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

EDSON ZAMBON MONTE

ANÁLISE DE COMPONENTES PRINCIPAIS EM SÉRIES TEMPORAIS MULTIVARIADAS COM HETEROSCEDASTICIDADE CONDICIONAL E OUTLIERS: UMA APLICAÇÃO PARA A POLUIÇÃO DO AR, NA REGIÃO DA GRANDE VITÓRIA, ESPÍRITO SANTO, BRASIL

VITÓRIA
2016

EDSON ZAMBON MONTE

ANÁLISE DE COMPONENTES PRINCIPAIS EM SÉRIES TEMPORAIS MULTIVARIADAS COM HETEROSCEDASTICIDADE CONDICIONAL E OUTLIERS: UMA APLICAÇÃO PARA A POLUIÇÃO DO AR, NA REGIÃO DA GRANDE VITÓRIA, ESPÍRITO SANTO, BRASIL

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

VITÓRIA
2016

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

Monte, Edson Zambon, 1982-

M772a Análise de componentes principais em séries temporais multivariadas com heteroscedasticidade condicional e outliers : uma aplicação para a poluição do ar, na Região da Grande Vitória, Espírito Santo, Brasil / Edson Zambon Monte. – 2016.

167 f. : il.

Orientador: Valdério Anselmo Reisen.

Tese (Doutorado em Engenharia Ambiental) – Universidade Federal do Espírito Santo, Centro Tecnológico.

1. Análise de componentes principais. 2. Valores estranhos (Estatística). 3. Outliers. 4. Estatística robusta. 5. Ar – Poluição. 6. Vitória, Região Metropolitana de (ES). 7. Heteroscedasticidade (Estatística). I. Reisen, Valdério Anselmo. II. Universidade Federal do Espírito Santo. Centro Tecnológico. III. Título.

CDU: 628

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

"Análise de componentes principais em séries multivariadas com heteroscedasticidade condicional e outliers: uma aplicação para a poluição do ar, na Região da Grande Vitória, Espírito Santo, Brasil".

EDSON ZAMBON MONTE

Banca Examinadora:

Prof. Dr. Valdério Anselmo Reisen
Orientador – DEST/CCE/UFES

Prof. Dr. Neyval Costa Reis Jr.
Examinador Interno – DEA/CT/UFES

Profa. Dra. Taciana Toledo de Almeida Albuquerque
Examinador Interno – PPGEA/CT/UFES

Prof. Dr. Celso José Munaro
Examinador Interno – DEE/CT/UFES

Prof. Dr. Pascal Bondon
Examinador Externo – Supélec-França

Prof. Dr. Márton Ispány
Examinador Externo – Debrecen/Hungria

Coordenador do PPGEA: Prof. Dr. Edmilson Costa Teixeira
UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO
Vitória, ES, 01 de abril de 2016.

Aos meus pais Dorcino e Carmen.

À minha mulher Renata.

À minha filha Maria.

Ao meu irmão Helder.

AGRADECIMENTOS

À Deus, por me dar força, saúde e humildade para enfrentar as dificuldades e conquistar meus objetivos.

Aos meus pais, Dorcino Monte e Carmen da Penha Zambon Monte, que sempre depositaram em mim toda confiança, e que propiciam as condições necessárias para o seguimento dos meus estudos. Ao meu irmão, Helder, que sempre me incentivou a caminhar em busca do melhor e pela grande amizade.

À minha querida esposa, Renata, pelo grande amor, companheirismo, compreensão, dedicação e por estar sempre presente ao meu lado, ajudando-me a trilhar pelos caminhos da vida. Obrigado pela paciência durante todos estes anos. À minha filha, Maria, que ainda está por vir, mas que já me proporciona tanta alegria.

Ao meu orientador, Valdério Anselmo Reisen, pelo estímulo dado ao desenvolvimento deste trabalho, pelos conhecimentos e orientações que possibilitaram a finalização do mesmo e pela amizade.

À todos os professores do Programa de Pós-Graduação em Engenharia Ambiental (PPGEA) da Universidade Federal do Espírito Santo, pela significante contribuição para minha formação profissional. Especialmente, aos professores da área de poluição do ar: Davidson Martins Moreira, Jane Meri Santos, Neyval Costa Reis Jr. e Taciana Toledo de Almeida Albuquerque. Também, a todos os funcionários do PPGEA, em especial, a secretaria Rose Mary Nunes Leão, pelos auxílios e esclarecimentos indispensáveis.

Aos professores da banca examinadora, Celso José Munaro, Márton Ispány e Pascal Bondon, pelas sugestões e críticas ao trabalho.

Aos amigos de doutorado e do NuMEs. Agradeço especialmente à Adriano Sgrâncio, Alessandro Sarnaglia, Angélica Rossow, Bartolomeu Zamprogno, Carla Maziero, Carlo Solci, Fátima Leite, Faradiba Serpa, Higor Cotta, Juliana Bottoni, Milena Machado, Paulo Prezotti, Wanderson Pinto e Wharley Borges pela ajuda direta ou indireta, conselhos e pelos momentos de descontração.

Aos professores Josu Arteche González (University of Basque Country, Espanha), Ruey S. Tsay (University of Chicago, Estados Unidos), Yu-Pin Hu (National Chi Nan University, Taiwan), Flávio Augusto Ziegelmann (Universidade Federal do Rio Grande do Sul, Brasil) e Fábio Alexander Fajardo Molinares (Universidade Federal do Espírito Santo, Brasil), pelos esclarecimentos importantíssimos para melhoria e finalização deste estudo.

Aos colegas, técnicos e professores, do Departamento de Economia da UFES, especialmente, a Romilda Alves da Silva e a Téthys Cysne Gobbi, pelo companheirismo de sempre.

Por fim, à todos que, direta ou indiretamente, contribuíram para a realização deste trabalho.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E/OU SIGLAS

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

LISTA DE ABREVIATURAS E/OU SIGLAS

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

LISTA DE ABREVIATURAS E/OU SIGLAS

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

RESUMO

As questões relativas à qualidade do ar têm se tornado cada vez mais importantes, uma vez que vários problemas de saúde decorrem da poluição atmosférica. Além disso, a poluição do ar contribui para a degradação do meio ambiente e, consequentemente, para o agravamento do efeito estufa. Dessa forma, diversos estudos adotando técnicas estatísticas têm sido realizados, com o intuito de contribuir na tomada de decisões dos agentes públicos e privados no que diz respeito ao combate à poluição, à prevenção de altas concentrações e à formulação de legislações para esse fim. Uma das metodologias estatísticas adotadas é a análise de componentes principais (ACP) clássica, sendo a mesma utilizada para o redimensionamento de rede, em análises de cluster, em análise de regressão, entre outros. No entanto, observa-se que, entre os estudos que têm adotado a ACP clássica, uma característica comum é negligenciar a heteroscedasticidade condicional e/ou a presença de outliers aditivos, que pode levar à resultados espúrios (enganosos), uma vez que a matriz de autocovariância estimada pode ser viesada (estimada incorretamente). Nota-se que as séries temporais relacionadas à poluição atmosférica tendem à apresentar heteroscedasticidade condicional e outliers aditivos. Assim, o primeiro artigo desta tese propôs aplicar um filtro multivariado VARFIMA-GARCH aos dados originais e utilizar a ACP clássica sobre os resíduos do modelo VARFIMA-GARCH. Com esse modelo, buscou-se filtrar, além da volatilidade, a correlação temporal e o comportamento de memória longa. A aplicação da ACP sobre os resíduos do modelo VARFIMA-GARCH mostrou-se mais coerente com as características ambientais da Região da Grande Vitória (RGV), Espírito Santo, Brasil, do que a aplicação usando os dados originais. No segundo artigo, que é a principal contribuição desta tese, a técnica de componentes principais com volatilidade (PVC), proposta por Hu e Tsay (2014), foi estendida para uma abordagem robusta (RPVC), a fim de capturar a volatilidade presente nos processos temporais multivariados, mas, levando-se em consideração os efeitos de outliers aditivos sobre a covariância condicional, uma vez que esses outliers podem mascarar (“esconder”) a heteroscedasticidade condicional ou, até mesmo, produzir efeitos voláteis espúrios, quando os dados não apresentarem volatilidade. O método RPVC proposto melhorou as predições dos picos de concentração de MP_{10}, na estação de Laranjeiras, RGV.

Palavras-chave: Análise de Componentes Principais. Heteroscedasticidade Condicional. Outliers. Robustez. Poluição do Ar.

ABSTRACT

Issues relating to air quality have become increasingly important, since many health problems come from air pollution. In addition, air pollution contributes to the degradation of the environment, contributing to the greenhouse effect. Thus, several studies adopting technical statistics have been conducted in order to contribute in the making of public and private actors with regard to combating pollution, prevention of high concentrations and formulation of laws for this purpose. The classical principal component analysis (PCA) is a statistical methodologies adopted. The PCA is used for dimensional reduction, cluster analysis, regression analysis, among others. However, among the studies that have adopted the classical PCA, a common feature is to neglect the conditional heteroscedasticity and/or the presence of additive outliers, which may lead to spurious results (misleading), since the estimated autocovariance matrix may be biased (estimated incorrectly). It is possible to note that the time series related to air pollution tend to present conditional heteroscedasticity and additive outliers. Then, the first paper of this thesis proposed to apply a multivariate filter VARFIMA-GARCH to the original data and use the classical PCA on residuals of the VARFIMA-GARCH model. Besides the volatility, this model was used to filter the temporal correlation and the long memory behavior. The application of the PCA on the residuals of the VARFIMA-GARCH model was more consistent with the environmental characteristics of the Greater Victoria Region (GVR), Espírito Santo, Brazil, than the application using the original data The second paper, that is the core of this thesis, the technique of principal volatility components (PVC), proposed by Hu e Tsay (2014), was extended for a robust approach (RPVC), in order to capture the volatility present in the multivariate time processes, but considering the effects of additive outliers on conditional covariance, since these outliers may mask ("hide") the conditional heteroscedasticity or even produce spurious volatility. The proposed RPVC improved the predictions of PM_{10} exceedance days in the Laranjeiras station, in the GVR.

Keywords: Principal Component Analysis. Conditional Heteroscedasticity. Outliers. Robustness. Air Pollution.

SUMÁRIO

1 INTRODUÇÃO

A intensificação do processo de industrialização ocorrida no século XIX, aliado ao crescimento populacional, especialmente, o crescimento da população urbana em detrimento da rural, vem aumentando as preocupações dos governos, sejam locais ou centrais, relacionadas à proteção do meio ambiente. Em relação à poluição do ar, de acordo com Vingarzan (2004) e Oltmans et al. (2006), em diversas partes do mundo essa vem crescendo em função, principalmente, da industrialização, da urbanização e da queima de combustíveis fósseis. Conforme Gramsch et al. (2006), dado que a poluição atmosférica é mais concentrada em áreas urbanas e industriais, os esforços de monitoramento da qualidade do ar são maiores nessas áreas ou regiões.

As questões relativas à qualidade do ar têm se tornado cada vez mais importantes, uma vez que vários problemas de saúde estão relacionados à poluição atmosférica, entre eles: asmas, renites, ardor nos olhos, cansaço, tosse seca, doenças cardiovasculares e pulmonares, insuficiência cardíaca, e, etc. Autores como Brunekreef e Holgate (2002), Maynard (2004), World Health Organization (2005), Curtis et al. (2006), entre outros, demonstraram a relação entre os poluentes legislados (partículas inaláveis com diâmetro menor que 10 micrômetros (MP_{10}), monóxido de carbono (CO), dióxido de enxofre (SO_2), óxidos de nitrogênio (NO_X) e ozônio (O_3)) e os problemas de saúde. No ano de 2012, por exemplo, a morte de 4,3 milhões pessoas foi atribuída à poluição atmosférica (WORLD HEALTH ORGANIZATION, 2014). Além disso, a poluição do ar contribui para a degradação do meio ambiente, contribuindo para o agravamento do efeito estufa.

Nesse contexto, vale dizer que a economia capixaba (na qual está inserida a área de estudo desta pesquisa) vem crescendo fortemente no decorrer dos últimos anos, especialmente, a partir de 2003, inclusive com taxas de crescimento do Produto Interno Bruto (PIB) superiores à média nacional. Com isso, diversas indústrias e empresas se instalaram ou ampliaram suas instalações no estado, principalmente, na Região da Grande Vitória[1] (RGV, região de estudo desta pesquisa), o que tende, consequentemente, a elevar o nível de poluição atmosférica, mesmo existindo diversas regulamentações impostas pelos órgãos de controle ambiental à essas indústrias e empresas. Além disso, o crescimento da frota de veículos, o maior consumo de energia, e, etc., também contribuem para a maior emissão de poluentes na região.

Destaca-se aqui que, no ano de 2010, a população do Espírito Santo era de 3.514.952 (INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA, 2014). Desse total, 1.687.704 estava residindo na região metropolitana do estado, que é composta pelos municípios da Grande Vitória, mais Fundão e Guarapari. Tomando-se somente a Grande Vitória, a população chegou a 1.565.393, o que representa cerca de 45% da população capixaba. Logo, uma vez que Gramsch

[1] A Região da Grande Vitória é formada por cinco municípios, a saber: Cariacica, Serra, Viana, Vila Velha e Vitória.

(2006) descreveu que a maior concentração de poluentes está nas áreas urbanas e industriais, aproximadamente 45% da polução capixaba foi fortemente afetada pelas emissões de poluição na atmosfera.

De acordo com o Inventário de Emissões Atmosféricas da Região da Grande Vitória[2] (ECOSOFT CONSULTORIA E SOFTWARES AMBIENTAIS, 2011), essa região possui diferentes fontes de poluentes tais como as originarias de: vias de tráfego, indústrias de diversos seguimentos, portos, aeroportos, emissões residenciais e comerciais, entre outras. Para monitorar a qualidade do ar da Grande Vitória existem oito estações de monitoramento (para ver a distribuição destas estações na RGV observar item 4.4), que são de propriedade do Instituto Estadual do Meio Ambiente e Recursos Hídricos (IEMA). Baseando-se nos padrões de qualidade do ar estabelecidos pela Resolução nº 03, de junho de 1990, do Conselho Nacional do Meio Ambiente (CONAMA, 1990), a Secretaria de Estado de Meio Ambiente e de Recursos Hídricos do Espírito Santo (SEAMA) faz o acompanhamento *online* da qualidade do ar da RGV.

O Governo do Estado do Espírito Santo, por meio do Decreto no 3463-R, de 16 de dezembro de 2013 (GOVERNO DO ESTADO DO ESPÍRITO SANTO, 2013), estabeleceu os padrões estaduais de qualidade do ar. Além dos padrões já descritos e estabelecidos na Resolução CONAMA n° 03/90 (excessão feita à fumaça), foram incluídos o material particulado com diâmetro aerodinâmico equivalente de corte igual a 2,5 $\mu g/m^3$ e as partículas sedimentadas (poeira sedimentada). Além disso, o decreto inseriu o conceito de Metas Intermediárias (MI), que são estabelecidas como valores temporários a serem cumpridos em etapas, visando à melhoria gradativa da qualidade do ar, e Padrões Finais (PF), que representam os alvos de longo prazo. No mais, foram criadas três MI que levam ao gradual atendimento dos padrões finais, estabelecidos com base nas diretrizes da Organização Mundial de Saúde (OMS) para os poluentes de interesse investigados por essa organização. Estratégia semelhante à adotada pelo estado de São Paulo, em abril de 2013 (para mais detalhes ver Instituto Estadual de Meio Ambiente e Recursos Hídricos do Estado do Espírito Santo (2014)). Os padrões nacionais e estaduais de qualidade do ar e as diretrizes da OMS podem ser verificados na Tabela 1, Subseção 3.1.

Nos estudos recentes relacionados com a poluição do ar, especial atenção tem sido dada aos modelos matemáticos chamados modelos receptores, que adotam medidas de concentrações dos poluentes nos receptores, com o intuito de identificar suas fontes e, consequentemente, estimar qual a contribuição de cada fonte identificada em termos de massa total, tais como PM_{10}, SO_2, entre outros (SEINFELD; PANDIS, 2006). Na literatura, os modelos receptores mais estudados são: balanço químico de massa (BQM), análise multivariada, análise de componentes principais (ACP), análise fatorial (AF), regressão linear múltipla, análise de cluster, fatoração da matriz positiva (FPM), entre outros (WATSON et al., 2002).

Em relação à ACP clássica, a ideia central é decompor um vetor aleatório k-dimensional em k componentes não correlacionados contemporaneamente, de acordo com a quantidade de variabilidade explicada por esses componentes. Além de ser adotada para a redução de dimensão de

[2]O inventário realizado refere-se ano de 2009.

um conjunto de dados, a ACP tem sido amplamente utilizada nas seguintes técnicas: análise fatorial, análise de cluster, análise de correlação canônica, análise discriminante, regressão linear e não-linear, entre outras. No entanto, entre os estudos que adotaram (ou adotam) a abordagem ACP, no domínio do tempo, uma característica comum é negligenciar a dependência dos dados[3]. Porém, em sua forma clássica, tal técnica pressupõe que os dados sejam independentes (ANDERSON, 2003; JOHNSON; WICHERN, 2007). Zamprogno (2013) demonstrou, empiricamente, que as análises estatísticas podem ser espúrias quando a técnica de ACP for aplicada à séries temporais multivariadas com forte correlação temporal. Assim, o autor propôs a utilização do modelo vetorial autorregressivo fracionário integrado de médias móveis (VARFIMA), para filtrar as séries temporais originais e, então, aplicar a técnica de ACP sobre resíduos (ruídos brancos) do modelo VARFIMA. O filtro permitiu corrigir e amenizar os problemas. No entanto, em relação à correlação cruzada entre os componentes, por exemplo, mesmo após a aplicação do filtro, permaneceram correlações significativas. Segundo Zamprogno (2013), isso pode ser decorrência da estrutura de heterogeneidade dos dados.

Vale destacar que, conforme Ding, Granger e Engle (1993), uma série temporal pode ser serialmente não correlacionada, porém ser dependente no tempo. Isso porque, de acordo com os autores, se uma série é um processo independente e identicamente distribuído (i.i.d.), qualquer transformação da mesma também deveria ser i.i.d. (por exemplo, as formas absoluta e quadrática do processo). Para exemplificar essa característica, Baillie, Bollerslev e Mikkelsen (1996) relataram que os retornos dos ativos em determinados mercados especulativos são pouco correlacionados, mas não independentes ao longo do tempo, dado que a maioria dos retornos tende a apresentar forte volatilidade temporal.

Conforme Hu e Tsay (2014), além da correlação temporal, outro ponto importante tem sido negligenciado nas análises de componentes principais, a saber: a heterocedasticidade condicional ou volatilidade[4]. Isto é, a técnica de ACP não leva em conta a dependência dinâmica entre os processos estocásticos com volatilidade. Conforme Matteson e Tsay (2011), os componentes principais são contemporaneamente não correlacionados. Entretanto, as correlações transversais defasadas podem ser diferentes de zero, as correlações condicionais podem ser diferentes de zero e as correlações cruzadas de transformações não lineares, tais como os processos quadráticos, podem ser diferentes de zero.

Nesse contexto, nota-se que as séries relacionadas à poluição atmosférica tendem a apresentar, além de correlação temporal, forte volatilidade ao longo do tempo, em função, por exemplo, das variações atmosféricas, que fazem com que os níveis de poluição oscilem fortemente durante o dia. Assim, a aplicação da ACP clássica ou, até mesmo, da técnica de ACP após a

[3]Cabe mencionar que, Milionis e Davies (1994) já destacavam que, à época, os estudos que utilizam séries temporais, no contexto da poluição do ar, ainda davam pouca atenção para algumas propriedades importantes que envolvem tais séries, como: estacionariedade, regressão espúria, quebras estruturais, causalidade, cointegração, dentre outros. Algumas dessas falhas foram corrigidas, porém, algumas persistem.

[4]A volatilidade pode ser entendida como o desvio padrão condicional da série (MATTESON; TSAY, 2011). De forma geral, a volatilidade pode ser vista como uma grande variabilidade das séries temporais em torno da sua média, sendo tal volatilidade condicional no tempo.

adoção do filtro VARFIMA, não seria suficiente para remover a heterocedasticidade condicional das séries, o que pode gerar resultados espúrios (ou enganosos). Dessa forma, o primeiro artigo desta tese, denominado *"Principal component analysis in multivariate time series with conditional heteroscedasticity and long memory: an application to air pollution in the Greater Vitória Region, Espírito Santo, Brazil"*, propôs aplicar um filtro multivariado VARFIMA-GARCH nos dados originais e utilizar a ACP clássica sobre os resíduos do modelo VARFIMA-GARCH. Com esse modelo, buscou-se filtrar, além da volatilidade, a correlação temporal e o comportamento de memória longa.

Ressalta-se que, os dados de poluição do ar podem conter observações influenciadas por eventos externos (níveis de concentração acima do padrão médio), conhecidas na literatura estatística como outliers[5]. Como exemplo, dado o posicionamento geográfico das grandes indústrias da RGV em relação às conglomerados urbanos, dependendo da direção do vento, uma pluma de poluição pode se deslocar e produzir altos níveis de poluição em certa localidade da região. Essas observações são importantes para explicar as características ambientais das séries temporais e não podem ser removidas nas análises estatísticas. Contudo, tais observações atípicas podem provocar sérios problemas para algumas funções estatísticas, como a média, o desvio-padrão e a variância condicional. No mais, uma vez que a estimativa dos modelos de séries temporais está conectado com essas funções, o modelo estimado final pode ser fortemente afetado por grandes valores da série.

Assim, no segundo artigo, denominado *"Robust principal volatility component analysis: an application to air pollution in the Greater Vitória Region, Espírito Santo, Brazil"*, e que é a principal contribuição desta tese, a técnica de componentes principais com volatilidade (PVC), proposta por Hu e Tsay (2014), foi estendida para uma abordagem robusta (RPVC), a fim de capturar a volatilidade presente nos processos temporais multivariados, mas, levando-se em consideração os efeitos de outliers aditivos sobre a covariância condicional, uma vez que esses outliers podem mascarar ("esconder") a heterocedasticidade condicional ou, até mesmo, produzir efeitos voláteis espúrios, quando os dados não apresentarem volatilidade (FRANKE, 2014; DIJK; FRANSES; LUCAS, 1999; CARNERO; PEÑA; RUIZ, 2007).

Por fim, destaca-se que, quatro trabalhos adicionais foram desenvolvidos e serviram de suporte para os dois principais artigos desta tese (descritos no parágrafo anterior). Esses trabalhos estão apresentados no Capítulo 8, "Apêndice: estudos adicionais". O primeiro, denominado "Previsão da concentração de ozônio na Região da Grande Vitória, Espírito Santo, Brasil, utilizando o modelo ARMAX-GARCH", foi publicado na Revista Brasileira de Meteorologia. O objetivo foi estimar e prever a concentração horária de ozônio na RGV, Espírito Santo, Brasil, utilizando o modelo ARMAX-GARCH, para o período 01/01/2011 a 31/12/2011. Foram utilizados dados da rede de monitoramento do IEMA, sendo escolhidas três estações: Laranjeiras, Enseada do Suá e Cariacica. Adotou-se alguns parâmetros medidos nas estações como variáveis explicativas da concentração de ozônio, a saber: temperatura, umidade relativa, velocidade do

[5]Em estatística, um outlier é uma observação que está distante das outras observações em um série de dados.

vento e concentração de dióxido de nitrogênio. Essas variáveis foram significativas e melhoraram a estimativa do modelo ajustado. As previsões horárias para o dia 31/12/2011 revelaram-se muito próximas dos valores observados, sendo que as estimativas, em geral, seguiram a trajetória diária da concentração de ozônio. No mais, em comparação aos modelos ARMA e ARMAX, o modelo ARMAX-GARCH revelou-se mais eficaz na predição de episódios de poluição de ozônio (concentração horária superior a 80 μg/m^3), reduziu o número de falsos alarmes estimados e apresentou menor taxa de ocorrência de episódios não detectados.

O segundo artigo adicional, denominado "Impactos das variáveis meteorológicas na qualidade do ar da Região da Grande Vitória, Espírito Santo, Brasil", foi aceito para publicação na Revista Brasileira de Meteorologia. Esse trabalho objetivou verificar os impactos das variáveis meteorológicas temperatura, umidade relativa, velocidade do vento e precipitação sobre a qualidade do ar, na RGV, Espírito Santo, Brasil, considerando o poluente MP_{10}, por meio do modelo Logit. O período de estudo foi de janeiro de 2005 a dezembro de 2010, onde a qualidade do ar foi classificada como "não boa" e "boa". Também foram estimados os efeitos dos dias da semana e das estações do ano sobre a probabilidade de ocorrência de qualidade do ar "não boa". Os resultados demonstraram que os fatores meteorológicos precipitação pluviométrica e velocidade do vento contribuíram significantemente para a redução da probabilidade de ocorrência de qualidade do ar "não boa". Além disso, os resultados simulados mostraram que, durante os finais de semana, as chances de ocorrer qualidade do ar "não boa" foram fortemente reduzidas e, nas estações do outono e do inverno, a probabilidade de se verificar qualidade do ar "não boa" caiu de maneira relevante.

Um terceiro artigo complementar, denominado "Inter-relações entre as concentrações de ozônio e de dióxido de nitrogênio na RGV, Espírito Santo, Brasil", foi aceito para publicação na revista Engenharia Sanitária e Ambiental. O artigo verificou as inter-relações entre as concentrações de O_3 e de NO_2, na RGV, Espírito Santo, Brasil. Adotou-se a metodologia VAR e o teste de causalidade de Granger. Os resultados revelaram que as concentrações de O_3 e de NO_2 da região (estação) de Laranjeiras foram as menos afetadas por concentrações de outras estações. Dada à localização, as concentrações de O_3 e NO_2 da Enseadá do Suá tiveram significativa influência de outras regiões, especialmente de Jardim Camburi, Ibes e Vitória-Centro. A concentração de ozônio na região do Ibes foi fortemente influenciada pelas concentrações de O_3 e de NO_2 da Enseada do Suá. Além disso, as concentrações de Cariacica sofreram impactos relevantes das concentrações da Enseada do Suá, provavelmente devido à direção do vento Norte/Nordeste, predominante na RGV.

O quarto artigo adicional, denominado *"Robust factor modeling for high-dimensional time series: an application to air pollution data"*, foi enviando para avaliação no journal Computational Statistics and Data Analysis. Esse artigo considerou a modelagem fatorial para séries tempotais de alta dimensão, na presença de outliers aditivos, propondo uma variante robusta dos métodos de estimação de Lam e Yao (2012). O estimador do número de fatores foi obtido pela análise de autovalores de uma matriz não-negativa definida robusta, isto é, uma matriz de covariância robusta. Algumas propriedades assintóticas dos valores próprios robustos foram de-

rivadas. Em particular, demonstrou-se que os autovalores da matriz covariância robusta estimativa tem a mesma taxa de convergência do que os autovalores da matriz de covariância clássica estimada. Simulações, para amostras finitas, foram realizadas para analisar a performance do estimador robusto do número de fatores, sob os cenários de séries temporais multivariadas contaminadas e não contaminadas. Como exemplo de aplicação, a análise factorial robusta foi utilizada para identificar o comportamento do poluente MP_{10}, na RGV, Brasil, com o objetivo de reduzir a dimensionalidade dos dados e produzir boas previsões para os níveis de poluição do MP_{10}.

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

2 OBJETIVOS

2.1 OBJETIVO GERAL

Este trabalho teve como objetivo geral estudar a técnica de componentes principais em séries temporais multivariadas com heteroscedasticidade condicional e outliers, com aplicação para a poluição do ar, na Região da Grande Vitória, Espírito Santo.

2.2 OBJETIVOS ESPECÍFICOS

De forma específica, pretendeu-se:

a) Estudar empiricamente as propriedades da técnica de ACP em séries temporais multivariadas com heterocedasticidade condicional e memória longa;
b) Estudar empiricamente as propriedades da técnica de PVC em séries temporais multivaridas na presença de outliers;
c) Aplicar o filtro VARFIMA-GARCH aos dados amostrais antes de aplicar a técnica de ACP clássica;
d) Aplicar a técnica de PVC considerando a estimação robusta na presença de outliers aditivos.

3 REVISÃO DE LITERATURA

3.1 POLUIÇÃO ATMOSFÉRICA

A Resolução nº 03, de junho de 1990, do Conselho Nacional do Meio Ambiente (CONAMA, 1990), caracteriza como poluente atmosférico qualquer forma de matéria ou energia com intensidade e em quantidade, concentração, tempo ou características em desacordo com níveis estabelecidos, e que tornem ou possam tornar o ar: i) impróprio, nocivo ou ofensivo à saúde; ii) inconveniente ao bem-estar público; iii) danoso aos materiais, à fauna e a flora; e, iv) prejudicial à segurança, ao uso e ao gozo da propriedade e às atividades normais da comunidade.

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

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

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

- Dióxido de enxofre (SO_2): gás tóxico e incolor, pode ser emitido por fontes naturais ou por fontes antropogênicas e pode reagir com outros compostos na atmosfera, formando material particulado de diâmetro reduzido;
- Dióxido de nitrogênio (NO_2): gás poluente altamente oxidante, sendo que sua presença na atmosfera é fator proponderante na formação do ozônio troposférico;
- Hidrocarbonetos (HC): compostos formados de carbono e hidrogênio e que podem se apresentar na forma de gases, partículas finas ou gotas;
- Material particulado (MP): mistura complexa de sólidos com diâmetro reduzido. Seus componentes apresentam características físicas e químicas variadas. Geralmente o material particulado é classificado de acordo com o diâmetro das partículas, em função da relação existente entre diâmetro e possibilidade de penetração no trato respiratório;
- Monóxido de carbono (CO): gás inodoro e incolor, formado no processo de queima de combustíveis;

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

Vale frisar que, em relação aos padrões de qualidade do ar existe uma diferença entre os valores estabelecidos pela Resolução CONAMA nº 03 e as diretrizes da Organização Mundial da Saúde (OMS), uma vez que as diretrizes da OMS (revisadas em 2005) levam em conta os diversos estudos científicos realizados a partir de 1990. Além disso, como descrito anteriormente, o Governo do Estado do Espírito Santo, por meio do Decreto no 3463-R, estabeleceu novos padrões de qualidade do ar para o Espírito Santo que também apresentam diferenças quando à Resolução CONAMA nº 03. Na Tabela 1 são apresentados os padrões nacionais e estaduais de qualidade do ar e as diretrizes da OMS. Pode-se verificar a grande discrepância que há entre os padrões nacionais e as diretrizes da OMS.

3.2 ESTADO DA ARTE SOBRE O USO DA TÉCNICA DE ACP NA POLUIÇÃO ATMOSFÉRICA

O aumento dos níveis de poluição atmosférica ocorrido nos últimos anos tem feito com que, cada vez mais, os pesquisadores voltem suas atenções para essa problemática, que afeta a população com um todo. Para isso, novos modelos para análise da poluição do ar têm surgido, sendo que esses podem se enquadram em três categorias: modelos receptores; modelos de dispersão; e, modelos fotoquímicos. Em relação aos modelos receptores, esses se caracterizam por adotarem medidas de concentrações dos poluentes nos receptores, com o intuito de identificar suas fontes e, consequentemente, estimar qual a contribuição de cada fonte identificada em termos de massa total.

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

			$MP_{2.5}$ [µg/m^3]	MP_{10} [µg/m^3]	PTS [µg/m^3]	PS [g/m^2. 30 dias]	SO_2 [µg/m^3]	NO_2 [µg/m^3]	O_3 [µg/m^3]	CO [µg/m^3]	FUMAÇA [µg/m^3]
Padrão nacional (CONAMA nº 03/1990)	Curta exposição	Padrão primário	-	150[a] (24h)	240[a] (24h)	-	365[a] (24h)	320[a] (1h)	160[a] (1h)	10.000[a] (8h) 40.000[a] (1h)	150[a] (24h)
		Padrão secundário	-	150[a] (24h)	150[a] (24h)	-	100[a] (24h)	190[a] (1h)	160[a] (1h)	10.000[a] (8h) 40.000[a] (1h)	100[a] (24h)
	Longa exposição	Padrão primário	-	50 (ano)[b]	80 (ano)[c]	-	80 (ano)[b]	100 (ano)[b]	-	-	60 (ano)[b]
		Padrão secundário	-	50 (ano)[b]	60 (ano)[c]	-	40 (ano)[b]	100 (ano)[b]	-	-	40 (ano)[b]
Metas e padrão estadual (Decreto nº 3463-R 2013)	Curta exposição	MI1-ES	-	120 (24h)	180 (24h)	14	60 (24h)	240 (1h)	140 (8h)	-	-
		MI2-ES	50 (24h)	80 (24h)	170 (24h)	-	40 (24h)	220 (1h)	120 (8h)	-	-
		MI3-ES	37 (24h)	60 (24h)	160 (24h)	-	30 (24h)	210 (1h)	110 (8h)	-	-
		PF-ES	25 (24h)	50 (24h)	150 (24h)	-	20 (24h)	200 (1h)	100 (8h)	10.000 (8h) 30.000 (1h)	-
	Longa exposição	MI1-ES	-	45 (ano)[b]	65 (ano)[c]	-	40 (ano)[b]	50 (ano)[b]	-	-	-
		MI2-ES	20 (ano)[b]	33 (ano)[b]	63 (ano)[c]	-	30 (ano)[b]	45 (ano)[b]	-	-	-
		MI3-ES	15 (ano)[b]	25 (ano)[b]	62 (ano)[c]	-	20 (ano)[b]	42 (ano)[b]	-	-	-
		PF-Es	10 (ano)[b]	20 (ano)[b]	60 (ano)[c]	-	-	40 (ano)[b]	-	-	-
Diretriz OMS	Curta exposição		25 (24h)	50 (24h)	-	-	20 (24h) 500 (10min)	200 (1h)	100 (8h)	10.000 (8h) 30.000 (1h)	-
	Longa exposição		10 (ano)[b]	20 (ano)[b]	-	-	-	40 (ano)[b]	-	-	-

Note: 1) O tempo de média considerado para o cálculo da concentração do poluente está indicado entre parênteses; e, 2) [a] Não pode ser excedido mais que uma vez por ano; [b] Média aritmética anual; [c] Média geométrica anual.
Fonte: Iema (2014).

Conforme Watson et al. (2002), os modelos receptores utilizam concentrações ambientais e a abundância de componentes químicos nas fontes de emissão para quantificar a contribuição de cada fonte. Ainda, de acordo com Watson et al. (2002), pode-se citar entre os modelos

receptores: balanço químico de massa (BQM); ACP; análise fatorial (AF); regressão linear múltipla; redes neurais; análise de cluster; entre outros. Uma vez que o objetivo desta pesquisa foi analisar a técnica de componentes principais na presença de heteroscedasticidade condicional, esta seção visa descrever alguns estudos que aplicaram a técnica de ACP na poluição atmosférica[6].

Henry e Hidy (1979) realizaram uma análise multivariada para dados de sulfato particulado, meteorológicos e de qualidade do ar, para as cidades de Los Angeles e Nova York, Estados Unidos. Os autores aplicaram ACP ao conjunto de dados (excluindo o sulfato) produzindo combinações lineares independentes dos dados originais. Verificou-se que somente três componentes foram necessários para explicar 50% da variabilidade de todas as variáveis. A variável sulfato foi então regredida contra os componentes independentes. Em Los Angeles, a ACP indicou que os fatores SO_2 e de dispersão são insignificantes para explicar a variabilidade do sulfato. Em contrapartida, os fatores associados à atividade fotoquímica e a umidade atmosférica explicaram mais da metade da variabilidade do sulfato. Em Nova York, um componente fotoquímico similar foi encontrado. No entanto, os fatores de dispersão e o dióxido de enxofre também contribuíram para explicar a variabilidade do sulfato.

