi n)^{-1} |\sum_{t=1}^{n} X_t e^{i\lambda t}|^2$ is an asymptotic unbiased and inconsistent estimator of the spectral density and it is the standard estimator used in time series modeling.

## 3.1 The OLS regression estimators

The fractional memory **OLS** estimators of the vector $\mathbf{d} = (d_1, d_2)'$ ( Eq. (7)) are the slope estimators of the multiple regression equation

$$\log I_{k,j} = a_k - 2d_1 \log X_{1,k,j} - 2d_2 \log X_{2,k,j} + V_{k,j} \quad , k = 0, 1, ..., [s'/2], \tag{8}$$

where $s' = \max(s_1, s_2)$, $j = 1, ..., m$ ( $m \in \mathbb{N}^*$) if $k = 0$, $j = -1, ..., -m$ if $k = s'/2$ ($s$ even) and $j = \pm 1, ..., \pm m$ otherwise, $I_{k,j} = I(2\pi k/s' + \lambda_j)$, $\lambda_j = 2\pi j/n$ is the Fourier frequency, $[\cdot]$

means the integer part and

$$\begin{aligned}
a_k &= \log f_{\nu,k} + E\left(\log \frac{I_{k,j}}{f_{k,j}}\right) \\
V_{k,j} &= U_{k,j} + \varepsilon_{k,j} \\
U_{k,j} &= \log \frac{I_{k,j}}{f_{k,j}} - E\left(\log \frac{I_{k,j}}{f_{k,j}}\right) \\
\varepsilon_{k,j} &= \log \frac{f_\nu\left(\frac{2\pi k}{s'} + \lambda_j\right)}{f_{\nu,k}} = b_k \lambda_j^\alpha + O(\lambda_j^{\alpha+\iota}) \\
X_{1,k,j} &= 2\sin\left(\frac{s_1}{2}\left[\frac{2\pi k}{s'} + \lambda_j\right]\right) \\
X_{2,k,j} &= 2\sin\left(\frac{s_2}{2}\left[\frac{2\pi k}{s'} + \lambda_j\right]\right)
\end{aligned}$$

for $f_{k,j} = f\left(\frac{2\pi k}{s'} + \lambda_j\right)$ and $b_k = c_k / f_{v,k}$. The regression equation (8) is easily derived from the expression of the $\log f(\lambda)$ where $f(\lambda)$ is the spectral density given in Proposition 1. To avoid the estimation of the constants $a_k$, the variables are locally centered such that the estimates are obtained by least squares in the regression model

$$Y_{k,j} = d_1 Z_{1,k,j} + d_2 Z_{2,k,j} + V^*_{k,j}, \tag{9}$$

where $V^*_{kj} = V_{k,j} - \frac{1}{m_k}\sum_j^* V_{k,j}$ for $m_k = \delta_k m$ with $\delta_k = 1$ for $k = 0, s'/2$ and $\delta_k = 2$ otherwise and the sum $\sum^*$ runs for $j = 1, ..., m$ if $k = 0$, $j = -1, ..., -m$ if $k = s'/2$ and $j = \pm 1, ..., \pm m$ otherwise. $Y_{k,j}$, $Z_{1,k,j}$ and $Z_{2,k,j}$ are the locally centered dependent variable and regressors in (8) similarly defined. The local centering is needed here because the regression model in (8) has different constants depending on the frequency bandwidth. A global centering can be used only if $a_1 = ... = a_{[s'/2]}$ which holds for example if $\nu_t$ is a white noise process with a constant spectral density function.

The estimation procedure based on the above regression equation is motivated by the pioneer regression estimator proposed by Geweke & Porter-Hudak (1983) for the ARFIMA model. Since the introduction of the method, it has became one of the most popular estimation procedures and its empirical and asymptotic properties have been well established. Robinson (1995) and Hurvich et al. (1998) proved that the GPH-estimator is consistent and asymptotically normal for Gaussian time series processes. Hurvich et al. (1998) also established that the optimal bandwidth is of order $O(n^{4/5})$. As example of applications of the GPH method and their variants see Franco & Reisen (2007), Palma (2007) among others.

When the model is a SARFIMA$(0, d, 0) \times (0, d_s, 0)_s$ process, Reisen et al. (2006*b*,*a*) proposed different estimation methods for $d_s$ and $d$. Basically, the regression estimators considered in their study are distinguished by the choice of the bandwidth when regressing $\log[I(\lambda)]$ on $\log[2\sin(\lambda s/2)]$ and $\log[2\sin(\lambda/2)]$. Following the same direction, their study is generalized

here in the case where the model has two seasonal fractional parameters $d_1$ and $d_2$ for the seasonal periods $s_1$ and $s_2$, respectively. Returning to the estimation of the fractional parameters in Eq (7), in what follows some asymptotical results are established.

**Assumption 2:** $s_1$ is a multiple of $s_2$.

**Assumption 3:** Let $m = m(n)$ is a sequence satisfying

$$\left(\frac{m}{n}\right)^{\iota} \log m + \frac{1}{m} \to 0 \quad \text{as} \quad n \to \infty$$

for some $\iota > 0$.

**Assumption 4:**

$$\frac{m^{\alpha^*+0.5}}{n^{\alpha^*}} \to K \quad \text{as} \quad n \to \infty$$

where $\alpha^* = \min(\alpha_1, \alpha_2)$.

**Assumption 5:** The process $X_t$ has Gaussian distribution.

Remark. Assumption 2 implies that both memory parameters affect the spectral density at some frequencies such that $d_2$ is only identifiable by identifiability of $d_1$. In that sense, Assumption 2 considers the most challenging case, at least for estimation of $d_2$, since that estimation is only possible after estimation of $d_1$. If assumption 2 is not satisfied both memory parameters are uniquely defined at some frequencies and thus joint estimation of both memory parameters is not necessary because separate log periodogram regressions can be run, one at frequencies affected only by $d_1$ and other at frequencies affected by $d_2$. Of course joint estimation is also a possibility in that case but is not necessary for identifiability of both parameters. Assumption 3 entails Condition 1 in Hurvich Hurvich et al. (1998) and permits the use of their results in our proofs. Assumptions 4 and 5 are needed to get an expression of the asymptotic bias in terms of $K$ and to use the results in Hurvich et al. (1998), respectively.

**Theorem 1.** *Under assumptions 1,2 and 3, as* $n \to \infty$,

$$E(\hat{\boldsymbol{d}}) - \boldsymbol{d} = Q^{-1} b_n (1 + o(1))$$

$$Var(\hat{\boldsymbol{d}}) = m^{-1} \frac{\pi^2}{6} Q^{-1} (1 + o(1)),$$

*where*

$$b_n = -2 \begin{pmatrix} \sum_{k=0}^{[s'/2]} b_k \delta_k (2\pi)^{\alpha_k} \frac{\alpha_k}{(\alpha_k+1)^2} \left(\frac{m}{n}\right)^{\alpha_k} \\ \sum_{k \in I_k} b_k \delta_k (2\pi)^{\alpha_k} \frac{\alpha_k}{(\alpha_k+1)^2} \left(\frac{m}{n}\right)^{\alpha_k} \end{pmatrix},$$

$$Q = 4 \begin{pmatrix} \sum_{k=0}^{[s'/2]} \delta_k & \sum_{k \in I_k} \delta_k \\ \sum_{k \in I_k} \delta_k & \sum_{k \in I_k} \delta_k \end{pmatrix},$$

*where* $I_k = \{0, k \text{ such that } ks_2 \text{ is a multiple of } s'\}$, $\delta_k = 1$ *for* $k = 0, s'/2$ *and* $\delta_k = 2$ *otherwise. In consequence,* $\hat{d}$ *is consistent.*

**Theorem 2.** *Under assumptions 1, 2, 3 , 4 and 5, as* $n \to \infty$,

$$\sqrt{m}(\hat{\boldsymbol{d}} - \boldsymbol{d}) \xrightarrow{d} N\left(Q^{-1}b, \frac{\pi^2}{6}Q^{-1}\right)$$

*for*

$$b = -2(2\pi)^{\alpha^*}\frac{\alpha^*}{(\alpha^*+1)^2}K\begin{pmatrix} \sum_{k\in J_k} b_k\delta_k \\ \sum_{k\in I_k \bigcap J_k} b_k\delta_k \end{pmatrix},$$

*where* $J_k = \{k \text{ such that } \alpha_k = \alpha^*\}$.

Proofs of the above theorems are in Appendix A.

As a particular case of Model 7, the statistical properties of the SARFIMA$(P, d, Q)_s$ model are now discussed.