Smeyers-Verbeke et al. (1984), por meio da ACP, buscaram demonstrar a variabilidade de um conjunto de poluentes atmosféricos orgânicos, em quatro cidades holandesas, Terschelling, Vlaardingen, Delft e Heilevoetsluis, no período de 1979 a 1981. Também verificaram a relação desses poluentes com alguns parâmetros meteorológicos, como: velocidade do vento, direção do vento, temperatura, estação do ano, e, etc. Os autores realizaram cerca de 400 medições para 26 compostos orgânicos gasosos. Em relação aos resultados, esses revelaram que, independente da cidade pesquisada, os 26 compostos orgânicos gasosos puderam ser trabalhados por meio de três componentes principais, que explicaram, aproximadamente, 62% da variabilidade total dos compostos orgânicos. No mais, relacionando os três componentes principais com os parâmetros meteorológicos, notou-se que os parâmetros mais importantes foram a direção do vento, a velocidade do vento e a estação do ano (inverno versus verão).

A pesquisa de Pio et al. (1991) objetivou detectar as possíveis fontes de aerossóis suspensos na atmosfera, na região costeira de Aveiro, Portugal. Os dados foram coletados em períodos amostrais de 24 horas, começando às oito horas da manhã, atendendo a um intervalo de três dias, para eliminar possíveis efeitos de fim de semana. No total foram obtidas 43 amostras de diversos aerossóis. A análise de componentes principais e a análise de cluster permitiram a identificação de seis grupos de fontes principais dos aerossóis atmosféricos: i) cinco regionais (emissão do solo, transporte, combustão de óleo diesel, poluentes secundários e água do mar); e, ii) uma, possivelmente local, que não foi identificada com precisão.

O trabalho de Statheropoulos, Vassiliadis e Pappa (1998) objetivou analisar as concentrações de CO, NO, NO_2, O_3, fumaça e SO_2, por meio da ACP, na cidade de Atenas, Grécia. Os dados

[6]Cabe mencionar que a técnica de componentes principais com volatilidade (PVC) foi desenvolvida recentemente por Hu e Tsay (2014). Dessa forma, salvo engano, não foram encontrados trabalhos com aplicação dessa técnica, principalmente no que se refere aos estudos sobre poluição atmosférica.

utilizados referem-se ao período de 1988 a 1992. A abordagem ACP também foi aplicada para as seguintes variáveis meteorológicas: umidade relativa, temperatura, radiação solar, velocidade do vento e direção do vento. As análises foram separadas para os períodos de verão e inverno. Os resultados revelaram que os principais componentes extraídos a partir da poluição do ar estavam relacionados à combustão da gasolina, à combustão do óleo diesel e às interações com o ozônio. Em relação às variáveis meteorológicas, os componentes mais importantes foram relacionados às condições de seca (período de verão) e aos ventos sudoeste de altas velocidades. Por fim, os autores utilizaram a análise de correlação canônica para determinar a relação entre os dois diferentes conjuntos de dados. A principal relação ocorreu entre a poluição total e as altas umidades relativas, em combinação com as baixas velocidades do vento.

Yu e Chang (2000) adotaram a técnica de ACP para avaliar a poluição do ar oriunda do ozônio, no sul de Taiwan, China. O período de análise foi de 01 de julho de 1993 a 30 de junho de 1998, sendo que foram analisadas 17 estações de monitoramento. O método de rotação varimax foi aplicado à análise, para verificar áreas com características homogêneas em termos de poluição de ozônio. Os resultados revelaram que o sul de Taiwan pode ser separado em quatro sub-regiões homogêneas (quatro componentes principais), no que tange à poluição via ozônio. Essas quatro sub-regiões responderam por 70,8% da variância da concentração de ozônio. Um ponto interessante é que a brisa marítima ocidental foi um fator dominante na produção de altas concentrações de ozônio em muitas das estações. As análises também revelaram uma similaridade dos padrões meteorológicos das estações nas mesmas sub-regiões.

Guo, Wang e Louie (2004) utilizaram o método de componentes principais/componentes principais absolutos (CP/CPS), para estudar a poluição atmosférica proveniente dos hidrocarbonetos não-metano (HCNM), no período de janeiro a dezembro de 2001, em dois locais de Hong Kong, China: Tsuen Wan (TW) e Centro/Ocidental (CW). A ACP identificou quatro maiores fontes de poluição em TW, a saber: emissões dos veículos; uso de solvente; gás liquefeito de petróleo (GLP) ou gás natural; e, fontes industriais, comerciais e domésticas. Para a região Centro/Ocidental foram identificadas cinco fontes principais: uso de solventes; emissões de veículos; gás liquefeito de petróleo (GLP) ou gás natural; fontes comerciais, domésticas e industriais; e, emissões biogênicas. No mais, nos dois locais pesquisados, o maior contribuinte para concentração de HCNM foi a emissão de poluentes veiculares. Na sequência vieram o uso de solventes, o gás natural e as fontes comerciais, domésticas e industriais, respectivamente.

Lu et al. (2004) propuseram um novo modelo de rede neural que combinou a análise de componentes principais com a função de base radial (FBR), para prever as concentrações de material particulado em suspensão, NO_X e NO_2, na área urbana de Mong Kok, Hong Kong, China. Os dados das concentrações horárias foram relativos ao ano de 2000. A técnica de componentes principais foi adotada para verificar a características correlacionadas dos poluentes e das variáveis meteorológicas, a fim de reduzir o número de variáveis de entrada do modelo de redes neurais. Os resultados demonstraram que a abordagem proposta foi mais viável, confiável e eficaz, quando comparada às abordagens de redes neurais tradicionais.

Song et al. (2006) determinaram quais foram as principais fontes de emissão de material particulado fino ($MP_{2,5}$), em Pequim (área metropolitana), China, por meio das técnicas de análise de componentes principais/componentes principais absolutos (ACP/APCS) e UNMIX. Os dados foram retirados de uma análise química de amostras de 24 horas, coletadas em intervalos de seis dias, nos meses de janeiro, abril, julho e outubro, de 2000. Ambos os modelos identificaram cinco fontes que contribuíram para formação do $MP_{2,5}$: sulfato secundário e nitrato secundário; uma fonte mista de combustão do carvão e da queima de biomassa; emissões industriais; escapamento de veículos a motor; e, poeira de estrada. Em média, a ACP/APCS e a UNMIX explicaram 73% e 85% da variância da concentração de massa, respectivamente. Esses valores foram similares às estimativas realizadas por meio da fatoração da matriz positiva (FMP) e do BQM. No mais, verificou-se que os aerossóis secundários e a queima de carvão e biomassa foram os principais determinantes da formação do $MP_{2,5}$.

Harkat, Mourot e Ragot (2006) sugeriram um sensor de detecção de falhas e isolamento, para avaliar a rede de monitoramento da qualidade do ar, na cidade de Lorraine, França. A finalidade foi detectar anomalias de funcionamento dos sensores existentes, principalmente, para a concentração de ozônio e para os óxidos de nitrogênio. A abordagem ACP foi adotada com a finalidade de reduzir a dimensão da rede e lidar com o elevado grau de correlação entre as variáveis consideradas. O modelo proposto pelos autores permitiu isolar os sensores com defeito e estimar as amplitudes das falhas.

Steven e Martin (2006) adotaram modelos de controle para as tendências de longo prazo e os efeitos do tempo, em conjunto com a ACP e a análise de componentes principais supervisionada (ACPS), para estimar a associação entre vários poluentes atmosféricos e mortalidade para nove cidades norte-americanas. A periodicidade dos dados foi diária, correspondendo ao período de 1987 a 2000. Os autores consideraram cinco poluentes: MP_{10}, O_3, SO_2, CO e NO_2. Também utilizaram algumas variáveis meteorológicas, a saber: temperatura ambiente e temperatura do ponto de orvalho. De acordo com os autores, a técnica de ACPS mostrou-se mais eficaz que a ACP na identificação correta dos poluentes associados à mortalidade de cada região, uma vez que a ACPS permite separar os poluentes em subconjuntos, ou seja, leva em conta somente os poluentes que são importantes na explicação da mortalidade de uma dada região.

Pires et al. (2009) objetivaram mostrar como a análise de componentes principais pode ser útil nas medições redundantes em redes de monitoramento de qualidade do ar. O número mínimo de estações de monitoramento de qualidade do ar, na região Metropolitana de Oporto, Portugal, foi avaliado por ACP, e então comparado com a legislação vigente. Nove estações de monitoramento de NO_2, O_3 e MP_{10} foram selecionadas, sendo que o período de estudo foi de janeiro de 2003 a dezembro de 2005. Para verificar a persistência dos resultados da ACP os autores fizeram uma análise trimestral para os anos de 2003 e 2004. Baseando-se no critério de que o número de componentes deve representar pelo menos 90% da variância dos dados originais, os autores concluíram que somente cinco estações para NO_2, três para O_3 e sete para MP_{10} foram necessárias para caracterizar a região em termos de concentração de poluentes. Assim, os analisadores de poluentes atmosféricos correspondentes às medições redundantes poderiam

ser instalados em regiões não monitoradas, permitindo a ampliação da rede de monitoramento da qualidade do ar.

Chavent et al. (2009) aplicaram a técnica de ACP, juntamente com a técnica FMP, para caracterizar e identificar quais elementos químicos influenciaram na emissão de $MP_{2,5}$. Para isso, 65 amostras de $MP_{2,5}$ foram coletadas na área urbana de Anglet, França, no mês de dezembro de 2005. A aplicação, tanto da ACP quanto da FMP, possibilitou a identificação de cinco principais fontes de $MP_{2,5}$, a saber: poeira do solo, veículos, indústria, combustão e mar. Também foi possível verificar que durante o inverno, a fonte combustão predominou sobre a poeira do solo, em termos de determinação das partículas finas.

O estudo de Liu (2009) buscou realizar simulações para a concentração média de MP_{10}, na cidade de Ta-Liao, China. O autor adotou o modelo de regressão com erros de séries temporais (RTSE), incluindo uma variável explicativa resultante da análise de componentes principais para completar a simulação de MP_{10} (denominada de "CP trigger"). Para melhorar a previsibilidade de altas concentrações de MP_{10}, foram construídos quatro tipos de modelos RTSE: RTSE sem ACP; ACP4S (levando em conta as cidades vizinhas a Ta-Liao); RTSE PCTL (com variáveis meteorológicas e co-poluentes em Ta-Liao); e, RTSE PCTL4S (uma combinação dos dois últimos modelos). As variáveis ozônio, temperatura do ponto de orvalho, óxido de nitrogênio, direção do vento e ACP foram significantes nos modelos RTSE na maior parte do tempo. Os resultados demonstraram que as predições são melhores quando da presença da ACP, sendo que para os modelos mais completos, PC4S ou PCTL4S, a acurácia das predições foi ainda maior. Estudos semelhentes foram realizados por Liu e Johnson (2002), Liu e Johnson (2003), Liu (2007) and Liu et al. (2013), no contexto de regressão múltipla e dos modelos Box-Jenkins de séries temporais.

Yoo et al. (2011) avaliaram e caracterizaram as emissões de *black carbon*, material particulado (MP_{10} e $MP_{2,5}$) e outros poluentes gasosos, em uma área industrial, da cidade de Incheon, China, por meio das técnicas de ACP, análise de cluster e análise de correlação. A análise de componentes principais produziu quatro componentes que revelaram informações a respeito das concentrações de poluentes e suas fontes de emissão. Por exemplo, o primeiro componente principal (ACP1) apresentou forte concentração de MP_{10}, $MP_{2.5}$, CO e benzeno, além de ter demonstrado forte associação com fontes de emissões veiculares e industriais. O ACP2 teve alta concentração de NO_2 e black carbon, tendo maior relação com as emissões de veículos tais como: ônibus, vans, táxis, carros, motocicletas e caminhões. O ACP3 mostrou elevada concentração de tolueno e xileno e o ACP4 de SO_2. Por fim, os resultados sugeriram que a gestão apropriada das emissões veiculares, juntamente com o controle da poluição industrial, torna-se fundamental para o controle das partículas de $MP_{2.5}$ e dos gases poluentes, incluindo benzeno, tolueno, etil-benzeno e xilenos.

Dominick et al. (2012) investigaram as possíveis fontes de poluentes atmosféricos e seus padrões espaciais em oito estações de monitoramento da qualidade do ar, na Malásia, para um banco de dados de dois anos (2008 e 2009). Os poluentes e as variáveis meteorológicas utilizadas foram: ozônio, monóxido de carbono, óxido de nitrogênio, dióxido de nitrogênio, dióxido

de enxofre, MP_{10}, temperatura ambiente, umidade relativa e velocidade do vento. Também se trabalhou com o Índice de Poluição do Ar (IPA) da Malásia. Utilizou-se a análise de cluster aglomerativo hierárquico (ACAH) para avaliar os padrões espaciais (construção dos clusters), a ACP para determinar as principais fontes de poluição do ar e a análise de regressão múltipla verificar o percentual de contribuição de cada poluente. A abordagem ACAH agrupou as oito estações em três clusters. A ACP revelou que as maiores fontes de poluição do ar foram: veículos a motor, aviões, indústrias e áreas de alta densidade populacional. A análise de regressão mostrou que o MP_{10} foi o poluente que mais contribui para a variabilidade do IPA, em todas as estações. Além disso, verificou-se que o monóxido de carbono foi o poluente que mais influenciou na alta concentração de MP_{10}. Os fatores meteorológicos também influenciaram na concentração de MP_{10}.

Rajab, MatJafri e Lim (2013) combinaram o modelo de regressão linear múltiplo e o método de ACP para obter uma equação de regressão da concentração de ozônio contra algumas variáveis preditoras, a saber: temperatura da superfície do ar (TSA), monóxido de carbono, metano (CH_4), vapor de água (H_2O-vapor), temperatura da superfície da pele (TSP), temperatura ambiente (T), umidade relativa (UR) e pressão de superfície média (PSM). O estudo foi realizado para algumas regiões da Malásia e considerou o período de janeiro de 2003 a dezembro de 2008. Nesse caso, o método de ACP, por meio da rotação varimax, foi utilizado para construir subconjuntos das variáveis preditoras, com o intuito de melhorar a acurácia das estimativas do modelo de regressão. Os resultados indicaram que as concentrações de ozônio foram correlacionadas negativamente com CH_4, H_2O-vapor, RH e PSM, e positivamente relacionadas com CO, TSA, TSP e TA. Além disso, os valores estimados para a concentração de ozônio foram muito próximos dos valores observados para o ano de 2009 (ano escolhido para verificar o ajuste dos modelos).

Ressalta-se aqui que, conforme descrito anteriormente, o método de ACP pressupõe, em sua forma clássica, que haja independência dos dados (ANDERSON, 2003; JOHNSON; WICHERN, 2007). No entanto, as variáveis relacionadas à poluição atmosférica apresentaram, em geral, correlação temporal. Nesse contexto, observa-se que as pesquisas apresentadas não levaram em conta esse aspecto na aplicação da ACP, o que pode ter comprometido os resultados encontrados. Zamprogno (2013), por exemplo, demonstrou, empiricamente, análises estatísticas espúrias quando a técnica de ACP foi aplicada à séries temporais multivariadas com forte correlação temporal. O autor propôs a utilização do modelo VARFIMA, para filtrar as séries temporais referentes à poluição atmosférica, na RGV, e aplicou a técnica de ACP aos resíduos (ruídos brancos) desse modelo. Souza et al. (2014) utilizaram o modelo aditivo generalizado, em combinação com a técnica de ACP sobre os dados filtrados pelo modelo VAR, para analisar a associação entre as concentrações de algunas poluentes atmosféricos e as internações de crianças com problemas respiratórios, na RGV. Melo et al. (2015) combinaram o modelo de regressão logística com a ACP, a fim de estimar o risco relativo entre as concentrações de material particulado e o incômodo percebido pela população da RGV. Os autores também fil-

traram a correlação tamporal, por meio no modelo VAR, antes de aplicar a técnica de PCA[7]. Por fim, mesmo que Zamprogno (2013), Souza et al. (2014) e Melo et al. (2015) tenham filtrado a correlação temporal dos dados, para posterior aplicação da ACP, conforme discutido na introdução, tal técnica também tem negligenciado a heterocedasticidade condicional dos dados (HU; TSAY, 2014), sendo esse o foco principal desta pesquisa.

[7]Vale frisar que Tsay (2005) e Matteson e Tsay (2011) já faziam menção à aplicação do filtro VARMA (ou apenas o filtro VAR) para remover a correlação serial e a correlação cruzada dos dados antes da aplicação da ACP.

4 MATERIAIS E MÉTODOS

4.1 REGIÃO DE ESTUDO E REDE DE MONITORAMENTO

A área de estudo compreendeu a RGV, Espírito Santo, Brasil, localizada na costa sul do oceano Atlântico [latitude 20°19 S (Sul), longitude 40°20 W (Oeste)]. Por estar situada na região litorânea, a RGV apresenta clima tropical quente (Aw), possuindo inverno ameno e seco, e verão chuvoso e quente. As temperaturas médias variam entre 24° C (Celsius) e 30° C, e os ventos predominantes são de Norte/Nordeste na primavera-verão, sofrendo alterações durante outono e inverno devido ao posicionamento do sistema de alta pressão (Alta Pressão Subtropical do Atlântico Sul-ASAS) mais próximo do continente, possibilitando alterações na direção predominante do vento, a qual passa a variar entre as direções Sul/Oeste.

O relevo da Grande Vitória é caracterizado por cadeias montanhosas nas porções Noroeste (Mestre Álvaro) e Oeste (Região Serrana). Planícies (aeroporto e manguezais) e planaltos (Planalto Serrano) na porção Norte. Planícies (Barra do Jucu) na porção Sul. Todas porções são intercaladas por maciços rochosos de pequeno e médio porte. As condições de relevo no geral são favoráveis em grande parte da região à circulação de ventos para dispersão de poluentes.

No ano de 2010, a população do Espírito Santo era de 3.514.952 (INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA, 2014). Deste total 1.687.704 estava residindo na região metropolitana do estado, que é composta pelos municípios da Grande Vitória, mais Fundão e Guarapari. Tomando-se somente a RGV, a população chegou a 1.565.393, o que representa cerca de 45% da população capixaba.

A qualidade do ar na RGV é afetada principalmente por veículos automotores, por empreendimentos industriais e pelas atividades da construção civil. Além disso, a RGV conta com um complexo sistema portuário. Conforme o Inventário de Emissões Atmosféricas da Região da Grande Vitória (ECOSOFT CONSULTORIA E SOFTWARES AMBIENTAIS, 2011), as principais fontes emissoras de partículas na RGV são, respectivamente: a) veículos automotores, que representam mais de 60% das emissões de partículas, estando ligados à ressuspensão de partículas em vias; b) setor industrial minero-siderúrgico; e, c) setor logístico que inclui portos e aeroportos. O inventário não contempla as atividades da construção civil.

Em relação ao poluente SO_2, as principais fontes poluidoras são: indústria minero-siderúrgica e setor de logística (portos e aeroportos). Já para o poluente CO, destacam-se as fontes: indústria minerosiderúrgica e veículos automotores. Quanto aos hidrocarbonetos não-metano (HCNM), também denominados compostos orgânicos voláteis não-metano (COVNM), as principais fontes são: veículos, seguidos por atividades residenciais e comerciais, seguidas por estocagem e comercialização de combustíveis, indústrias química e minero-siderúrgicas. No que se refere ao O_3, esse é um poluente secundário, proveniente das reações fotoquímicas dos óxidos de nitrogênio na atmosfera, bem como dos HCNM e CO. Assim, a fontes emissoras de NO_X,

HC e CO são responsáveis indiretas pela presença de O_3 no ambiente (SEINFELD; PANDIS, 2006).

O monitoramente da qualidade do ar na RGV é feito pela Rede Automática de Monitoramento da Qualidade do Ar da Grande Vitória - RAMQAr, que entrou em funcionamento no ano de 2000, sendo de propriedade e responsabilidade do IEMA. Para o período de estudo, a rede era formada por oito estações de monitoramento distribuídas nos municípios da RGV. A Figura 1 apresenta a distribuição espacial das estações de monitoramento na região da Grande Vitória, que são descriminadas a seguir:

- Estação 1: Laranjeiras; Hospital Dório Silva;
- Estação 2: Carapina; ArcelorMittal-Av. Brigadeiro Eduardo Gomes, s/n;
- Estação 3: Jardim Camburi; Unidade de Saúde de Jardim Camburi;
- Estação 4: Enseada do Suá; Corpo de Bombeiros;
- Estação 5: Vitória-Centro; Prédio do Ministério da Fazenda;
- Estação 6: Vila Velha-Ibes; 4° Batalhão da Polícia Militar;
- Estação 7: Vila Velha-Centro; Av. Champagnat n° 911;
- Estação 8: Cariacica; CDA (CEASA).

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

Vale lembrar que a RAMQAr monitora os seguintes poluentes: PTS; MP_{10}; SO_2; CO; NO_X; HC; e, O_3. Existe, também, o monitoramento de alguns parâmetros meteorológicos, a saber: direção do vento (DV); velocidade do vento (VV); umidade relativa (UR); precipitação pluviométrica (PP); pressão atmosférica (P); temperatura (T); e, radiação solar (R). Um resumo dos poluentes e parâmetros meteorológicos que são medidos em cada estação pode ser visto na Tabela 2.

Tabela 2: Poluentes e parâmetros meteorológicos monitorados nas estações da RAMQAR

Estação	PTS	MP_{10}	SO_2	CO	NO_X	HC	O_3	Meteorologia
Laranjeiras	X	X	X	X	X		X	
Carapina	X	X						DV,VV,UR,PP,P,T,R
Jardim Camburi	X	X	X		X			
Enseada do Suá	X	X	X	X	X	X	X	DV,VV
Vitória-Centro	X	X	X	X	X	X		
Vila Velha-Ibes	X	X	X	X	X	X	X	DV,VV
Vila Velha-Centro		X	X					
Cariacica	X	X	X	X	X		X	DV,VV,T

4.2 DADOS

Uma vez que as características estatísticas estudadas neste trabalho podem ocorrer em diversos poluentes atmosféricos medidos temporalmente, a teoria estatística proposta poderia ser aplicada a qualquer poluente que apresente as mesmas. Neste caso específico, para fins de aplicação, adotou-se: i) no primeiro artigo: as concentrações de MP_{10} das oitos estações de monitoramento da RGV; e, ii) segundo artigo: as concentrações de MP_{10} das estações de Laranjeiras, Carapina e Camburi; as variáveis meteorológicas, temperatura, humidade relativa, preciptação, velocidade do vento e direção do vento; e, outros poluentes: O_3 e NO_2. Para o primeiro artigo as concentrações referem-se ao período de janeiro de 2005 a dezembro de 2009 e, para o segundo, ao período de janeiro de 2005 a dezembro de 2012, sendo a periodicidade diária e medida em $\mu g/m^3$.

4.3 *SOFTWARE* ESTATÍSTICO

A metodologia proposta e toda análise efetuada foi realizada por meio do *software* R (R Development Core Team, 2014). O R possui um grande número de procedimentos estatísticos convencionais, entre eles estão os modelos lineares, modelos de regressão não linear, análise de séries temporais, testes estatísticos paramétricos e não paramétricos, análise multivariada, etc. O *software* tem uma grande quantidade de funções para o desenvolvimento de ambiente gráfico e criação de diversos tipos de apresentação de dados. Alguns códigos necessários para o desenvolvimento do trabalho não estavam (e não estão) disponíveis no *software* R. Assim, quando necessário, novos códigos foram criados e podem ser adquiridos junto ao autor.

5 RESULTADOS E DISCUSSÕES

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

Principal component analysis in multivariate time series with conditional heteroscedasticity and long memory: an application to air pollution in the Greater Vitória Region, Espírito Santo, Brazill

Edson Zambon Monte[1], Valdério Anselmo Reisen[2], Josu Arteche González[3]

[1]Graduate Program in Environmental Engineering, Federal University of Espírito Santo, Espírito Santo, Brazil; e-mail: edsonzambon@yahoo.com.br

[2]Department of Statistics, Federal University of Espírito Santo, Espírito Santo, Brazil; e-mail: valderioanselmoreisen@gmail.com.

[3]Department of Economics, University of the Basque Country, Bilbao, Spain; e-mail: josu.arteche@ehu.es.

Abstract

This paper consider the classical principal component analysis (PCA) in multivariate time series with conditional heteroscedasticity and long memory, since ignore these features may lead to erroneous conclusions, if the PCA is adopted in techniques such as dimensionality reduction, factor analysis, cluster analysis, source identification, detection of outliers, linear and non-linear regression, among others. To avoid these problems, this paper proposes applying the PCA on the residuals of the VARFIMA-GARCH model. The proposed method is analyzed by means of Monte Carlo simulations and it is applied to the pollutant PM_{10}, for dimensional reduction and cluster analysis, in the Greater Vitória Region, Espírito Santo, Brazil.

Keywords: principal component analysis; temporal correlation; conditional heteroscedasticity; long memory; air pollution.

1 Introduction

Researches related to air quality have become increasingly important, because of the many health problems resulting from air pollution, such as asthma, rhinitis, eyes burning, fatigue, dry cough, heart and lung diseases, heart failure, etc. Authors as Brunekreef e Holgate (2002), Maynard (2004), WHO-World Health Organization (2005), Curtis et al. (2006), among others have shown the relationship between some pollutants (particulate matter with diameter smaller than 10 micrometers (PM_{10}), carbon monoxide (CO), sulfur dioxide (SO_2), nitrogen oxides (NO_X) and ozone (O_3)) and health problems. In 2012, for example, the deaths of 4.3 million people were attributed to air pollution (WHO-World Health Organization, 2014). In addition, air pollution contributes to the degradation of the environment, leading to the greenhouse effect.

In the recent studies related to air pollution much attention has been paid to the mathematical models named receptor models, which attempt to measure and analyses concentrations at their sources from a given site without reconstructing the dispersion patterns of the pollutants, such as PM_{10}, SO_2, among others. These models have mathematical and statistical tools which are

mainly used to provide the identification of the pollutant emission sources from chemical characteristics of the particles on the receiver and the pollutant emission sources (SEINFELD; PANDIS, 2006). In the literature, the majority of receptor models studied are: chemical balance of mass (CBM), multivariate analysis, principal component analysis (PCA), factor analysis (AF), multiple linear regression, cluster analysis and factoring positive matrix (FPM) (WATSON et al., 2002).

Regarding classical PCA, the central idea is to reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much of the variation in the data set as possible. This is achieved by transforming to a new set of variables, the principal components (PCs), which are uncorrelated, and which are ordered so that the first few retain most of the variation present in the original variables (JOLLIFFE, 2002). The PCA is based on algebraic theory of vectors by means of the eigenvalues and eigenvectors of the covariance or correlation matrix of the data. Apart from the purpose of dimensionality reduction, the PCA has been widely used in the following techniques: factor analysis, cluster analysis, canonical correlation analysis, discriminant analysis, linear and non-linear regression, among others.

In the context of the atmospheric pollution, for example, the PCA was adopted for: i) Pires et al. (2009) adopted PCA in the network redimensioning. This also was the goal of Zamprogno (2013); ii) Pio et al. (1991) used the PCA as cluster analysis to identify the main sources of atmospheric aerosol; iii) Liu (2009) adopted the PCA to detect outliers in the predicting of particulate matter concentrations (PM_{10}); iv) the sources identification is also part of the applications of PCA, for example: Guo, Wang e Louie (2004), Song et al. (2006) e Chavent et al. (2009); v) in linear and non-linear regression models, the principal components are used as predictors or as the response variable: Henry e Hidy (1979), Dominick et al. (2012) e Rajab, MatJafri e Lim (2013); and, vi) Souza et al. (2014) used the generalized additive model together with PCA to analyze the association between air pollutants and hospital admissions of children with respiratory problems.

It should be noted that, among the studies that adopted the PCA approach in the time domain, a common feature is to neglect the dependence of the data. However, in its classic form, this technique assumes that the data are independent (Anderson (2003) and Johnson e Wichern (2007)). According to Jolliffe (2002), the use of the PCA in multivariate time series requires some caution in its application, especially, if more than very weak dependence is present in the series. Zamprogno (2013), for instance, showed, empirically, spurious statistical analysis when applying PCA specifically in multivariate time series with more than a weak dependence property. In addition, the simulations demonstrated which, due to temporal correlation (serial and cross-correlation), the percentage of explanation migrated to the first principal component, causing a misleading reduction of the data set. Since the PCs are a linear combination of the original variables, the temporal correlation (serial and cross-correlation) of these series will translate to the PCs.

In order to solve this problem, Zamprogno (2013), in a study for the Greater Vitória Region (GVR), Espírito Santo, Brazil, used the vector fractionally integrated autoregressive moving

average (VARFIMA) to filter the time series and applied the PCA technique to the residuals (white noise) of the VARFIMA model. This idea is similar to that proposed by Matteson e Tsay (2011) and Hu e Tsay (2014). The filter allowed the correction and alleviation of the problems generated by temporal correlation. However, regarding the cross-correlation between the components, for example, even after applying the filter remained significant correlations. According to Zamprogno (2013), this may be due to the heterogeneity of the data structure.

In addition to temporal correlation, another important point has been overlooked in the PCA analysis, namely the conditional heteroscedasticity (HU; TSAY, 2014). In the analysis of multivariate time series with conditional heteroscedasticity, the results of PCA may be misleading, since the estimated autocovariance matrix and the eigenvalues and eigenvectors generated from the spectral decomposition tend to be biased (estimated incorrectly). To Matteson e Tsay (2011), in multivariate time series, the principal components are contemporaneously uncorrelated. However, lagged cross-correlations may be nonzero, conditional correlations may be nonzero, and cross-correlations of nonlinear transformations, such as the square process, may be non-zero. Therefore, the VARFIMA filter would not be enough to correct the problems of volatility on the classical PCA.

Furthermore, it is important to emphasize that, recently, several authors have studied the long dependency phenomenon (long memory) of time series, especially for univariate time series, as in Hosking (1981), Granger (1980), Granger (1981), Granger e Joyeux (1980) and Sowell (1992a), Sowell (1992b). In the time domain, the long dependency is characterized by a slow and significant decay of the autocorrelations, even for observations separated by long periods of time. This phenomenon has also received attention in multivariate time series ((SOWELL, 1989; CHUNG, 2001; CHUNG, 2002; SELA; HURVICH, 2008), among others) and may influence the results of principal component analysis, since it may affect the estimated autocovariance matrix.

The main goal of this paper was to investigate the use of the PCA technique in multivariate time series with conditional heteroscedasticity and long memory behavior, since ignoring these features may lead to erroneous conclusions, when the PCA is adopted in techniques such as dimensionality reduction, factor analysis, cluster analysis, identification of source, detection of outliers, linear and non-linear regression, among others. To solve these problems, this paper proposes to apply a multivariate VARFIMA-GARCH (generalized autoregressive conditional heteroscedasticity) filter to the data and then to use the PCA on the residuals of VARFIMA-GARCH model. Monte Carlo simulations were conducted to corroborate the results. Finally, given the features of temporal correlation, conditional heteroscedasticity and long memory of the air pollution series and the increasing adoption of principal component analysis to these time series, it is proposed that this new method be applied to air pollution series of the Greater Vitória Region (GVR), Espírito Santo, Brazil.

The rest of this paper is organized as follows. Section 2 presents the PCA technique in time series with conditional heteroscedasticity and long memory. Simulations of PCA in the presence of temporal correlation, conditional heteroscedasticity and long memory are shown

in Section 3. In Section 4, an application to air pollution data from GVR is presented. Some conclusions are described in Section 5.

2 PCA in multivariate time series with conditional heteroscedasticity and long memory

This section discusses the multivariate times series with conditional heteroscedasticity and long memory with the aim of using this process in PCA frameworks.

2.1 PCA in multivariate time series with conditional heteroscedasticity

2.1.1 Multivariate time series with conditional heteroscedasticity

Let $\boldsymbol{X}_t = \{X_{1,t}, \ldots, X_{k,t}\}^{'}$, $t \in \mathbb{Z}$, be a k-dimensional linear vector process given by the form

$$\boldsymbol{X}_t = \boldsymbol{\mu} + \sum_{j=0}^{\infty} \boldsymbol{\Psi}_j \boldsymbol{\epsilon}_{t-j}, \tag{1}$$

where $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_k]'$ is the mean vector; $\boldsymbol{\Psi}_0$ is the identity matrix of $k \times k$ dimension; $\boldsymbol{\Psi}_j$, $j = 1, ..., \infty$, are $k \times k$ coefficient matrices satisfying $\sum_{j=0}^{\infty} ||\boldsymbol{\Psi}_j||^2 < \infty$, wherein $||\boldsymbol{F}||$ denotes a norm for the matrix $\boldsymbol{F}$ such that $||\boldsymbol{F}||^2 = tr(\boldsymbol{F}'\boldsymbol{F})$. The matrix of coefficients $\boldsymbol{\Psi}_j$ are often referred as impulse responses. $\boldsymbol{\epsilon}_t = [\epsilon_{1,t}, \ldots, \epsilon_{k,t}]'$ is a vector white noise process which satisfies

$$\boldsymbol{\epsilon}_t = \boldsymbol{\Sigma}_t^{1/2} \boldsymbol{a}_t, \tag{2}$$

where $\boldsymbol{\Sigma}_t = Cov(\boldsymbol{\epsilon}_t | F_{t-1})$ is the conditional covariance matrix of $\boldsymbol{\epsilon}_t$; F_{t-1} is the sample information available at $t-1$; $\boldsymbol{a}_t$ is independent and identically distributed such that $E(\boldsymbol{a}_t) = \boldsymbol{0}$, $Cov(\boldsymbol{a}_t) = \boldsymbol{I}$ ($k \times k$ identity matrix); and, $\boldsymbol{\Sigma}_t^{1/2}$ denotes a positive square matrix of $\boldsymbol{\Sigma}_t$. If $E(\boldsymbol{\epsilon}_t \boldsymbol{\epsilon}_t' | F_{t-1}) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} > 0$, the process given in Equation (1) becomes a multivariate process with time-invariant covariance.

The process $\boldsymbol{X}_t$ has autocovariance matrix given by

$$\boldsymbol{\Gamma}_{\boldsymbol{X}}(h) = \sum_{j=-\infty}^{\infty} \boldsymbol{\Psi}_j \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \boldsymbol{\Psi}_{j+h}^{'},$$

where the *(i,j)*-th element of matrix $\boldsymbol{\Gamma}_{\boldsymbol{X}}(h)$ is $\gamma_{ij}(h) = E[(X_{i,t} - \mu_i)(X_{j,t+h} - \mu_j)]$, $i, j = 1, \ldots, k$; and, h is equal to the number of lags. $\boldsymbol{\Gamma}_{\boldsymbol{X}}(h)$ is absolutely summable if individually each of its elements form an absolutely summable sequence (HAMILTON, 1994).

The parametric vector autoregressive moving average process with generalized autoregressive conditional heteroscedasticity (VARMA-GARCH) is a particular case of the process defined by Equation (1) (for assumptions on VARMA-GARCH models, consult Ling e McAleer

(2003)). For example, if $\boldsymbol{X}_t$ follows a VAR(1)-GARCH$(1,1)$ process, its autocovariance matrix at lag zero is given by (more details are given in Remark 3, Section 2.1.2)

$$vec[\boldsymbol{\Gamma}_{\boldsymbol{X}}(0)] = [\boldsymbol{I}_{k^2} - (\boldsymbol{\Phi}_1 \otimes \boldsymbol{\Phi}_1)]^{-1}[\boldsymbol{I}_{k^2} - (\boldsymbol{A}_1 \otimes \boldsymbol{A}_1) - (\boldsymbol{B}_1 \otimes \boldsymbol{B}_1)]^{-1} vec[\boldsymbol{H}\boldsymbol{H}'], \quad (3)$$

where $vec(\boldsymbol{F})$ denotes the column-stacking vector of the matrix $\boldsymbol{F}$; $\boldsymbol{H}$, $\boldsymbol{A}_i$ and $\boldsymbol{B}_i$ are matrices $k \times K$, being $\boldsymbol{H}$ a lower triangular matrix; $\boldsymbol{I}$ is the k^2-dimensional identity matrix; and, $\otimes$ is the Kronecker product (for details, see Engle e Kroner (1995), Laurent, Bauwens e Rombouts (2006) and Lütkepohl (2005)). Here, $\boldsymbol{\Gamma}_{\boldsymbol{X}}(0)$ is based on BEKK method, where $\boldsymbol{\Sigma}_t$ is obtained as follows:

$$\boldsymbol{\Sigma}_t = \boldsymbol{H}\boldsymbol{H}' + \boldsymbol{A}_1 \boldsymbol{\epsilon}_{t-1} \boldsymbol{\epsilon}_{t-1}' \boldsymbol{A}_1' + \boldsymbol{B}_1 \boldsymbol{\Sigma}_{t-1} \boldsymbol{B}_1'. \quad (4)$$

Note that, from Equations (3) and (4), $\boldsymbol{\Gamma}_{\boldsymbol{X}}(0)$ is also a function of the matrices $\boldsymbol{A}_1$ and $\boldsymbol{B}_1$ which have the coefficients of the time-varying conditional covariance matrix $\boldsymbol{\Sigma}_t$. Therefore, the eigenvalues and eigenvectors of this covariance matrix are directed connected to the volatility of the process. Thus, in practical problems, the feature of time-varying conditional variance can not be ignored in the interpretation and analysis of the PCA or in any other multivariate technique. Ignoring the volatility in these tools may lead to erroneous conclusions and model misspecification such as wrong regressors, measurement errors, dimension reduction, among other problems in multivariate time series data. The use of the PCA technique in multivariate time series with time-varying conditional variance is discussed in the next section.

2.1.2 PCA considering conditional heteroscedasticity

The following remarks discuss some PCA properties when applied to multivariate time series.

Remark 1. Let $\boldsymbol{X}_t = \{X_{1,t}, \ldots, X_{k,t}\}'$ be the process defined in Equation (1), with $\boldsymbol{\epsilon}_t$ given by Equation (2), and consider the pairs of eigenvalues and eigenvectors $(\lambda_1, \boldsymbol{e}_1), (\lambda_2, \boldsymbol{e}_2), \ldots, (\lambda_k, \boldsymbol{e}_k)$ of the autocovariance matrix $\boldsymbol{\Gamma}_{\boldsymbol{X}}(0)$, where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k \geq 0$. Then, the i-th principal component is given by

$$Y_{i,t} = \boldsymbol{e}_i' \boldsymbol{X}_t = \boldsymbol{e}_{i,1} X_{1,t} + \boldsymbol{e}_{i,2} X_{2,t} + \cdots + \boldsymbol{e}_{i,k} X_{k,t}, \quad i = 1, 2, \ldots, k, \quad t \in \mathbb{Z}, \quad (5)$$

where, for $\forall\ h = 0$,

a) $Var(Y_{i,t}) = \boldsymbol{e}_i' \boldsymbol{\Gamma}_{\boldsymbol{X}}(0) \boldsymbol{e}_i = \lambda_i, \quad i = 1, 2, \ldots, k,$

b) $Cov(Y_{i,t}, Y_{j,t}) = \boldsymbol{e}_i' \boldsymbol{\Gamma}_{\boldsymbol{X}}(0) \boldsymbol{e}_j = 0, \quad i \neq j,$

and $\forall\ h \neq 0$,

c) $Cov(Y_{i,t}, Y_{i,t+h}) = \boldsymbol{e}_i' \boldsymbol{\Gamma}_{\boldsymbol{X}}(h) \boldsymbol{e}_i, \quad i = 1, 2, \ldots, k,$

d) $Cov(Y_{i,t}, Y_{j,t+h}) = \boldsymbol{e}_i' \boldsymbol{\Gamma}_{\boldsymbol{X}}(h) \boldsymbol{e}_j, \quad i \neq j.$

The results in (a) and (b) are the standard results of the PCA analysis, that is, the PCs are contemporaneously uncorrelated (no multicollinearity) (see, Anderson (2003) and Johnson e Wichern (2007)). However, items (c) and (d) show that the PCA preserves the autocorrelation structures present in the multivariate time series, that is, lagged conditional correlations and cross-correlations may be nonzero (see, also, Matteson e Tsay (2011)). These features are studied empirically in Section 3.

Remark 2. If some λ_i are equal, the choices of the corresponding coefficient vectors, $\boldsymbol{e}_i$, and hence, Y_{it}, are not unique (JOHNSON; WICHERN, 2007).

Remark 3. Consider that $\boldsymbol{X}_t$ is a stationary process VAR(1)-GARCH$(1,1)$, given by $\boldsymbol{X}_t = \boldsymbol{\Phi}_1 \boldsymbol{X}_{t-1} + \boldsymbol{\epsilon}_t$, with $\boldsymbol{\epsilon}_t = \boldsymbol{\Sigma}_t^{1/2} \boldsymbol{a}_t$ and $\boldsymbol{\Sigma}_t$ equal to Equation (4). Furthermore, suppose $\boldsymbol{Y}_t = \boldsymbol{e}' \boldsymbol{X}_t$ (vector of components), where $\boldsymbol{\Lambda}$ e $\boldsymbol{e}$ are the matrices of eigenvalues (diagonal matrix) and eigenvectors of $\boldsymbol{\Gamma}_{\boldsymbol{X}}(0)$, respectively. The autocovariance matrix of $\boldsymbol{X}_t$ can be represented recursively by (REINSEL, 2003; RAO et al., 2008)

$$\boldsymbol{\Gamma}_{\boldsymbol{X}}(h) = \begin{cases} \boldsymbol{\Gamma}_{\boldsymbol{X}}(0) = \boldsymbol{\Phi}_1 \boldsymbol{\Gamma}_{\boldsymbol{X}}'(1) + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \Rightarrow \boldsymbol{\Gamma}_{\boldsymbol{X}}(0) = \boldsymbol{\Phi}_1 \boldsymbol{\Gamma}_{\boldsymbol{X}}(0) \boldsymbol{\Phi}_1' + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}, & h = 0, \\ \boldsymbol{\Gamma}_{\boldsymbol{X}}(h) = \boldsymbol{\Gamma}_{\boldsymbol{X}}(h-1) \boldsymbol{\Phi}_1' \Rightarrow \boldsymbol{\Gamma}_{\boldsymbol{X}}(h) = \boldsymbol{\Gamma}_{\boldsymbol{X}}(0) \boldsymbol{\Phi}_1'^{h}, & h > 0. \end{cases} \tag{6}$$

Regarding $\boldsymbol{\Gamma}_{\boldsymbol{X}}(0)$, based on the Equation (4), and after some algebraic procedures, it is possible to get the following equation:

$$vec[\boldsymbol{\Gamma}_{\boldsymbol{X}}(0)] = [\boldsymbol{I}_{k^2} - (\boldsymbol{\Phi}_1 \otimes \boldsymbol{\Phi}_1)]^{-1} vec[\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}]. \tag{7}$$

Since $\boldsymbol{\Sigma}_t$ is based on a BEKK process, the term $vec[\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}]$ of Equation (7) is given by

$$vec[\Sigma_{\boldsymbol{\epsilon}}] = [\boldsymbol{I}_{k^2} - (\boldsymbol{A}_1 \otimes \boldsymbol{A}_1) - (\boldsymbol{B}_1 \otimes \boldsymbol{B}_1)]^{-1} vec[\boldsymbol{H}\boldsymbol{H}']. \tag{8}$$

Thus,

$$vec[\boldsymbol{\Gamma}_{\boldsymbol{X}}(0)] = [\boldsymbol{I}_{k^2} - (\boldsymbol{\Phi}_1 \otimes \boldsymbol{\Phi}_1)]^{-1} [\boldsymbol{I}_{k^2} - (\boldsymbol{A}_1 \otimes \boldsymbol{A}_1) - (\boldsymbol{B}_1 \otimes \boldsymbol{B}_1)]^{-1} vec[\boldsymbol{H}\boldsymbol{H}'], \tag{9}$$

i.e., the unconditional autocovariance matrix of $\boldsymbol{X}_t$, $\boldsymbol{\Gamma}_{\boldsymbol{X}}(0)$, is a function of the matrix of coefficients $\boldsymbol{\Phi}_1$, $\boldsymbol{A}_1$ and $\boldsymbol{B}_1$.

As,

$$\boldsymbol{\Gamma}_{\boldsymbol{Y}}(0) = \boldsymbol{e}' \boldsymbol{\Gamma}_{\boldsymbol{X}}(0) \boldsymbol{e} = \boldsymbol{e}' \boldsymbol{e} \boldsymbol{\Lambda} \boldsymbol{e}' \boldsymbol{e} = \boldsymbol{\Lambda}, \tag{10}$$

since the eigenvectors are normalized, i.e., $\boldsymbol{e}_i' \boldsymbol{e}_i = 1, \ i = 1, ..., k$, the temporal correlation and the conditional heteroscedasticity directly effect the autocovariance matrix of the principal

components, $\boldsymbol{\Gamma_Y}(0)$, changing, especially, the variances of the PCs, given by $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k \geq 0$. Besides, as $\boldsymbol{\Gamma_X}(0)\boldsymbol{\Phi}_1'^h$, the matrix $\boldsymbol{\Gamma_Y}(h)$ also depends of $\boldsymbol{\Phi}_1$, $\boldsymbol{A}_1$ and $\boldsymbol{B}_1$.

Remark 4. Given the Equation (1), suppose that the eigenvalues $(\lambda_1^*, \lambda_2^*, \ldots, \lambda_k^*)$ are from a white noise process, without conditional heteroscedasticity, such that $\boldsymbol{\Gamma_\epsilon}(0) = \boldsymbol{\Sigma_\epsilon}$. Define $(\lambda_1, \lambda_2, \ldots, \lambda_k)$ as the eigenvalues arising from a multivariate time series with conditional heteroscedasticity, $\boldsymbol{X}_t$, with autocovariance matrix $\boldsymbol{\Gamma_X}(0)$. Then, $\sum_{j=0}^{k} \lambda_i \geq \sum_{j=0}^{k} \lambda_i^*$, given that $Var(\boldsymbol{X}_t) \geq Var(\boldsymbol{\epsilon}_t)$ (see Result 8.2, Johnson e Wichern (2007, p. 432). Thus, the principal components of $\boldsymbol{X}_t$ have higher variability than the components generated by process $\boldsymbol{\epsilon}_t$.

In order to illustrate Remark 4, without loss of generality, consider a VAR(1)-GARCH$(1,0)$, with two variables ($k = 2$), and compare it with a white noise process with no conditional heteroscedasticity.

Let $\boldsymbol{\epsilon}_t = \{\epsilon_{1,t}', \epsilon_{2,t}'\}$ be a bi-dimensional white noise process, with time-invariant covariance matrix equal to

$$\boldsymbol{\Sigma_\epsilon} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}.$$

The characteristic polynomial $\boldsymbol{\Sigma_\epsilon}$ is given by

$$\lambda^2 - (\sigma_{11} + \sigma_{22})\lambda + \sigma_{11}\sigma_{22} - \sigma_{12}\sigma_{21}. \tag{11}$$

Thereby, according to Result 8.2, Johnson e Wichern (2007, p. 432), the total variance of the white noise process without volatility is given by $S_{\boldsymbol{\Sigma_\epsilon}} = \sigma_{11} + \sigma_{22}$.

Now, consider the VAR(1)-GARCH(1,0) model, given by $\boldsymbol{X}_t = \boldsymbol{\Phi}_1\boldsymbol{X}_{t-1} + \boldsymbol{\epsilon}_t$, with $\boldsymbol{\epsilon}_t = \boldsymbol{\Sigma}_t^{1/2}\boldsymbol{a}_t$, that is stationary and bi-dimensional, where $vec[\boldsymbol{\Gamma_X}(0)] = [\boldsymbol{I}_{k^2} - (\boldsymbol{\Phi}_1 \otimes \boldsymbol{\Phi}_1)]^{-1}[\boldsymbol{I}_{k^2} - (\boldsymbol{A}_1 \otimes \boldsymbol{A}_1)]^{-1}vec[\boldsymbol{HH'}]$. The matrices of the coefficients $\boldsymbol{\Phi}_1$ and $\boldsymbol{A}_1$ are given by:

$$\boldsymbol{\Phi}_1 = \begin{bmatrix} \phi_{11} & 0 \\ 0 & \phi_{22} \end{bmatrix}, \qquad \boldsymbol{A}_1 = \begin{bmatrix} a_{12} & 0 \\ 0 & a_{22} \end{bmatrix}.$$

Note that, in $vec[\boldsymbol{\Gamma_X}(0)]$, if the coefficients of the matrices $\boldsymbol{\Phi}_1$ and $\boldsymbol{A}_1$ are all equal to zero, then there is a situation with no temporal correlation and no conditional heteroscedasticity. Thus, $vec[\boldsymbol{\Gamma_X}(0)] = vec[\boldsymbol{HH'}]$. As $\boldsymbol{HH'}$, in this particular case, is featured as white noise process with time-invariant covariance, it is considered, for purpose of comparison, that $\boldsymbol{HH'} = \boldsymbol{\Sigma_\epsilon}$. Thus, $vec[\boldsymbol{\Gamma_X}(0)] = [\boldsymbol{I}_{k^2} - (\boldsymbol{\Phi}_1 \otimes \boldsymbol{\Phi}_1)]^{-1}[\boldsymbol{I}_{k^2} - (\boldsymbol{A}_1 \otimes \boldsymbol{A}_1)]^{-1}vec[\boldsymbol{\Sigma_\epsilon}]$, and

$$\boldsymbol{\Gamma_X}(0) = \begin{bmatrix} \frac{\sigma_{11}}{(1-\phi_{11}^2)(1-a_{11}^2)} & \frac{\sigma_{12}}{(1-\phi_{11}\phi_{22})(1-a_{11}a_{22})} \\ \frac{\sigma_{21}}{(1-\phi_{22}\phi_{11})(1-a_{22}a_{11})} & \frac{\sigma_{22}}{(1-\phi_{22}^2)(1-a_{22}^2)} \end{bmatrix}. \tag{12}$$

The characteristic polynomial $\boldsymbol{\Gamma}_{\boldsymbol{X}}(0)$ is given by

$$\lambda^2-\left(\frac{\sigma_{11}}{(1-\phi_{11}^2)(1-a_{11}^2)}+\frac{\sigma_{22}}{(1-\phi_{22}^2)(1-a_{22}^2)}\right)\lambda+\left(\frac{\sigma_{11}}{(1-\phi_{11}^2)(1-a_{11}^2)}\frac{\sigma_{22}}{(1-\phi_{22}^2)(1-a_{22}^2)}\right)-\left(\frac{\sigma_{12}}{(1-\phi_{11}\phi_{22})(1-a_{11}a_{22})}\right)\left(\frac{\sigma_{21}}{(1-\phi_{22}\phi_{11})(1-a_{22}a_{11})}\right). \tag{13}$$

The total variance of the process $\boldsymbol{X}_t$ is equal to $S_{\boldsymbol{\Gamma}_{\boldsymbol{X}}(0)} = \frac{\sigma_{11}}{(1-\phi_{11}^2)(1-a_{11}^2)} + \frac{\sigma_{22}}{(1-\phi_{22}^2)(1-a_{22}^2)}$. Thus, it is possible verify that $S_{\boldsymbol{\Gamma}_{\boldsymbol{X}}(0)} > S_{\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}}$. Additionally, observe that the coefficients that characterize the temporal correlation and the conditional heteroscedasticity directly influence the calculation of the eigenvalues of $\boldsymbol{\Gamma}_{\boldsymbol{X}}(0)$ and, consequently, the variance of each principal component (see Equation (10)).

2.2 PCA in multivariate time series with long memory

In this section some aspects of multivariate time series with long memory, along with their effects on the PCA, are presented. Here, the conditional volatility is not considered. The VARFIMA process is written based on (CHUNG, 2002) and the VARFIMA covariances were described according to Sela e Hurvich (2008).

2.2.1 VARFIMA processes

According to (CHUNG, 2002), the linear process $\boldsymbol{X}_t = \{X_{1,t}, \ldots, X_{k,t}\}'$, with $E(\boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_t'|F_{t-1}) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} > 0$ (see Equation (1)), presents long memory behavior if the impulses responses ($\boldsymbol{\Psi}_j$) converge at slow hyperbolic rates as $j \longrightarrow \infty$. More precisely, there are k memory parameters $d_1, d_2, \ldots, d_k$, whose values lie in $(0, 0.5)$ such that the impulse responses can be approximated by

$$\boldsymbol{\Psi}_j \backsim \boldsymbol{D}\left[\frac{1}{\Gamma(d)}j^{d-1}\right]\boldsymbol{\Pi}, \text{ as } j \longrightarrow \infty\,, \tag{14}$$

where $\Gamma(\cdot)$ is a gamma function and $\boldsymbol{\Pi}$ is a nonsingular $k \times k$ matrix of constants that are independent of j and may be functions of a smaller set of unknown parameters. The notation $\boldsymbol{D}[j^{d-1}/\Gamma(d)]$ represents $k \times k$ diagonal matrix with $j^{d_1-1}/\Gamma(d_1), \ldots, j^{d_k-1}/\Gamma(d_k)$ on the diagonal. In fact, for any univariate function f of a single variable, the notation $\boldsymbol{D}[f(d)]$ represents $k \times k$ diagonal matrix with $f(d_1), \ldots, f(d_k)$ on the diagonal. Also, the notation $\sim$ is defined as follows: given two sequences of matrices $\boldsymbol{U}_j$ and $\boldsymbol{V}_j$ of the same dimensions, $\boldsymbol{U}_j \sim \boldsymbol{V}_j$, as $j \longrightarrow \infty$, if $u_{ik,j}/v_{ik,j} \longrightarrow \infty$, as $j \longrightarrow \infty$, for i and k, where $u_{ik,j}$ and $v_{ik,j}$ are the (i,k)th elements of $\boldsymbol{U}_j$ and of $\boldsymbol{V}_j$, respectively. Let $\boldsymbol{\psi}_{i,j}'$ and $\boldsymbol{\pi}_i$ be the ith rows of $\boldsymbol{\Psi}_j$ and $\boldsymbol{\Pi}$, respectively; then, Equation (14) implies that $\boldsymbol{\psi}_{i,j}' \sim j^{d_1-1}\Gamma(d_i)^{-1}\boldsymbol{\pi}_i'$, as $j \longrightarrow \infty$, for all $i = 1, \ldots, k$. Note that the conditions on the memory parameters $d_i \in (0, 0.5)$, for $i = 1, \ldots, k$, ensure that the impulse responses are square-summable and the infinite sum in Equation (1) exists.

As in the case of univariate ARFIMA process, the type of dependence of the VARFIMA process follows the general definition of the memory of a univariate stationary time series process $y_t, t \in \mathbb{Z}$, with finite variance (see, for example, Taqqu (2003)). The following are common

definitions of short, long and intermediate dependency:

a) The process y_t has absolutely summable autocovariance. That is, the ACF decays at geometrical rate in the sense that there is an upper bound $\mid \rho_y(h) \mid \leq ba^h$, such that, $0 < b < \infty$, $0 < a < 1$ are constants. This implies that the process belongs to the short-range dependence class and the spectral density satisfies $0 < f_y(0) < \infty$. In the case of ARFIMA process, when $d_i = 0$ the resulting process is short-memory. If this is valid for all i, then the VARFIMA model becomes an VARMA model (short-memory);

b) The process does not have absolutely summable autocovariance. That is, the ACF decays at hyperbolic rate, given by $\rho_y(h) \simeq h^{-\alpha}$ (for some $0 < \alpha < 1$). In this situation, the process is defined to have long-memory property and $f_y(0) = \infty$. In the ARFIMA case, the long-memory is defined when $0 < d < 0.5$, whereas for the VARFIMA model there is at least one i such that $0 < d_i < 0.5$;

c) The intermediate memory of an ARFIMA model is defined when the memory parameter is in $[-0.5, 0)$.

The typical stochastic process that represent the long memory behavior of $\boldsymbol{X}_t$ is the k-dimensional stationary and invertible VARFIMA process, defined by

$$\boldsymbol{\Phi}(B)\boldsymbol{D}[(1-B)^d](\boldsymbol{X}_t - \boldsymbol{\mu}) = \boldsymbol{\Theta}(B)\boldsymbol{\epsilon}_t, \tag{15}$$

where B is the backshift operator; $\boldsymbol{\Phi}(B) = \boldsymbol{I} - \sum_{i=1}^{p} \boldsymbol{\Phi}_i B^i$ and $\boldsymbol{\Theta}(B) = \boldsymbol{I} + \sum_{i=1}^{q} \boldsymbol{\Theta}_i B^i$ are polynomial matrices with order p and q, respectively; $\boldsymbol{\epsilon}_t$ is a vector white noise process, with $E(\boldsymbol{\epsilon}_t \boldsymbol{\epsilon}_t' | F_{t-1}) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} > 0$; and, the operator $\boldsymbol{D}[(1-B)^d]$ is a $k \times k$ diagonal matrix with $(1-B)^{d_1} \ldots, (1-B)^{d_k}$ on the diagonal. (HOSKING, 1981) showed that if $-0.5 < d_i < 0.5$, for all $i = 1, \ldots, k$, then the process $\boldsymbol{X}_t$ is both stationary and invertible. The process $\boldsymbol{X}_t$ shall be called of VARFIMA$(d_1, \ldots, d_k)$ process to stress the central role of the fractional differencing parameters $d_1, d_2, \ldots, d_k$.

Also, the univariate version with $k = 1$ is referred to as the ARFIMA(p, d_1, q) process, which has been extensively studied by Hosking (1981), Granger (1980), Granger (1981), Granger e Joyeux (1980) and Sowell (1992a), Sowell (1992b). In Equation (15), when $\boldsymbol{\Phi}(B)$, $\boldsymbol{\Theta}(B)$ and $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ are both diagonal, the individual series, X_t, are uncorrelated univariate ARFIMA series. A stationary and invertible VARFIMA$(d_1, \ldots, d_k)$ process $\boldsymbol{X}_t$ has an infinite-order moving average representation as in Equation (1). It is straightforward to derive $\boldsymbol{\Psi}_j$ for the VARFIMA$(d_1, \ldots, d_k)$ process, and Chung (2001) has demonstrated the approximation Equation (14) with $\boldsymbol{\Pi} = \boldsymbol{\Phi}(1)^{-1}\boldsymbol{\Theta}(1)$.

Given the hyperbolic convergence rates of the impulse responses (Equation (14)), the cumulative impulse responses also progress at hyperbolic rates as indicated in the Lemma 1, of

Chung (2002), i.e.,

$$\sum_{k=0}^{j} \boldsymbol{\Psi}_k = \boldsymbol{D}\left[\frac{1}{\Gamma(d+1)} j^d\right]\boldsymbol{\Pi}, \text{ as } j \longrightarrow \infty \,. \tag{16}$$

Note that it is because the cumulative impulse response $\sum_{k=0}^{j} \boldsymbol{\Psi}_k$ diverges hyperbolically that the process $\boldsymbol{X}_t$ has long memory. These hyperbolically divergent cumulative impulse responses are very different from the geometrically convergent cumulative impulse responses that characterize the stationary and invertible vector ARMA model.

Furthermore, according to Chung (2002), the autocovariances $\boldsymbol{\Gamma}_{\boldsymbol{X}}(h) \equiv Cov(\boldsymbol{X}_t, \boldsymbol{X}_{t+h})$ of the long memory process $\boldsymbol{X}_t$ must also converge at hyperbolic rates. Hence, not only do the autocovariances $\gamma_{i,i}(h)$ of each $x_{i,t}$ die out slowly at a hyperbolic rate, the covariances $\gamma_{i,k}(h)$ between the current $x_{i,t}$ and the future $x_{k,t+h}$, for $i \neq k$, also taper off at hyperbolic rates. Hosking (1996) presents the result for the univariate case. More details about multivariate long memory process can be consulted in (CHUNG, 2002).

2.2.2 PCA considering VARFIMA covariances

Sela e Hurvich (2008) computed the autocovariances of VARFIMA provesses by means of methods presented by Bertelli e Caporin (2002). This authors computed the autocovariance sequence, $\gamma_X(j)$, of an univariate ARFIMA(p, d, q) model, writing the covariances as the infinite convolution of the autocovariances, $\xi(h)$, of a ARMA(p, q) process, and the autocovariances, $\varphi(j)$, of an ARFIMA$(0, d, 0)$. Thus, the ARFIMA(p, d, q) autocovariances can be written as:

$$\gamma_X(j) = \sum_{h=-\infty}^{\infty} \xi(h)\varphi(j-h). \tag{17}$$

As the autocovariances of an ARMA model decay exponentially fast, (BERTELLI; CAPORIN, 2002) recommended setting $\xi(h)$ to zero for $|h| > M$ for large M.

Based on the idea of Bertelli e Caporin (2002) the autocovariances of a multivariate long memory process can be derived. Firstly, the vector ARMA process is defined so that $\boldsymbol{\Phi}(B)\boldsymbol{Z}_t = \boldsymbol{\Theta}(B)\boldsymbol{\epsilon}_t$, with $Cov(\boldsymbol{\epsilon}_t) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$. The polynomials $\boldsymbol{\Phi}(B)$ and $\boldsymbol{\Theta}(B)$ have all of their roots outside the unit circle. Let $\boldsymbol{\xi}(h) = E(\boldsymbol{Z}_{t+h}\boldsymbol{Z}_t')$ be the autocovariance sequence of $\boldsymbol{Z}_t$. The VARFIMA model, $\boldsymbol{\Phi}(B)\boldsymbol{D}[(1-B)^d]\boldsymbol{X}_t = \boldsymbol{\Theta}(B)\boldsymbol{\epsilon}_t$, can be written as $\boldsymbol{D}[(1-B)^d]\boldsymbol{X}_t = \boldsymbol{Z}_t$ (moving average expansion). Then, the process $\boldsymbol{X}_t$ may be described with $\boldsymbol{X}_t = \sum_{j=0}^{\infty} \boldsymbol{C}_j \boldsymbol{Z}_{t-j}$, where $\boldsymbol{C}_j$ is a diagonal matrix with (k, k) elements equal to

$$\zeta(j, d_k) = \frac{\Gamma(j+d_k)}{\Gamma(j+1)\Gamma(d_k)}, \tag{18}$$

and $\Gamma(\cdot)$ is the gamma function. By means of the moving average expansion it is possible to find an expression for the autocovariances of $\boldsymbol{X}_t$:

$$\boldsymbol{\Gamma}_{\boldsymbol{X}}(h) = Cov(\boldsymbol{X}_t, \boldsymbol{X}_{t-h}),$$

$$\boldsymbol{\Gamma}_{\boldsymbol{X}}(h) = Cov\left(\sum_{i=0}^{\infty} \boldsymbol{C}_i \boldsymbol{Z}_{t-i} \sum_{j=0}^{\infty} \boldsymbol{C}_j \boldsymbol{Z}_{t-j-h}\right),$$

$$\boldsymbol{\Gamma}_{\boldsymbol{X}}(h) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \boldsymbol{C}_i Cov(\boldsymbol{Z}_{t-i}, \boldsymbol{Z}_{t-j-h}) \boldsymbol{C}_j',$$

$$\boldsymbol{\Gamma}_{\boldsymbol{X}}(h) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \boldsymbol{C}_i \boldsymbol{\xi}(h+j-i) \boldsymbol{C}_j'. \tag{19}$$

Now the focus is on the (k,l) entry of $\boldsymbol{C}_i\boldsymbol{\xi}(h+j-i)\boldsymbol{C}_j'$. Let $\xi_{k,l}(h)$ be the (k,l) entry of $\boldsymbol{\xi}(h)$, that is, $\xi_{k,l}(h) = E(Z_{k,t+h}Z_{l,t})$. Since $\boldsymbol{C}_i$ and $\boldsymbol{C}_j$ are both diagonal matrices, the (k,l) entry of $\boldsymbol{C}_i\boldsymbol{\xi}(h+j-i)\boldsymbol{C}_j'$ is $\zeta(i,d_k)\zeta(j,d_l)\xi_{k,l}(h+j-i)$. Adopting this, the (k,l) entry of $\boldsymbol{\Gamma}_{\boldsymbol{X}}(h)$ is given by

$$\gamma_{k,l}(h) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \zeta(i,d_k)\zeta(j,d_l)\xi_{k,l}(h+j-i),$$

$$\gamma_{k,l}(h) = \sum_{m=0}^{\infty}\sum_{j=m}^{\infty} \zeta(j-m,d_k)\zeta(j,d_l)\xi_{k,l}(h+m),$$

$$\gamma_{k,l}(h) = \sum_{m=0}^{\infty} \xi_{k,l}(h+m)\left(\sum_{j=m}^{\infty} \zeta(j,d_l)\zeta(j-m,d_k)\right), \tag{20}$$

where the second equality follows from the substitution $m = j - i$ and the interchange of the order summation. The inner sum is the cross-covariance of an ARFIMA$(0,d_k,0)$ process and an ARFIMA$(0,d_l,0)$ process that are driven by a common white noise. Writing this cross-covariance in terms of the integral of the cross-spectrum, the following can be found

$$\sum_{j=m}^{\infty} \zeta(j,d_l)\zeta(j-m,d_k) = \frac{1}{2\pi}\int_0^{2\pi} (1-e^{-i\lambda})^{-d_k}(1-e^{-i\lambda})^{-d_l}\, e^{i\lambda m}\, d\lambda,$$

$$\sum_{j=m}^{\infty} \zeta(j,d_l)\zeta(j-m,d_k) = \frac{\Gamma(1-d_k-d_l)(-1)^m}{\Gamma(1-d_k-m)\Gamma(1-d_l+m)},$$

$$\sum_{j=m}^{\infty} \zeta(j,d_l)\zeta(j-m,d_k) = \frac{\Gamma(1-d_k-d_l)\Gamma(d_k+m)}{\Gamma(d_k)\Gamma(1-d_k)\Gamma(1-d_l+m)}, \tag{21}$$

where the two equations follow from Sowell (1989). Observe that this agrees with the usual expression for the autocovariance of a ARFIMA$(0,d_k,0)$ process when $d_k = d_l$ (see, for instance, Theorem 13.2.1, Brockwell e Davis (1993)). For notational convenience, Equation (21) can be

written as

$$\varphi_{l,k}(h) = \frac{\Gamma(1-d_k-d_l)\Gamma(d_k+h)}{\Gamma(d_k)\Gamma(1-d_k)\Gamma(1-d_l+h)}. \tag{22}$$

Note that $\varphi_{k,l}(h) = \varphi_{l,k}(-h)$, as must be true for any cross-covariances.

Based on Bertelli e Caporin (2002), it is considered the finite approximation to the outer sum in Equation (20), by setting $\xi(m) = 0$ for all $|m| > M$. As the autocovariance sequence of a vector ARMA decays exponentially fast, it can choose a relatively small M to approximate the process to a given degree of accuracy. The choice of M depends on the parameters of the ARMA process; if $\boldsymbol{\xi}(h)$ is the autocovariance sequence of an MA(q) process, then $M = q$ can chosen to compute the autocovariances exactly. Otherwise, an M which accounts for how quickly the autocovariances of the vector ARMA process decay must be chosen.

Remark 5. Let $\boldsymbol{X}_t$ be a long memory process VARFIMA(p, d, q), $\boldsymbol{\Phi}(B)\boldsymbol{D}[(1-B)^d]\boldsymbol{X}_t = \boldsymbol{\Theta}(B)\boldsymbol{\epsilon}_t$, that follows the properties described in Subsection 2.2.1. Furthermore, assume that $h = 0$ in Equation (20), that is, $\boldsymbol{\Gamma}_{\boldsymbol{X}}(0) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\boldsymbol{C}_i\boldsymbol{\xi}(j-i)\boldsymbol{C}_j'$. As the classic PCA analysis is based in the spectral decomposition of the autocovariance matrix at lag zero, that depends of $\boldsymbol{C}$ (diagonal matrix formed by elements of Equation (18)), the long memory phenomenon influences the autocovariance matrix and hence the PCA analysis.

Proof. Without loss of generality, based on Equation (15), consider a simple case where $\boldsymbol{X}_t$ is generated by a VARFIMA$(0, d_k, 0)$ model,

$$\boldsymbol{X}_t = \boldsymbol{\mu} + \boldsymbol{D}^{-1}[(1-B)^{d_k}]\boldsymbol{\epsilon}_t, \tag{23}$$

where $\boldsymbol{X}_t$, $\boldsymbol{\mu}$, $\boldsymbol{D}[(1-B)^{d_k}]$ and $\boldsymbol{\epsilon}_t$ satisfy the same conditions of Equation (15).

Equation (23) can be described as

$$\boldsymbol{X}_t = \boldsymbol{\mu} + \sum_{j=0}^{\infty}\boldsymbol{C}_j\boldsymbol{\epsilon}_{t-j}, \tag{24}$$

that can be called moving average expansion of a VARFIMA$(0, d_k, 0)$, because $\boldsymbol{\epsilon}_t$ is a white noise process. Therefore, $\boldsymbol{\Gamma}_{\boldsymbol{X}}(h) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\boldsymbol{C}_i\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}\boldsymbol{C}_j'$. The (k,l) entry of $\boldsymbol{\Gamma}_{\boldsymbol{X}}(h)$ for a VARFIMA $(0, d_k, 0)$ process can be represented by $\gamma_{kl}(h) = \sigma_{kl}\varphi_{kl}(h)$. Then, denoting $\boldsymbol{\Gamma}_{\boldsymbol{X}}(h) = E(\boldsymbol{X}_{t+h}\boldsymbol{X}_t')$, $h = -T+1, -T+2, \ldots, 0, 1, 2, \ldots, T-2, T-1$, the autocovariance function of $\boldsymbol{X}_t$ of Equation (23) at lag h is

$$\boldsymbol{\Gamma}_{\boldsymbol{X}}(h) \equiv \begin{bmatrix} \sigma_{11}\varphi_{11}(h) & \cdots & \sigma_{1l}\varphi_{1l}(h) \\ \sigma_{21}\varphi_{21}(h) & \cdots & \sigma_{2l}\varphi_{2l}(h) \\ \vdots & \ddots & \vdots \\ \sigma_{k1}\varphi_{k1}(h) & \cdots & \sigma_{kk}\varphi_{kk}(h) \end{bmatrix}, \tag{25}$$

where $\sigma_{k,l}$ represents the (k,l)-th element of $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$; and, in mean diagonal $k = l$. For more

details, see Tsay (2010). In a particular case, Hosking (1981) presented the effects of the long dependence on the autocovariances of the univariate ARIMA$(0, d, 0)$ process.

Now, consider $\mathbf{\Gamma}_{\boldsymbol{X}}(0)$ with two variables, $k = 2$. Then,

$$\mathbf{\Gamma}_{\boldsymbol{X}}(0) \equiv \begin{bmatrix} \sigma_{11}\varphi_{11}(0) & \sigma_{12}\varphi_{12}(0) \\ \sigma_{21}\varphi_{21}(0) & \sigma_{22}\varphi_{22}(0) \end{bmatrix}. \tag{26}$$

Therefore, the characteristic polynomial $\mathbf{\Gamma}_{\boldsymbol{X}}(0)$ is given by

$$\lambda^2 - [\sigma_{11}\varphi_{11}(0) + \sigma_{22}\varphi_{22}(0)]\lambda + \sigma_{11}\varphi_{11}(0) \times \sigma_{22}\varphi_{22}(0) - \sigma_{12}\varphi_{12}(0) \times \sigma_{21}\varphi_{21}(0). \tag{27}$$

Note that, based on Result 8.2, Johnson e Wichern (2007, p. 432), the total variance of the VARFIMA$(0, d_k, 0)$ is given by $S_{\mathbf{\Gamma}_{\boldsymbol{X}}(0)} = \sigma_{11}\varphi_{11}(0) + \sigma_{22}\varphi_{22}(0)$. It is possible verify that if in Equation (24), $j > 0$, then $S_{\mathbf{\Gamma}_{\boldsymbol{X}}(0)} > S_{\mathbf{\Sigma}_{\boldsymbol{\epsilon}}}$, which demonstrates the increased of variability of the principal components in the presence of long memory. Furthermore, the long memory parameters directly effect the autocovariance matrix and hence the calculation of the eigenvalues of $\mathbf{\Gamma}_{\boldsymbol{X}}(0)$ and, consequently, the variance of each principal component. □

3 Simulations

This section aimed at demonstrating, through simulations, some PCA results in multivariate time series with conditional heteroscedasticity (volatility) and long memory, using Monte Carlo simulations. All simulations were generated three-dimensional time series (i.e., k = 3).