The OLS estimator of $d$ is given by

$$\widehat{d} = (-0.5)\frac{\sum_{k=0}^{[\frac{s}{2}]}\sum_{j=1}^{m}(X_{1,k,j} - \bar{X_1})\log I_{k,j}}{\sum_{k=0}^{[\frac{s}{2}]}\sum_{j=1}^{m}(X_{1,k,j} - \overline{X_1})^2}, \tag{10}$$

where $X_{1,k,j} = \log\{2\sin((s\lambda_{k,j}/2)\}$. By simple algebra, the following expression is reached.

$$\widehat{d} - d \approx -\frac{1}{2S_{X_1X_1}}\sum_{k=0}^{[\frac{s}{2}]}\sum_{j=1}^{m}(X_{1,k,j} - \overline{X_1})U_{k,j}, \tag{11}$$

where $S_{X_1X_1} = \sum_{k=0}^{[\frac{s}{2}]}\sum_{j}^{m}(X_{1,k,j} - \overline{X_1})^2$ and $j$ is defined as in (8).

**Collorary 1.** *Let* $X_t$ *be a SARFIMA*$(P, d, Q)_s$ *model and* $\widehat{d}$ *is the OLS estimator of* $d$ *provided by (10). Under assumptions 1 to 4, as* $n \to \infty$,

*(a)*

$$E(\widehat{d}) \approx d.$$

*(b) The variance of the estimator is given by*

$$Var(\widehat{d}) \approx \frac{\pi^2}{24sm}. \tag{12}$$

*(c) The estimate* $\widehat{d}$ *satisfies*

$$\sqrt{m}\,(\widehat{d} - d) \xrightarrow{d} N\left(0, \frac{\pi^2}{24s}\right).$$

*Demonstração.* The above results are particular cases of Theorems 1 and 2 and they coincide with Theorems 1 and 2 in Hurvich et al. (1998) when $s = 1$. Note that the variance of the estimator suggested in Porter-Hudak (1990) and Ray (1993) is approximately $4s^2Var(\widehat{d})$ which is much larger than the one proposed here. □

# 4 Finite sample investigation

The finite sample performance of the estimator discussed previously is investigated in this section through Monte Carlo experiments for different structures of Model 7 where $\nu_t$ follows a SARMA model. To generate the models, the procedure used is the one suggested in Hosking (1984) with i.i.d innovations from a N(0,1) distribution. The models are: SARFIMA$(p, d_1, q)_{s_1}(P, d_2, Q)_{s_2}$ with $p = P = 0, 1$, $s_1 = 4$, 12, $s_2 = 1$,4 and the AR non-seasonal ($\phi_1$) and seasonal ($\phi_s$) parameters with values $\phi_1$=$\phi_s$= 0.0, 0.3 and 0.8. The parameters are also displayed in the tables. The empirical investigations were based on sample size $n = 1080$, and the sample quantities mean, correlation and mean squared error ($mse$) of the estimators were calculated over 2,000 replications. Other cases with smaller sample sizes ($n = 480, 600$) were also investigated. The performance of the methods were quite similar to the one presented here and they are available upon-request. The calculations were carried out by means of an Ox program in an AMD Athlon XP 1800 computer.

Since the models also involve short-range dynamics, the regression estimators were obtained by using different bandwidths. In the case where the model has not AR contribution, the bandwidth $m = [\frac{n-1}{\max(s_1,s_2)}]$ was fixed. In this context where all Fourier frequencies are used the regression estimator ($GPH_T$) becomes a parametric procedure. For the models with short-memory dynamics, the two bandwidths $m = n^{\alpha_i}$, $\alpha_1 = 0.5$ and $\alpha_2 = 0.3$ were used, and the estimators are denoted as $GPH_1$ and $GPH_2$, respectively. The bandwidth $n^{0.5}$ is here considered because this specification has been widely used in the case of ARFIMA models with short-memory components, while the choice of $n^{0.3}$ is based on the empirical investigation discussed below.

Table 1 summarizes the results for the SARFIMA model with $d_1 = 0.3$ ($s_1 = 4$). The first part of this table shows the performance of the regression methods when there is no seasonal AR contribution. For the case of $\phi_1 = 0$, $GPH_T$ has the best performance among the GPH based ones, which is an expected result since the method uses all non-seasonal frequencies. The effect of the bandwidth is also a motivation of this study. The reduction of the bandwidth causes an increase in the $mse$, especially when $\phi_1$= 0. This is a not surprising result, since the AR contribution is mainly concentrated at zero frequency. The absence of short-memory component allows a wider bandwidth because the bias is quite controlled. Thus, a reduction of $m$ implies a larger variance and does not reduce the bias.

In the second part of the table, the estimates were computed when the model has the AR

contribution at the seasonal period $s = 4$. From this, the GPH estimates are more affected by the AR component than the previous case, the bias is strongly positive and the $mse$ also increases. In this case the AR component has spectral power not only at frequency zero but also at the seasonal ones, affecting to a greater extent the estimation of $d_1$. The small value of the bandwidth mitigates the effect of the short-memory parameters. This is clearer when $\phi_4$ changes from 0.3 to 0.8. Hence, in this context, the decrease of the bandwidth produces reduction on the size of the bias and the $mse$.

Tables 2 and 3 present the estimates when the models have more than one fractional parameter. Thus, the sample correlations between the estimates were also calculated.

Tabela 1: Results for the seasonal ARFIMA model with $d_1 = 0.3$ ($s_1 = 4$) and $\phi_s$, $s = 1, 4$, $n = 1080$.

| $\phi_s$ | estimators | $\hat{d}_1$ | |
|---|---|---|---|
| | | mean | mse |
| $\phi_1 = 0.0$ | $\text{GPH}_T$ | 0.3004 | 0.0012 |
| | $\text{GPH}_1$ | 0.2988 | 0.0046 |
| | $\text{GPH}_2$ | 0.2984 | 0.0311 |
| $\phi_1 = 0.3$ | $\text{GPH}_1$ | 0.3002 | 0.0041 |
| | $\text{GPH}_2$ | 0.3074 | 0.0255 |
| $\phi_1 = 0.8$ | $\text{GPH}_1$ | 0.3097 | 0.0054 |
| | $\text{GPH}_2$ | 0.3085 | 0.0242 |
| $\phi_4 = 0.3$ | $\text{GPH}_1$ | 0.3459 | 0.0068 |
| | $\text{GPH}_2$ | 0.3114 | 0.0278 |
| $\phi_4 = 0.8$ | $\text{GPH}_1$ | 0.7281 | 0.1877 |
| | $\text{GPH}_2$ | 0.4144 | 0.0426 |

Table 2 displays the result when the models are SARFIMA $(1, d_1, 0)_{s_1}$ $(1, d_2, 0)_{s_2}$ with $d_1 = 0.3(s_1 = 4)$, $d_2 = 0.1(s_2 = 1)$, $\phi_1 = 0.0$, 0.3, 0.8 and $\phi_4 = 0.3$, 0.8 whereas Table 3 shows the performance of the estimates when the SARFIMA model has seasonal periods $s_1 = 12$ and $s_2 = 4$. From Table 2 it should be noted that the contribution of the parameter $d_2$ is mainly at zero frequency. Hence, in general, the semiparametric estimators perform similarly to the previous case that is, the estimate of $\mathbf{d}$ depends on the the values of the bandwidth and of the AR counterpart. The memory parameters are estimated simultaneously, thus there is a balance effect between the two estimates $\hat{d}_1$ and $\hat{d}_2$ which justifies the negative correlation values between them. In addition, the estimates are balanced to have the value of $\hat{d}_1 + \hat{d}_2$ approximately equal to $d_1 + d_2$ which is the total memory at zero frequency. The correlations between the GPH estimates increases with the bandwidth and the AR coefficients.

Although the model in Table 3 has fractional parameters at seasonality periods 4 and 12, similar conclusions of the performance of the estimates to the previous cases are observed.

As an additional illustrative tool to observe the method's performance, the box-plots in Figures 1 and 2 show the variation of the estimates for the model in Table 3 with $\phi_1 = 0.0$ and $\phi_1 = 0.3$, respectively. As previously observed, in the model with no AR part the regression

Tabela 2: Results for models $d_1 = 0.3$ ($s_1 = 4$), $d_2 = 0.1$ ($s_2 = 1$) and $\phi_s$, $s = 1, 4$, case $n = 1080$.

| $\phi_s$ | estimators | $\hat{d}_2$ | | corr. | $\hat{d}_1$ | |
|---|---|---|---|---|---|---|
| | | mean | mse | | mean | mse |
| $\phi_1 = 0.0$ | $\text{GPH}_T$ | 0.1043 | 0.0018 | −0.2228 | 0.3010 | 0.0013 |
| | $\text{GPH}_1$ | 0.1135 | 0.0290 | −0.5144 | 0.2995 | 0.0053 |
| | $\text{GPH}_2$ | 0.0818 | 0.1327 | −0.4164 | 0.3121 | 0.0388 |
| $\phi_1 = 0.3$ | $\text{GPH}_1$ | 0.1166 | 0.0215 | −0.3397 | 0.3098 | 0.0045 |
| | $\text{GPH}_2$ | 0.1463 | 0.1553 | −0.4927 | 0.3083 | 0.0308 |
| $\phi_1 = 0.8$ | $\text{GPH}_1$ | 0.2208 | 0.0388 | −0.5415 | 0.3004 | 0.0062 |
| | $\text{GPH}_2$ | 0.1257 | 0.1607 | −0.5190 | 0.3148 | 0.0375 |
| $\phi_4 = 0.3$ | $\text{GPH}_1$ | 0.1074 | 0.0249 | −0.4751 | 0.3404 | 0.0083 |
| | $\text{GPH}_2$ | 0.1107 | 0.1299 | −0.4233 | 0.3037 | 0.0348 |
| $\phi_4 = 0.8$ | $\text{GPH}_1$ | 0.1045 | 0.0285 | −0.5011 | 0.7269 | 0.1874 |
| | $\text{GPH}_2$ | 0.0445 | 0.1591 | −0.4773 | 0.4251 | 0.0549 |

Tabela 3: Results for the SARFIMA model with $d_1 = 0.3$ ($s_1 = 12$), $d_2 = 0.1$ ($s_2 = 4$), $\phi_s$, $s = 1, 4, 12$ and $n = 1080$.

| $\phi_s$ | estimators | $\hat{d}_2$ | | Corr. | $\hat{d}_1$ | |
|---|---|---|---|---|---|---|
| | | mean | mse | | mean | mse |
| $\phi_1 = 0.0$ | $\text{GPH}_T$ | 0.1047 | 0.0018 | −0.3194 | 0.3065 | 0.0015 |
| | $\text{GPH}_1$ | 0.0994 | 0.0052 | −0.4405 | 0.3071 | 0.0021 |
| | $\text{GPH}_2$ | 0.0723 | 0.0442 | −0.5529 | 0.3095 | 0.0113 |
| $\phi_1 = 0.3$ | $\text{GPH}_1$ | 0.1063 | 0.0061 | −0.5230 | 0.3017 | 0.0022 |
| | $\text{GPH}_2$ | 0.0797 | 0.0433 | −0.5446 | 0.2954 | 0.0119 |
| $\phi_1 = 0.8$ | $\text{GPH}_1$ | 0.1468 | 0.0067 | −0.5212 | 0.2861 | 0.0030 |
| | $\text{GPH}_2$ | 0.1177 | 0.0374 | −0.5767 | 0.2861 | 0.0136 |
| $\phi_4 = 0.3$ | $\text{GPH}_1$ | 0.2043 | 0.0160 | −0.4707 | 0.2813 | 0.0020 |
| | $\text{GPH}_2$ | 0.1146 | 0.0338 | −0.4223 | 0.3019 | 0.0084 |
| $\phi_4 = 0.8$ | $\text{GPH}_1$ | 0.5962 | 0.2528 | −0.4792 | 0.2618 | 0.0038 |
| | $\text{GPH}_2$ | 0.2634 | 0.0625 | −0.5317 | 0.2912 | 0.0130 |
| $\phi_{12} = 0.3$ | $\text{GPH}_1$ | 0.1055 | 0.0062 | −0.5344 | 0.5044 | 0.0439 |
| | $\text{GPH}_2$ | 0.1255 | 0.0481 | −0.6629 | 0.3480 | 0.0161 |
| $\phi_{12} = 0.8$ | $\text{GPH}_1$ | 0.0863 | 0.0063 | −0.4587 | 1.0064 | 0.5010 |
| | $\text{GPH}_2$ | 0.1089 | 0.0354 | −0.5466 | 0.7553 | 0.2187 |

estimators obtained with larger amount of Fourier frequencies are the dominators in the sense of presenting better empirical statistical properties, as the case of the methods $GPH_T$ and $GPH_1$. The former has small variability. An introduction of a short-run term may changes the scenario and it is not simple to make one general conclusion of the behavior of estimates. It will depend on the size and the sign of the short-run parameter and, also, in which seasonal component the this is associated. In addition, the empirical properties of the regression estimator will also depend on the bandwidth used in the regression equation (see also the

tables). More complex the model becomes, it will be more difficult to have interpretation of the empirical behavior of the estimates. For all methods, in general, the bias increases with the size of the short-memory coefficients. The asymptotic distribution given in Theorem

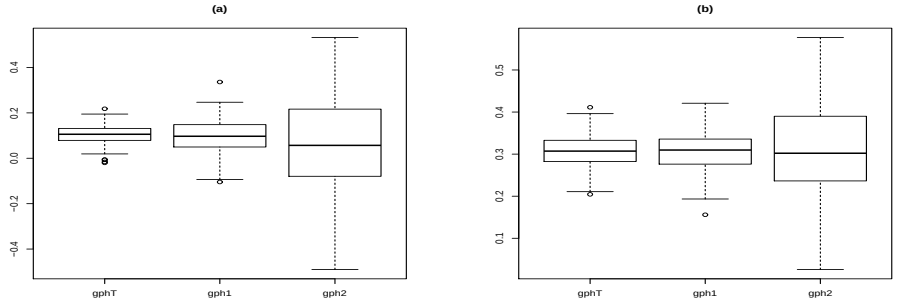



Figura 1: Box-plots of the estimates of $d_1$ (a) and $d_2$ (b) for the SARFIMA model with $d_1 = 0.3(s_1 = 12)$, $d_2 = 0.1(s_2 = 4)$ and $\phi = 0.0$.

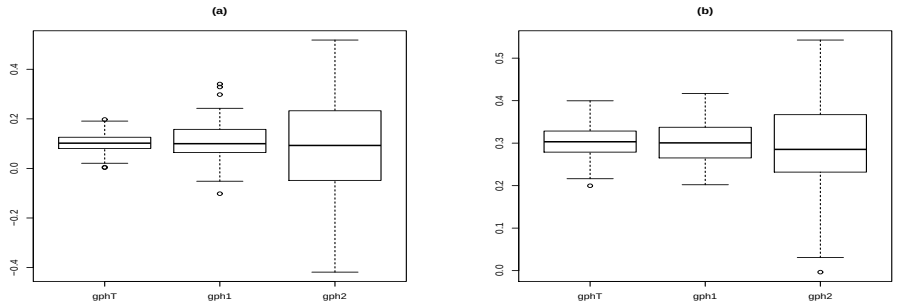



Figura 2: Box-plots of the estimates of $d_1$ (a) and $d_2$ (b) for the SARFIMA model with $d_1 = 0.3(s_1 = 12)$, $d_2 = 0.1(s_2 = 4)$ and $\phi_1 = 0.3$.

2 is also empirically investigated for the model in Table 3 with $\phi = 0$, and the results are depicted in Figure 3 which presents the empirical densities of the standardized GPH estimates of a SARFIMA model. These figures are examples to support the claim given in Theorem 2. The empirical densities of the estimates appear to be fairly close to the density of N(0,1) distribution.

# 5 Examples of Application

This section illustrates and discusses the semiparametric fractional estimator previously discussed using two series observed in Brazil, the daily average $PM_{10}$ concentrations and the hourly electricity demands. Since the full model estimation of the series is beyond of the

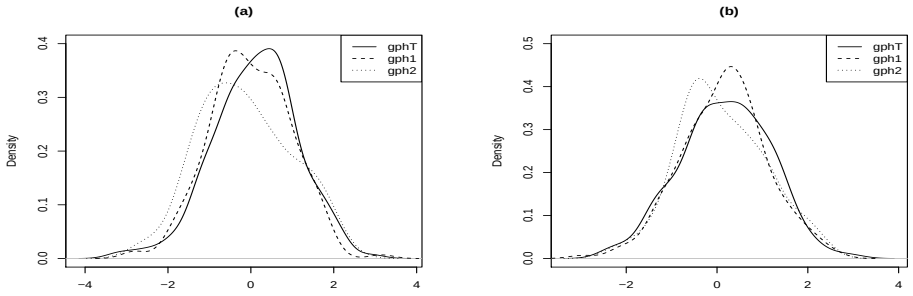



Figura 3: Empirical densities of the standardized GPH estimates of $d_1$ (a) and $d_2$ (b) for the SARFIMA model with $d_1 = 0.3(s_1 = 12)$ and $d_2 = 0.1(s_2 = 4)$ .

scop of this section, the examples are only explored in the context of the seasonal fractional parameter estimation.