3.1 Simulations considering temporal correlation

The purpose of this subsection is simulate the results of the PCA for series that present serial correlation and cross-correlation. In all simulations, the covariance matrix for the white noise process $\boldsymbol{\epsilon}_t$ ($(\mathbf{\Gamma}_{\boldsymbol{\epsilon}}(0))$, considering the absence of volatility, and denoted by Model 1, is given by Table 1.

Table 1: Covariance matrix for white noise process without volatility (Model 1)

127,41	30,59	47,44
30,59	58,79	33,89
47,44	33,89	64,18

Three vector autoregressive models of order one (VAR(1)) were simulated (Model 2, Model 3 and Model 4), whose matrices of coefficients are presented in Table 2. Model 2 presents weak serial correlation. Model 3 is a process with strong serial correlation (diagonal) and low cross-correlation (off-diagonal). Finally, Model 4 considers a strong structure of serial correlation and cross-correlation, for example, $\phi_{23} = 0,8$. To see assumptions about VAR models to consult Lütkepohl (2005), for example. According to coefficients in Table 2, random samples of a size

equal to 1000 were generated. The random samples were replicated 1000 times for each model (2, 3 and 4).

Table 2: Matrices of $\mathbf{\Phi}_1$ for VAR(1) processes

$\mathbf{\Phi}_1$ (Model 2)			$\mathbf{\Phi}_1$ (Model 3)			$\mathbf{\Phi}_1$ (Model 4)		
0,3	0,0	0,1	0,7	0,0	0,0	0,7	0,0	0,3
0,0	0,2	0,0	0,0	0,3	0,0	0,0	0,3	0,0
0,0	0,0	0,1	0,2	0,0	0,5	0,2	0,8	0,6

Table 3 shows the autocovariance matrices ($\mathbf{\Gamma}_{\boldsymbol{X}}(0)$) for each VAR$(1)$ model. Note that, for Model 2, where there is weak temporal correlation, the autocovariances increased, but they were similar to those of the Model 1 (white noise). However, for Model 4, where the temporal correlations are strong in the diagonal and off-diagonal, the autocovariances rose significantly, as expected.

Table 3: Autocovariances ($\mathbf{\Gamma}_{\boldsymbol{X}}(0)$) for VAR(1) processes

$\mathbf{\Gamma}_{\boldsymbol{X}}(0)$ (Model 2)			$\mathbf{\Gamma}_{\boldsymbol{X}}(0)$ (Model 3)			$\mathbf{\Gamma}_{\boldsymbol{X}}(0)$ (Model 4)		
137,81	29,52	44,84	254,19	41,92	130,29	661,71	240,06	626,98
29,52	60,13	30,89	41,92	66,56	44,42	240,06	149,71	277,51
44,84	30,89	63,23	130,29	44,42	136,84	626,98	277,51	783,58

The results in Table 4 are related to the eigenvalues for Models 1, 2, 3 and 4. The table also presents the percentages of variability of each eigenvalue in the total variability of each model, given by: $\frac{\lambda_i}{\sum_{i=1}^{k}\lambda_i}$, where λ_i is a eigenvalue of $\mathbf{\Gamma}_{\boldsymbol{X}}(0)$. It is observed that the eigenvalues and the percentages of variability of Model 2 were very close to Model 1 (white noise). This is due to the fact that the coefficients of the Model 1 are relatively small (weak temporal correlation).

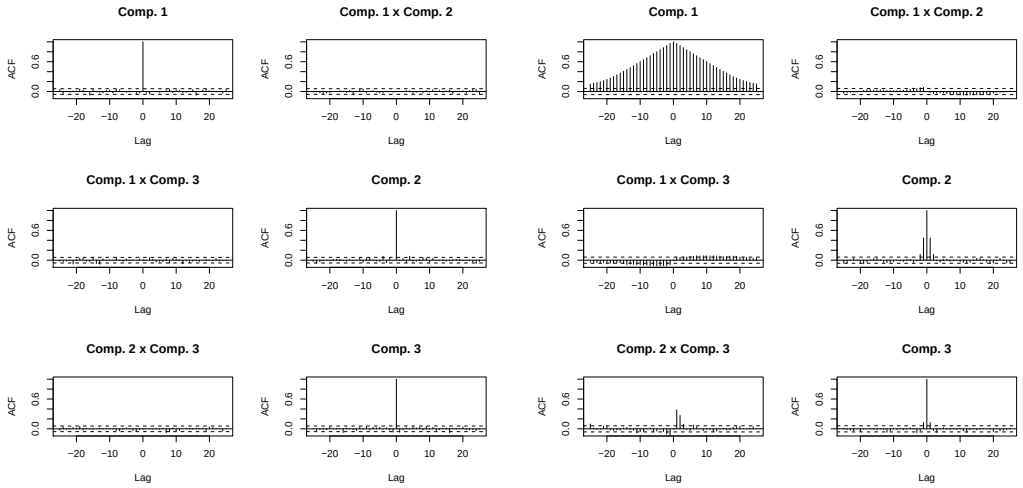
In the case of Model 4, which has the strongest correlation structure, it was found that the eigenvalues had a large increase, especially with respect to the first eigenvalue, λ_1. Furthermore, it is possible note that the percentage of variation explained by the first principal component was significantly increased when a more complex temporal correlation structure was considered. Thus, the temporal correlation tends to drive much of the variability of a data set for the first principal component, which causes a spurious (misleading) reduction of the dimension of the space generated by multivariate temporal processes. These results were similar to those of Zamprogno (2013).

Table 4: Eigenvalues of $\mathbf{\Gamma}_{\boldsymbol{X}}(0)$ for VAR(1) processes and percentege of variability

Models	λ_1	λ_2	λ_3	% λ_1	% λ_2	% λ_3
1	169,72	54,82	25,83	67,78	21,90	10,32
2	173,12	58,70	29,34	66,28	22,48	11,24
3	350,82	67,64	39,12	76,67	14,78	8,55
4	1455,62	94,20	45,18	91,26	5,91	2,83

Finally, Figures 1(a) and 1(b) show the autocorrelation and cross-correlation structures of

the principal components generated by Models 1 (white noise) and 4 (strong temporal correlation), respectively. As expected, in the case of the Model 1, the principal components were not autocorrelated and the cross-correlation was null. However, for Model 4, it is noted that the temporal correlation structure of multivariate stochastic process was transferred to the principal components.



(a) White noise (b) Model 4

Figure 1: Autocorrelation and cross-correlation of the principal components generated by white noise process and by Model 4.

3.2 Simulations considering temporal correlation and conditional heteroscedasticity

As previously described, the classical PCA does not capture the conditional heteroscedasticity (volatility) of the time series, because this technique assumes that the variance of the innovations (random shocks) is constant over time. In addition, the conditional heteroscedasticity, besides effecting the conditional variance, impacts in the unconditional variance, $\mathbf{\Gamma}_{\boldsymbol{X}}(0)$. Therefore, in this subsection, the Monte Carlo simulations are performed considering processes with temporal correlation and conditional heteroscedasticity, namely here VAR-GARCH.

To simulate the VAR-GARCH processes, the presence of temporal correlation and conditional heteroscedasticity (i.e., $\boldsymbol{\epsilon}_t = \mathbf{\Sigma}_t^{1/2}\boldsymbol{a}_t$) in the model presented in Equation (1) was considered . The temporal correlation was modeled by a VAR(1) process with the coefficients of $\mathbf{\Phi}_1$ equal to Model 2, Table 2. To model the conditional heteroscedasticity, a GARCH (1,1), with $\mathbf{\Sigma}_t$ described by a BEKK, was adopted as Equation (4). Moreover, the matrix of the white noise process $\boldsymbol{a}_t$, $(\mathbf{\Gamma}_{\boldsymbol{a}}(0))$, was equal to Table 1.

The coefficients of $\boldsymbol{A}_1$ and $\boldsymbol{B}_1$ of the GARCH-BEKK part (see Equation (4)) are presented in Table 5. From these coefficients were simulated three models, denoted by: Model 5, Model 6 and Model 7. Random samples of a size equal to 1000 were generated, and replicated 1000

times for each model. Model 6 presents the weaker volatility structure, whereas Model 7 has a high volatility in or outside the main diagonal.

Table 5: Matrices $\boldsymbol{A}_1$ and $\boldsymbol{B}_1$ for the part GARCH(1,1)

Model 5						Model 6						Model 7					
$\boldsymbol{A}_1$			$\boldsymbol{B}_1$			$\boldsymbol{A}_1$			$\boldsymbol{B}_1$			$\boldsymbol{A}_1$			$\boldsymbol{B}_1$		
0,3	0,0	0,0	0,2	0,0	0,0	0,5	0,0	0,4	0,3	0,0	0,0	0,6	0,0	0,0	0,2	0,0	0,1
0,0	0,1	0,0	0,0	0,1	0,0	0,0	0,3	0,0	0,0	0,2	0,0	0,0	0,7	0,8	0,8	0,2	0,0
0,0	0,0	0,2	0,0	0,0	0,2	0,0	0,0	0,4	0,0	0,0	0,5	0,2	0,0	0,4	0,0	0,0	0,3

In Table 6, the autocovariance matrices ($\boldsymbol{\Gamma}_{\boldsymbol{X}}(0)$) of each VAR$(1)$-GARCH$(1,1)$ model are shown. Observe that, with the increasing of the volatility, the autocovariances rose when comparing to Models 1 (white noise without volatility) and 2 (process with temporal correlation), according to the described in Subsection 2.1.1.

Table 6: Autocovariances ($\boldsymbol{\Gamma}_{\boldsymbol{X}}(0)$) of VAR$(1)$-GARCH$(1,1)$ processes

$\boldsymbol{\Gamma}_{\boldsymbol{X}}(0)$ (Model 5)			$\boldsymbol{\Gamma}_{\boldsymbol{X}}(0)$ (Model 6)			$\boldsymbol{\Gamma}_{\boldsymbol{X}}(0)$ (Model 7)		
282,41	45,11	143,44	542,52	65,74	271,40	250,96	303,61	109,45
45,11	74,43	48,67	65,74	76,49	59,38	303,61	1996,66	270,94
143,44	48,67	150,22	271,40	59,38	250,91	109,45	270,94	117,49

The eigenvalues for Models 5, 6 and 7 and the percentage of variability of each eigenvalue in the total variability of each model are demonstrated in Table 7. It can observed that even when the model already presents temporal correlation, the inclusion of volatility increased the values of the eigenvalues. Besides, the percentage of variability of the model was strongly directed for the first principal component (especially for Model 7, with strong volatility) .

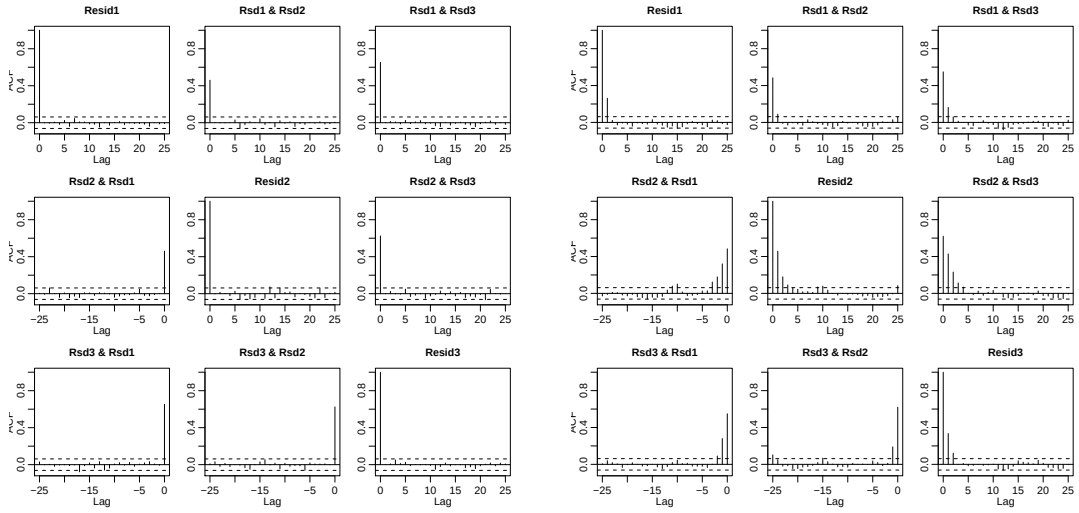
Table 7: Autocovariances $\boldsymbol{\Gamma}_{\boldsymbol{X}}(0)$ for VAR(1)-GARCH$(1,1)$ processes and percentage of variability

Models	λ_1	λ_2	λ_3	% λ_1	% λ_2	% λ_3
5	387,47	76,19	43,41	76,41	15,03	8,56
6	716,61	97,50	55,81	81,38	11,21	6,42
7	2089,30	225,48	50,33	88,34	9,53	2,13

Since, for models with serial correlation and conditional heteroscedasticity, a large part of the variability of simulated dataset was directed for the first principal component, similar to Zamprogno (2013), the VAR(1) filter was applied to the Model 7, in order to eliminate this problem. Figure 2(a) presents the autocorrelation and cross-correlation functions of the residuals of Model 7. As can be seen, the VAR(1) filter eliminated the serial correlation and the cross-correlation, this is, the residuals of Model 7 presented temporal correlation structure similar to a multivariate white noise process.

However, it is possible to note by Figure 2(b) that the squared residuals of Model 7 had several significant correlations, i.e., outside the confidence interval. In addition, the ARCH-LM

test revealed conditional heteroscedasticity in all residuals of Model 7. This shows that the VAR filter eliminated the temporal correlation, but the volatility was not filtered. Therefore, as the PCA does not take into account the conditional heteroscedasticity, the interpretation of the results may be misleading, even after using the VAR filter (VAR(1)). Is is important to say that the PCA was applied in the residuals of Model 7 (results available upon request), but the percentage of explanation of the first principal component remained unchanged, at 88%. Thus, in Subsection 2.1.2 will only be showed the PCA results in the presence of conditional heteroscedasticity.



(a) Residuals of Model 7.

(b) Squared residuals of Model 7.

Figure 2: Autocorrelation and cross-correlation of residuals and squared residuals of Model 7.

3.3 Simulations considering conditional heteroscedasticity

In this subsection the goal is to verify the impacts of conditional heteroscedasticity on the classical PCA, when the multivariate process is said to be a white noise. Thus, the coefficients of the matrix of the VAR(1) process, $\mathbf{\Phi}_1$, were considered null and the conditional heteroscedasticity of the innovations $\boldsymbol{\epsilon}_t$ were modeled via GARCH$(1,1)$ process, by BEKK method. The coefficients of $\boldsymbol{A}_1$ and $\boldsymbol{B}_1$ were equal to those in Table 5. Model 8 has the weaker volatility structure. Model 10 presents a higher volatility within or outside the main diagonal. Random samples of size a equal to 1000 were generated, and replicated 1000 times for each model.

Table 8: Matrices $\boldsymbol{A}_1$ and $\boldsymbol{B}_1$ for the part GARCH$(1,1)$

Model 8						Model 9						Model 10					
$\boldsymbol{A}_1$			$\boldsymbol{B}_1$			$\boldsymbol{A}_1$			$\boldsymbol{B}_1$			$\boldsymbol{A}_1$			$\boldsymbol{B}_1$		
0,3	0,0	0,0	0,2	0,0	0,0	0,5	0,0	0,4	0,3	0,0	0,0	0,6	0,0	0,0	0,2	0,0	0,1
0,0	0,1	0,0	0,0	0,1	0,0	0,0	0,3	0,0	0,0	0,2	0,0	0,0	0,7	0,8	0,8	0,2	0,0
0,0	0,0	0,2	0,0	0,0	0,2	0,0	0,0	0,4	0,0	0,0	0,5	0,2	0,0	0,4	0,0	0,0	0,3

Table 9 demonstrates the covariance matrices ($\mathbf{\Gamma}_{\boldsymbol{X}}(0)$) for each GARCH$(1,1)$ model. It is worth noting that, in this case, there was an increase in the values of the autocovariances when compared to the case of white noise and no volatility (Model 1).

Table 9: Autocovarinces ($\mathbf{\Gamma}_{\boldsymbol{X}}(0)$) for GARCH$(1,1)$ processes

$\mathbf{\Gamma}_{\boldsymbol{X}}(0)$ (Model 8)			$\mathbf{\Gamma}_{\boldsymbol{X}}(0)$ (Model 9)			$\mathbf{\Gamma}_{\boldsymbol{X}}(0)$ (Model 10)		
153,57	33,37	51,19	321,29	49,34	102,23	208,09	251,21	100,11
33,37	61,75	34,60	49,34	68,78	42,11	251,21	1756,61	250,98
51,19	34,60	67,34	102,23	42,11	104,10	100,11	250,98	117,49

In Table 10, it can be observed the eigenvalues for Models 8, 9 and 10, as well as the percentage of variability of each eigenvalue in the total variability of each model. It is found that even when the process followed the multivariate white noise, the presence of conditional heteroscedasticity increased the value of the eigenvalues and again directed much of the variability of the data set for first principal component.

Table 10: Eigenvalues of $\mathbf{\Gamma}_{\boldsymbol{X}}(0)$ for GARCH$(1,1)$ models and pencentage of variability

Models	λ_1	λ_2	λ_3	% λ_1	% λ_2	% λ_3
8	192,95	61,28	28,43	68,26	21,68	10,06
9	394,24	70,81	29,12	79,78	14,33	5,89
10	1836,74	196,84	46,36	88,31	9,46	2,23

3.4 Simulations considering long memory

Finally, the aim of this subsection is to briefly verify the results of the PCA in time series with long memory phenomenon. For simulations, the following were considered: a) a white noise model (without volatility) equal to Table 1, but with long memory (Model 11, in Table 11); and, b) VARFIMA$(1,d,0)$ models, without conditional heteroscedasticity (Models 12, 13 and 14). The VAR(1) part adopted the coefficients of $\mathbf{\Phi}_1$ equal to the models presented in Table 2, respectively. The fractionally integrated parameters were equal to $d = (0.4, 0.3, 0.2)$ for all Models (11, 12, 13 and 14). Random samples of a size equal to 1000 were generated, and replicated 1000 times for each model.

The results revealed an increase of the values of the autocovariances (results available upon request) when compared with all models of Tables 1 and 2. Again, there was an increase of the eigenvalues and the variability was directed for first principal component (Table 11). Furthermore, the long memory seems to strengthen the effects of temporal correlation with respect to direct the percentage of variability for the first principal component.

Based on simulations, the temporal correlation, the conditional heteroscedasticity and the long memory may cause two main problems for the use of principal component analysis: a) the percentage of variability of the multivariate stochastic process may be transferred for the first

Table 11: Eigenvalues of $\mathbf{\Gamma}_{\boldsymbol{X}}(0)$ for long memory models and pencentage of variability

Models	λ_1	λ_2	λ_3	% λ_1	% λ_2	% λ_3
11	244.66	71.66	30.85	70.47	20.64	8.89
12	317.42	95.64	36.97	70.53	21.25	8.22
13	1189.42	133.28	50.00	86.65	9.71	3.64
14	1811.04	75.32	27.22	94.64	3.94	1.42

principal component; and, b) the principal components may present serial correlation and cross-correlation. To solve these problems, this paper proposes to apply a multivariate VARFIMA-GARCH filter to the data and then to use the PCA on the residuals of VARFIMA-GARCH model.

It is important to mention that, for models with temporal correlation, conditional heteroscedasticity and long memory simulated in this paper, as expected, the VARFIMA-GARCH filter was effective, removing these features from the models. The application of PCA on the residuals of VARFIMA-GARCH models generated results consistent with the assumptions of the classical PCA theory. In the next section the results of an application for the pollutant PM_{10} are demonstrated, in the GVR, Espírito Santo, Brazil.

4 Application

4.1 Study area

The study area included the GVR, Espírito Santo, Brazil, located on the south coast of the Atlantic Ocean [latitude 20°19 S (South), longitude 40°20 W (West)]. Because it is situated in the coastal region, the GVR has hot tropical climate (Aw), with mild and dry winters, and rainy and hot summers. Average temperatures range between 24°C and 30°C. The prevailing winds are from North/Northeast in the spring-summer, undergoing changes during autumn and winter due to the positioning of the high pressure system (South Atlantic Subtropical High Pressure) closer to the continent, allowing changes in the direction of the prevailing wind, which starts to vary between the South and West directions.

The PM_{10} concentrations, with daily frequency (daily average of 24 hours in $\mu g/m^3$), were collected from Air Quality Automatic Monitoring Network (AQAMN), belonging to the State Institute of Environment and Water Resources (IEMA) and refer to eight monitoring stations, namely: Laranjeiras, Carapina, Jardim Camburi (Camburi), Enseada do Suá (Sua), Vitória-Centro (VixCentro), Vila Velha-Ibes (Ibes), Vila Velha-Centro (VVCentro) and Cariacica. The period of analysis was from January 2005 to December 2009.

It is important to remember that the AQAMN, in addition to PM_{10}, measures the following pollutants: total suspended particles (TPS); SO_2; CO; NO_X; hydrocarbons (HC); and, O_3. The AQAMN also monitors some meteorological parameters, such as: wind direction (DV); wind speed (VV); relative humidity (RH); precipitation (PP); atmospheric pressure (P); temperature

(T); and, solar radiation (R).

4.2 Application to air pollution

This section presents an application of the methodology discussed in Section 2 for the pollutant PM_{10}. In Figure 3, the time series of the pollutant PM_{10} are shown. It is possible note that for all stations there is a great variability in the data throughout the study period. As the series presented atypical observations (named here as outliers), based on Ma e Genton (2000), the sample robust autocorrelation functions can be seen in Figure 4. It appears that there are significant autocorrelations, even to distant lags. The slow decay of autocorrelation functions indicate a long-memory behavior of the series (REISEN et al., 2014). In addition, the sample robust ACFs show possible seasonal pattern in the period equal to seven.

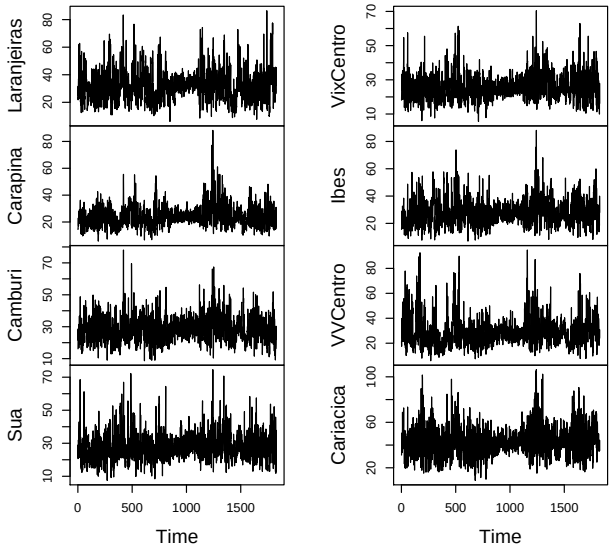


Figure 3: PM_{10} concentrations, in $(\mu g/m^3)$, from 2005.01.01 to 2009.12.31.

Furthermore, the robust ACFs of the squared series and the ARCH-LM test (test to verify the presence of volatility) revealed that there is conditional heteroscedasticity (volatility) of PM_{10} concentrations, a feature expected for time series of air pollution (results available upon request).

The main focus of this research is the multivariate time series with conditional heteroscedasticity and long memory. However, given the characteristics of seasonality of the pollutant PM_{10}, in the application, before use the PCA, the seasonal VARFIMA-GARCH filter was adopted on the original data. Reisen et al. (2014) explored 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. As the PM_{10} concentrations present atypical observations (here, considered as outliers) and the outliers affect the autocorrelation structure of the series, the

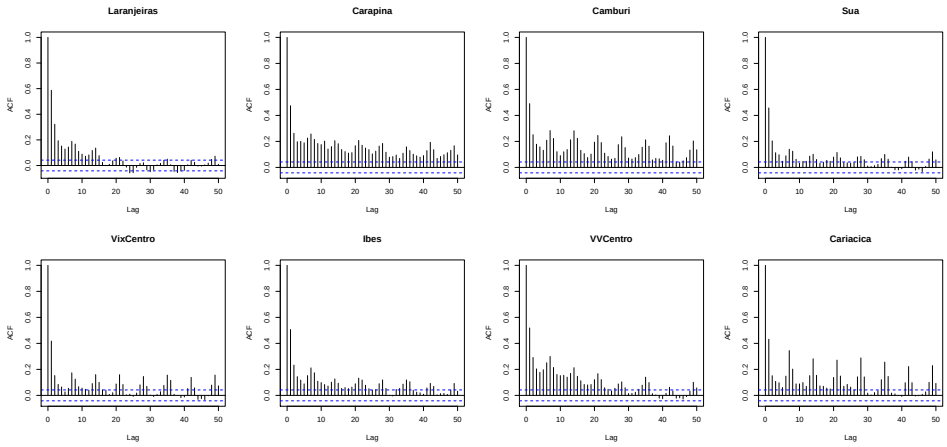


Figure 4: Robust ACF of PM_{10} concentrations.

robust periodogram was used to calculate the non-seasonal and seasonal memory parameters, such as Molinares, Reisen e Cribari-Neto (2009) and Reisen et al. (2015)).

The estimates of robust d and D were obtained considering different values of bandwidths, $M = n^{\alpha}$, where $0 < \alpha < 1$ and n is the size of the series. In the semiparametric approach, the choice of estimates depends on the size of the bandwidth. For example, large M gives more bias to the fractional estimators, when there are short-run components in the model. It can be noted that increasing M leads to fractional estimators with less power. In this work, α was chosen to be equal to 0.86. In Table 12, there are the estimated values for d and D and its respective standard deviations. Note that, the stationarity condition was guaranteed for all stations (series), since $0 < |\hat{d} + \hat{D}| < 0.5$ and $|\hat{D}| < 0.5$.

Table 12: Estimates of fractional robust parameters d and D for different stations (series)

Stations	$\hat{d}$	$sd(\hat{d})$	$\hat{D}$	$sd(\hat{D})$
Laranjeiras	0.3113	0.0403	*	*
Carapina	0.3108	0.0297	0.1572	0.0398
Camburi	0.2685	0.0315	0.0924	0.0381
Sua	0.2949	0.0347	0.1572	0.0558
VixCentro	0.2486	0.0316	0.0747	0.0425
Ibes	0.2166	0.0346	*	*
VVCentro	0.3096	0.0343	0.1198	0.0479
Cariacica	0.2478	0.0288	0.1679	0.0387

Note: * In this case, H_0: $D = 0$ was not rejected.

In Table 13 are presented the PCA results considering two situations: a) autocovariance matrix and, consequently, eigenvalues and eigenvectors, were obtained from the original PM_{10} concentrations; and, b) autocovariance matrix and, consequently, the eigenvalues and eigenvectors, were performed after applying the VARFIMA$(7, d, 0)(0, D, 0)$-GARCH$(1, 1)$ filter. In the latter case, given the presence of aberrant observations (outliers), the robust covariance matrix of Cotta e Reisen (2015) to estimate the eigenvalues and eigenvectors was used. Regarding

the original data, it can be seen that the first four principal components explained 84.97% of the variability of the data set, while, for the filtered data, as expected, this percentage dropped to 79.97%. It is important to say that the percentage of explanation of each component was reduced by about five percentage points when the filtered data were used.

Table 13: PCA results for PM_{10} concentrations

	PCA on original data				PCA on filtered data			
Stations	1	2	3	4	1	2	3	4
Laranjeiras	-0.3003	**0.7194**	-0.1757	0.1461	-0.2990	**0.7531**	0.1684	0.0062
Carapina	-0.3555	**-0.4005**	0.2628	0.1750	-0.3583	**-0.4229**	0.0737	-0.1974
Camburi	-0.3473	0.1700	0.0503	**0.7020**	-0.321	0.2340	-0.0100	**-0.7495**
Sua	-0.3632	0.2164	0.0406	**-0.6119**	-0.3644	0.0672	0.0591	**-0.5533**
VixCentro	**-0.3864**	-0.2265	-0.1026	-0.1629	**-0.3858**	-0.2335	-0.1228	0.0536
Ibes	**-0.3869**	0.1787	0.2359	-0.2271	**-0.3898**	0.1012	0.2474	0.2789
VVCentro	-0.3055	-0.2943	**-0.8391**	0.0142	-0.3222	0.0222	**-0.8674**	0.0639
Cariacica	**-0.3822**	-0.2767	0.3543	0.0508	-0.3677	-0.3598	-0.3661	-0.0917
Eigenvalue	4.8972	0.7744	0.6282	0.4974	4.4537	0.7266	0.5985	0.5109
Proportion (%)	61.22	9.68	7.85	6.22	56.32	9.18	7.56	6.91
Cumulative (%)	61.22	70.90	78.75	84.97	56.32	65.50	73.06	79.97

Besides the use for dimensionality reduction, the PCA technique can be used for clustering the variables of a data matrix. Cadima e Jolliffe (1995) discussed in their paper the clustering of variables by means of the eigenvectors of PCA. In this case, the grouping refers to the choice of variables that have similar values for its eigenvectors (in module) and which are highly correlated with the principal components.

Considering some environmental characteristics of the monitoring stations, the value 0.38, in module, was chosen as the cutoff point. Thus, for the original data the grouping was given by: i) the first principal component presented a similarity in the variability of the concentrations of VixCentro, Ibes and Cariacica; ii) the group of the second component was formed by Laranjeiras and Carapina; iii) the third component consisted of only VVCentro; and, iv) Camburi and Sua formed the group for the fourth component.

It is worth mentioning here that the features and properties of the PM_{10} concentrations of the Cariacica station are different from the properties of VixCentro and Ibes. Thus, the grouping of such stations may be due to the application of the PCA to the original data, which has temporal correlation, conditional heteroscedasticity and long memory, which violates some assumptions of this methodology. Santos e Reis Jr (2011) showed the chemical composition of the particles sedimented to each monitoring station of the GVR, highlighting the mass fraction of each chemical element. For the VixCentro station the composition was given by: silicon (29%), iron (16%), aluminum (12%) and others (43%). In the case of Ibes station the composition was: silicon (27%), iron (27%), aluminum (9%) and others (37%). The Cariacica station had the following composition: silicon (36%), aluminum (16%), organic carbon (15%) and others (33%).

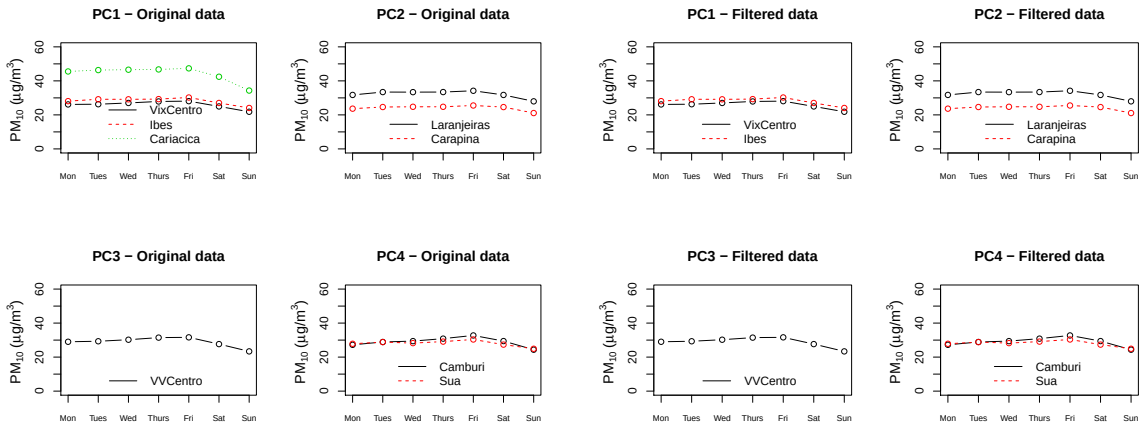
Furthermore, the Cariacica station usually has levels of PM_{10} concentration higher than the other stations of the GVR. It might be a result of the constant vehicular emissions on the site, resulting from exhausts or soil resuspension, especially of heavy vehicles that travel daily in the Supply Center of the Espírito Santo (CEASA-ES), where the station is located. Another factor

that contributes to this is its proximity to roads with high traffic of vehicles, namely, Contour Highway (BR-101) and BR-262. In addition, the Cariacica station is far away from VixCentro and Ibes stations.

In the case of PCA results for the filtered data, the groupings were: i) the first component grouped the VixCentro and Ibes stations; ii) Laranjeiras and Carapina were grouped in the second component; iii) in the third component consisted of only VVCentro; and, iv) the fourth component was the group formed by Camburi and Sua. Importantly, given the particular characteristics of Cariacica, this station was not grouped with other stations when the filtered data were adopted and four principal components was analyzed, which seems to be more coherent than when applying PCA to the original data.

To complement the cluster analysis in Figure 5 are shown the averages of the PM_{10} concentrations per day, according to the groupings presented in Table 13. It is noted that the grouping performed by the filtered data appear to be more realistic than those made with original data.

It is important to say that, according to Jolliffe (2002), the number of principal components required to represent a data set may be greater or smaller than the number indicated by the estimated PCA model. In the atmospheric sciences, for example, where the number of components can be very large, it may be of interest to stay restricted only to the first few dominant and physically interpretable patterns of variation, even though the number of components is smaller than those associated with the PCA rules. "The main message is that different objectives for a PCA lead to different requirements concerning how many PCs to retain" (JOLLIFFE, 2002, p. 132).



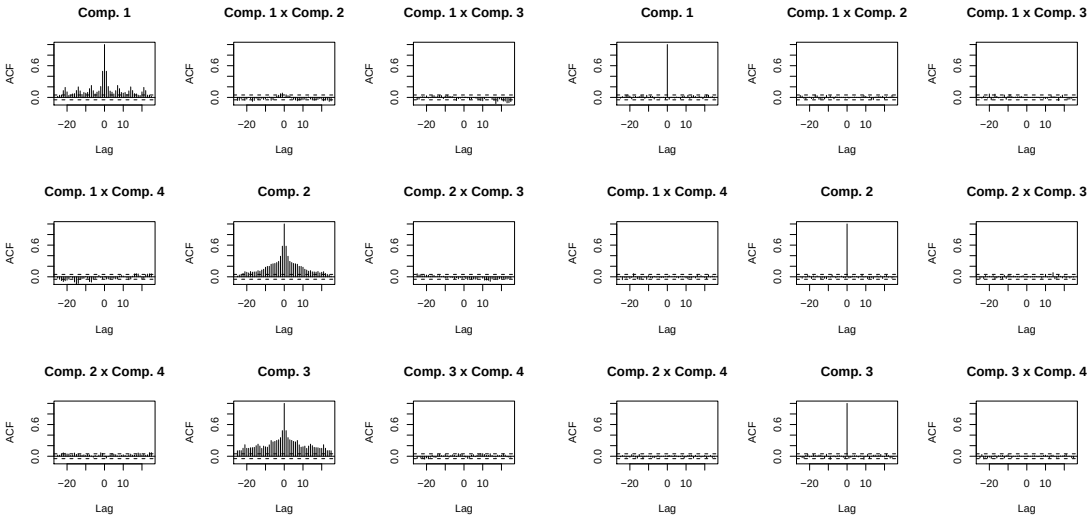
(a) Original data. (b) Filtered data.

Figure 5: Daily average of PM_{10} concentrations by grouping of stations.