## 5.1 Daily average $PM_{10}$ concentration

The daily average Particulate Matter ($PM_{10}$) concentration is expressed in $\mu g/m^3$ and it was observed in the Metropolitan Region of Greater Vitória (RGV) in Brazil. RGV is comprised of five cities with a population of approximately 1.7 million inhabitants in an area of 1,437 $km^2$. The region is situated on the South Atlantic coast of Brazil (latitude 20°19S, longitude 40°20W) and has a tropical humid climate, with average temperatures ranging from $24^oC$ to $30^oC$. The rainfall is fairly distributed throughout the entire year (average precipitation of 98.3 $mm$ per month during the period of study), but with drier periods from June to August (average precipitation of 60.8 $mm$ per month) and more heavier precipitation from October to January (average precipitation of 158.3 $mm$ per month).

The raw series has a sample size of 2037 observations, measured from the 1st of January 2001 to 2nd of August 2006. The sample autocorrelation (ACF) function is shown in Figure 4. From this plot a strong seasonal component in the series is evident, which was an expected property due to the characteristic of such a physical phenomena. It is also observed that the seasonality behavior has period $s = 7$, which is also an expected data behavior since the series was observed daily.

An interesting feature observed from the sample ACF is the slow decay of the correlations in the first lags, in the lags multiple of 7 and in the lags between the seasonal periods. The ACF plot strongly indicates that the process has fractional memory parameters in the lung-run and in the seasonal periods. This empirical evidence indicates the use of a particular case of the SARFIMA model defined previously (Model (7)) with $s_1 = 7$, $s_2 = 1$.

Table 4 displays the results of the memory estimates obtained from different values of the

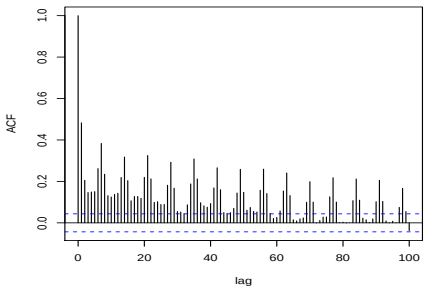



Figura 4: Sample ACF of $PM_{10}$.

bandwidth $m = n^{\alpha}$, $0 < \alpha < 1$ and the values in parenthesis are the corresponding number of the frequency used in the regression. From this table, it can be seen that the values of the estimates of $d_1$ and $d_2$ are stable for $0.52 < \alpha < 0.65(max.)$, and they are in the range $0 < d_1, d_2 < 0.5$. The estimated standard errors of $\hat{d_1}$ are relatively small and two-sided confidence intervals for $d_1$ and $d_2$ are correspondingly tight. Therefore, for $\alpha > 0.52$ the null hypotheses that $H_0 : d_1 = 0$ and $H_0 : d_2 = 0$ are rejected. Also, for all values of the bandwidths given in the table, F test was performed for the null hypothesis $H_0 : \mathbf{d} = \mathbf{0}$, and it indicated that at least one fractional parameter is different from zero. The stable value of the estimate of $d_2$ in the the range $0.52 < \alpha < 0.65(max.)$ gives an empirical evidence that if there is any non-seasonal short-memory part in the model, the parameter is not large enough to make a significant contribution in the regression estimators. A similar conclusion is also observed in the case of the seasonal fractional estimate $\hat{d_1}$. Therefore, $\alpha = 0.54$ was chosen to estimate the memory parameters. The vector $\hat{\mathbf{d}} = (0.1798,\ 0.1918)'$ shows that the model satisfies the stationarity, invertibility and long-memory properties.

Tabela 4: Estimates of the fractional parameters of $PM_{10}$.

| Estim. | $\alpha$ | | | | | |
|---|---|---|---|---|---|---|
| | 0.52 (361) | 0.54 (424) | 0.56 (494) | 0.58 (578) | 0.6 (669) | 0.65(max.) (1016) |
| $\hat{d_1}$ | 0.1954 | 0.1798 | 0.1575 | 0.1443 | 0.1548 | 0.0806 |
| $\hat{d_2}$ | 0.1004 | 0.1918 | 0.1787 | 0.2071 | 0.1645 | 0.2534 |
| Var $\hat{d_1}$ | 0.0016 | 0.0013 | 0.0011 | 0.0009 | 0.0008 | 0.0002 |
| Var $\hat{d_2}$ | 0.0110 | 0.0090 | 0.0072 | 0.0060 | 0.0049 | 0.0014 |

## 5.2 Hourly electricity demand

The electricity series corresponds to 30 weeks of hourly observations for electricity demand in Rio de Janeiro (Brazil), from Sunday 5th of May to 30th of November 1996. This data set was used by Taylor et al. (2006) with the main to compare univariate methods for forecasting electricity demand up to a day ahead. Among the models they fitted to the data, the double seasonal $ARIMA(3,0,3)(3,0,3)_{s_1}(3,0,3)_{s_2}$ model was chosen with $s_1 = 168$ to model the within-day seasonal cycle of 24 hours and $s_2 = 24$ to model the within-week cycle of 168 hours. Differently of the model property that they suggested, the sample autocorrelation function (Figure 5) appears to decay hyperbolically at seasonal frequencies indicating long-memory property close or inside the non-stationary region. Such phenomenon in the data was not considered by the authors. Hence, the series is here investigated by allowing that $s_1 = 168$ and $s_2 = 24$ are seasonal periods with fractional differencing parameters.

For this purpose, we consider two possibilities: Model I has fractional parameters at seasonal frequencies only in the form $(1 - B^{s_1})^{d_1}(1 - B^{s_2})^{d_2}$ (Table 5); and Model II also includes one non-seasonal fractional parameter with the fractional difference operator $(1-B)^d$ (Table 6). Although the focus of the paper is on the SARFIMA model with two fractional parameters, the estimated model displayed in Table 6 is a simple example that, besides its model adequacy, shows that the proposed estimation method can be easily extended to classes of SARFIMA models with a number of parameters memory larger than two.

The results displayed in Tables 5 and 6 indicate that all fractional estimates strongly suggested that differencing fractional seasonal parameters should be also considered in the model. In addition, the results also indicated that the series has non-stationary long-memory phenomenon since the estimate of the total memory (the sum of the memory parameters) at the zero frequency and at the seasonal periods are bigger than 0.5 (see Proposition 1). Although the paper has discussed the theory in the stationary case only, the estimation procedure can also be applied in the non-stationary region. This can be justified by the fact that the estimation of the standard ARFIMA models, in the stationary and non-stationary cases, is today, theoretically established and empirically justified in the literature of long-memory time series.

The estimate of the fractional parameter in Model II at zero frequency seems to be significative, however, the fractional seasonal estimates in Models I and II are quite similar. This preliminary model estimation suggests that SARFIMA model with fractional differencing parameters, the model discussed in this paper, could be also used as an alternative univariate method to modeling and forecasting the hourly electricity demands. By using the fractional degrees of seasonal differencing parameter at different seasonal lags, the seasonal fluctuation can be removed while avoiding over-differencing. In addition, if the underling series has strongly persistence, the use of standard seasonal ARMA model will yield poorer forecast, specially for long-term prediction (see, for example, Reisen & Lopes (1999) and Man (2003).

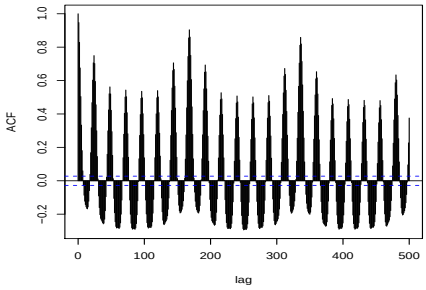



Figura 5: The ACF function of the hourly electricity demands.

Tabela 5: Estimates of the fractional parameters of Model I.

| Estimates | $\alpha$ | | | |
|---|---|---|---|---|
| | 0.26 (1577) | 0.28 (1743) | 0.3 (2075) | 0.317 (2407) |
| $\hat{d_1}$, s=168 | 0.0884 | 0.0890 | 0.1144 | 0.1136 |
| $\hat{d_2}$, s=24 | 0.7848 | 0.7744 | 0.7093 | 0.6354 |
| Var $\hat{d_1}$ | 0.0006 | 0.0005 | 0.0005 | 0.0004 |
| Var $\hat{d_2}$ | 0.0035 | 0.0029 | 0.0022 | 0.0017 |

Tabela 6: Estimates of the fractional parameters of Model II

| Estimates | $\alpha$ | | | |
|---|---|---|---|---|
| | 0.26 (1577) | 0.28 (1743) | 0.3 (2075) | 0.317 (2407) |
| $\hat{d_1}, s = 168$ | 0.0884 | 0.0890 | 0.1142 | 0.1133 |
| $\hat{d_2}, s = 24$ | 0.7803 | 0.7665 | 0.6975 | 0.6265 |
| $\hat{d}, s = 1$ | 0.1369 | 0.2466 | 0.3822 | 0.3024 |
| Var $\hat{d_1}$ | 0.0006 | 0.0005 | 0.0005 | 0.0004 |
| Var $\hat{d_2}$ | 0.0036 | 0.0029 | 0.0023 | 0.0017 |
| Var $\hat{d}$ | 0.0888 | 0.0736 | 0.0587 | 0.0453 |

# 6 Conclusions

The paper deals with the seasonal ARFIMA model with two seasonal fractional parameters. Properties and model estimation are discussed. To estimate the parameters, a multilinear

regression method is suggested. The Monte Carlo experiment evidenced that, in general, the estimator presented very good accuracy for sample size equal to 1080. The method is very easy to be implemented and does not require sophisticated computer capacities. The usefulness of the SARFIMA model and the semiparametric fractional estimator was exemplified using daily average $PM_{10}$ concentration and electricity demand series.

## Acknowledgements

V. A. Reisen and B. Zamprogno gratefully acknowledge partial financial support from CAPES, PIBIC-UFES, FAPES-ES and CNPq/Brazil. W. Palma was partially supported by Fondecyt grant number 1085239 to 1120758.. J. Arteche acknowledges financial support from the Spanish Ministerio de Ciencia y Tecnología and ERDF grant SEJ2007-61362/ECON. The authors thanks L. M de Meneses for providing the series of hourly demand. The authors also thanks the referees for helpful comments.

## APPENDIX A

**Proof of Theorem 1:** The asymptotic mean and variance of $\hat{\mathbf{d}}$ follows as in Theorem 1 in Hurvich et al. (1998). Denote $Z$ the $\sum_k m_k \times 2$ matrix of regressors in (9) and similarly the vector $V$ for the disturbances. Then

$$\hat{\mathbf{d}} - \mathbf{d} = (Z'Z)^{-1}Z'V.$$

Note now that

$$\begin{aligned} 2\log X_{1,k,j} &= 2\log\left\{2\sin\left(\pi k + \frac{\pi j s'}{n}\right)\right\} = 2\log|\lambda_{js'}| + O(|\lambda_{js'}|^2) \\ 2\log X_{2,k,j} &= 2\log\left\{2\sin\left(\frac{\pi k s_2}{s'} + \frac{\pi j s_2}{n}\right)\right\} = 2\log|\lambda_{js_2}| + O(|\lambda_{js'}|^2) \end{aligned}$$

if $k \in I_k$ and

$$2\log X_{2,k,j} = 2\log\left\{2\sin\left(\frac{\pi k s_2}{s'}\right)\left[1 + \frac{\cos(\pi k s_2/s')}{\sin(\pi k s_2/s')}\frac{\pi|j|s_2}{n} + O(\lambda_{js_2}^2)\right]\right\} \qquad (13)$$

if $k \notin I_k$. Then

$$\begin{aligned}
\sum_{k=0}^{[s'/2]} \sum_{j}^{*} Z_{1,k,j}^2 &= 4 \sum_{k=0}^{[s'/2]} m_k(1+o(1)) \\
\sum_{k=0}^{[s'/2]} \sum_{j}^{*} Z_{2,k,j}^2 &= 4 \sum_{k\in I_k} m_k(1+o(1)) \\
\sum_{k=0}^{[s'/2]} \sum_{j}^{*} Z_{1,k,j} Z_{2,k,j} &= 4 \sum_{k\in I_k} m_k(1+o(1))
\end{aligned}$$

This result follows from the fact that for those $k \notin I_k$, $\sum_j^* Z_{2,k,j}^2 = O(m^3 n^{-2}) = o(m)$ and $\sum_j^* Z_{2,k,j} Z_{1,k,j} = O(m^2 n^{-1} \log m) = o(m)$. Then

$$Z'Z = mQ(1+o(1)) \tag{14}$$

Denoting now $\varepsilon$ the vector with elements $\varepsilon_{k,j}$ we have that

$$Z'\varepsilon = mb_n(1+o(1)) \tag{15}$$

since

$$\begin{aligned}
\sum_{k=0}^{[s'/2]} \sum_{j}^{*} Z_{1,k,j}\varepsilon_{k,j} &= -2m \sum_{k=0}^{[s'/2]} b_k \delta_k (2\pi)^{\alpha_k} \frac{\alpha_k}{(\alpha_k+1)^2} \left(\frac{m}{n}\right)^{\alpha_k} \left(1 + O\left[\log m \left(\frac{m}{n}\right)^{\iota}\right]\right) \\
\sum_{k=0}^{[s'/2]} \sum_{j}^{*} Z_{2,k,j}\varepsilon_{k,j} &= -2m \sum_{k\in I_k} b_k \delta_k (2\pi)^{\alpha_k} \frac{\alpha_k}{(\alpha_k+1)^2} \left(\frac{m}{n}\right)^{\alpha_k} \left(1 + O\left[\log m \left(\frac{m}{n}\right)^{\iota}\right]\right)
\end{aligned}$$

because for $k \notin I_k$

$$\sum_{k\notin I_k} \sum_{j}^{*} Z_{2,k,j}\varepsilon_{k,j} = O\left(\sum_{j}^{*} |\lambda_j|^{\alpha_k+1}\right) = o\left(m\left[\frac{m}{n}\right]^{\alpha_k}\right)$$

The rest of the proof follows as in Hurvich et al. (1998).

**Proof of Theorem 2:** The proof follows as in Hurvich et al. (1998) applying Lemma 4 in Sun & Phillips (2003) which holds in our multiple log periodogram regression context. Since

$$\sqrt{m}(\hat{\mathbf{d}} - \mathbf{d}) = (m^{-1}Z'Z)^{-1} m^{-1/2} Z'\varepsilon + (m^{-1}Z'Z)^{-1} m^{-1/2} Z'U$$

by using (14) and (15) it only remains to show that $m^{-1/2}v'Z'U \xrightarrow{d} N(0, \pi^2 v'Qv/6)$ for any vector $v = (v_1, v_2)$. As in Hurvich et al. (1998)

$$\frac{1}{\sqrt{m}}v'Z'U = o_p(1) + \frac{1}{\sqrt{m}}\sum_{k=0}^{[s'/2]}{\sum_{|j|>m^{0.5+\delta}}}^{*} g_{k,j}U_{k,j}$$

for some $0.5 > \delta > 0$ and $g_{k,j} = v_1 Z_{1,k,j} + v_2 Z_{2,k,j}$. Now $\max_{j,k}|g_{k,j}| = O(\log m)$ and $\sum^*_{|j|>m^{0.5+\delta}} |g_{k,j}|^p = O(m)$ for all $p \geq 1$ (see formula (A18) in Hurvich et al. (1998) for $Z_{1,k,j}$ and $Z_{2,k,j}$ for $ks_2 \in I_k$ and use (5) for $ks_2 \notin I_k$). Since by equation (14) $\sum^*_{|j|>m^{0.5+\delta}} g^2_{k,j} = mv'Qv(1 + o(1))$, we can apply Lemma 4 in Sun & Phillips (2003) to get the desired result.

## Referências


Arteche, J. (2002), 'Semiparametric robust tests on seasonal or cyclical long memory time series', *J. Time Series Anal.* **23**(3), 251–285.

Arteche, J. & Robinson, P. (1999), *Seasonal and cyclical long memory. In Asymptotics Nonparametrics and Time Series (ed. S. Ghosh)*, Marcel Dekke, Inc., New York.

Arteche, J. & Robinson, P. M. (2000), 'Semiparametric inference in seasonal and cyclical long memory processes', *J. Time Series Anal.* **21**(1), 1–25.

Arteche, J. & Velasco, C. (2005), 'Trimming and tapering semi-parametric estimates in asymmetric long memory time series', *J. Time Series Anal.* **26**(4), 581–611.

Ferrara, L. & Guegan, D. (2006), 'Fractional seasonality: models and applications to economic activity in the euro area', *Working Paper, Center d´Economie de la Sorbonne,Paris* .

Franco, G. & Reisen, V. (2007), 'Bootstrap approaches and confidence intervals for stationary and non-stationary long-range dependence processes', *Phys. A.* **375**, 546–562.

Geweke, J. & Porter-Hudak, S. (1983), 'The estimation and application of long memory time series models', *J. Time Series Anal.* **4**(4), 221–238.

Gil-Alana, L. (2001), 'Testing stochastic cycles in macroeconomic time series', *J. Time Series Anal.* **22**(4), 411–430.

Giraitis, L. & Leipus, R. (1995), 'A generalized fractionally differencing approach in long-memory modeling', *Lith. Math. J.* **35**(1), 53–65.

Gould, P., Koehler, A., Ord, J., Snyder, R., Hyndman, R. & Vahid-Araghi, F. (2008), 'Forecasting time series with multiple seasonal patterns', *European J. Oper. Res.* **191**(1), 207–222.