Finally, Figure 6 demonstrates the robust autocorrelations and cross correlations of the principal components generated by the original and filtered data, respectively. It is possible to note that the principal components generated from original data presented autocorrelation and cross-correlation. This problem was completely eliminated when the PCA was used on the filtered data. Also, by means of the robust autocorrelations and cross-correlations of the squared PCAs,

it was observed that the principal components obtained from the original data also showed conditional heteroscedasticity, which was not seen for the components generated from the filtered data.

Therefore, in light of the problems presented in this research, to adopt the PCA technique in data with temporal correlation, conditional heteroscedasticity and long memory may negatively impact the analysis of dimensionality reduction and the following methods: factor analysis, cluster analysis, identification of sources, detection of outliers, linear and nonlinear regression, among others.



(a) Original data. (b) Filtered data.

Figure 6: Robust autocorrelations and cross-correlations of the components of PM_{10} series.

5 Conclusions

The main goal of this paper was to investigate the use of the PCA technique in multivariate time series with conditional heteroscedasticity and long memory, since ignoring these features may lead to erroneous conclusions, if the PCA is adopted in techniques such as dimensionality reduction, factor analysis, cluster analysis, identification of source, detection of outliers, linear and non-linear regression, among others. To avoid these problems, this paper proposed the application of the PCA on the residuals of the VARFIMA-GARCH model.

The results showed that the temporal correlation, the conditional heteroscedasticity and the long memory may affect the autocovariance matrix at lag zero, with reflections on the eigenvalues and the eigenvectors. In the case of the eigenvalues, a large percentage of the explanation of the variability of the data set was directed to the first principal component. Furthermore, the principal components generated from the conventional method presented serial correlation and cross-correlation. The adoption of the VARFIMA-GARCH filter, with subsequent application of the robust PCA on its residuals, allowed the correction of the problems described. This was

corroborated by an application to the pollutant PM_{10}, in the Greater Vitória Region, Espírito Santo, Brazil, but can be applied in other real situations in various areas of study.

6 Acknowledgements

The authors acknowledge partial financial support from FAPES/ES and CNPq/Brazil.

References

ANDERSON, T. W. **An introduction to multivariate statistical analysis**. 3rd. ed. New York: John Wiley & Sons, 2003. 752 p.

BERTELLI, S.; CAPORIN, M. A note on calculating autocovariances of long-memory processes. **Journal of Time Series Analysis**, v. 23, n. 5, p. 503–508, 2002.

BROCKWELL, P. J.; DAVIS, R. A. **Time Series: theory and methods**. 2nd. ed. [S.l.]: Springer Series in Statistics, 1993.

BRUNEKREEF, B.; HOLGATE, S. T. Air pollution and health. **The Lancet**, v. 360, n. 9341, p. 1233–1242, 2002.

CADIMA, J.; JOLLIFFE, I. T. Loading and correlations in the interpretation of principle components. **Journal of Applied Statistics**, Taylor & Francis, v. 22, n. 2, p. 203–214, 1995.

CHAVENT, M. et al. PCA and PMF based methodology for air pollution sources identification and apportionment. **Environmetrics**, John Wiley & Sons, Ltd., v. 20, n. 8, p. 928-942, 2009. ISSN 1099-095X. Disponível em: <http://dx.doi.org/10.1002/env.963>.

CHUNG, C. Calculating and analyzing impulse responses for the vector ARFIMA model. **Economics Letters**, v. 71, n. 1, p. 17–25, 2001. ISSN 0165-1765.

CHUNG, C. Sample means, sample autocovariances, and linear regression of stationary multivariate long memory processes. **Econometric Theory**, v. 18, p. 51–78, 2 2002. ISSN 1469-4360.

COTTA, H. H. A.; REISEN, V. A. Robust principal component analysis with air pollution data: an application to the clustering of RAMQAr. Unpublished manuscript. 2015.

CURTIS, L. et al. Adverse health effects of outdoor air pollutants. **Environment International**, v. 32, n. 6, p. 815–830, 2006. ISSN 0160-4120. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0160412006000444>.

DOMINICK, D. et al. Spatial assessment of air quality patterns in malaysia using multivariate analysis. **Atmospheric Environment**, v. 60, n. 0, p. 172–181, 2012. ISSN 1352-2310. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1352231012005729>.

ENGLE, R. F.; KRONER, K. F. Multivariate simultaneous generalized arch. **Econometric Theory**, v. 11, p. 122–150, 2 1995. ISSN 1469-4360. Disponível em: <http://journals.cambridge.org/article_S0266466600009063>.

GRANGER, C. Long memory relationships and the aggregation of dynamic models. **Journal of Econometrics**, v. 14, n. 2, p. 227–238, 1980. ISSN 0304-4076.

GRANGER, C. Some properties of time series data and their use in econometric model specification. **Journal of Econometrics**, v. 16, n. 1, p. 121–130, 1981. ISSN 0304-4076.

GRANGER, C. W. J.; JOYEUX, R. An introduction to long-memory times series models and fractional differencing. **Journal of Time Series Analysis**, v. 1, p. 15–29, 1980.

GUO, H.; WANG, T.; LOUIE, P. Source apportionment of ambient non-methane hydrocarbons in hong kong: Application of a principal component analysis/absolute principal component scores ($pca/apcs$) receptor model. **Environmental Pollution**, v. 129, n. 3, p. 489–498, 2004. ISSN 0269-7491. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0269749103004561>.

HAMILTON, J. M. **Time Series Analysis**. Princeton: Princeton University Press, 1994. 820 p.

HENRY, R.; HIDY, G. Multivariate analysis of particulate sulfate and other air quality variables by principal components-part i: annual data from los angeles and new york. **Atmospheric Environment (1967)**, v. 13, n. 11, p. 1581–1596, 1979. ISSN 0004-6981. Disponível em: <http://www.sciencedirect.com/science/article/pii/0004698179900684>.

HOSKING, J. R. Fractional differencing. **Biometrika**, v. 68, p. 165–176, 1981.

HOSKING, J. R. Asymptotic distributions of the sample mean, autocovariances, and autocorrelations of long-memory time series. **Journal of Econometrics**, v. 73, n. 1, p. 261–284, 1996. ISSN 0304-4076.

HU, Y.-P.; TSAY, R. S. Principal volatility component analysis. **Journal of Business & Economic Statistics**, v. 32, n. 2, p. 153–164, 2014. Disponível em: <http://dx.doi.org/10.1080/07350015.2013.818006>.

JOHNSON, R.; WICHERN, D. **Applied multivariate statistical analysis**. 6rd. ed. New Jersey: Prentice Hall, 2007. 800 p.

JOLLIFFE, I. T. **Principal component analysis**. 2th. ed. New York: Springer, 2002. 488 p.

LAURENT, S.; BAUWENS, L.; ROMBOUTS, J. V. K. Multivariate GARCH models: a survey. **Journal of Applied Econometrics**, v. 21, n. 1, p. 79–109, 2006. Disponível em: <http://ideas.repec.org/a/jae/japmet/v21y2006i1p79-109.html>.

LING, S.; MCALEER, M. Asymptotic theory for a vector arma-garch model. **Econometric Theory**, null, p. 280–310, 4 2003. ISSN 1469-4360. Disponível em: <http://journals.cambridge.org/article_S0266466603192092>.

LIU, P.-W. G. Simulation of the daily average PM_{10} concentrations at ta-liao with box-jenkins time series models and multivariate analysis. **Atmospheric Environment**, v. 43, n. 13, p. 2104–2113, 2009. ISSN 1352-2310. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1352231009000247>.

LüTKEPOHL, H. **New introduction to multiple time series analysis**. Berlin: Springer-Verlag, 2005. 764 p.

MA, Y.; GENTON, M. G. Highly robust estimation of the autocovariance function. **Journal of Time Series Analysis**, Blackwell Publishers Ltd, v. 21, n. 6, p. 663–684, 2000. ISSN 1467-9892. Disponível em: <http://dx.doi.org/10.1111/1467-9892.00203>.

MATTESON, D. S.; TSAY, R. S. Dynamic orthogonal components for multivariate time series. **Journal of the American Statistical Association**, v. 106, n. 496, p. 1450–1463, 2011. Disponível em: <http://dx.doi.org/10.1198/jasa.2011.tm10616>.

MAYNARD, R. Key airborne pollutants: the impact on health. **Science of The Total Environment**, v. 334-335, n. 0, p. 9–13, 2004. ISSN 0048-9697. Highway and Urban Pollution. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0048969704003493>.

MOLINARES, F. F.; REISEN, V. A.; CRIBARI-NETO, F. Robust estimation in long-memory processes under additive outliers. **Journal of Statistical Planning and Inference**, v. 139, n. 8, p. 2511–2525, 2009. ISSN 0378-3758.

PIO, C. et al. Particulate and gaseous air pollutant levels at the portuguese west coast. **Atmospheric Environment**, v. 25, n. 3-4, p. 669–680, 1991. ISSN 0960-1686. Disponível em: <http://www.sciencedirect.com/science/article/pii/096016869190065F>.

PIRES, J. et al. Identification of redundant air quality measurements through the use of principal component analysis. **Atmospheric Environment**, v. 43, n. 25, p. 3837–3842, 2009. ISSN 1352-2310. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1352231009004208>.

RAJAB, J. M.; MATJAFRI, M.; LIM, H. Combining multiple regression and principal component analysis for accurate predictions for column ozone in peninsular malaysia. **Atmospheric Environment**, v. 71, n. 0, p. 36–43, 2013. ISSN 1352-2310. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1352231013000447>.

RAO, N. R. et al. Statistical eigen-inference from large wishart matrices. **Ann. Statist.**, The Institute of Mathematical Statistics, v. 36, n. 6, p. 2850–2885, 12 2008. Disponível em: <http://dx.doi.org/10.1214/07-AOS583>.

REINSEL, G. **Elements of multivariate time series analysis**. New York: Springer, 2003. 358 p.

REISEN, V. A. et al. Fractional seasonal process with outliers to model and forecast daily average SO_2 concentrations. Preprint submitted to European Journal of Operational Research. 2015.

REISEN, V. A. et al. A semiparametric approach to estimate two seasonal fractional parameters in the SARFIMA model. **Mathematics and Computers in Simulation**, v. 98, p. 1–17, 2014. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0378475413002863>.

SANTOS, J. M.; Reis Jr, N. C. **Caracterização e quantificação de partículas sedimentáveis na Região da Grande Vitória**. [S.l.], 2011.

SEINFELD, J. H.; PANDIS, S. N. **Atmospheric chemistry and physics: from air pollution to climate change**. New York: J. Wiley, 2006.

SELA, R. J.; HURVICH, C. M. **Computationally Efficient Gaussian Maximum Likelihood Methods for Vector ARFIMA Models**. [S.l.], 2008. Disponível em: <http://ssrn.com/abstract=1301944>.

SONG, Y. et al. Source apportionment of $PM_{2.5}$ in beijing using principal component analysis/absolute principal component scores and UNMIX. **Science of The Total Environment**, v. 372, n. 1, p. 278–286, 2006. ISSN 0048-9697. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0048969706006760>.

SOUZA, J. B. d. et al. Principal components and generalized linear modeling in the correlation between hospital admissions and air pollution. **Revista de Saúde Pública**, scielo, v. 48, p. 451–458, 06 2014. ISSN 0034-8910. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0034-89102014000300451&nrm=iso>.

SOWELL, F. Maximum likelihood estimation of fractionally integrated time series models. Unpublished manuscript. 1989. Disponível em: <http://fsowell.tepper.cmu.edu>.

SOWELL, F. Maximum likelihood estimation of stationary univariate fractionally integrated time series models. **Journal of Econometrics**, v. 53, n. 1-3, p. 165–188, 1992. ISSN 0304-4076. Disponível em: <http://www.sciencedirect.com/science/article/pii/0304407692900845>.

SOWELL, F. Modeling long-run behavior with the fractional ARIMA model. **Journal of Monetary Economics**, v. 29, n. 2, p. 277–302, 1992. ISSN 0304-3932. Disponível em: <http://www.sciencedirect.com/science/article/pii/030439329290016U>.

TAQQU, M. S. Fractional brownian motion and long-range dependence. In: DOUKHAN, P.; OPPENHEIM, G.; TAQQU, M. (Ed.). **Theory and applications of long-range dependence**. Boston: Birkhäuser, 2003.

TSAY, W. Maximum likelihood estimation of stationary multivariate ARFIMA processes. **Journal of Statistical Computation and Simulation**, v. 80, n. 7, p. 729–745, 2010. Disponível em: <http://dx.doi.org/10.1080/00949650902773536>.

WATSON, J. G. et al. Receptor modeling application framework for particle source apportionment. **Chemosphere**, v. 49, n. 9, p. 1093–1136, 2002. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0045653502002436>.

WHO-World Health Organization. **WHO air quality guidelines global update 2005. Report on a working group meeting, Bonn/Germany**. [S.l.], 2005. Disponível em: <http://www.euro.who.int/__data/assets/pdf_file/0008/147851/E87950.pdf>.

WHO-World Health Organization. **Air pollution estimates**. [S.l.], 2014. Disponível em: <http://www.who.int/phe/health_topics/outdoorair/databases/FINAL_HAP_AAP_BoD_24March2014.pdf?ua=1>.

ZAMPROGNO, B. **Uso e interpretação de análise de componentes principais, em séries temporais, com enfoque no gerenciamento da qualidade do ar**. 2013. 107 f. Tese (Doutorado em Engenharia Ambiental) — Programa de Pós-graduação em Engenharia Ambiental, Universidade Federal do Espírito Santo, Vitória, 2013.

Robust principal volatility component analysis: an application to air pollution in the Greater Vitória Region, Espírito Santo, Brazil

Edson Zambon Monte[1], Valdério Anselmo Reisen[2]

[1]Graduate Program in Environmental Engineering, Federal University of Espírito Santo, Espírito Santo, Brazil; e-mail: edsonzambon@yahoo.com.br

[2]Department of Statistics, Federal University of Espírito Santo, Espírito Santo, Brazil; e-mail: valderioanselmoreisen@gmail.com.

Abstract

This paper consideres the principal volatility component (PVC) analysis, proposed by Hu e Tsay (2014), in the presence of additive outliers. A cumulative generalized robust kurtosis matrix to summarize the volatility dependence of multivariate time series is defined. Spectral analysis of this generalized robust kurtosis matrix is adopted to define robust principal volatility components (RPVCs). In addition, a robust generalized Ling-Li test statistic for dimensional reduction is proposed. The methodology is analyzed by means of Monte Carlo simulations. As an example of application, the RPVC analysis was utilized to improve the predictions of PM_{10} exceedance days in the Laranjeiras station, Greater Vitória Region (GVR), Espírito Santo, Brazil.

Keywords: principal volatility component analysis; conditional heteroscedasticity; outliers; robustness; air pollution.

1 Introduction

Issues relating to air quality have become increasingly important because many health problems such as asthma, rhinitis, eyes burning, fatigue, dry cough, heart and lung diseases, heart failure, and etc, come from air pollution. Authors as Brunekreef e Holgate (2002), Maynard (2004), WHO (2005), Curtis et al. (2006), among others, showed the relationship between the legislated pollutants (particulate matter with diameter smaller than 10 micrometers (PM_{10}), carbon monoxide (CO), sulfur dioxide (SO_2)) and health problems. In 2012, for example, the deaths of 4.3 million people were attributed to air pollution WHO (2014). In addition, air pollution contributes to the degradation of the environment, contributing to the greenhouse effect.

In recent studies related to air pollution, much attention has been paid to the mathematical models named receptor models, which attempt to measure and analyse the concentrations at their sources from a given site without reconstructing the dispersion patterns of the pollutants, such as PM_{10} and SO_2, among others. These models have mathematical and statistical tools which are mainly used to provide the identification of the pollutant emission sources from chemical characteristics of the particles on the receiver and the pollutant emission sources Seinfeld e Pandis (2006). The majority of receptor models studied are: chemical balance of mass

(CBM), multivariate analysis, principal component analysis (PCA), factor analysis (FA), multiple linear regression, cluster analysis and factoring positive matrix (FPM) Watson et al. (2002).

Regarding classical PCA, this method is adopted in techniques such as dimensionality reduction, factor analysis, cluster analysis, identification of source, detection of outliers, linear and non-linear regression, among others (for details, see Jolliffe (2002)). However, it should be noted that, according to Hu e Tsay (2014), among the studies that used the classical PCA, in addition to temporal correlation, another important point has been overlooked, namely the conditional heteroscedasticity or volatility. For a good explanation of volatility in time series, consult Engle (1982), Engle e Kroner (1995), Laurent, Bauwens e Rombouts (2006), among others. That is, the PCA technique does not take into account the dynamic dependency between the stochastic processes with volatility. To Matteson e Tsay (2011), the principal components are contemporaneously uncorrelated. However, lagged cross-correlations may be nonzero, conditional correlations may be nonzero, and cross-correlations of nonlinear transformations, such as the square process, may be non-zero.

Then, Hu e Tsay (2014), to deal with the curse of dimensionality (for a k dimensional time series, there are $k(k+1)/2$ conditional variance and covariance processes) and the difficulty in maintaining the positive definiteness of the estimated volatility matrices, generalized the idea of PCA to principal volatility component (PVC) analysis[1]. It is based on the cumulative generalized kurtosis matrix and it is used to capture the conditional heteroscedasticity of the data set. As mentioned by Hu e Tsay (2014), the PVC is adopted to identify common volatility factors. In addition, the authors developed a test statistic, called the generalized Ling-Li test statistic, to verify if a detected linear combinations does not have conditional heteroscedasticity. This test is used for dimensional reduction in multivariate modeling.

As described above, the method of PVC is based on the cumulative generalized kurtosis matrix, which is derived of the generalized classical covariance matrix. However, according to Franke (2014), the development of a robust PVC analysis is very important, since the additive outliers (AOs)[2] may mask the conditional heteroscedasticity which is common to all the component time series and that is a main target of PVC. Additionally, as mentioned by Dijk, Franses e Lucas (1999), the presence of additive outliers can produce two situations: i) spurious ARCH effects, when the data do not have volatility; and, ii) hide true conditional heteroscedasticity, when the data have true volatility. According to Carnero, Peña e Ruiz (2007), additive outliers in uncorrelated stationary time series bias the sample autocorrelations of squares in the same direction, regardless of whether the generating process is homoscedastic or heteroscedastic. Also, the outliers also bias the estimators of the parameters of the conditional variance as well as their

[1]Recently, Li et al. (2015) proposed to model the conditional variance and covariance by latent common factors that, according to authors, can be viewed as a generalized version of the principal volatility components of Hu e Tsay (2014).

[2]With respect to outliers in nonlinear GARCH models, some authors distinguish between additive and innovational (or innovations) outliers. The former type is classified in two categories: a) additive levels outliers (LO), which affects just the level of the series and has no effect on the conditional variance; and, b) additive volatility outliers (VO), that affects both, the level and the variance of the series. For details, consult Hotta e Tsay (1998), Dijk, Franses e Lucas (1999), Carnero, Peña e Ruiz (2001) and Grané e Veiga (2014).

standard deviations.

Moreover, for a fully robust version of PVC, a robust method for generalized Ling-Li test statistic should be taken into account. As previously described, additive outliers in uncorrelated GARCH series may generate spurious conditional heteroscedasticity or hide the legitimate conditional heteroscedasticity. Consequently, both the size and power of tests for conditional heteroscedasticity may be distroyed in the presence of outliers. For example, Dijk, Franses e Lucas (1999) analyzed the properties of the Lagrange Multiplier (LM) test for ARCH models in the presence of additive outliers and showed that both the size and the power are adversely affected if AOs are neglected: the test rejects the null hypothesis of homoscedasticity too often when it is in fact true, while the test has difficulty detecting genuine ARCH effects.

The environmental time series are often of a high dimension due to a large number of indexes monitored across many different locations. These data may also present interesting phenomena from being considered from applied and theoretical statistical points of view. Due to the high variability of environmental time series, especially, of air pollution concentrations, it is usually found to have a time-varying conditional variance (see, (CHELANI; DEVOTTA, 2006; REISEN et al., 2014), among others).

In addition, environmental data may have high level observations which can be defined as outliers from the statistical point of view. These observations are important to explain the environmental features of the series and can not be removed in the statistical analyzes. However, as is well known (see, for example, Chang, Tiao e Chen (1988), Tsay (1988), Chen e Liu (1993) and the references therein), the outliers can destroy the statistical properties of sample functions. Since the estimation of time series models is connected with these sample functions, the final estimated model can be strongly affected by large values of the series. One way to deal with model estimation with outliers is to use the robust ACF function based on the robust scale function $Q_n(\cdot)$ proposed by Rousseeuw e Croux (1993). The extension of this statistics for multivariate ARMA (VARMA) model is the recent paper by Cotta e Reisen (2015)). These authors proposed to adopt robust covariance matrix in the classical PCA, considering independent data set.

The goal of this paper is to extend the idea of PVC Hu e Tsay (2014) to robust principal volatility component (RPVC) analysis, considering the presence of possible atypical observations (outliers). Furthermore, a robust generalized Ling-Li test statistic was proposed. For this, the robust method given in Cotta e Reisen (2015), which make uses of $Q_n(\cdot)$ proposed by Rousseeuw e Croux (1993) and considered in Ma e Genton (2000), was used. The article is mainly based on empirical results. The theoretical results are being developed and will be presented in the future. The usefulness of the proposed methodology was applied to the data observed at the Air Quality Automatic Monitoring Network (AQAMN), Greater Vitória Region (GVR), Brazil.

The rest of this paper is organized as follows. Section 2 presents the principal volatility component analysis and the generalized Ling-Li test statistic of Hu e Tsay (2014). Section 3 shows the proposed robust principal volatility component analysis and the robust generalized

Ling-Li test statistic. Simulations are used in Section 4 to demonstrate the performance of the RPVC and robust generalized Ling-Li test statistic proposed. In Section 5, an application to air pollution data from GVR is presented. Some conclusions are considered in Section 6.

2 Principal volatility components analysis

This section summarize some results given in Hu e Tsay (2014), for the PVC in multivariate time series, which will be the basis for the robust principal volatility component (RPVC) analysis proposes here (Section 3).

2.1 Conditional heteroscedasticity and generalized kurtosis matrix

Let $\boldsymbol{y}_t = \{y_{1t}, \ldots, y_{kt}\}'$ be a k-dimensional weakly stationary time series with fourth moment finite. Let $F_{t-1} = \sigma[\boldsymbol{y}_{t-1}, \boldsymbol{y}_{t-2}, ...]$ denoting the information available at time $t-1$. As the focus is on volatility, it is assumed that $E(\boldsymbol{y_t}|F_{t-1}) = \mathbf{0}$. The volatility matrix $\boldsymbol{\Sigma}_t$ of $\boldsymbol{y}_t$ is assumed as $\boldsymbol{\Sigma}_t \equiv Cov(\boldsymbol{y}_t|F_{t-1}) = E(\boldsymbol{y}_t\boldsymbol{y}_t^{'}|F_{t-1})$. According to Hu e Tsay (2014), this volatility matrix can be written as

$$vec(\boldsymbol{\Sigma}_t) = \boldsymbol{c}_0 + \sum_{i=1}^{\infty} \boldsymbol{C}_i vec(\boldsymbol{y}_{t-i}\boldsymbol{y}_{t-i}^{'}), \tag{1}$$

where $vec(D)$ denotes the column-stacking vector of the matrix D; $\boldsymbol{c}_0$ is a k^2-dimensional positive constant vector and $\boldsymbol{C}_i$ are $k^2 \times k^2$ constant matrices for $i > 0$. The vector $\boldsymbol{c}_0$ and the matrices $\boldsymbol{C}_i$ must satisfy certain conditions to ensure that $\boldsymbol{\Sigma}_t$ be positive definite for all t. From Equation (1), the process $\boldsymbol{y}_t$ has conditional heteroscedasticity if and only if $\boldsymbol{C}_i \neq 0$ for some $i > 0$. For multivariate autoregressive conditional heteroscedasticity (ARCH) models, the summation in Equation (1) is truncated at a finite lag. The vector time series $\boldsymbol{y}_t$ has ARCH effects or conditional heteroscedasticity if $\boldsymbol{C}_i \neq 0$ for some $i > 0$.

From Equation (1), the existence of ARCH effects in $\boldsymbol{y}_t$ implies that $\boldsymbol{y}_t\boldsymbol{y}_t'$ is correlated with $\boldsymbol{y}_{t-i}\boldsymbol{y}_{t-i}'$ for some $i > 0$. This motivates the use of the lag-ℓ generalized kurtosis matrix $\boldsymbol{\gamma}_\ell$ of $\boldsymbol{y}_t$ as

$$\boldsymbol{\gamma}_\ell = \sum_{i=1}^{k}\sum_{j=i}^{k} cov^2(\boldsymbol{y}_t\boldsymbol{y}_t', x_{ij,t-\ell}) = \sum_{i=1}^{k}\sum_{j=i}^{k} \boldsymbol{\gamma}_{\ell,ij}\boldsymbol{\gamma}_{\ell,ij}', \quad \ell > 0, \tag{2}$$

where $x_{ij,t-\ell}$ is a function of $y_{i,t-\ell}y_{j,t-\ell}$, para $1 \leq i, j \leq k$,

$$\boldsymbol{\gamma}_{\ell,ij} = cov(\boldsymbol{y}_t\boldsymbol{y}_t', x_{ij,t-\ell}) = E[(\boldsymbol{y}_t\boldsymbol{y}_t' - \boldsymbol{\Sigma})(x_{ij,t-\ell} - E(x_{ij}))], \tag{3}$$

and $\boldsymbol{\Sigma} = E(\boldsymbol{y}_t\boldsymbol{y}_t')$ is the unconditional covariance matrix of $\boldsymbol{y}_t$.

The matrix $\boldsymbol{\gamma}_{\ell,ij}$ of Equation (3) is the generalized covariance matrix. It is a $k \times k$ symmetric matrix, but might be negative definite. However, its square matrix, which is equivalent

to $\boldsymbol{\gamma}_{\ell,ij}\boldsymbol{\gamma}'_{\ell,ij}$, is semipositive definite. This justifies the use of square in Equation (2). From the definition, the lag-ℓ generalized kurtosis matrix $\boldsymbol{\gamma}_\ell$ is symmetric and semipositive definite, because it is the sum of $k(k+1)/2$ symmetric and semipositive definite matrices. An important property of $\boldsymbol{\gamma}_\ell$ is that $\boldsymbol{\gamma}_\ell = 0$ if and only if $\boldsymbol{y}_t\boldsymbol{y}'_t$ is not correlated with $\boldsymbol{y}_{t-i}\boldsymbol{y}'_{t-i}$ for all i and j.

For a given positive integer m, Hu e Tsay (2014) defined the cumulative generalized kurtosis matrix as

$$\boldsymbol{\Gamma}_m = \sum_{\ell=1}^{m} \boldsymbol{\gamma}_\ell. \tag{4}$$

This cumulative matrix is symmetric and semipositive definite, and it is used to measure the ARCH(m) effects in $\boldsymbol{y}_t$. For the general multivariate GARCH-type models, it is considered

$$\boldsymbol{\Gamma}_\infty = \sum_{\ell=1}^{\infty} \boldsymbol{\gamma}_\ell. \tag{5}$$

which is assumed to exist. The $\boldsymbol{\Gamma}_\infty$ is the cumulative generalized kurtosis matrix of $\boldsymbol{y}_t$. $\boldsymbol{\Gamma}_\infty$ is symmetric and semipositive definite.

2.1.1 Properties of generalized kurtosis matrix

Let $\boldsymbol{M}$ be a $k \times k$ nontrivial linear transformation matrix so that $\boldsymbol{z}_t = \boldsymbol{M}'\boldsymbol{y}_t$ is a transformed series (HU; TSAY, 2014). Let x_{t-1} be a scalar function of F_{t-1}, for example, $x_{t-1} = y_{i,t-h}y_{j,t-h}$ for some $h > 0$. It is easy to see that the following lemma holds.

Lemma 1. ((HU; TSAY, 2014)) For a constant $k \times k$ matrix $\boldsymbol{M}$, let $\boldsymbol{z}_t = \boldsymbol{M}'\boldsymbol{y}_t$. Then, $cov(\boldsymbol{z}_t\boldsymbol{z}'_t, x_{t-1}) = cov(\boldsymbol{M}'\boldsymbol{y}_t\boldsymbol{y}'_t\boldsymbol{M}, x_{t-1}) = \boldsymbol{M}'cov(\boldsymbol{y}_t\boldsymbol{y}'_t, x_{t-1})\boldsymbol{M}$.

Let $\boldsymbol{m} = (m_{1v}, \ldots, m_{kv})'$ be the vth column of $\boldsymbol{M}$. If $\boldsymbol{z}_t = \boldsymbol{M}'\boldsymbol{y}_t$ is a linear combination of $\boldsymbol{y}_t$ that has no ARCH effects, then $E(z_{vt}^2|F_{t-1}) = c_v^2$, which is a constant. This implies that z_{vt}^2 in not correlated with $y_{i,t-\ell}y_{j,t-\ell}$ for $\ell > 0$ and $1 \leq i \leq j \leq k$. Using Lemma 1, it can seen that $\boldsymbol{\gamma}_{\ell,ij}$ is singular for all ℓ and $1 \leq i \leq j \leq k$ and, hence, $\boldsymbol{\gamma}_\ell$ is singular for all ℓ. Consequently, $\boldsymbol{\Gamma}_\infty$ in also singular.

On the other hand, assume that $\boldsymbol{\Gamma}_\infty$ is singular and $\boldsymbol{m}_v$ is in its null space. That is, $\boldsymbol{\Gamma}_\infty\boldsymbol{m}_v = \boldsymbol{0}$. Since $\boldsymbol{\gamma}_\infty$ is semipositive definite, it follows that $\boldsymbol{m}'_v\boldsymbol{\gamma}_\ell\boldsymbol{m}_v = 0$ for all ℓ. This in turn shows that $\boldsymbol{m}'_v\boldsymbol{\gamma}^2_{\ell,ij}\boldsymbol{m}_v = 0$ for all ℓ and $1 \leq i \leq j \leq k$. Again, adopting Lemma 1, it is possible to see that $\boldsymbol{z}^2_{vt}$ is not correlated with $y_{i,t-\ell}y_{j,t-\ell}$ for all ℓ and $1 \leq i \leq j \leq k$, where $z_{vt} = \boldsymbol{m}'\boldsymbol{y}_t$. This implies that $E(z_{vt}^2|F_{t-1})$ is not time varying. In other words, z_{vt} does not have conditional heteroscedasticity. The above discussion shows that an eigenvector of $\boldsymbol{\Gamma}_\infty$ associated with a zero eigenvalue gives rise to a linear combination of $\boldsymbol{y}_t$ that has no ARCH effect. The Theorem 1 summarize this result.

Theorem 1. (HU; TSAY, 2014) Consider a weakly stationary process $\boldsymbol{y}_t$ with finite fourth moment and satisfying Equation (1). Let $\boldsymbol{\Gamma}_\infty$ be the cumulative generalized kurtosis matrix

defined in Equation (5), where $x_{ij,t-\ell} = y_{i,t-\ell}y_{j,t-\ell}$ in Equation (3). Then, there exist $k - m$ linearly independent linear combinations of $\boldsymbol{y}_t$ that have no ARCH effects if and only if rank($\boldsymbol{\Gamma}_\infty$) $= m$.

2.2 Principal volatility components

Based on Hu e Tsay (2014), consider the spectral decomposition of the cumulative generalized kurtosis matrix $\boldsymbol{\Gamma}_\infty$, say $\boldsymbol{\Gamma}_\infty \boldsymbol{M} = \boldsymbol{M}\boldsymbol{\Lambda}$, where $\boldsymbol{\Lambda} = diag(\lambda_1^2 \geq \lambda_2^2 \geq \cdots \geq \lambda_k^2)$ is the diagonal matrix of ordered eigenvalues and $\boldsymbol{M} = [\boldsymbol{m}_1, \ldots, \boldsymbol{m}_k]$ is the matrix of eigenvectors. Here, the notation λ_v^2 is used to denote eigenvalues because $\boldsymbol{\Gamma}_\infty$ is semipositive definite. It is assumed that the columns $\boldsymbol{m}_v$ are normalized with $\|\boldsymbol{m}_v\| = 1$.

The v-th PVC of $\boldsymbol{y}_t$ is defined as $z_{vt} = \boldsymbol{m}_v'\boldsymbol{y}_t$. From the definition and spectral decomposition of $\boldsymbol{\Gamma}_\infty$, the follows equation can be derived

$$\sum_{\ell=1}^{\infty}\sum_{i=i}^{k}\sum_{j=1}^{k} \boldsymbol{m}_v'\boldsymbol{\gamma}_{\ell,ij}^2\boldsymbol{m}_v = \lambda_v^2, v = 1, \ldots, k. \tag{6}$$

Let $\boldsymbol{\gamma}_{\ell,ij}\boldsymbol{m}_v = \boldsymbol{\omega}_{\ell,ij,v}$. Then, it can define $\sum_{\ell=1}^{\infty}\sum_{i=i}^{k}\sum_{j=1}^{k}\boldsymbol{\omega}_{\ell,ij,v}'\boldsymbol{\omega}_{\ell,ij,v} = \lambda_v^2$. Using Lemma 1, follows that

$$\boldsymbol{m}_v'\boldsymbol{\omega}_{\ell,ij,v} = \boldsymbol{m}_v'\boldsymbol{\gamma}_{\ell,ij}\boldsymbol{m}_v = cov(z_{vt}^2, x_{ij,t-\ell}). \tag{7}$$

This result indicates that $\boldsymbol{m}_v'\boldsymbol{\gamma}_{\ell,ij}\boldsymbol{m}_v$ can be regarded as a measure of the dependence of volatility of the portfolio z_{vt} on the lagged cross-product term $x_{ij,t-\ell}$. In practice, this quantity can be negative so that squared matrices are used in Equation (7) to construct a nonnegative dependence measure.

From Equation (6), λ_v^2 summarizes the dependence measure in Equation (7) over all combinations of i and j and over all lags. As such, it can be considered as an approximate measure of volatility dependence of the portfolio z_{vt}. A larger λ_v is indicative of a stronger volatility dependence. Therefore, z_{vt} is called of vth PVC[3].

Since $\boldsymbol{\Gamma}_\infty$ is semipositive definite, its eigenvectors are orthogonal provided that the associated eigenvalues are distinct. Consequently, any two PVCs, $z_{vt} = \boldsymbol{m}_v'\boldsymbol{y}_t$ and $z_{ut} = \boldsymbol{m}_u'\boldsymbol{y}_t$, are uncorrelated if $\lambda_v^2 \neq \lambda_u^2$. On the other hand, for PVC z_{vt} associated with a nonzero λ_v^2, z_{vt}^2 may still be correlated with lagged values of z_{ut}^2. However, as described by Hu e Tsay (2014), such correlations, if exist, are of smaller magnitudes compared with those of the observed y_{it}^2 series.

According to Hu e Tsay (2014), like the traditional PCA, an important application of the proposed PVC analysis is to reduce the dimension in volatility modeling. To this end, the number of zero eigenvalues of $\boldsymbol{\Gamma}_\infty$ is of special interest. Following Theorem 1, there are common volatility components if the cumulative generalized kurtosis matrix, $\boldsymbol{\Gamma}_\infty$, is not of full rank.

[3] According Hu e Tsay (2014), the summation in Equation (6) distinguishes the PVC from the traditional principal components (PCA) of $\boldsymbol{y}_t$, which depends on the covariance matrix alone. PVC analysis is designed to consider simultaneously volatility dependence at all past lags.

2.3 Sample principal volatility components

Based on Hu e Tsay (2014), in this section, the estimation of the generalized kurtosis matrices is considered and the sample PVCs is obtained. The authors established some consistency properties of the sample estimate of $\mathbf{\Gamma}_\infty$. Furthermore, to verify that a given volatile component does not have ARCH effects, a test statistic is presented, with its asymptotic distribution.