Gray, H., Zhang, N. & Woodward, W. (1989), 'On generalized fractional processes', *J. Time Series Anal.* **10**(3), 233–257.

Gray, H., Zhang, N. & Woodward, W. (1994), 'Correction to on generalized fractional processes', *J. Time Series Anal.* **15**(5), 561–562.

Hassler, U. (1993), 'Regression of spectral estimators with fractionally integrated time series', *J. Time Series Anal.* **14**(4), 369–380.

Hassler, U. (1994), '(mis)specification of long memory in seasonal time series', *J. Time Series Anal.* **15**(1), 19–30.

Hosking, J. (1981), 'Fractional differencing', *Biometrika* **68**(1), 165–176.

Hosking, J. (1984), 'Modeling persistence in hydrological time series using fractional differencing', *Water Resour Res* **20**(12), 1898–1908.

Hurvich, C., Deo, R. & Brodsky, J. (1998), 'The mean squared error of geweke and porter-hudak's estimator of the memory parameter of a long-memory time series', *J. Time Series Anal.* **19**(1), 19–46.

Leipus, R. & Viano, M. (2000), 'Modelling long-memory time series with finite or infinite variance: a general approach', *J. Time Series Anal.* **21**(1), 61–74.

Man, K. (2003), 'Long memory time series and short term forecasts', *Int. J. Forecasting* **19**(3), 477–491.

Marques, G. (2011), 'Empirical aspects of the whittle-based maximum likelihood method in jointly estimating seasonal and non-seasonal fractional integration parameters', *Phys. A.* **390**(1), 8–17.

Ooms, M. & Franses, P. (2001), 'A seasonal periodic long memory model for monthly river flows', *Environ Modell Softw* **16**(6), 559–569.

Palma, W. (2007), *Long-memory time series: theory and methods*, John Wiley & Sons, New Jersey.

Palma, W. & Chan, N. (2005), 'Efficient estimation of seasonal long-range-dependent processes', *J. Time Series Anal.* **2**(6), 863–892.

Porter-Hudak, S. (1990), 'An application of the seasonal fractionally differenced model to the monetary aggregates', *J. Amer. Statist. Assoc.* **85**(410), 338–344.

Ray, B. (1993), 'Long-range forecasting of ibm product revenues using a seasonal fractionally differenced arma model', *Int. J. Forecasting* **9**(2), 255–269.

Reisen, V. & Lopes, S. (1999), 'Some simulations and applications of forecasting long-memory time-series models', *J. Statist. Plann. Inference* **80**(1-2), 269–287.

Reisen, V., Moulines, E., Soulier, P. & Franco, G. (2010), 'On the properties of the periodogram of a stationary long-memory process over different epochs with applications', *J. Time Series Anal.* **31**(1), 20–36.

Reisen, V., Rodrigues, A. & Palma, W. (2006*a*), 'Estimating seasonal long-memory processes: a monte carlo study', *J. Stat. Comput. Simul.* **76**(4), 305–316.

Reisen, V., Rodrigues, A. & Palma, W. (2006*b*), 'Estimation of seasonal fractionally integrated processes', *Comput Stat Data An* **50**(2), 568–582.

Robinson, P. (1995), 'Log-periodogram regression of time series with long range dependence', *Ann. Statist.* **23**(3), 1048–1072.

Sun, Y. & Phillips, P. (2003), 'Nonlinear log-periodogram regression for perturbed fractional processes', *J. Econometrics* **115**(2), 355–389.

Taylor, J. (2010), 'Triple seasonal methods for short-term electricity demand forecasting', *European J. Oper. Res.* **204**(1), 139–152.

Taylor, J., De Menezes, L. & McSharry, P. (2006), 'A comparison of univariate methods for forecasting electricity demand up to a day ahead', *Int. J. Forecasting* **22**(1), 1–16.

Viano, M., Deniau, C. & Oppenheim, G. (1995), 'Long-range dependence and mixing for discrete time fractionary processes', *J. Time Series Anal.* **16**(3), 323–338.

Woodward, W., Cheng, Q. & Gray, H. (1998), 'A k-factor garma long-memory model', *J. Time Series Anal.* **19**(4), 485–504.

# 5 Discussão Geral

As pesquisas que geraram os resultados desta tese foram motivadas pela aplicação da análise de componentes principais em diferentes contextos da área poluição do ar, em especial, no uso da técnica no gerenciamento de rede de monitoramento da poluição e, em particular, na aplicação em séries observadas na rede da Região da Grande Vitória (RGV). Essa metodologia, em termos práticos, produz informações com precisões estatística para análise e inferência de dados multivariados e, no contexto desta tese, para tomada de decisões importantes no gerenciamento de rede de monitoramento de controle e da qualidade do ar, entre outras vertentes de aplicação, como por exemplo, no estudo da relação entre poluentes e os efeitos deletérios à saude.

A inferência estatística obtida por meio da técnica ACP basea-se no pressuposto de variáveis independentes, geradas por uma distribuição normal multivariada, propriedades praticamente não observadas em muitas áreas de aplicação, em especial, em variáveis da poluição do ar. Os resultados desta pesquisa mostram que as análises oriundas da técnica aplicada em processos multivariados podem levar em interpretações espúrias. Por exemplo, autocorrelações fortemente positivas das variáveis dos poluentes atmosféricos podem acarretar no subdimensionamento da rede de monitoramento da poluição do ar.

A tese está dividida em artigos e cada um trata particularidades do uso de ACP em séries temporais com aplicação à poluição do ar. O primeiro artigo apresenta, no domínio do tempo, fundamentação teórica e empírica da técnica ACP e suas propriedades fundamentais que devem ser obedecidas em situações práticas, tais como análise da estrutura de dependência e garantia da estacionariedade para existência da matriz de covariância. Em termos teóricos foi mostrado que as componentes principais oriundas do método usual de ACP são autocorrelacionadas e podem apresentar correlação cruzada significativa nos *lags* diferentes de zero. Nesse caso o uso de componentes principais correlacionadas em modelos de regressão pode provocar efeito distorcido nas estimativas dos coeficientes, isto é, ocorrer uma inflação das variâncias dos estimadores e essas propriedades não desejadas podem inviabilizar a aplicação dessa técnica nesse contexto. Essa problemática é exemplificada por meio do estudo da associação entre a exposição atmosférica dos poluentes $PM_{10}$, $SO_2$, $NO_2$, $O_3$ e CO e o número de atendimentos por doenças respiratórias, em crianças menores de 6 anos, na região da Grande Vitória, ES, Brasil.

De forma empírica, verificou-se que a estrutura de correlação presente nas variáveis altera as estimativas dos autovalores e, consequentemente, os percentuais de explicação dessas quantidades para cada componente. Isso, em termos práticos, modifica resultados e conclusões de diversos problemas que fazem o uso dessa técnica. O foco central dessa problemática se concentrou nos estudos de identificação de fonte poluidora e do gerenciamento da rede de monitoramento. Nesse contexto, as análises avaliadas com essa metodologia fazem uso dos coeficientes (autovetores) das componentes principais e de suas variabilidades (percentual dos autovalores) e, portanto, séries com propriedades de forte autocorrelação podem levar em conclusões equivocadas. Quando a estrutura de autocorrelação é fraca, o efeito da autocorrelação é praticamente nulo, dessa forma a técnica pode ser empregada sem maiores problemas.

Para minimizar o efeito da autocorrelação nas estimativas das componentes principais, o artigo propõe, sob algumas condições de regularidade, o uso do filtro linear VARMA. Como mostrado empiricamente, sob a ordem correta do modelo, a correlação cruzada entre componentes é eliminada e, consequentemente, os percentuais de explicação e as estimativas das

componentes não apresentam o efeito temporal gerado pela estrutura da séries originais. Em adição, o uso desse procedimento evidenciou análises mais compreensiva e precisas na resolução dos problemas investigados relacionados à poluição do ar. Outro fator importante do processo de filtragem é que torna possível utilizar ACP em processos com raiz unitária, isto é, processos com matriz de covariância não definida.

Em geral, análise de sinais, mais especificamente de séries temporais, pode ser feita tanto no domínio do tempo quanto no da frequência. A relação matemática que une essas duas áreas de séries temporais é baseada na integral de Fourier-Stieljes, onde a função de autocovariancia do processo pode ser escrita por meio da transformada de Fourier da função espectral (ver, por exemplo, Priestley (1983)).