To simplify the moment restrictions of $\boldsymbol{y}_t$ for making statistical inference of $\mathbf{\Gamma}_\infty$ or $\mathbf{\Gamma}_m$, Hu e Tsay (2014) followed the idea of Matteson e Tsay (2011), by adopting the Huber's function of the cross-product variable $y_{i,t-\ell}y_{j,t-\ell}$, for some $0 < c < \infty$. As seen in Matteson e Tsay (2011), several other functions with similar properties, such as those associated with M-estimators, also may be considered.

2.3.1 Estimation

Following Hu e Tsay (2014), consider the data $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ of a stationary process $\boldsymbol{y}_t$. Let

$$\widehat{cov}(\boldsymbol{y}_t\boldsymbol{y}_t', x_{ij,t-\ell}) = \frac{1}{n}\sum_{t=\ell+1}^{n}(\boldsymbol{y}_t\boldsymbol{y}_t' - \bar{\boldsymbol{Y}})(x_{ij,t-\ell} - \bar{x}_{ij}) \tag{8}$$

where $\bar{\boldsymbol{Y}}$ and $\bar{x}_{ij}$ are the sample mean of $\boldsymbol{y}_t\boldsymbol{y}_t'$ and $x_{ij,t}$, respectively.

$\mathbf{\Gamma}_m$ function can be estimated by

$$\widehat{\mathbf{\Gamma}}_m = \sum_{\ell=1}^{m}\sum_{i=1}^{k}\sum_{j=i}^{k}\left(1 - \frac{\ell}{n}\right)\widehat{cov}^2(\boldsymbol{y}_t\boldsymbol{y}_t', x_{ij,t-\ell}). \tag{9}$$

Suppose which $\boldsymbol{y}_t$ is stationary with finite sixth moment and Assumption A.1 of Hu e Tsay (2014, pg. 162) holds. Then, for a fixed positive integer $m < \infty$,

$$\widehat{\mathbf{\Gamma}}_m = \mathbf{\Gamma}_m + O_p\left(\frac{1}{\sqrt{n}}\right). \tag{10}$$

$\widehat{\mathbf{\Gamma}}_\infty$ is estimated as follows

$$\widehat{\mathbf{\Gamma}}_n = \sum_{\ell=1}^{n-1}\sum_{i=1}^{k}\sum_{j=i}^{k}\omega^*\left(\frac{\ell}{m_n}\right)\widehat{cov}^2(\boldsymbol{y}_t\boldsymbol{y}_t', x_{ij,t-\ell}). \tag{11}$$

where $\omega^*(\ell/m_n)$ is a prespecified smoothing function and m_n is a positive real number depending on n.

Considering the regularity conditions used in spectral density estimation Hannan (1970), and $m_n/n \longrightarrow 0$, $m_n \longrightarrow \infty$ as $n \longrightarrow \infty$, then

$$\widehat{\mathbf{\Gamma}}_n = \mathbf{\Gamma}_\infty + O_p(a_n), \tag{12}$$

where a_n is a function of m_n and n. Some guidelines for choosing m_n are as follows: the convergence rate of $\widehat{\mathbf{\Gamma}}_n$ of Equation (12) depends on the minimal mean squared error $E(\|\widehat{\mathbf{\Gamma}}_n -$

$\boldsymbol{\Gamma}_\infty\|)^2$. For a specific smoothing function $\omega^*(\cdot)$, m_n can be selected carefully to achieve the minimal mean squared error. For example, if ω^* is the Bartlett's smoothing function, the best choice of m_n is $m_n = n^{1/3}$ such that $a_n = n^{-1/3}$, and if ω^* is the Daniell's smoothing function, one can choose $m_n = n^{1/5}$ such that $a_n = n^{-2/5}$. For more details see Hannan (1970).

2.3.2 Testing

The PVC analysis proposed by Hu e Tsay (2014) can be used for dimension reduction. Here, dimension reduction means finding linear combinations of $\boldsymbol{y}_t$ that have no ARCH effects. Let $\widehat{\boldsymbol{M}}_1$ be a $k \times s$ matrix consisting of eigenvectors associated with the s smallest eigenvalues of the $\widehat{\boldsymbol{\Gamma}}_m$ (or $\widehat{\boldsymbol{\Gamma}}_\infty$) matrix. In other words, $\widehat{\boldsymbol{M}}_1$ gives rise to the $(k-s+1)$-th to the k-th PVCs of $\boldsymbol{y}_t$. The aim is to verify if the transformed series $\hat{\boldsymbol{e}}_t = \widehat{\boldsymbol{M}}_1'\boldsymbol{y}_t$ indeed has no ARCH effects.

Many data-generating processes (DGPs) of $\boldsymbol{y}_t$ can lead to linear combinations of $\boldsymbol{y}_t$ that have no ARCH effects. Consider, for instance, the case of common volatility components and assume that

$$\boldsymbol{y}_t = \boldsymbol{H}\boldsymbol{f}_t + \boldsymbol{\epsilon}_t, \tag{13}$$

where $\boldsymbol{H}$ is a $k \times k$ real-valued matrix of rank r; $\boldsymbol{f}_t = (f_{1t}, \ldots, f_{rt})'$ consists of r independent conditional heteroscedastic processes; $\{\boldsymbol{\epsilon}_t\}$ is a sequence of independent and identically distributed random vectors with mean zero and constant positive-definite covariance matrix $\boldsymbol{\Sigma}_\epsilon$; and, $\boldsymbol{\epsilon}_t$ is independent of $\boldsymbol{f}_t$.

Based on Equation (1), each volatility component var$(f_{it}|F_{t-1})$ is a nontrivial function of elements of $\{\boldsymbol{y}_{t-j}\boldsymbol{y}_{t-j}'|j > 0\}$. For this particular DGP, if $r < k$, then the volatility of $\boldsymbol{y}_t$ is driven by the r-dimensional common volatility components in $\boldsymbol{f}_t$. Let $\boldsymbol{M}_1$ be a $k \times (k-r)$ real-valued matrix such that $\boldsymbol{M}_1'\boldsymbol{H} = \boldsymbol{0}$. Let $\hat{\boldsymbol{e}}_t = \widehat{\boldsymbol{M}}_1'\boldsymbol{y}_t$. Is is easy to see that $\hat{\boldsymbol{e}}_t = \widehat{\boldsymbol{M}}_1'\boldsymbol{\epsilon}_t$ and, hence, it has no ARCH effects.

As described by Hu e Tsay (2014), there are several tests available in the literature to check for multivariate ARCH effects. The authors generalized the results of Ling e Li (1997), Duchesne e Lalancette (2003), Duchesne e Roy (2004), and Hong (1996) to derive two test statistics that are applicable to the classical PVC analysis: i) Ling-Li test statistic $(T_{d,s})$; and, ii) generalized Ling-Li test statistic $(G_{p_n,s})$. According Hu e Tsay (2014), the $T_{d,s}$ statistic is designed to detect the serial volatility dependence in the first d lags. However, in empirical applications, a test statistic that can account for volatility dependence in all past lags is more interesting . Thus, it is more convenient to use the $G_{p_n,s}$ test to check the ARCH dependence of the sample PVCs.

2.6.2.1. Generalized test statistic

Following Ling e Li (1997), Hu e Tsay (2014) defined

$$\hat{\epsilon}_t = \frac{\hat{\boldsymbol{e}}_t'\widehat{\boldsymbol{V}}^{-1}\hat{\boldsymbol{e}}_t - s}{\sqrt{\sum_{t=1}^{n}(\hat{\boldsymbol{e}}_t'\widehat{\boldsymbol{V}}^{-1}\hat{\boldsymbol{e}}_t - s)^2/n}}, \tag{14}$$

$$\hat{x}_{t-j} = \frac{\hat{h}_{\boldsymbol{y},t-j} - \bar{h}_{\boldsymbol{y}}}{\sqrt{\sum_{t=1}^{n}(\hat{h}_{\boldsymbol{y},t} - \bar{h}_{\boldsymbol{y}})^2/n}}. \tag{15}$$

and, the correlation between $\hat{\epsilon}_t$ and $\hat{x}_{t-j}$, given by

$$\hat{\rho}_{j,s} = \frac{1}{n}\sum_{t=j+1}^{n}\hat{\epsilon}_t\hat{x}_{t-j} = \frac{1/n\sum_{t=j+1}^{n}(\hat{\boldsymbol{e}}_t'\widehat{\boldsymbol{V}}^{-1}\hat{\boldsymbol{e}}_t - s)(\hat{h}_{\boldsymbol{y},t-j} - \bar{h}_{\boldsymbol{y}})}{\sqrt{\sum_{t=1}^{n}(\hat{\boldsymbol{e}}_t'\widehat{\boldsymbol{V}}^{-1}\hat{\boldsymbol{e}}_t - s)^2/n}\sqrt{\sum_{t=1}^{n}(\hat{h}_{\boldsymbol{y},t} - \bar{h}_{\boldsymbol{y}})^2/n}}, \tag{16}$$

where

$$\hat{h}_{\boldsymbol{y},t} = \begin{cases} \hat{\boldsymbol{y}}_t'\widehat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{y}}_t/k, & \text{if } \hat{\boldsymbol{y}}_t'\widehat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{y}}_t/k \leq c^2, \\ 2c\sqrt{\hat{\boldsymbol{y}}_t'\widehat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{y}}_t/k} - c^2, & \text{if } \hat{\boldsymbol{y}}_t'\widehat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{y}}_t/k > c^2, \end{cases} \tag{17}$$

and $\bar{h}_{\boldsymbol{y}} = (1/n)\sum_{t=1}^{n}\bar{h}_{\boldsymbol{y},t}$; and, $\widehat{\boldsymbol{V}}$ and $\widehat{\boldsymbol{\Sigma}}$ are the sample covariance matrix of $\hat{\boldsymbol{e}}_t$ and $\boldsymbol{y}_t$, respectively. Furthermore, let

$$\epsilon_t = \frac{\boldsymbol{e}_t'\boldsymbol{V}^{-1}\boldsymbol{e}_t - s}{\sigma_\epsilon} \quad \text{and} \quad x_{t-j} = \frac{h_{\boldsymbol{y},t-j} - \bar{h}_{\boldsymbol{y}}}{\sigma_x},$$

where $\sigma_\epsilon^2 = E(\boldsymbol{e}_t'\boldsymbol{V}^{-1}\boldsymbol{e}_t - s)^2$ and $\sigma_x^2 = E(h_{\boldsymbol{y},t-j} - \bar{h}_{\boldsymbol{y}})^2$. Hence, $\boldsymbol{e}_t$ and $h_{\boldsymbol{y},t}$ are the theoretical counterparts of $\hat{\boldsymbol{e}}_t$ and $\hat{h}_{\boldsymbol{y},t}$, respectively.

Then, Hu e Tsay (2014), adopting the idea of Hong (1996), proposed a generalized Ling-Li test statistic, defined by

$$G_{p_n,s} = \frac{n\sum_{j=1}^{n-1}\omega^2(j/p_n)\hat{\rho}_{j,s} - M_n(\omega)}{[2\Delta V_n(\omega)]^{1/2}}, \tag{18}$$

where $M_n(\omega) = \sum_{j=1}^{n-1}(1-j/n)\omega^2(j/p_n)$; $V_n(\omega) = \sum_{j=1}^{n-2}(1-j/n)[1-(j+1)/n]\omega^4(j/p_n)$; $\Delta = 1 + 2\sum_{h=1}^{\infty}cov^2(x_t, x_{t-h})$; p_n is a function of n such that $p_n \longrightarrow \infty$ and, $p_n/n \longrightarrow 0$ as $n \longrightarrow \infty$; and, $\omega(\cdot)$ is a symmetric kernel function.

As mentioned by Hu e Tsay (2014), if $\boldsymbol{y}_t$ is a k-dimensional process of independent and identically distributed random variables, then $s = k$. In this case, $G_{p_n,s}$ reduces to the Hong's statistic in which $\Delta = 1$, because $cov^2(x_t, x_{t-h}) = 0$, for all $h > 0$. In this sense, Δ is used to adjust for ARCH effects in the $\boldsymbol{y}_t$ series. This quantity can be estimated by a smoothing method such as

$$\widehat{\Delta} = 1 + 2\sum_{h=1}^{n-1}k(h/s_n)\widehat{cov}^2(\hat{x}_t, \hat{x}_{t-h}) \tag{19}$$

where $k(\cdot)$ is a kernel function satisfying some regularity conditions such that $\Delta^* = 1+2\sum_{h=1}^{\infty}k(h/s_n)\widehat{cov}^2(x_t, x_{t-h})$ is a consistent estimate of Δ; and, S_n is a function of n such that $S_n \longrightarrow \infty$ and $S_n/n \longrightarrow 0$ as $n \longrightarrow \infty$. Here, the Daniell function, $g(z) = sin(\pi z)/(\pi z)$, was used.

Theorem 2. (HU; TSAY, 2014) Suppose that $\boldsymbol{y}_t$ is a k-dimensional weakly stationary process with ARCH effects governed by Equation (1) and has finite sixth moment. Let $\boldsymbol{M}_1$ be a constant full-rank $k \times s$ transformation matrix such that $\boldsymbol{e}_t = \boldsymbol{M}_1'\boldsymbol{y}_t$ has no ARCH effects. Assume that $\widehat{\boldsymbol{M}}_1$ is a consistent estimate of $\boldsymbol{M}_1$ and $\hat{\boldsymbol{e}}_t = \widehat{\boldsymbol{M}}_1'\boldsymbol{y}_t$. Under the Assumptions A.1 and A.2 of Hu e Tsay (2014, pg. 162), if $p_n \longrightarrow \infty$, $p_n/n \longrightarrow 0$, $\frac{na_n^4}{\sqrt{p_n}} \longrightarrow 0$, and $P_n a_n^2 \longrightarrow 0$ as $n \longrightarrow \infty$, then $G_{p_n,s} \xrightarrow{d} N(0,1)$, where $G_{p_n,s}$ is the test statistic present in Equation (18) and a_n in given in Equation (10).

2.6.2.2. Generalized test for ARCH effects

According Hu e Tsay (2014), for a given data set, let $\widehat{\boldsymbol{\Gamma}}_m$ be the sample estimate of the cumulative generalized kurtosis matrix. Further, let $\widehat{\boldsymbol{M}} = [\widehat{\boldsymbol{m}}_1, \ldots, \widehat{\boldsymbol{m}}_k]$ be the matrix of standardized eigenvectors such that $\widehat{\boldsymbol{\Gamma}}_m\widehat{\boldsymbol{m}}_v = \lambda_v^2\widehat{\boldsymbol{m}}_v$, where $\lambda_1^2 \geq \lambda_2^2 \geq \cdots \geq \lambda_k^2$ are the eigenvalues. Since zero eigenvalues of $\widehat{\boldsymbol{\Gamma}}_m$ give rise to components without ARCH effects, the following test procedure is used to detect linear combinations of $\boldsymbol{y}_t$ that have no ARCH effects. Let s be the number of linear combinations of $\boldsymbol{y}_t$ that have no ARCH effects. Let $\widehat{\boldsymbol{M}_1} = [\widehat{\boldsymbol{m}}_k, \ldots, \widehat{\boldsymbol{m}}_{k-s+1}]$ be the $k \times s$ matrix consisting of the last s columns of $\widehat{\boldsymbol{M}}$, which consists of the standardized eigenvectors of $\widehat{\boldsymbol{\Gamma}}_m$ corresponding to the s smallest eigenvalues. Let $\hat{\boldsymbol{e}}_t = \widehat{\boldsymbol{M}}_1'\boldsymbol{y}_t$ be the s-dimensional transformed process consisting of the last s PVC series of $\boldsymbol{y}_t$. The generalized Ling-Li test statistic of Theorem 2 is applied to test the null hypothesis that $\hat{\boldsymbol{e}}_t$ has no ARCH effects. In practice, the test is performed for $s = 1, \ldots, k$. If the null hypothesis $H_0 : s = s_*$ is not rejected, but the null hypothesis $H_0 : s = s_* + 1$ is rejected, then there is s_* linear combinations of $\boldsymbol{y}_t$ that have no ARCH effects. In other words, the hypothesis that the smallest s eigenvalues of $\widehat{\boldsymbol{\Gamma}}_m$ are zero is tested sequentially.

3 Robust principal volatility component (RPVC) analysis

As mentioned in the introduction, the additive outliers may mask the conditional heteroscedasticity that is a main target of the classical PVC or generate spurious ARCH effects. The additive volatility outliers (AVOs) are defined follows. Let $\boldsymbol{y}_t^*,\ t = 1, \ldots, t\ \in \mathbb{Z}$ be a vector process contaminated by additive outliers, with contaminated volatility matrix ($\boldsymbol{\Sigma}_t^*$) following a ARCH(1) specification. This can be also generalized for a more complex representation. Then,

$$\boldsymbol{y}_t^* = \boldsymbol{y}_t + \boldsymbol{\omega} \circ \boldsymbol{\delta}_t, \tag{20}$$

$$\boldsymbol{\Sigma}_t^* = \boldsymbol{c}_0 + \boldsymbol{C}_1\boldsymbol{y}_{t-1}^*\boldsymbol{y}_{t-1}^{*'}\boldsymbol{C}_1', \tag{21}$$

where "$\circ$" is the Hadamard product (JOHNSON, 1989); $\boldsymbol{\omega} = \{\omega_1, ..., \omega_k\}'$ is a magnitude vector of additive outliers; $\boldsymbol{\delta}_t = \{\delta_{1t}, ..., \delta_{kt}\}'$ is a random vector indicating the occurrence of an outlier at time t, in variable k, such as $P(\delta_{k,t} = -1) = P(\delta_{k,t} = 1) = p/2$ and $P(\delta_{k,t} = 0) = 1 - p$,

where $E[\delta_{k,t}] = 0$ and $E[\delta_{k,t}^2] = Var(\delta_{k,t}) = p$. The model described above assumes that $\{\mathbf{y}_t^*\}$ and $\{\boldsymbol{\delta}_t\}$ are independent processes. Also, it is assumed that $E(\boldsymbol{\delta}_t\boldsymbol{\delta}_t^{'}) = \Sigma_{\boldsymbol{\delta}} = diag(p, ..., p)$ and $E(\boldsymbol{\delta}_t\boldsymbol{\delta}_{t+\ell}^{'}) \neq 0$ for $\ell \neq 0$. Finally, $\boldsymbol{\Sigma}_t^*$ denote the conditional covariance matrix for the contaminated process $\boldsymbol{y}_t^*$.

Remark 1. $\boldsymbol{\delta}_{kt}$ is the product of $Bernoulli(p)$ random variable with $Rademacher$ random variable; the latter equals 1 or -1, both with probability 1/2.

Remark 2. In a particular case, at time t, when the outlier occurs, $\boldsymbol{y}_t^* = \boldsymbol{y}_t + \boldsymbol{\omega}$. Then, after some mathematical operations, the contaminated conditional covariance matrix by additive volatility outliers is represented as

$$\boldsymbol{\Sigma}_t^* = \boldsymbol{c}_0 + \boldsymbol{C}_1\boldsymbol{y}_{t-1}\boldsymbol{y}_{t-1}'\boldsymbol{C}_1' + \boldsymbol{C}_1(\boldsymbol{y}_{t-1}\boldsymbol{\omega}' + \boldsymbol{\omega}\boldsymbol{y}_{t-1}' + \boldsymbol{\omega}\boldsymbol{\omega}')C_1', \tag{22}$$

or

$$\boldsymbol{\Sigma}_t^* = \boldsymbol{\Sigma}_t + \boldsymbol{C}_1(\boldsymbol{y}_{t-1}\boldsymbol{\omega}' + \boldsymbol{\omega}\boldsymbol{y}_{t-1}' + \boldsymbol{\omega}\boldsymbol{\omega}')\boldsymbol{C}_1', \tag{23}$$

Note that the conditional covariance matrix of the contaminated process $\boldsymbol{y}_t^*$ is different from the uncontaminated process $\boldsymbol{y}_t$. This difference is given by term $\boldsymbol{C}_1(\boldsymbol{y}_{t-1}\boldsymbol{\omega}' + \boldsymbol{\omega}\boldsymbol{y}_{t-1}' + \boldsymbol{\omega}\boldsymbol{\omega}')\boldsymbol{C}_1'$. Thus, the presence of outliers directly affects the conditional covariance matrix $\boldsymbol{\Sigma}_t$.

Remark 3. If the vector of additive outliers $\boldsymbol{\omega} = [\omega_1, ..., \omega_k]' = \mathbf{0}$, for all $i = 1, \ldots, k$, then $\boldsymbol{\Sigma}_t^* = \boldsymbol{\Sigma}_t$, i.e., the case studied by Hu e Tsay (2014).

Based on Rousseeuw e Croux (1993), Ma e Genton (2000) and Cotta e Reisen (2015), this paper proposes a robust principal volatility component (RPVC) analysis and a robust generalized Ling-Li test statistic ($G_{Q_n,p_n,s}$). Regarding the RPVC, the lag-ℓ generalized robust kurtosis matrix $\boldsymbol{\gamma}_{Q_n,\ell}$ suggested here is

$$\boldsymbol{\gamma}_{Q_n,\ell} = \sum_{i=1}^{k}\sum_{j=i}^{k} cov_{Q_n}^2(\boldsymbol{y}_t^*\boldsymbol{y}_t^{*'}, x_{ij,t-\ell}^*) = \sum_{i=1}^{k}\sum_{j=i}^{k} \boldsymbol{\gamma}_{Q_n,\ell,ij}\boldsymbol{\gamma}_{Q_n,\ell,ij}', \quad \ell > 0, \tag{24}$$

where $cov_{Q_n} = \boldsymbol{\gamma}_{Q_n,\ell,ij}$ is the generalized robust covariance matrix, which is the multivariate extension of the one proposed by Ma e Genton (2000), and $x_{ij,t-\ell}^*$ is a function of $y_{i,t-\ell}^* y_{j,t-\ell}^*$, for $1 \leq i, j \leq k$. $\boldsymbol{\gamma}_{Q_n,\ell,ij}$ is a symmetric matrix, but might be negative definite. However $\boldsymbol{\gamma}_{Q_n,\ell,ij}\boldsymbol{\gamma}_{Q_n,\ell,ij}'$ (square matrix) is semipositive definite. From this, the lag-ℓ generalized robust kurtosis matrix $\boldsymbol{\gamma}_{Q_n,\ell}$ is symmetric and semipositive definite.

Specifically, for a positive integer m, the cumulative generalized robust kurtosis matrix can be defined by Equation (25). This matrix is symmetric and semipositive definite, and is used to measure the ARCH(m) effects in the contaminated process $\boldsymbol{y}_t^*$.

$$\boldsymbol{\Gamma}_{Q_n,m} = \sum_{\ell=1}^{m} \boldsymbol{\gamma}_{Q_n,\ell}. \tag{25}$$

The $\mathbf{\Gamma}_{Q_n,m}$ can be estimated by

$$\widehat{\mathbf{\Gamma}}_{Q_n,m} = \sum_{\ell=1}^{m}\sum_{i=1}^{k}\sum_{j=i}^{k}\left(1-\frac{\ell}{n}\right)\widehat{cov}^2_{Q_n}(\boldsymbol{y}^*_t\boldsymbol{y}^{*'}_t, x^*_{ij,t-\ell}). \tag{26}$$

For simplicity, the estimate $\widehat{\mathbf{\Gamma}}_{Q_n,m}$ is used in the simulations and application. The stability of the results is checked by means of several choices of m. For a given estimate $\widehat{\mathbf{\Gamma}}_{Q_n,m}$, the eigenvalue-eigenvector analysis is performed to obtain the sample RPVCs. Specifically, the v-th sample robust volatility component is $\widehat{z}_{vt} = \widehat{\boldsymbol{m}}'_v\boldsymbol{y}_t$, where $\boldsymbol{m}_v$ is the normalized eigenvector of the v-th eigenvalue of $\widehat{\mathbf{\Gamma}}_{Q_n,m}$.

Finally, a robust hypothesis testing to check the ARCH dependence of the sample RPVCs at all lags is presented. The robust generalized Ling-Li test statistic is defined as

$$G_{Q_n,p_n,s} = \frac{n\sum_{j=1}^{n-1}\omega^2(j/p_n)\hat{\rho}_{Q_n,j,s} - M_n(\omega)}{[2\Delta_{Q_n}V_n(\omega)]^{1/2}}, \tag{27}$$

where $\hat{\rho}_{Q_n,j,s}$ is the robust correlation between $\hat{\epsilon}^*_t$ and $\hat{x}^*_{t-j}$, based on the robust autocorrelation estimator proposed by Ma e Genton (2000); and, $\Delta_{Q_n} = 1 + 2\sum_{h=1}^{\infty} cov^2_{Q_n}(x^*_t, x^*_{t-h})$, that can be estimated by a smoothing method, as in Equation 19, but here considering the robust covariance (cov_{Q_n}).

The RPVC method and the robust generalized Ling-Li test statistic proposed in this paper were concentrated on empirical investigations presented in Section 4. Therefore, the asymptotic properties of generalized robust kurtosis matrix and the proofs relating, specifically, to the robust generalized Ling-Li statistic test, remain open problems and these are within the current research topics.

4 Simulations

Several simulations were conducted to study the finite-sample performance of the proposed RPVC analysis and the robust test statistic $G_{Q_n,p_n,s}$ of Equation (27), in the presence of additive outliers. Firstly, the accuracy of the RPVC in the estimation of the no-ARCH component was investigated. Secondly, the effects of the choice of tuning parameters p_n and s_n on the behavior of $G_{Q_n,p_n,s}$ were evaluated. The performance of the RPVC and $G_{Q_n,p_n,s}$ were compared with the classical PVC and the $G_{p_n,s}$ proposed by Hu e Tsay (2014), in the presence of additive outliers.

The DGP employed to generate the multivariate series without outliers was the model described in Equation (13), with $\boldsymbol{y}_t$ being a five-dimensional series (i.e., $k = 5$) and $\boldsymbol{f}_t = (f_{1t}, \ldots, f_{rt})'$ such that $f_{it} = \sigma_{it}e_{it}$, where

$$\sigma^2_{1,t} = 1 + 0.9f^2_{1,t-1}, \quad \sigma^2_{2,t} = 2 + 0.8f^2_{2,t-1},$$

$$\sigma^2_{3,t} = 3 + 0.7f^2_{3,t-1}, \quad \sigma^2_{4,t} = 1 + 0.95f^2_{4,t-1},$$

and $\{e_{it}\}$ are sequences of independent and identically distributed (i.i.d.) standard normal random variables and $\{e_{it}\}$ and $\{e_{jt}\}$ are independent for $i \neq j$. The loading matrix is $\boldsymbol{H} = [\boldsymbol{h}_1, \boldsymbol{h}_2, \boldsymbol{h}_3, \boldsymbol{h}_4]$ with $\boldsymbol{h}_1 = (1,1,1,1,1)'$, $\boldsymbol{h}_2 = (1,0,0,0,0)'$, $\boldsymbol{h}_3 = (0,1,0,-1,0)'$, and $\boldsymbol{h}_4 = (0,0,-1,0,1)'$, and the noise term $\boldsymbol{\epsilon}_t$ is a sequence of i.i.d. standard multivariate normal random vectors.

The contaminated process $\boldsymbol{y}_t^*$ was simulated according to Equation (20). The probability of an outlier occurring at the time t was $p = 0.01$ and, without loss of generality, it is also assumed that $\boldsymbol{\omega} = [10\sigma_1, 10\sigma_2, 10\sigma_3, 0, 0]'$, where $\sigma_i = \sqrt{y_{i,t}}, i = 1,2,3$. Thus, $y_{1,t}^*$, $y_{2,t}^*$ and $y_{3,t}^*$, $t = 1, \ldots, n$, are the tree processes in $\boldsymbol{y}_t^* = \{y_{1,t}^*, y_{2,t}^*, y_{3,t}^*, y_{4,t}^*, y_{5,t}^*\}'$ contaminated with additive outliers. More details and other methods to include outliers in time series with ARCH effects can be seen in Carnero, Peña e Ruiz (2001), Carnero, Peña e Ruiz (2012), among others.

4.1 Performance of the RPVC

In order to analysis the performance of the RPVC and to compare with the classical PVC, the vector $\boldsymbol{M}_1 = (0,-1,1,-1,1)'$ was considered, that gives rise to the no-ARCH component, i.e., $\boldsymbol{M}_1'\boldsymbol{H} = \boldsymbol{0}$. Furthermore, let $\widehat{\boldsymbol{m}}_1$ be the normalized eigenvector corresponding to the smallest eigenvalue of the $\widehat{\boldsymbol{\Gamma}}_{Q_n,m}$ matrix. For the tradicional PVC the $\widehat{\boldsymbol{\Gamma}}_m$ matrix was adopted; see Equation (4). Under the proposed RPVC analysis, $\widehat{\boldsymbol{m}}_1$ is a consistent estimator of $\boldsymbol{M}_1$. Based on Hu e Tsay (2014), two statistics were used to measure the performance of the RPVC

$$R_1 = \frac{|\widehat{\boldsymbol{m}}_1'\boldsymbol{H}(\boldsymbol{H}'\boldsymbol{H})^{-1}\boldsymbol{H}'\widehat{\boldsymbol{m}}_1|}{|\widehat{\boldsymbol{m}}_1'\widehat{\boldsymbol{m}}_1|}, \quad R_2 = cor^2(\widehat{\boldsymbol{m}}_1'\boldsymbol{y}_t^*, \boldsymbol{M}_1\boldsymbol{y}_t^*), \tag{28}$$

which are expected to be close to zero $(R_1 \approx 0)$ and one $(R_2 \approx 1)$, respectively. For the classical PVC, the uncontaminated process $\boldsymbol{y}_t$ was considered for the R_2 statistic.

In simulations, different choices of m $(m \in \{5, 10, 20\})$ for the cumulative generalized robust kurtosis matrix $\widehat{\boldsymbol{\Gamma}}_{Q_n,m}$ of Equation (26) was considered and sample size ranging from 250 to 1000. For a given sample size, 1000 data sets of $\boldsymbol{y}_t$ (without outliers) and $\boldsymbol{y}_t^*$ (with outliers) were generated, and the proposed RPVC and the classical PVC were applied. For comparison, the performance of the traditional PVC (HU; TSAY, 2014) is reported in Table 1 and the performance of the RPVC is presented in Table 2.

The results in Table 1 show that the classical PVC works well when the data set have no outliers, especially for large sample sizes. In this case, the R_1 and R_2 statistics are close to zero and one, respectively. However, in the presence of additive outliers, the accuracy of the classical PVC is destroyed. In particular, the R_1 criterion increases and the R_2 criterion falls. This was expected, since the presence of outliers masks or generates the volatility of data and leads to spurious results of the classical estimated generalized kurtosis matrix. As can be seen in Table 2, when the proposed robust principal volatility component is adopted, the estimates were satisfactory for both cases: without and with outliers. This fact indicates that the robust methodology (RPVC) can be used when one is uncertain of the presence of outliers in the series.

In general, the choice of m does not influence the results.

Table 1: Summary statistics of the performance of the classical PVC, without and with outliers

	Without outliers					
	$m = 5$		10		20	
Sample size	R_1	R_2	R_1	R_2	R_1	R_2
250	0.023	0.834	0.020	0.840	0.025	0.835
500	0.019	0.911	0.024	0.909	0.034	0.912
1000	0.007	0.958	0.005	0.958	0.006	0.960
	With outliers					
	$m = 5$		10		20	
Sample size	R_1	R_2	R_1	R_2	R_1	R_2
250	0.350	0.260	0.330	0.243	0.368	0.207
500	0.307	0.301	0.327	0.267	0.319	0.242
1000	0.273	0.373	0.279	0.331	0.266	0.317

Note: based on Hu e Tsay (2014), the classical cumulative generalized kurtosis matrix, $\widehat{\mathbf{\Gamma}}_m$, was used. In the estimates was adopted the Huber's function, with $c = 2.5$.

Table 2: Summary statistics of the performance of the RPVC, without and with outliers

	Without outliers					
	$m = 5$		10		20	
Sample size	R_1	R_2	R_1	R_2	R_1	R_2
250	0.013	0.877	0.014	0.872	0.014	0.874
500	0.006	0.929	0.007	0.922	0.005	0.936
1000	0.004	0.952	0.003	0.958	0.003	0.962
	With outliers					
	$m = 5$		10		20	
Sample size	R_1	R_2	R_1	R_2	R_1	R_2
250	0.019	0.929	0.018	0.934	0.018	0.933
500	0.010	0.958	0.010	0.956	0.010	0.958
1000	0.006	0.972	0.007	0.978	0.006	0.978

Note: the cumulative generalized robust kurtosis matrix, $\widehat{\mathbf{\Gamma}}_{Q_n,m}$, was used according to Equation (26). Huber's function was not used.

4.2 Performance of the robust generalized test statistic ($G_{Q_n,p_n,s}$)

The simulations were made considering, in Equations (18) and (27), $p_n \in \{5, 10, 20\}$[4], $s_n \in \{5, 10, 20\}$ and sample size ranging from 250 to 2000. Again, for a given sample size, the DGP to generate 1000 data sets of $\boldsymbol{y}_t$ (without outliers) and $\boldsymbol{y}_t^*$ (with outliers) were used, and the classical PVC and the proposed RPVC were applied. In order to compare the performance of the $G_{p_n,s}$, proposed by Hu e Tsay (2014), and $G_{Q_n,p_n,s}$, proposed in this article, the analysis

[4]As described in (LING; LI, 1997), if p_n is too large, the power of the statistic is usually lower. In contrast, if p_n is too small, the test may miss some high order dependencies.

was divided in two forms: i) Tables 3 and 4 present the size and the power of the classical $G_{p_n,s}$, for data without and with outliers, respectively, using the classical PVC; and, ii) Table 5 shows the size and the power of the robust generalized Ling-Li test statistic ($G_{Q_n,p_n,s}$), for time series with outliers, in this case, using the RPVC. In addition, the $\widehat{\Delta}$, for $G_{p_n,s}$, and $\widehat{\Delta}_{Q_n}$, for $G_{Q_n,p_n,s}$, that are used to adjust for the ARCH effects in the data, is reported. The results of robust generalized Ling-Li test statistic, for time series without outliers, are available upon request.

Based on results the of Table 1, the $\widehat{\mathbf{\Gamma}}_{10}$, with $c = 2.5$ in the Huber's function, was adopted to verify the performance of the $G_{p_n,s}$. According to Table 2, the performance of $G_{Q_n,p_n,s}$ was made taking the $\widehat{\mathbf{\Gamma}}_{Q_n,10}$. Since in the simulations there are five-dimensional series ($k = 5$) and four common volatility factors ($r = 4$), the size and the power of the statistical tests are obtained from the hypotheses $s = 1$ and $s = 2$, respectively.

In Table 3, it can be seen that the classical $G_{p_n,s}$ does not fare well for small sample sizes, even for data without outliers. However, the test statistic works well when the sample size is large ($n = 2000$). Furthermore, the choice of p_n has relatively significant effects on size and power. The choice of s_n has little impact on size and power. The adjusted term $\widehat{\Delta}$ is far away from one, indicating strong ARCH effects in the data.

Table 3: Size and power of the generalized Ling-Li test statistic, $G_{p_n,s}$, considering the classical PVC, in series without outliers

			Size					
			Type I: 0.05			Type II: 0.10		
Sample size	s_n	$\widehat{\Delta}$	$p_n = 5$	10	20	5	10	20
250	5	1.494	0.040	0.052	0.062	0.063	0.078	0.096
500		1.610	0.041	0.055	0.067	0.058	0.068	0.086
2000		1.861	0.058	0.057	0.065	0.085	0.077	0.089
250	10	1.576	0.036	0.051	0.059	0.063	0.077	0.093
500		1.712	0.040	0.054	0.065	0.055	0.068	0.083
2000		2.019	0.050	0.054	0.061	0.080	0.073	0.085
250	20	1.639	0.034	0.049	0.058	0.062	0.076	0.090
500		1.765	0.039	0.054	0.065	0.055	0.067	0.082
2000		2.077	0.047	0.050	0.060	0.077	0.073	0.085
			Power					
250	5		0.292	0.243	0.209	0.336	0.293	0.262
500			0.612	0.548	0.501	0.659	0.595	0.553
2000			0.992	0.983	0.966	0.995	0.990	0.977
250	10		0.287	0.238	0.205	0.330	0.289	0.259
500			0.606	0.542	0.491	0.652	0.587	0.544
2000			0.992	0.982	0.965	0.995	0.990	0.975
250	20		0.284	0.236	0.200	0.325	0.283	0.252
500			0.604	0.538	0.490	0.647	0.587	0.542
2000			0.992	0.982	0.965	0.995	0.988	0.975

Note: based on Hu e Tsay (2014), the classical cumulative generalized kurtosis matrix, $\widehat{\mathbf{\Gamma}}_{10}$, was used. In the estimates of the $\widehat{\mathbf{\Gamma}}_{10}$ the Huber's function was adopted. For generalized Ling-Li test statistic the Huber's function was also used, as in Hu e Tsay (2014). In both cases, $c = 2.5$.

The results of Table 4 demonstrate that in the presence of outliers, the size of the generalized Ling-Li test statistic is overestimated at the 5% and 10% levels. That is, there is a over-rejection of the null hypothesis, independently of the sample size. Thus, the outliers are able to generate spurious conditional heteroscedasticity, increasing the rejection of the null hypothesis of homoscedasticity. Regarding the power, it is possible to observe that the generalized Ling-Li test does not work well in all situations. That is, the outliers hide genuine conditional heteroscedasticity. Thus, the test rejects the null hypothesis of homoscedasticity too often when it is in fact true, while the test has difficulty detecting genuine ARCH effects. Furthermore, the adjusted term $\widehat{\Delta}$ is reduced to close from one, indicating that the ARCH effects, in data with outliers, are dissipated, i.e., the outliers mask the conditional heteroscedasticity (volatility) of the series.

Table 4: Size and power of the generalized Ling-Li test statistic, $G_{p_n,s}$, considering the classical PVC, in series with outliers

			Size					
			Type I: 0.05			Type II: 0.10		
Sample size	s_n	$\widehat{\Delta}$	$p_n = 5$	10	20	5	10	20
250	5	1.039	0.122	0.099	0.098	0.152	0.131	0.121
500		1.050	0.211	0.184	0.156	0.251	0.219	0.200
2000		1.097	0.668	0.592	0.508	0.700	0.640	0.555
250	10	1.060	0.120	0.099	0.098	0.150	0.131	0.120
500		1.066	0.211	0.184	0.156	0.250	0.219	0.199
2000		1.118	0.667	0.592	0.503	0.698	0.638	0.555
250	20	1.096	0.119	0.098	0.094	0.147	0.129	0.118
500		1.087	0.210	0.182	0.156	0.247	0.215	0.199
2000		1.129	0.667	0.590	0.502	0.698	0.638	0.555
			Power					
250	5		0.160	0.141	0.114	0.184	0.154	0.128
500			0.231	0.193	0.157	0.264	0.224	0.187
2000			0.647	0.576	0.494	0.697	0.628	0.548
250	10		0.159	0.140	0.113	0.181	0.154	0.128
500			0.231	0.192	0.156	0.264	0.223	0.187
2000			0.647	0.573	0.494	0.695	0.626	0.544
250	20		0.157	0.135	0.111	0.180	0.154	0.125
500			0.230	0.191	0.157	0.263	0.223	0.184
2000			0.644	0.573	0.492	0.694	0.629	0.541

Note: based on Hu e Tsay (2014), the classical cumulative generalized kurtosis matrix, $\widehat{\mathbf{\Gamma}}_{10}$, was used. In the estimates of the $\widehat{\mathbf{\Gamma}}_{10}$ the Huber's function was adopted. For generalized Ling-Li test statistic the Huber's function was also used, as in Hu e Tsay (2014). In both cases, $c = 2.5$.

Finally, the size and power of the robust generalized Ling-Li test, $G_{Q_n,p_n,s}$, are presented in Table 5. As can be seen, there is some over-rejection of the null hypothesis when the size of the test is considered, especially, for small sample sizes. However, for large sample size the robust test works well. In addition, it is notable that the $G_{Q_n,p_n,s}$ presented good power for $n = 2000$.

In general, comparing Tables 3 and 5, it can observed that the power of the classical $G_{p_n,s}$, in series without outliers, is larger than the power of the robust $G_{Q_n,p_n,s}$, in data with outliers.

This can be considered as a kind of "insurance premium" one has to pay in order to be protected against the detrimental effects of outliers since, in the presence of outliers, the power of the classical $G_{p_n,s}$ drops significantly (see, Table 4)[5]. Besides, the adjusted robust term $\widehat{\Delta}_{Q_n}$ is far away from one, indicating strong ARCH effects in the data, when the $G_{Q_n,p_n,s}$ is used.

Table 5: Size and power of the robust generalized Ling-Li test statistic, $G_{Q_n,p_n,s}$, considering the RPVC, in series with outliers

			Size					
			Type I: 0.05			Type II: 0.10		
Sample size	s_n	$\widehat{\Delta}$	$p_n = 5$	10	20	5	10	20
250	5	1.309	0.084	0.083	0.081	0.132	0.127	0.125
500		1.413	0.083	0.082	0.082	0.125	0.122	0.116
2000		1.628	0.067	0.065	0.062	0.117	0.111	0.105
250	10	1.440	0.071	0.070	0.069	0.115	0.112	0.110
500		1.552	0.083	0.080	0.076	0.123	0.120	0.114
2000		1.823	0.066	0.063	0.057	0.112	0.106	0.102
250	20	1.550	0.066	0.063	0.056	0.101	0.099	0.098
500		1.657	0.079	0.078	0.073	0.121	0.114	0.106
2000		1.896	0.062	0.060	0.055	0.109	0.106	0.103
			Power					
250	5		0.287	0.293	0.283	0.342	0.335	0.338
500			0.393	0.389	0.400	0.478	0.460	0.479
2000			0.880	0.848	0.829	0.910	0.892	0.860
250	10		0.272	0.277	0.271	0.381	0.356	0.387
500			0.385	0.384	0.388	0.441	0.449	0.449
2000			0.878	0.846	0.818	0.910	0.898	0.858
250	20		0.255	0.257	0.259	0.366	0.334	0.366
500			0.375	0.368	0.369	0.434	0.443	0.430
2000			0.889	0.846	0.813	0.913	0.893	0.857

Note: based on Section 3, the cumulative generalized robust kurtosis matrix, $\widehat{\mathbf{\Gamma}}_{Q_n,10}$, was used according to Equation (26). For robust generalized Ling-Li test statistic the $G_{Q_n,p_n,s}$ presented in Equation (27) was adopted. Huber's function was not used in both cases.

5 Application

5.1 Study area

The study area included the GVR, Espírito Santo, Brazil, located on the south coast of the Atlantic Ocean [latitude 20°19 S (South), longitude 40°20 W (West)]. Because it is situated in the coastal region, the GVR has hot tropical climate (Aw), with a mild and dry winter and rainy and hot summer. Average temperatures range between 24°C and 30°C. The prevailing winds are from North/Northeast in the spring-summer, undergoing changes during autumn and winter

[5]Note that the classical $G_{p_n,s}$ presents the size closer of the nominal levels, 5% and 10%, than the robust $G_{Q_n,p_n,s}$. Again, an "insurance premium" against the effects of outliers. As mentioned by Franses, Dijk e Lucas (1998), the protection against aberrant observations sometimes comes at a cost.

due to the positioning of the high pressure system (South Atlantic Subtropical High Pressure) closer to the continent, allowing changes in the prevailing wind direction, which starts to vary between the South and West directions.

In the GVR there are eight monitoring stations belonging to the State Institute of Environment and Water Resources (IEMA), namely: Laranjeiras, Carapina, Jardim Camburi (Camburi), Enseada do Suá (Sua), Vitória-Centro (VixCentro), Vila Velha-Ibes (Ibes), Vila Velha-Centro (VVCentro) and Cariacica. The stations are part of the Automatic Air Quality Monitoring Network (AQAMN). The AQAMN measures the following pollutants: PM_{10}, total suspended particles (TPS); SO_2; CO; NO_X; hydrocarbons (HC); and, O_3. In addition, the AQAMN monitors some meteorological parameters, such as: wind direction (WD); wind speed (WS); relative humidity (RH); precipitation (PP); atmospheric pressure (P); temperature (T); and, solar radiation (R).

5.2 Application to air pollution

5.2.1 General aspects

The application of this paper was based on Liu e Johnson (2002), Liu e Johnson (2003), Liu (2007), Liu (2009) and Liu et al. (2013). These authors adopted the traditional PCA in the context of multiple regression and Box-Jenkins time series models. According to Liu (2009), although traditional methodologies such as multiple regression and ARIMA have been used to simulation and forecast air pollutants (for example, O_3 and PM_{10}), the problem of under predictions of exceedances (high levels of pollution) was still unsolved. Then, the authors applied an alternative method that use the principal component analysis to create an extra explanatory variable in order to trigger the peak of air pollutant concentrations. In general, the variable PC trigger was designed to summarize atmospheric circumstances when the concentration of a pollutant is larger than a certain level. In all cases, the results were better when the authors used the PC trigger than without.

It should be noted that a common feature of these studies was to neglect the dependence of the data in building the PC trigger variable. In addition, the PC trigger does not consider the possible presence of outliers in the data set, that may affect the classical PCA (the results may be spurious). Then, as the environmental time series adopted here present volatility and outliers (which can be defined as outliers from the statistical point of view), this paper proposes to apply the robust principal volatility component analysis to create the variable RPVC trigger, instead of the traditional PCA.

The period of analysis was from January 2005 to December 2012 (daily data). Regarding air pollution data, the PM_{10} concentrations of Laranjeiras station was chosen to estimate the linear regression model, because this station showed the highest PM_{10} exceedances of the level of 50 μg/m^3 (the guidelines of the World Health Organization (WHO) for short exposure is equal to 50 μg/m^3) among the eight monitoring stations of GVR. Here, the Cariacica station

was not considered, since it presented a largest number of invalid and missing data. As both the classical PVC (HU; TSAY, 2014) and the RPVC work best when the sample size is greater than 500 ($n \geq 500$), 40 μg/m^3 was selected to be the threshold in this study.

In addition to PM_{10} concentrations, to create the trigger variable and to estimate the regression model, the following variables were used: i) to represent meteorological conditions the temperature, relative humidity, precipitation, wind direction and wind speed were adopted. According to Seinfeld e Pandis (2006), in general, the temperature affects fuel usage and ambient chemical reactions; and, precipitation and relative humidity largely remove pollutants from the atmosphere. In relation to wind speed, if the region has more stationary sources and less dust on the ground, for example, the increase in wind speed disperses the pollutant emitted at source, reducing the PM_{10} concentrations. Unlike, due to the resuspension effects of the wind and the potential of winds to transport particulates between regions, in the places with a large amount of soil dust resuspension the correlation between wind speed and PM_{10} tends to be positive. Furthermore, if the wind direction is relatively constant, the same area is exposed continuously to high pollution levels. On the other hand, if the wind direction is changed constantly, the pollutants are dispersed over a larger area and the concentrations of any exposed area are smaller (LIU; LIPTAK, 1997); ii) Particulate matter can also form in the atmosphere from gases such as SO_2, nitrogen oxides (NO_X), NO_2 and volatile organic compounds (VOCs), which are emitted mainly due to combustion , turning into particles as a result of chemical reactions in the air. Thus, based on Liu (2009), the pollutants O_3 and NO_2 were used[6]; and, iii) given the correlation and the proximity between Laranjeiras, Carapina and Camburi stations, the PM_{10} concentrations of Carapina and Camburi were also adopted to create the trigger variables, named PVC and RPVC. Briefly, the variables were:

- Daily average PM_{10} of Laranjeiras station (μg/m^3): $LARPM_{10}$;
- Daily average PM_{10} of Carapina station (μg/m^3): $CARPM_{10}$;
- Daily average PM_{10} of Camburi station (μg/m^3): $CAMPM_{10}$;
- Daily average temperature ($^\circ$C): T;
- Daily average relative humidity (%): RH;
- Daily average precipitation (mm): PP;
- Daily average scalar wind direction ($^\circ$): WD;
- Daily average scalar wind speed (m/s): WS;
- Daily average ozone (μg/m^3): O_3;
- Daily average nitrogen dioxide (μg/m^3): NO_2;

[6]The pollutant SO_2 was not considered due to several missing data in its series.

- Principal volatility component of larger volatility, according to Hu e Tsay (2014): PVC. The most volatile component was adopted.
- Robust principal volatility component of larger volatility: RPVC. The most robust volatile component was adopted.

5.2.2 Results and discussions

Figure 1 displays the plots of the environmental variables from January 2005 to December 2012. As can be observed, the series present high levels which can be identified, from statistical point of view, as being outliers (additive). In Table 6, the descriptive statistics of the variables are summarized. From this table, it can be noted that the maximum values are much larger than the third quartile quantities, except for temperature (T). This may be an indication that the data present aberrant observations.

In addition, there is considerable evidence that the conditional variance is not constant over time, so that conditional heteroscedastic models seem to be appropriate choice for capturing the time-varying volatility in the level of the series. The robust ACFs of the squared series and the ARCH-LM test (test to verify the presence of volatility) revealed that there is conditional heteroscedasticity (volatility) in the variables, a feature expected for air pollution and meteorological time series (results available upon request). Thus, this empirical evidence justifies the use of the RPVC analysis to create the variable RPVC trigger, instead of the classical PCA.

Table 6: The descriptive statistics of the variables

	$LARPM_{10}$	$CARPM_{10}$	$CAMPM_{10}$	T	RH
Min.	4.27	2.71	3.54	17.00	59.40
1st Qu.	24.46	17.91	20.88	22.41	72.87
Median	31.54	21.56	25.63	24.34	77.42
Mean	32.79	22.96	26.48	24.35	77.68
3rd Qu.	39.42	26.71	31.21	26.37	82.12
Max.	106.88	88.25	78.08	30.89	97.65
	PP	WD	WS	O_3	NO_2
Min.	0.00	20.90	0.86	8.33	3.40
1st Qu.	0.00	93.32	1.60	26.48	17.36
Median	0.00	152.03	1.99	30.79	22.09
Mean	0.17	144.95	2.05	32.44	22.32
3rd Qu.	0.11	200.78	2.43	37.34	26.56
Max.	6.74	262.85	5.49	74.18	49.43

To compare the performance of the RPVC trigger proposed in this paper, three models are estimated (Model 1, Model 2 and Model 3). In all models, due to a very good fit to the log-normal distribution (LIU, 2007; LIU; JOHNSON, 2003), logarithmic-transformed PM_{10} concentrations was used as the dependent variable in the multiple regression. The PVC and RPVC triggers and the multiple regressions estimated are presented below. The significant explanatory variables of the regression models were listed in Table 7 that also presents the adjusted coefficient of determination (R^2 adjusted) and the Root Mean Squared Error (RMSE).

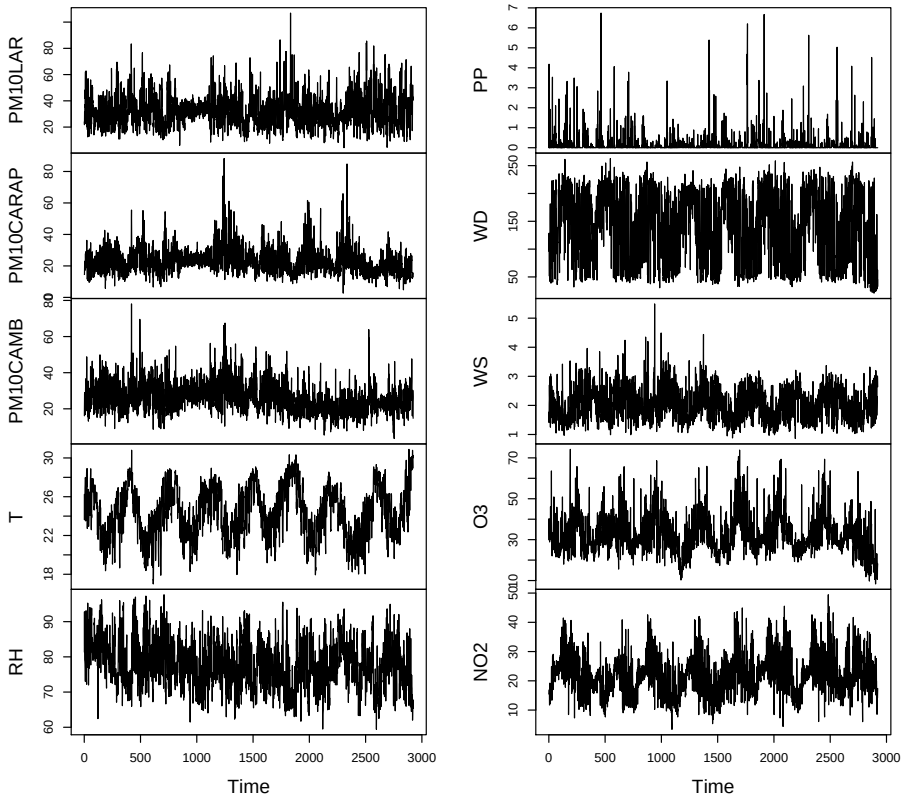


Figure 1: Temporal behaviour of the variables.

Given the volatility of the PM_{10} concentrations, besides modeling the conditional average, the GARCH (GARCH$(1,1)$) technique was adopted to model the volatility (conditional variance) of the residuals. The Box-Pierce and Ljung-Box statistics demonstrated that the sample residuals are not time correlated and which the squares of the residuals presented no conditional heteroscedasticity.

Model 1: multiple regression model for multivariate time series without RPC, PVC and RPVC trigger. In order to demonstrate the usefulness of the RPC, PVC and RPVC trigger, a multiple regression model without RPC, PVC and RPVC trigger was firstly constructed and tested.

$$\begin{aligned} LOG(PM_{10})_t = 1.881 + 0.548LOG(LARPM_{10})_{t-1} + 0.078T_t \\ -0.070T_{t-1} - 0.007RH_t - 0.077PP_t + 0.054WS_t \\ -0.043WS_{t-1} + 0.012NO_{2,t} - 0.007NO_{2,t-1} \end{aligned} \tag{29}$$

Model 2: multiple regression model for multivariate time series with PVC trigger, adopting the PVC analysis proposed by Hu e Tsay (2014), applied to the residuals of the VARFI(3,d) model. In this case, the GARCH filter was not used because the PVC analysis proposed by Hu e Tsay (2014) capture the conditional heteroscedasticity of the data set. The tests showed that there are volatile components. The most volatile component was used in the multiple regression

model. A multiple regression model with PVC trigger.

$$\begin{aligned} PVC_t = &-0.017RESCARPM_{10,t} + 0.030RESCAMPM_{10,t} \\ &+0.315REST_t - 0.053RESRH_t + 0.942RESPP_t \\ &+0.027RESWS_t + 0.088RESNO_{2,t} \end{aligned} \tag{30}$$

$$\begin{aligned} LOG(PMLAR_{10})_t = &1.544 + 0.513LOG(LARPM_{10})_{t-1} + 0.086T_t - 0.062T_{t-1} \\ &-0.006RH_t - 0.088PP_t + 0.066WS_t - 0.039WS_{t-1} \\ &+0.07NO_{2,t} - 0.005NO_{2,t-1} + 0.002PVC_t \end{aligned} \tag{31}$$

Model 3: multiple regression model for multivariate time series with RPVC trigger, adopting the robust PVC analysis proposed in Section 3, applied to the residuals of the VARFI(3,d) model. The tests showed that there are volatile components. The most volatile component was used in the multiple regression model.

$$\begin{aligned} RPVC_t = &0.013RESCARPM_{10,t} - 0.051RESCAMPM_{10,t} \\ &+0.116REST_t - 0.009RESRH_t + 0.974RESPP_t \\ &+0.184RESWS_t + 0.022RESNO_{2,t} \end{aligned} \tag{32}$$

$$\begin{aligned} LOG(PMLAR_{10})_t = &1.935 + 0.565LOG(LARPM_{10})_{t-1} + 0.076T_t - 0.072T_{t-1} \\ &-0.008RH_t - 0.081PP_t + 0.046WS_t - 0.0405WS_{t-1} \\ &+0.010NO_{2,t} - 0.006NO_{2,t-1} + 0.098RPVC_t \end{aligned} \tag{33}$$

Based on Ryan (1995) and Liu e Johnson (2003), some statistics (Table 8) were calculated to compare the capacity of predictions of PM_{10} exceedance days of Models 1, 2 and 3. The statistics measured were: the false alarm rate (FAR), which measures the tendency to overestimate the prediction of PM_{10} exceedance days; the probability of detection (POD), which measures the probability of the model correctly estimates the PM_{10} exceedance days, i.e., predict PM_{10} exceedance days when they actually occurred; and, the loss rate (MISS), which refers to rate at which PM_{10} exceedance days occurred, but were not predicted.

The results in Table 8 demonstrate that Model 3 (that use RPVC trigger) presented better performance in detected the PM_{10} exceedance days. Model 3 estimated the PVC trigger according to Hu e Tsay (2014). As the time series present peaks of concentrations (here, called outliers), these outliers masked (or hid) the volatility of the data set. Model 3 adopted the RPVC (proposed in this paper) to estimate the RPVC trigger. As the RPVC consider (or preserve) the conditional heteroscedasticity in the presence of outliers, the results were expected to be better than Models 1 and 2.

For the purpose of exemplification of the performance measures (FAR, POD and MISS), it is possible to note that Model 3 presented a FAR equal to 0.18, which means that Model 3

Table 7: Parameter statistics significance in the four models

	Model 1		Model 2		Model 3	
Variables	Coef.	S.e.	Coef.	S.e.	Coef.	S.e.
C (CONSTANT)	1.881	0.148	1.544	0.134	1.935	0.140
LOG(PM10LAR)$_{t-1}$	0.548	0.017	0.513	0.016	0.564	0.015
T$_t$	0.078	0.006	0.086	0.005	0.079	0.006
T$_{t-1}$	-0.070	0.006	-0.062	0.006	-0.072	0.006
RH$_t$	-0.007	0.001	-0.006	0.001	-0.008	0.001
RH$_{t-1}$	-	-	-	-	-	-
PP$_t$	-0.077	0.017	-0.088	0.020	-0.081	0.020
PP$_{t-1}$	-	-	-	-	-	-
WD$_t$	-	-	-	-	-	-
WD$_{t-1}$	-	-	-	-	-	-
WS$_t$	0.054	0.016	0.066	0.014	0.046	0.014
WS$_{t-1}$	-0.043	0.016	-0.039	0.013	-0.040	0.013
O$_{3,t}$	-	-	-	-	-	-
O$_{3,t-1}$	-	-	-	-	-	-
NO$_{2,t-1}$	0.012	0.001	0.070	0.001	0.010	0.001
NO$_{2,t-1}$	-0.007	0.001	-0.005	0.001	-0.006	0.001
PVC$_t$	*	*	0.002	0.007	*	*
RPVC$_t$	*	*	*	*	0.098	0.025
R^2 adjusted	0.497		0.643		0.517	
RMSE	0.264		0.237		0.273	

Note: 1) Coef.: coefficients; 2) S.e.: standard error; 3) trace represents the variables that were not statistically significant; and, 4) * represents the variables that were not used in the model.

Table 8: Performance of statistical models to predict PM_{10} concentrations

Model	1	2	3
FAR	0.29	0.21	0.18
POD	0.40	0.55	0.62
MISS	0.20	0.17	0.14
Correlation coefficient	0.70	0.72	0.75

Note: correlation coefficient was calculated for observed and predicted PM_{10} values.

estimated false alarms in 18% of the time, while for Model 1 the rate was equal to 29%. In the case of the POD, it is observed that the probability of correctly detecting PM_{10} exceedances was larger for Model 3 than for Model 1. In this case, 62% of the time the Model 3 successfully estimated the high PM_{10} concentrations. Moreover, the rate (MISS) for the occurrence of episodes which were not detected was smaller for Model 3 than for Model 1.

Finally, the series was divided into two parts: learning and forecasting sets. The observations from January 1st, 2005 to July 31st, 2012 were considered as learning set and the remaining observations were considered for the forecasting study. Figure 2 presents the visual analysis of the one-step-ahead forecast values of the Model 3, that is, from August 1st, 2012 to December 31st, 2012. It can be observed that Model 3 presented a reasonably good performance, including high levels of PM_{10} concentrations.

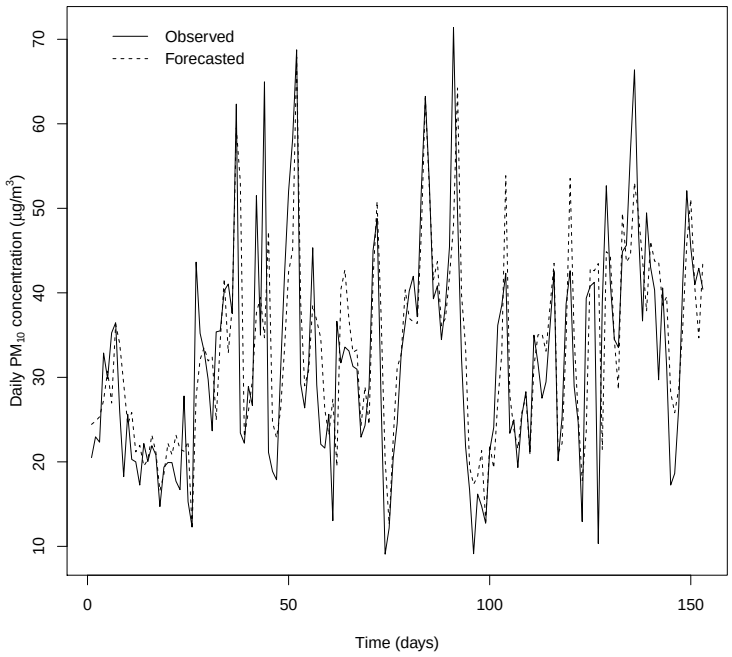


Figure 2: Observed and forecasted PM_{10} concentrations ($\mu g/m^3$) from August 1st, 2012 to December 31st, 2012, one-step-ahead, for Model 3.

6 Conclusions

In this paper a robust principal volatility component (RPVC) and a robust generalized Ling-Li test statistic ($G_{Q_n,p_n,s}$) for high-dimensional time series with conditional heteroscedasticity and additive outliers were proposed. Some results are discussed and these were empirically investigated by Monte Carlo experiments under different scenarios, which showed evidence that the presence of additive outliers affects the performance of the classical principal volatility component and the generalized Ling-Li test statistic proposed by Hu e Tsay (2014). The proposed robust principal volatility component and the robust generalized Ling-Li test statistic performed quite well and this indicates that these robust methods can be very useful in practical applications where there is any evidence of volatility and aberrant observations, such as, high levels of environmental time series. The proposed RPVC was used to improve the predictions of PM_{10} exceedance days in the Laranjeiras station, in the Greater Vitória Region, Espírito Santo, Brazil, which can be very useful for the management of the air quality network. The results in this paper will hopefully stimulate further research on this theme.

Finally, future works are suggested in the following research lines: i) the robust PVC method and the robust generalized Ling-Li test statistic proposed in this paper were concentrated on empirical investigations. Therefore, the asymptotic properties of generalized robust kurtosis matrix and the proofs relating, specifically, to the robust generalized Ling-Li statistic test, remain open problems and these are within the current research topics; and, ii) extending the PVC technique to cases where the temporal series have long memory behavior in volatility.

7 Acknowledgements

The authors acknowledge partial financial support from FAPES/ES and CNPq/Brazil.

References

BRUNEKREEF, B.; HOLGATE, S. T. Air pollution and health. **The Lancet**, v. 360, n. 9341, p. 1233–1242, 2002.

CARNERO, M. A.; PEñA, D.; RUIZ, E. **Outliers and conditional autoregressive heteroscedasticity in time series**. [S.l.], 2001. Disponível em: <http://ideas.repec.org/p/cte/wsrepe/ws010704.html>.

CARNERO, M. A.; PEñA, D.; RUIZ, E. Effects of outliers on the identification and estimation of GARCH models. **Journal of Time Series Analysis**, v. 28, n. 4, p. 471–497, 2007. Disponível em: <http://dx.doi.org/10.1111/j.1467-9892.2006.00519.x>.

CARNERO, M. A.; PEñA, D.; RUIZ, E. Estimating GARCH volatility in the presence of outliers. **Economics Letters**, v. 114, n. 1, p. 86–90, 2012. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0165176511003521>.

CHANG, I.; TIAO, G. C.; CHEN, C. Estimation of time series parameters in the presence of outliers. **Technometrics**, Taylor & Francis, v. 30, n. 2, p. 193–204, 1988.

CHELANI, A. B.; DEVOTTA, S. Air quality forecasting using a hybrid autoregressive and nonlinear model. **Atmospheric Environment**, v. 40, n. 10, p. 1774–1780, 2006. ISSN 1352-2310. Disponível em: <http://www.sciencedirect.com/science/article/pii/S135223100501071X>.

CHEN, C.; LIU, L.-M. Joint estimation of model parameters and outlier effects in time series. **Journal of the American Statistical Association**, Taylor & Francis Group, v. 88, n. 421, p. 284–297, 1993.

COTTA, H. H. A.; REISEN, V. A. Robust principal component analysis with air pollution data: an application to the clustering of RAMQAr. Unpublished manuscript. 2015.

CURTIS, L. et al. Adverse health effects of outdoor air pollutants. **Environment International**, v. 32, n. 6, p. 815–830, 2006. ISSN 0160-4120. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0160412006000444>.

DIJK, D. V.; FRANSES, P. H.; LUCAS, A. Testing for arch in the presence of additive outliers. **Journal of Applied Econometrics**, v. 14, n. 5, p. 539–562, 1999. Disponível em: <http://dx.doi.org/10.1002/(SICI)1099-1255(199909/10)14:5<539::AID-JAE526>3.0.CO;2-W>.

DUCHESNE, P.; LALANCETTE, S. On testing for multivariate arch effects in vector time series models. **Canadian Journal of Statistics**, v. 31, n. 3, p. 275–292, 2003. Disponível em: <http://dx.doi.org/10.2307/3316087>.

DUCHESNE, P.; ROY, R. On consistent testing for serial correlation of unknown form in vector time series models. **Journal of Multivariate Analysis**, v. 89, n. 1, p. 148–180, 2004. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0047259X0300126X>.

ENGLE, R. F. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. **Econometrica**, v. 50, n. 4, p. 987–1007, 1982. Disponível em: <http://ideas.repec.org/a/ecm/emetrp/v50y1982i4p987-1007.html>.

ENGLE, R. F.; KRONER, K. F. Multivariate simultaneous generalized ARCH. **Econometric Theory**, v. 11, p. 122–150, 2 1995. ISSN 1469-4360. Disponível em: <http://journals.cambridge.org/article_S0266466600009063>.

FRANKE, J. Comment. **Journal of Business & Economic Statistics**, v. 32, n. 2, p. 171–172, 2014. Disponível em: <http://dx.doi.org/10.1080/07350015.2014.903652>.

FRANSES, P. H.; DIJK, D. van; LUCAS, A. **Short patches of outliers, ARCH and volatility modeling**. [S.l.], 1998. Disponível em: <https://ideas.repec.org/p/tin/wpaper/19980057.html>.

GRANé, A.; VEIGA, H. Outliers,GARCH-type models and risk measures: a comparison of several approaches. **Journal of Empirical Finance**, v. 26, p. 26–40, 2014. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0927539814000061>.

HANNAN, E. J. **Multiple time series**. New Jersey: Wiley, 1970.

HONG, Y. Consistent testing for serial correlation of unknown form. **Econometrica**, v. 64, n. 4, p. 837–864, 1996. Disponível em: <http://www.jstor.org/stable/2171847>.

HOTTA, L.; TSAY, R. **Outliers in GARCH processes**. Chicago, 1998.

HU, Y.-P.; TSAY, R. S. Principal volatility component analysis. **Journal of Business & Economic Statistics**, v. 32, n. 2, p. 153–164, 2014. Disponível em: <http://dx.doi.org/10.1080/07350015.2013.818006>.

JOHNSON, C. R. (Ed.). **Proceedings of symposia in applied mathematics**. [S.l.]: American Mathematical Society, 1989.

JOLLIFFE, I. T. **Principal component analysis**. 2th. ed. [S.l.]: Prentice Hall, 2002.

LAURENT, S.; BAUWENS, L.; ROMBOUTS, J. V. K. Multivariate GARCH models: a survey. **Journal of Applied Econometrics**, v. 21, n. 1, p. 79–109, 2006. Disponível em: <http://ideas.repec.org/a/jae/japmet/v21y2006i1p79-109.html>.

LI, W. et al. Modelling multivariate volatilities via latent common factors. To appear in Journal of Business & Economic Statistics. 2015.

LING, S.; LI, W. K. Diagnostic checking of nonlinear multivariate time series with multivariate arch errors. **Journal of Time Series Analysis**, v. 18, n. 5, p. 447–464, 1997. Disponível em: <http://dx.doi.org/10.1111/1467-9892.00061>.

LIU, D. H.; LIPTAK, B. G. **Environmental engineers' handbook**. 2th. ed. New York: CRC Press, 1997. 1454 p.

LIU, P.-W. G. Establishment of a Box-Jenkins multivariate time-series model to simulate ground-level peak daily one-hour ozone concentrations at Ta-Liao in Taiwan. **Journal of the Air & Waste Management Association**, v. 57, n. 9, p. 1078–1090, 2007. Disponível em: <http://dx.doi.org/10.3155/1047-3289.57.9.1078>.

LIU, P.-W. G. Simulation of the daily average PM_{10} concentrations at ta-liao with box-jenkins time series models and multivariate analysis. **Atmospheric Environment**, v. 43, n. 13, p. 2104–2113, 2009. ISSN 1352-2310. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1352231009000247>.

LIU, P.-W. G.; JOHNSON, R. Forecasting peak daily ozone levels-I. A regression with time series errors model having a principal component trigger to fit 1991 ozone levels. **Journal of the Air & Waste Management Association**, v. 52, n. 9, p. 1064–1074, 2002. Disponível em: <http://dx.doi.org/10.1080/10473289.2002.10470841>.

LIU, P.-W. G.; JOHNSON, R. Forecasting peak daily ozone levels: part 2- A regression with time series errors model having a principal component trigger to forecast 1999 and 2002 ozone levels. **Journal of the Air & Waste Management Association**, v. 53, n. 12, p. 1472–1489, 2003. Disponível em: <http://www.tandfonline.com/doi/abs/10.1080/10473289.2003.10466321>.

LIU, P.-W. G. et al. Establishing multiple regression models for ozone sensitivity analysis to temperature variation in Taiwan. **Atmospheric Environment**, v. 79, p. 225–235, 2013. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1352231013004561>.

MA, Y.; GENTON, M. G. Highly robust estimation of the autocovariance function. **Journal of Time Series Analysis**, Blackwell Publishers Ltd, v. 21, n. 6, p. 663–684, 2000. ISSN 1467-9892. Disponível em: <http://dx.doi.org/10.1111/1467-9892.00203>.

MATTESON, D. S.; TSAY, R. S. Dynamic orthogonal components for multivariate time series. **Journal of the American Statistical Association**, v. 106, n. 496, p. 1450–1463, 2011. Disponível em: <http://dx.doi.org/10.1198/jasa.2011.tm10616>.

MAYNARD, R. Key airborne pollutants: the impact on health. **Science of The Total Environment**, v. 334-335, n. 0, p. 9–13, 2004. ISSN 0048-9697. Highway and Urban Pollution. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0048969704003493>.

REISEN, V. A. et al. Modeling and forecasting daily average PM_10 concentrations by a seasonal long-memory model with volatility. **Environmental Modelling & Software**, v. 51, p. 286–295, 2014. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1364815213002260>.

ROUSSEEUW, P. J.; CROUX, C. Alternatives to the median absolute deviation. **Journal of the American Statistical Association**, Taylor & Francis Group, v. 88, n. 424, p. 1273–1283, 1993.