Não diferentemente da teoria e de aplicações de modelos de séries temporais baseadas na análise espectral de processos, isto é, teoria do domínio da frequência, as inferências e interpretações obtidas por meio da técnica ACP, em séries temporais multivariadas, podem ser também elegantemente obtidas em função dessa vertente matemática. A técnica ACP, no domínio da frequência, é também explorada em diversos estudos, por exemplo, Brillinger (1969), Stoffer et al. (1993) e Boudou & Viguier-Pla (2006) entre outros, mas ainda não devidamente considerada na área da poluição do ar. Possíveis justificativas da pouca popularidade de estudos com essa ferramenta na análise e interpretação das componentes principais em varáveis da poluição do ar podem ser devido a dificuldade matemática da metodologia e a falta de programas computacionais específicos para a técnica. Entretanto, essas possíveis barreiras do uso da técnica devem ser ultrapassadas pois, como é amplamente conhecido na literatura de séries temporais, a contribuição da análise espectral nos estudos de diferentes áreas têm um papel importantíssimo devido à elegância e a riqueza matemática, das interpretações, das análises físicas e das inferências estatísticas obtidas por meio dessa ferramenta. Ressalta-se que a técnica ACP, no domínio da frequência, complementa as aplicações oriundas da teoria de séries temporais no domínio do tempo, o que torna possível com a técnica identificar ciclos e análises em frequências específicas. No entanto, a mesma não supre a análise no domínio do tempo e, portanto, um pararelo interessante a considerar.

A vertente de pesquisa do uso da ACP no domínio da frequência é a motivação do Artigo 2 desta tese. O resultados avaliam, de forma empírica e aplicada, o uso da técnica ACP e suas aplicações em séries autocorrelacionadas. A metodologia considerada é baseada na proposta de Brillinger (1969, cap. 9), mas além de abordar processos de memória curta (tipo modelos ARMA) também invoca processos de memória longa (modelos ARFIMA). Em termos de aplicação, a metodologia da ACP, no domínio da frequência, é considerada no problema de gerenciamento de rede de monitoramento do ar e comparada com a técnica estudada no Artigo 1, isto é, ACP no domínio do tempo. Em geral, os resultados simulados mostraram que a metodologia de Brillinger pode ser estendida para processos com memória longa e, assim, ser empregada nos dados de concentrações de poluentes monitorados pela rede de monitoramento da qualidade do ar da Região da Grande Vitória, variáveis identificadas com o fenõmeno de memória longa (ver aplicação do quarto artigo desta tese). Nesse contexto, em termos práticos, foi verificado que é possível reduzir de oito para duas o número de estações de monitoramento do poluente $PM_{10}$. No entanto, apesar de grande parte de variabilidade ficar concentrada em duas componentes, essa redução não levou a melhores resultados do que aqueles encontrados com ACP, no domínio do tempo, com filtro adequado.

Uma outra problemática dos dados da concentrações de poluentes do ar são os valores atípicos. Esses valores podem ocorrer durante pelo menos um dia da semana e mais frequentemente, em certos horários, e podem ser caracterizados como nível de poluição que fica exageradamente distante dos níveis normais, que não necessariamente estão fora dos padrões definidos em resoluções. Como uma contribuição de aplicação em gerenciamento de rede de monitoramento da qualidade do ar, o terceiro artigo desta tese invoca a questão da presença de

*outliers*, valores que podem ocasionar sérios problemas na estimação da matriz de covariância e, assim, modificar as estimativas dos autovalores, dos autovetores e, consequentemente, das componentes principais. O objetivo dessa aplicação foi analisar os efeitos de dados atípicos na análise e interpretação da técnica e os resultados obtidos indicaram uma leve vantagem do uso de ACP robusto, baseado no domínio do tempo, sobre a usual metodologia discutida no primeiro Artigo.

Como o foco principal desta tese é a técnica multivariada ACP em séries temporais com diferente estruturas de memória aplicado em problemas atmosféricos, o Artigo 4 apresenta de forma precisa conceitos de séries temporais sazonais com estrutura de memoria curta e longa (modelo SARFIMA), as propriedades e a estimação dos parâmetros. Além das fundamentações teóricas do modelo proposto e do método de estimação dos parâmetros, o artigo apresenta estudo simulado para evidenciar as propriedades empíricas do estimador dos parâmetros fracionários (parâmetros de memória longa). Como já observado na literatura, varíaveis da poluição do ar mostram bastantes evidências de apresentarem propriedades do modelo SARFIMA. Assim, como forma do corroborar os estudos téoricos e empíricos propostos no artigo, dois exemplos reais foram considerados, em especial, série de concentrações da poluição do ar observada na Região da Grande Vitória. Como esperado, as estimativas evidenciaram nesse exemplo, de forma significativa, a presença de memória longa em períodos sazonais.

# 6 Conclusão e trabalhos futuros

Esta tese além de atingir os objetivos, geral e específicos, descritos na Seção 2, promoveu, por meio de análise componentes principais (ACP) e de derivações, novas vertentes de pesquisa referentes a análise e a interpretação de fenômenos da poluição do ar, esses observados pelas variáveis da poluição, entre outras. As descrições dos principais resultados da tese e as propostas para futuros trabalhos estão sumarizados nos próximos parágrafos.

O estudo, justificado de forma teórica, empírica e aplicada, permitiu avançar de forma significativa na compreensão, na interpretação, na inferência entre outras questões relacionadas ao uso da técnica ACP, no domínio do tempo e da frequência, e suas aplicações na poluição do ar, no contexto de séries temporais multivariadas. O problema central do objetivo desta pesquisa está relacionado no efeito das estruturas de autocorrelação de séries temporais nas aplicações da ACP, em especial, no gerenciamento de rede de monitoramento do ar, na identificação de fontes poluidoras, no estudo entre relações lineares e não lineraes de variáveis (associação da concentração de poluentes ao número de internações por problemas respiratórios), na análise de método robusto para esses problemas, entre outros.

A investigação mostrou que negligenciar a estrutura temporal das observações pode acarretar em análises e interpretações totalmente equivocadas ou, até mesmo, inviabilizar o cálculo das componentes quando o processo vetorial não é estacionário. Para a práxis dessas questões, foram utilizadas séries coletadas na rede de monitoramento da Região da Grande Vitória (RGV).

No domínio do tempo, em especial, observou-se que a ACP, em séries temporais, deve ser utilizada sob certa cautela, pois em estruturas de forte dependência (forte autocorrelações), essa metodologia pode levar a resultados espúrios, como por exemplo, pode provocar um subdimensionamento da rede de monitoramento do ar. Nessa direção, os estudos mostraram que as componentes principais são correlacionadas. Se o processo tiver uma matriz de correlação fraca, a técnica ACP pode ser utilizada sem grandes problemas de interpretação, de inferência e de teste. Nesse sentido, em termos práticos, a pesquisa propõe um procedimento de teste estatístico para verificar se a estrutura temporal das séries são forte o suficiente para que possa causar análise e interpretação espúrias, por meio das componentes principais.

Como forma alternativa de usar a ACP no domínio do tempo, a tese sugere o procedimento de filtros lineares para reduzir ou eliminar a problemática descrita no parágrafo anterior, isto é, remover a estrutura temporal. Dessa maneira, a técnica é aplicada e suas interpretações seguem de acordo com teoria clássica da ACP baseada em observações independentes.

No domínio da frequência, o emprego do método sugerido em Brillinger (1983, Cap. 9) foi avaliado em processos com memória longa. Diante das dificuldades de interpretar e lidar com a metodologia, a aplicação ficou restrita ao problema de gerenciamento de rede. Entretanto, essa técnica tem riquezas matemáticas que podem ser melhor exploradas.

Um aspecto importante observado nas concentrações dos poluentes atmosféricos da Região da Grande Vitória foi que a tendência do processo pode ser capturada por um processo de memória longa. Portanto, ACP em modelos com memória longa (SARFIMA) também foi objeto de estudo nas questões abordas acima.

Como tópicos para trabalhos futuros, sabe-se que os dados da poluição do ar apresentam outras complexidades além das consideradas. Os *outliers* sugerem o uso de métodos robustos os quais podem ser explorados em ambos domínios, principalmente, no da frequência. A volatilidade e a não normalidade (gaussianidade) também são propriedades estocásticas comuns nesses dados. Dessa forma, estudar métodos não paramétricos de ACP pode ser uma alternativa interessante a ser considerada.

Outro ponto a explorar é a técnica *bootstrap* no intuito de calcular autovalores e realizar testes de hipóteses. Então, essas questões são linhas de estudos a serem investigadas em

virtude das complexidades e interpretação do fenômeno em questão, isto é, poluição do ar.

Além das vertentes de pesquisa mencionadas no parágrafo anterior, as contribuições desta tese são alicerce para outras metodologias que utilizam ACP, tal como análise fatorial, método amplamente empregado em identificação de fontes poluidoras, entre outros problemas relacionados com a análise e a interpretação das variáveis da poluição do ar e correlatas.

## Referências


Ahn, H. & James, R. T. (1999), 'Outlier detection in phosphorus dry deposition rates measured in South Florida', *Atmospheric Environment* **33**, 5123–5131.

Belis, C. A., Karagulian, F., Larsen, B. & Hopke, P. K. (2013), 'Critical review and meta-analysis of ambient particulate matter source apportionment using receptor models in Europe', *Atmospheric Environment* **69**, 94–108.

Boudou, A. & Viguier-Pla, S. (2006), 'On proximity between PCA in the frequency domain and usual PCA', *Statistics* **40**(5), 447–464.

Brillinger, D. R. (1969), The canonical analysis of stationary time series, *in* P. R. Krishnaiah, ed., 'Multivariate analysis', Vol. II, New York: Academic Press, pp. 311–350.

Brillinger, D. R. (1983), *Time series: data analysis and theory*, Holden Day.

Cohen, S. J. (1983), 'Classification of 500 mb height anomalies using obliquely rotated principal components', *J. Climate Appl. Meteorol.* **22**, 1975–1988.

Cosemans, G., Kretzschmar, J. & Mensink, C. (2008), 'Pollutant roses for daily averaged ambient air pollutant concentrations', *Atmospheric Environment* **42**, 6982–6991.

Croux, C. & Haesbroeck, G. (2000), 'Principal component analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies', *Biometrika* **87**, 603–618.

Ezcurra, A., Sáenz, J., Ibarra-Berastegi, G. & Areitio, J. (2008), 'Electrical storm rainfall yield characteristics observed in the Spanish Basque Country area during the period 1992-1996', *Atmospheric Research* **89**, 233–242.

Furrer, R. (2005), 'Covariance estimation under spacial dependence', *Journal of Multivariate Analysis* **94**, 366–381.

Guo, H., Wang, T., Simpson, I., Blake, D., Yu, X., Kwok, Y. & Li, Y. (2004), 'Source contributions to ambient vocs and co at a rural site in eastern China', *Atmospheric Environment* **38**, 4551–4560.

Ibarra-Berastegi, G., Elias, A., Barona, A., Saenz, J., Ezcurra, A. & Diaz de Argandonã, J. (2008), 'From diagnosis to prognosis for forecasting air pollution using neural networks: air pollution monitoring in Bilbao', *Environmental Modelling and Software* **23**, 622–637.

Johnson, R. A. & Wichern, D. W. (1998), *Applied Multivariate Statistical Analysis*, 6th edn, Prentice Hall.

Jollife, I. T. (2002), *Principal component analysis*, 2th edn, Prentice Hall.

Jorquera, H., Palma, W. & Tapia, J. (2000), 'An intervention analysis of air quality data at Santiago, Chile', *Atmospheric Environment* **34**, 4073–4084.

Juneng, L., Latif, M. T., Tangang, F. T. & Mansor, H. (2009), 'Spatio-temporal characteristics of $PM_{10}$ concentration across Malaysia', *Atmospheric Research* **43**, 4584–4594.

Karar, K. & Gupta, A. (2007), 'Source apportionment of $PM_{10}$ at residential and industrial sites of an urban region of Kolkata, India', *Atmospheric Research* **84**, 30–41.

Keller, M. (2000), Hauptkomponentenanalyse für intensivmedizinische Zeitreihen (in German), PhD thesis, Department of Statistics, University of Dortmund, Germany.

Lau, J., Hung, H. & Cheung, C. (2009), 'Interpretation of air quality in relation to monitoring station's surroundings', *Atmospheric Environment* **43**, 769–777.

Lehman, J., Swinton, K., Bortnick, S., Hamilton, C., Baldridge, E., Eder, B. & Cox, B. (2004), 'Spatio-temporal characterization of tropospheric ozone across the eastern United States', *Atmospheric Environment* **38**, 4357–4369.

Liu, P.-W. G. (2009), 'Simulation of the daily average $PM_{10}$ concentrations at Ta-Liao with Box-Jenkins time series models and multivariate analysis', *Atmospheric Environment* **43**, 2101–2113.

Lowell, L., Leonard, O. & Flocchini, R. (1984), 'A principal component analysis of sulphur concentrations in the western United States', *Atmospheric Environment* **18**, 783–791.

McGregor, G. R. (1996), 'Identification of air quality affinity areas in Birmingham (UK)', *Applied Geography* **16**, 109–122.

Oanh, N., Chutimon, P., Ekbordin, W. & Supat, W. (2005), 'Meteorological pattern classification and application for forecasting air pollution episode potential in a mountain-valley area', *Atmospheric Environment* **39**, 1211–1225.

Palma, W. (2007), *Long-Memory Time Series: Theory and Methods*, Wiley-Interscience.

Pires, J. C. M., Pereira, M. C., Alvim-Ferraz, M. C. M. & Martins, F. G. (2009), 'Identification of redundant air quality measurements through the use of principal component analysis', *Atmospheric Environment* **43**, 3837–3842.

Pires, J. C. M., Souza, S. I. V., Pereira, M. C., Alvim-Ferraz, M. C. M. & Martins, F. G. (2008*a*), 'Management of air quality monitoring using principal component and cluster analysis — part I: $SO_2$ and $PM_{10}$', *Atmospheric Environment* **42**, 1249–1260.

Pires, J. C. M., Souza, S. I. V., Pereira, M. C., Alvim-Ferraz, M. C. M. & Martins, F. G. (2008*b*), 'Management of air quality monitoring using principal component and cluster analysis — part II: CO, $NO_2$ and $O_3$', *Atmospheric Environment* **42**, 1261–1274.

Priestley, M. B. (1983), *Spectral Analysis and Time Series, Vol. I e II*, Academic Press.

Reisen, V. A., Zamprogno, B., Palma, W. & Arteche, J. (2010), 'Seasonal fractional long-memory processes. A semiparametric estimation approach', *arXiv:1011.5631* .

Richman, M. B. (1986), 'A principal component analysis of sulphur concentrations in the western United States', *Atmospheric Environment* **20**, 606–607.

Romero, R., Ramis, C., Guijarro, J. A. & Sumner, G. (1999), 'Daily rainfall affinity areas in Mediterranean Spain', *Int. J. Climatol* **19**, 557–578.

Sanchez, M. L., Casanova, J., Ramos, M. & Sanchez, J. (1996), 'Studying the spatial and temporal distribution of $SO_2$ in an urban area by principal component factor analysis', *Atmospheric Research* **20**, 53–65.

Shi, G.-L., Li, X., Yin-Chang, F., Wang, Y.-Q., Wu, J.-H., Jun, L. & Tan, Z. (2009), 'Combined source apportionment, using positive matrix factorization-chemical mass balance and principal component analysis/multiple linear regression-chemical mass balance models', *Atmospheric Environment* **43**, 2929–2937.

Soares, I. P. (2011), Avaliação do uso de diferentes modelos receptores para determinação da contribuição das fontes de partículas totais em suspensão, Master's thesis, Departamento de Engenharia Ambiental, Universidade Federal do Espírito Santo, Vitória.

Souza, J. B. (2013), Análise de componentes principais e a modelagem linear generalizada: uma associação entre o número de atendimentos hospitalares por causas respiratórias e a qualidade do ar, na região da grande vitória, es, Master's thesis, Departamento de Engenharia Ambiental, Universidade Federal do Espírito Santo, Vitória.

Statheropoulos, M., Vassiliadis, N. & Pappa, A. (1998), 'Principal component and canonical correlation analysis for examining air pollution and meteorological data', *Atmospheric Environment* **32**(6), 1087–1095.

Stefaniak, I. (2009), Multivariate statistical process control applications for autocorrelated data, Master's thesis, Dissertação (Mestrado em Engenharia), Technical University of Denmark, Kongens Lyngby, Denmark.

Stoffer, D. S., Tyler, D. E. & McDougall, A. J. (1993), 'Spectral analysis for categorical time series: Scaling and the spectral envelope', *Biometrika* **80**, 611–622.

Viana, M., Pandolfi, M., Minguillón, M., Querol, X., Alastuey, A., Monfort, E. & Celades, I. (2008), 'Inter-comparison of receptor models for PM source apportionment: Case study in an industrial area', *Atmospheric Environment* **42**, 3820–3832.

White, D., Richman, M. & Yarnal, B. (1991), 'Climate regionalization and rotation of principal components', *Int. J. Climatol.* **11**, 1–25.

Yu, T.-Y. & Chang, I.-C. (2006), 'Spatiotemporal features of severe air pollution in Northern Taiwan', *Environmental Science and Pollution Research* **13**, 268–275.

Yu, T.-Y. & Yu, T.-K. (2004), 'Spatial and temporal features of ambient air-quality over Taiwan', *Environmental Science and Pollution Research* **11**, 3–6.