RYAN, W. F. Forecasting severe ozone episodes in the Baltimore metropolitan area. **Atmospheric Environment**, v. 29, n. 17, p. 2387–2398, 1995. Disponível em: <http://www.sciencedirect.com/science/article/pii/1352231094003022>.

SEINFELD, J. H.; PANDIS, S. N. **Atmospheric chemistry and physics: from air pollution to climate change**. 2th. ed. New York: Wiley, 2006. 1232 p.

TSAY, R. S. Outliers, level shifts, and variance changes in time series. **Journal of forecasting**, v. 7, n. 1, p. 1–20, 1988.

WATSON, J. G. et al. Receptor modeling application framework for particle source apportionment. **Chemosphere**, v. 49, n. 9, p. 1093–1136, 2002. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0045653502002436>.

WHO. **WHO air quality guidelines global update 2005. Report on a working group meeting, Bonn/Germany**. [S.l.], 2005. Disponível em: <http://www.euro.who.int/__data/assets/pdf_file/0008/147851/E87950.pdf>.

WHO. **Air pollution estimates**. [S.l.], 2014. Disponível em: <http://www.who.int/phe/health_topics/outdoorair/databases/FINAL_HAP_AAP_BoD_24March2014.pdf?ua=1>.

6 CONCLUSÕES GERAIS

Uma vez que vários problemas de saúde estão relacionados à poluição atmosférica, as questões referentes à qualidade do ar têm se tornado cada vez mais importantes. Assim, diversos estudos adotando técnicas estatísticas têm sido realizados, com o intuito de contribuir na tomada de decisões dos agentes públicos e privados no que diz respeito ao combate à poluição, à prevenção de altas concentrações e à formulação de legislações para esse fim. Entre essa técnicas, está a análise de componentes principais clássica, sendo a mesma adotada no redimensionamento de rede, em análises de cluster, em análise de regressão, entre outros. Entretando, nota-se que, diversos estudos que adotaram a técnica de ACP tem negligenciado a dependência dos dados e a presença de observações atípicas (outliers). A aplicação da técnica de ACP em séries temporais multivariadas com heterocedasticidade conditional ou outliers, por exemplo, pode levar as resultados espúrios (enganosos), uma vez que a matriz de autocovariância estimada pode ser viesada (estimada incorretamente).

Assim, este trabalho objetivou estudar a técnica de componentes principais em séries temporais multivariadas com heteroscedasticidade condicional e outliers. Duas linhas de pesquisa foram propostas: i) aplicar um filtro multivariado VARFIMA-GARCH aos dados originais e utilizar a ACP robusta sobre os resíduos do modelo VARFIMA-GARCH. Com esse modelo, buscou-se filtrar, além da volatilidade, a correlação temporal e o comportamento de memória longa; e, ii) estender a técnica de componentes principais com volatilidade (PVC) proposta por Hu e Tsay (2014) para uma abordagem robusta, a fim de captar a volatilidade presente nos processos temporais multivariados, mas, levando-se em consideração os efeitos de outliers sobre a variância condicional. As duas linhas de pesquisa deram origem à dois artigos principais, estando suas conclusões descritas a seguir.

No primeiro artigo, o objetivo principal foi verificar o uso da técnica de ACP em séries temporais multivariadas com heteroscedasticidade condicional e memória longa. Os resultados demostraram que a correlação temporal, a volatilidade e a memória longa podem afetar a matriz de autocovariância, com reflexos sobre seus autovalores e autovetores. No caso dos autovalores, uma grande porcentagem da explicação da variabilidade do conjunto de dados foi direcionada para o primeiro componente principal. Além disso, os componentes principais gerados a partir do método convencional de PCA apresentaram correlação serial e correlação cruzada. A adoção do filtro VARFIMA-GARCH sazonal, com posterior aplicação da ACP robusta em seus resíduos, permitiu a correção de tais problemas. Isso foi corroborado por meio de uma aplicação ao poluente MP_{10}, na RGV, Espírito Santo, Brasil, mas pode ser aplicado em situações reais em várias áreas de estudo.

Em relação ao segundo artigo, foram propostos: i) método de PVC robusto; e, ii) teste estatístico de Ling-Li generalizado robusto para redução de dimensão em séries temporais multivaridas com volatilidade estocástica e outliers aditivos. Os resultados empíricos mostraram evidências de que a presença de outliers aditivos afeta o desempenho do método de PVC

clássico e do teste estatístico de Ling-Li generalizado proposto por Hu e Tsay (2014). Os métodos propostos no artigo, PVC robusto e teste estatístico de Ling-Li generalizado robusto, tiveram bom desempenho na presença ou não de outliers, o que indica que esses métodos robustos podem ser muito útil em aplicações práticas, onde não há qualquer evidência de observações aberrantes, tais como, os altos níveis de concentrações da poluição ar. A análise de PVC robusta foi usada para melhorar as predições das concentrações de MP_{10} quando as mesmas ultrapassaram determinado nível de poluição, na estação de Laranjeiras, RGV, Espírito Santo, Brasil, o que pode ser muito útil para a gestão da rede de qualidade do ar da região.

Para trabalhos futuros sugere-se: i) no caso do primeiro artigo, a aplicação da técnica de PCA sobre os dados filtrados considerou um período de tempo contínuo. Uma avaliação interessante seria fragmentar o período de tempo, em trimestres ou por estações do ano, por exemplo, para verificar se os resultados são semelhantes; ii) o método de PVC robusto e o teste estatístico de Ling-Li generalizado robusto propostos neste trabalho foram concentrados em investigações empíricas. Portanto, as propriedades assintóticas da matriz de curtose robusta generalizada e as provas relativas, especialmente, ao teste estatístico de Ling-Li generalizado robusto, permanecem como problemas abertos e esses estão dentro dos correntes temas de pesquisa; e, iii) estender a técnica de PVC para os casos em que as séries temporais apresentam comportamento de memória longa em volatilidade.

7 REFERÊNCIAS

ANDERSON, T. W. **An introduction to multivariate statistical analysis**. 3rd. ed. New York: John Wiley & Sons, 2003. 752 p.

BAILLIE, R. T.; BOLLERSLEV, T.; MIKKELSEN, H. O. Fractionally integrated generalized autoregressive conditional heteroskedasticity. **Journal of Econometrics**, v. 74, n. 1, p. 3–30, 1996.

BRUNEKREEF, B.; HOLGATE, S. T. Air pollution and health. **The Lancet**, v. 360, n. 9341, p. 1233–1242, 2002.

CARNERO, M. A.; PEÑA, D.; RUIZ, E. Effects of outliers on the identification and estimation of GARCH models. **Journal of Time Series Analysis**, v. 28, n. 4, p. 471–497, 2007.

CHAVENT, M. et al. PCA and PMF based methodology for air pollution sources identification and apportionment. **Environmetrics**, v. 20, n. 8, p. 928–942, 2009.

CONAMA. Dispõe sobre padrões de qualidade do ar, previstos no PRONAR. **Diário Oficial da República Federativa do Brasil**, Brasília, DF, p. 15937–15939, junho 1990.

CURTIS, L. et al. Adverse health effects of outdoor air pollutants. **Environment International**, v. 32, n. 6, p. 815–830, 2006.

DALLAROSA, J. B. **Estudo da formação e dispersão de ozônio troposférico em áreas de atividade de processamento de carvão aplicando modelos numéricos**. 127 f. Dissertação (Mestrado em Sensoriamento Remoto) — Programa de Pós-Graduação em Sensoriamento Remoto, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2005.

DIJK, D. V.; FRANSES, P. H.; LUCAS, A. Testing for arch in the presence of additive outliers. **Journal of Applied Econometrics**, v. 14, n. 5, p. 539–562, 1999.

DING, Z.; GRANGER, C. W.; ENGLE, R. F. A long memory property of stock market returns and a new model. **Journal of Empirical Finance**, v. 1, n. 1, p. 83–106, 1993.

DOMINICK, D. et al. Spatial assessment of air quality patterns in Malaysia using multivariate analysis. **Atmospheric Environment**, v. 60, n. 0, p. 172–181, 2012.

ECOSOFT CONSULTORIA E SOFTWARES AMBIENTAIS. **Inventário de emissões atmosféricas da Região da Grande Vitória**. Vitória, 2011. Disponível em: ⟨http://www.es.gov.br/Banco%20de%20Documentos/PDF/Maio/100511/RTC10131-R1.pdf⟩. Acesso em: 20 de mar. de 2014.

FRANKE, J. Comment. **Journal of Business & Economic Statistics**, v. 32, n. 2, p. 171–172, 2014.

GODISH, T. **Air quality**. 2. ed. Chelsea, Michigan: Lewis, 1991. 422 p.

GOVERNO DO ESTADO DO ESPÍRITO SANTO. Estabelece novos padrões de qualidade do ar e dá providências correlatas. **Diário Oficial do Estado do Espírito Santo**, Vitória, ES, dezembro 2013.

GRAMSCH, E. et al. Examination of pollution trends in Santiago de Chile with cluster analysis of PM_{10} and ozone data. **Atmospheric Environment**, v. 40, n. 28, p. 5464–5475, 2006.

GUO, H.; WANG, T.; LOUIE, P. Source apportionment of ambient non-methane hydrocarbons in Hong Kong: application of a principal component analysis/absolute principal component scores (PCA/APCS) receptor model. **Environmental Pollution**, v. 129, n. 3, p. 489–498, 2004.

HARKAT, M.-F.; MOUROT, G.; RAGOT, J. An improved PCA scheme for sensor FDI: application to an air quality monitoring network. **Journal of Process Control**, v. 16, n. 6, p. 625–634, 2006.

HENRY, R.; HIDY, G. Multivariate analysis of particulate sulfate and other air quality variables by principal components-part i: annual data from Los Angeles and New York. **Atmospheric Environment (1967)**, v. 13, n. 11, p. 1581–1596, 1979.

HU, Y.-P.; TSAY, R. S. Principal volatility component analysis. **Journal of Business &**

Economic Statistics, v. 32, n. 2, p. 153–164, 2014.
INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Banco de dados. Cidades**. Rio de Janeiro, 2014. Disponível em: ⟨http://www.cidades.ibge.gov.br/xtras/home.php⟩. Acesso em: 20 de mar. de 2014.
INSTITUTO ESTADUAL DE MEIO AMBIENTE E RECURSOS HÍDRICOS DO ESTADO DO ESPÍRITO SANTO. **Relatório da qualidade do ar da Região da Grande Vitória**. Vitória, 2014. Disponível em: ⟨http://www.meioambiente.es.gov.br/download/Relat%C3%B3rio_Anual_de_Qualidade_do_Ar_2013.pdf⟩. Acesso em: 27 de jun. de 2015.
JOHNSON, R.; WICHERN, D. **Applied multivariate statistical analysis**. 6rd. ed. New Jersey: Prentice Hall, 2007. 800 p.
LAM, C.; YAO, Q. Factor modeling for high-dimensional time series: Inference for the number of factors. **Ann. Statist.**, The Institute of Mathematical Statistics, v. 40, n. 2, p. 694–726, 04 2012.
LIU, P.-W. G. Establishment of a Box-Jenkins multivariate time-series model to simulate ground-level peak daily one-hour ozone concentrations at Ta-Liao in Taiwan. **Journal of the Air & Waste Management Association**, v. 57, n. 9, p. 1078–1090, 2007.
LIU, P.-W. G. Simulation of the daily average PM_{10} concentrations at Ta-Liao with Box-Jenkins time series models and multivariate analysis. **Atmospheric Environment**, v. 43, n. 13, p. 2104–2113, 2009.
LIU, P.-W. G.; JOHNSON, R. Forecasting peak daily ozone levels-I. A regression with time series errors model having a principal component trigger to fit 1991 ozone levels. **Journal of the Air & Waste Management Association**, v. 52, n. 9, p. 1064–1074, 2002.
LIU, P.-W. G.; JOHNSON, R. Forecasting peak daily ozone levels: part 2- A regression with time series errors model having a principal component trigger to forecast 1999 and 2002 ozone levels. **Journal of the Air & Waste Management Association**, v. 53, n. 12, p. 1472–1489, 2003.
LIU, P.-W. G. et al. Establishing multiple regression models for ozone sensitivity analysis to temperature variation in Taiwan. **Atmospheric Environment**, v. 79, p. 225–235, 2013.
LU, W.-Z. et al. Potential assessment of a neural network model with PCA/RBF approach for forecasting pollutant trends in Mong Kok urban air, Hong Kong. **Environmental Research**, v. 96, n. 1, p. 79–87, 2004.
MATTESON, D. S.; TSAY, R. S. Dynamic orthogonal components for multivariate time series. **Journal of the American Statistical Association**, v. 106, n. 496, p. 1450–1463, 2011.
MAYNARD, R. Key airborne pollutants: the impact on health. **Science of The Total Environment**, v. 334-335, n. 0, p. 9–13, 2004.
MELO, M. M. et al. Application of principal component analysis and logistic regression to investigate the relationship between annoyance and the combined effect of particulate matter. Unpublished manuscript. 2015.
MILIONIS, A.; DAVIES, T. Regression and stochastic models for air pollution-I. review, comments and suggestions. **Atmospheric Environment**, v. 28, n. 17, p. 2801–2810, 1994.
OLTMANS, S. et al. Long-term changes in tropospheric ozone. **Atmospheric Environment**, v. 40, n. 17, p. 3156–3173, 2006.
PIO, C. et al. Particulate and gaseous air pollutant levels at the portuguese west coast. **Atmospheric Environment**, v. 25, n. 3-4, p. 669–680, 1991.
PIRES, J. et al. Identification of redundant air quality measurements through the use of principal component analysis. **Atmospheric Environment**, v. 43, n. 25, p. 3837–3842, 2009.
R Development Core Team. **R: A language and environment for statistical computing**. Vienna, Austria, 2014. Disponível em: ⟨http://www.r-project.org/⟩.
RAJAB, J. M.; MATJAFRI, M.; LIM, H. Combining multiple regression and principal

component analysis for accurate predictions for column ozone in Peninsular Malaysia. **Atmospheric Environment**, v. 71, n. 0, p. 36–43, 2013.
SEINFELD, J. H.; PANDIS, S. N. **Atmospheric chemistry and physics: from air pollution to climate change**. New York: J. Wiley, 2006.
SMEYERS-VERBEKE, J. et al. The use of principal components analysis for the investigation of an organic air pollutants data set. **Atmospheric Environment (1967)**, v. 18, n. 11, p. 2471–2478, 1984.
SONG, Y. et al. Source apportionment of $PM_{2.5}$ in Beijing using principal component analysis/absolute principal component scores and UNMIX. **Science of The Total Environment**, v. 372, n. 1, p. 278–286, 2006.
SOUZA, J. B. et al. Componentes principais e modelagem linear generalizada na associação entre atendimento hospitalar e poluição do ar. **Revista de Saúde Pública**, v. 48, p. 451–458, 2014.
STATHEROPOULOS, M.; VASSILIADIS, N.; PAPPA, A. Principal component and canonical correlation analysis for examining air pollution and meteorological data. **Atmospheric Environment**, v. 32, n. 6, p. 1087–1095, 1998.
STEVEN, R.; MARTIN, M. A. Using supervised principal components analysis to assess multiple pollutant effects. **Environmental Health Perspectives**, v. 114, n. 12, p. 1877–1882, 2006.
TSAY, R. S. **Analysis of financial time series**. 2rd. ed. New Jersey: Wiley-Interscience, 2005.
VINGARZAN, R. A review of surface ozone background levels and trends. **Atmospheric Environment**, v. 38, n. 21, p. 3431–3442, 2004.
WATSON, J. G. et al. Receptor modeling application framework for particle source apportionment. **Chemosphere**, v. 49, n. 9, p. 1093–1136, 2002.
WORLD HEALTH ORGANIZATION. **WHO air quality guidelines global update 2005. Report on a working group meeting, Bonn/Germany**. Copenhagen, 2005. Disponível em: ⟨http://www.euro.who.int/__data/assets/pdf_file/0008/147851/E87950.pdf⟩. Acesso em: 20 de mar. de 2014.
WORLD HEALTH ORGANIZATION. **Air pollution estimates**. Copenhagen, 2014. Disponível em: ⟨http://www.who.int/phe/health_topics/outdoorair/databases/FINAL_HAP_AAP_BoD_24March2014.pdf?ua=1⟩. Acesso em: 26 de mar. de 2014.
YOO, H.-J. et al. Analysis of black carbon, particulate matter, and gaseous pollutants in an industrial area in Korea. **Atmospheric Environment**, v. 45, n. 40, p. 7698–7704, 2011.
YU, T.-Y.; CHANG, L.-F. W. Selection of the scenarios of ozone pollution at southern Taiwan area utilizing principal component analysis. **Atmospheric Environment**, v. 34, n. 26, p. 4499–4509, 2000.
ZAMPROGNO, B. **Uso e interpretação de análise de componentes principais, em séries temporais, com enfoque no gerenciamento da qualidade do ar**. 2013. 107 f. Tese (Doutorado em Engenharia Ambiental) — Programa de Pós-graduação em Engenharia Ambiental, Universidade Federal do Espírito Santo, Vitória, 2013.

8 APÊNDICE: ESTUDOS ADICIONAIS

Neste apêndice encontram-se os quatro artigos adicionais desta tese, que estão diretamente ligados ao tema central da pesquisa. Os artigos estão formatados de acordo com as normas das revistas as quais os artigos foram aceitos ou submetidos.

PREVISÃO DA CONCENTRAÇÃO DE OZÔNIO NA REGIÃO DA GRANDE VITÓRIA, ESPÍRITO SANTO, BRASIL, UTILIZANDO O MODELO ARMAX-GARCH[1]

OZONE CONCENTRATION FORECAST IN THE REGION OF GRANDE VITÓRIA, ESPÍRITO SANTO, BRAZIL, USING THE ARMAX-GARCH MODEL

Edson Zambon Monte[1]
Taciana Toledo de Almeida Albuquerque[2]
Valdério Anselmo Reisen[3]

[1] Universidade Federal do Espírito Santo (UFES), Departamento de Economia, Vitória, ES, Brasil e Doutorando do Programa de Pós-Graduação em Engenharia Ambiental, UFES, e-mail: edsonzambon@yahoo.com.br.
[2] Universidade Federal de Minas Gerais (UFMG), Departamento de Engenharia Sanitária e Ambiental, Belo Horizonte, MG, Brasil e Programa de Pós-Graduação em Engenharia Ambiental, UFES, e-mail: tacianatoledo26@gmail.com.
[3] Universidade Federal do Espírito Santo (UFES), Departamento de Estatística, Vitória, ES, Brasil e Programa de Pós-Graduação em Engenharia Ambiental, UFES, e-mail: valderioanselmoreisen@gmail.com.

Resumo:
O objetivo deste trabalho foi estimar e prever a concentração horária de ozônio na Região da Grande Vitória, Espírito Santo, Brasil, utilizando um modelo ARMAX-GARCH, para o período 01/01/2011 a 31/12/2011. Foram utilizados dados da rede de monitoramento do Instituto Estadual de Meio Ambiente e Recursos Hídricos (IEMA), sendo escolhidas três estações: Laranjeiras, Enseada do Suá e Cariacica. Adotou-se alguns parâmetros medidos nas estações como variáveis explicativas da concentração de ozônio, a saber: temperatura, umidade relativa, velocidade do vento e concentração de dióxido de nitrogênio. Estas foram significativas e melhoraram a estimativa do modelo ajustado. As previsões horárias para o dia 31/12/2011 revelaram-se muito próximas dos valores observados, sendo que as estimativas, em geral, seguiram a trajetória diária da concentração de ozônio. No mais, em comparação

[1] Artigo publicado na Revista Brasileira de Meteorologia.

aos modelos ARMA e ARMAX, o modelo ARMAX-GARCH revelou-se mais eficaz na predição de episódios de poluição de ozônio (concentração horária superior a 80 µg/m^3), reduziu o número de falsos alarmes estimados e apresentou menor taxa de ocorrência de episódios não detectados.

Palavras-chaves: ozônio; poluição do ar; séries temporais, ARMAX; GARCH.

***Abstract*:**

The objective of this study was to estimate and forecast the hourly ozone concentration in the Region of Grande Vitória, Espírito Santo, Brazil, using the ARMAX-GARCH model, for the period from 2011/01/01 to 2011/12/31. Data set from the State Environmental Institute (IEMA) was used. The models has run for three local stations: Laranjeiras, Enseadá do Suá and Cariacica. Some parameters measured at the stations were taken as explanatory variables of ozone concentration. These variables significantly improved the model estimated. The hourly forecasts for 2011/12/31 (chosen to verify the model accuracy) were very close to the observed values and the estimated ones generally followed the path of daily ozone concentration. When compared with the ARMA and ARMAX models, ARMAX-GARCH model proved to be more effective in the prediction of ozone pollution episodes (hourly concentration higher than 80 μg /m^3), reduction on the number of false alarms and lowering the rate of undetected episodes.

***Key-words*:** ozone; air pollution; time series, ARMAX; GARCH.

1. INTRODUÇÃO

A intensificação do processo de industrialização ocorrida no século XIX, aliado ao crescimento populacional, especialmente, o crescimento da população urbana em detrimento da rural, vem aumentando as preocupações relacionadas à proteção do meio ambiente. Logo, cada vez mais tem se dado atenção para os efeitos adversos que a poluição atmosférica pode causar na saúde humana, tais como: irritação dos olhos, problemas pulmonares, alergias, etc. Para Liu et al. (2013), os dois poluentes atmosféricos que mais preocupam em relação à saúde humana são o ozônio (O_3)[2] e o material particulado.

[2] Vale ressaltar que os males causados pelo ozônio ocorrem na faixa de ar perto da superfície terrestre. Conforme Seinfeld e Pandis (2006), o ozônio apresenta um duplo paradoxo na atmosfera, pois, o mesmo tem um papel benéfico na estratosfera e maléfico na troposfera.

Quanto ao ozônio, esse é um poluente secundário formado na troposfera, por meio de reações fotoquímicas sobre os óxidos de nitrogênio e os compostos orgânicos voláteis, sendo a radiação solar um dos forçantes que contribui para sua formação. Conforme Seinfeld e Pandis (2006), a formação do ozônio depende de diversos fatores químicos e físicos, que variam no espaço e no tempo de forma não linear. Destaca-se que os hidrocarbonetos são emitidos por fontes naturais e antropogênicas, sendo que estas últimas incluem fontes móveis (automóveis) e estacionárias (usos industriais). Assim, a concentração de O_3 tende a ser maior nos grandes centros urbanos, onde se concentram as grandes indústrias e o maior volume de automóveis. Além disso, como a formação de ozônio requer radiação ultravioleta, bem como a presença de precursores como óxidos de nitrogênio e compostos orgânicos voláteis, a concentração de O_3 alcança o máximo durante os meses de verão (Ryan et al., 1999).

Ressalta-se que, de acordo com Moreira et al. (2008), as condições meteorológicas desempenham um papel importantíssimo na dispersão ou acumulação de poluentes. Liu e Johnson (2002) descreveram que a poluição do ar, particularmente a concentração de ozônio, é altamente correlacionada no tempo, estando associada, geralmente, a fatores como temperatura, umidade relativa, velocidade e direção do vento, dentre outros.

Cabe mencionar que, para o ozônio, por ser um poluente secundário, tornam-se mais difíceis as modelagens e as previsões a respeito de sua formação (Carvalho, 2006). No entanto, a natureza da concentração de ozônio em ambientes, principalmente urbanos, tem sido objetivo de diversos estudos estatísticos, especialmente no que tange a predição e a previsão das concentrações[3]. Segundo Liu et al. (2013), alguns métodos utilizados são: análise de regressão e classificação em árvore (CART), modelos de redes neurais, modelos de séries temporais ARIMA (Box-Jenkins) e modelos de regressão.

Ryan (1995), por exemplo, realizou previsões para as altas concentrações diárias de ozônio (episódios), na região de Baltimore, Estados Unidos, utilizando diversas abordagens, dentre elas: o método CART e a análise de regressão. Jorquera et al. (1998) realizaram previsões para o nível máximo de concentração de ozônio diário, na cidade de Santigo, Chile, utilizando modelos de séries temporais (ARMAX), de redes neurais e o modelo fuzzy. Já Liu e Johnson (2002) fizeram previsões para picos diários de concentração de ozônio, em Milwaukee, Estados Unidos, no período de 1987 a 1993, por meio do modelo regressão com erros de séries temporais (RTSE), com a inclusão, dentre as variáveis exógenas, do que os

[3] Em estatística, predições são realizadas dentro da amostra considerada e, previsões, fora da amostra considerada.

autores denominaram de principal componente (PC) com gatilho. Liu e Johnson (2003) estudaram os picos diários de concentração de ozônio, para Milwaukee, Estados Unidos, no período de 1999 a 2002, utilizando o PC com gatilho na abordagem de Box-Jenkins com RTSE.

Nota-se que, independente do método econométrico de análise, alguns destes estudos, ao fazerem previsões ou predições da concentração de ozônio, têm ignorado a questão da heterocedasticidade (volatilidade)[4] temporal (Kumar e Ridder, 2010). Assim, esses autores estimaram um modelo de heterocedasticidade condicional autorregressivo generalizado (GARCH) associado com o método FFT-ARIMA (transformada rápida de Fourier–autorregressivo integrado de média móvel), para prever os episódios de concentração de ozônio em duas cidades europeias, Bruxelas e Londres. Os resultados revelaram que modelar a concentração de ozônio por meio do modelo GARCH, além de melhorar os intervalos de confiança das previsões de curto prazo, também proporcionou maior acurácia na probabilidade de previsão de episódios críticos de O_3. Reisen et al. (2014) modelaram a média diária de concentração de material particulado inalável (PM_{10}), na cidade de Cariacica, Espírito Santo, Brasil, utilizando um processo integrado fracionado sazonal, com volatilidade.

Neste contexto, este trabalho objetivou estimar e prever a concentração de ozônio horária na Região da Grande Vitória (RGV), Espírito Santo, Brasil, utilizando o modelo ARMAX-GARCH, para o período 01/01/2011 a 31/12/2011. Mesmo não tendo ultrapassado os padrões primário e secundário (160 $\mu g/m^3$) estabelecidos pela Resolução CONAMA 03, de 1990 (CONAMA, 1990), no período de estudo, em diversos momentos a concentração ultrapassou o valor de 80 $\mu g/m^3$ e, até mesmo, o de 100 $\mu g/m^3$, sendo que as maiores concentrações ocorreram no verão. Estes padrões estabelecem uma qualidade do ar regular (entre 80 e 160 $\mu g/m^3$), em termos de efeitos prejudiciais sobre a saúde, principalmente para população mais sensível, como idosos e crianças (Companhia Ambiental do Estado de São Paulo – CETESB, 2013). Lembrando que esta qualidade do ar regular é baseada no Índice de Qualidade do AR (IQA), da CETESB. Logo, esta pesquisa torna-se importante no que diz respeito, especialmente, à formulação de medidas preventivas por parte dos órgãos competentes, uma vez que a concentração de ozônio, na Grande Vitória, embora não tenha atingido níveis alarmantes, tem-se elevado nos últimos anos.

[4] Conforme Matteson e Tsay (2011), a volatilidade pode ser entendida como o desvio padrão condicional da série. Séries com alta variabilidade ao longo do tempo tendem a apresentar esta característica. A concentração horária de ozônio é um exemplo de série volátil.

O presente artigo está estruturado da seguinte forma. Além desta introdução, a seção 2 traz uma descrição da região de estudo, as variáveis utilizadas e os modelos estatísticos adotados. Na seção 3 apresentam-se as estimativas do modelo ARMAX-GARCH, os testes de diagnóstico e as previsões para a concentração de ozônio. Por fim, as conclusões são apresentadas na seção 4.

2. MATERIAL E MÉTODOS

2.1. Região de estudo e apresentação das variáveis

Os dados deste estudo foram do tipo séries temporais, abrangendo variáveis relacionadas à poluição atmosférica (concentração de ozônio e dióxido de nitrogênio) e às condições meteorológicas (temperatura, umidade relativa e velocidade do vento), para a Região da Grande Vitória, Espírito Santo, Brasil. A RGV é composta por cinco municípios, localizando-se na costa sul do oceano Atlântico (latitude 20°19S, longitude 40°20W). O clima é tropical quente, com temperaturas médias variando entre 24° C e 30° C.

O período de análise foi de janeiro a dezembro de 2011, sendo os dados tomados de forma horária (8.760 observações) e coletados através do banco de dados do IEMA. Uma vez que alguns dados horários não estavam disponíveis, algumas observações foram inseridas utilizando o "pacote mtsdi" (*multivariate time series data imputation*) (Junger e Leon, 2012), presente no *software* R 3.0.2. Na Tabela 1 são apresentadas as variáveis utilizadas, unidades, siglas e fontes.

Vale ressaltar que, atualmente, a Região da Grande Vitória possui oito estações de monitoramento de qualidade do ar, a saber: Laranjeiras; Carapina; Jardim Camburi; Enseada do Suá; Vitória – Centro; Vila Velha – Ibes; Vila Velha – Centro; e, Cariacica (ver Figura 1). A concentração de ozônio é medida nas estações de Laranjeiras, Enseada do Suá, Vila Velha – Ibes e Cariacica. Dado o grande percentual de dados faltantes e a alta porcentagem de dados invalidados (seja pelo sistema, equipamento, usuário, etc.), na estação Vila Velha – IBES, trabalhou-se, somente, com os dados de concentração de ozônio das estações de Laranjeiras, Enseada do Suá e Cariacica (foi estimado um modelo para cada estação). Consequentemente, para a concentração de dióxido de nitrogênio, também se adotou os dados destas três estações.

Tabela 1: Variáveis, unidades, siglas e fontes

Variáveis	**Unidades**	**Siglas**	**Fontes**
Concentração de ozônio em Laranjeiras – Frequência horária com amostra de uma hora a três metros.	$\mu g/m^3$	O3LAR	IEMA
Concentração de ozônio em Enseada do Suá – Frequência horária com amostra de uma hora a três metros.	$\mu g/m^3$	O3SUA	IEMA
Concentração de ozônio em Cariacica – Frequência horária com amostra de uma hora a três metros.	$\mu g/m^3$	O3CAR	IEMA
Concentração de dióxido de nitrogênio em Laranjeiras – Frequência horária com amostra de uma hora a três metros.	$\mu g/m^3$	NO2LAR	IEMA
Concentração de dióxido de nitrogênio em Enseada do Suá – Frequência horária com amostra de uma hora a três metros.	$\mu g/m^3$	NO2SUA	IEMA
Concentração de dióxido de nitrogênio em Cariacica – Frequência horária com amostra de uma hora a três metros.	$\mu g/m^3$	NO2CAR	IEMA
Temperatura média – Frequência horária com amostra de uma hora a três metros.	°C	T	IEMA
Umidade relativa – Frequência horária com amostra de uma hora a três metros.	%	UR	IEMA
Velocidade escalar média do vento – Frequência horária com amostra de 1 hora a 10 metros.	m/s	VV	IEMA

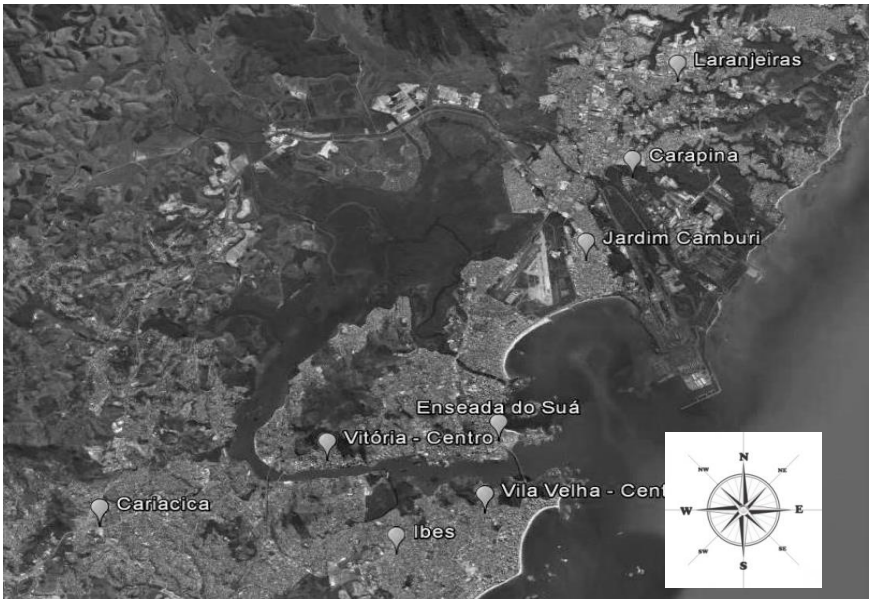


Figura 1: Estações de monitoramento da qualidade do ar na Grande Vitória.
Fonte: Google Earth (2014).

As variáveis temperatura, umidade relativa e velocidade do vento foram utilizadas conforme descrição da Tabela 2. Cabe mencionar aqui que foram testadas outras variáveis exógenas, como radiação solar e direção do vento. Porém, estas não se revelaram significativas ou comprometeram o ajuste do modelo.

Tabela 2: Descrição das variáveis temperatura, umidade relativa e velocidade do vento

Variáveis	Descrição
Temperatura	Média entre as estações de Carapina e Cariacica, únicas que possuem medições para tal variável.
Umidade Relativa	Existem medições para as estações de Carapina e Cariacica. Entretanto, como existem muitos dados faltantes para a estação de Cariacica, optou-se por trabalhar apenas com a umidade da estação de Carapina.
Velocidade do Vento	Mensurada nas estações de Carapina, Enseada do Suá, Vila Velha – Ibes e Cariacica. Dado que a estação de Carapina não atende a alguns padrões importantes para captação correta da velocidade do vento, adotou-se uma média entre as estações da Enseada do Suá, de Vila Velha – Ibes e de Cariacica.

2.2. Modelo ARMAX-GARCH[5]

Um modelo ARMA(p,q) é uma combinação de um processo autorregressivo (AR) e um processo de médias móveis (MA) e pode ser expresso por:

$$Y_t = c + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \cdots + \varphi_p Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}, \quad (1)$$

$$\left(1 - \varphi_1 L - \varphi_2 L^2 - \cdots - \varphi_p L^p\right) Y_t = \left(1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q\right) \varepsilon_t, \quad (2)$$

em que Y_t é processo estocástico a ser modelado; $\varphi_1, \varphi_2, \cdots, \varphi_p$, coeficientes do processo autorregressivo; $\theta_1, \theta_2, \cdots, \theta_q$, coeficientes do processo de médias móveis; L, operador de defasagem; e, $\varepsilon_t \sim RB\left(0, \sigma^2\right)$. Caso as raízes de $1 - \varphi_1 L - \varphi_2 L^2 - \cdots - \varphi_p L^p = 0$ estejam fora do círculo unitário, o processo estocástico é dito estacionário.

Cabe ressaltar que as ordens de p e q podem ser determinadas, respectivamente, pela função de autocorrelação parcial (FACP) e pela função de autocorrelação (FAC). Adicionalmente, critérios mais objetivos podem ser utilizados para identificar as ordens corretas de p e q, a saber: Critério de Informação de Akaike (AIC); Critério de Informação de Schwarz (SC); Critério de Informação de Hannan-Quinn (HQ); e, Erro de Predição Final (FPE) (Brockwell e Davis, 2002).

Já o modelo ARMAX é uma extensão do modelo ARMA, utilizando outras séries temporais como variáveis de entrada. Tal modelo pode ser descrito como:

$$Y_t = c + \sum_{i=1}^{p} \varphi_i Y_{t-i} + \varepsilon_t + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \sum_{k=1}^{n} \lambda_k X(t,k), \quad (3)$$

[5] Este item está baseado em Hamilton (1994).

em que λ é o vetor de coeficientes da matriz de componentes regressivos (X); e, $\varepsilon_t \sim RB(0,\sigma^2)$. Na prática, os coeficientes $(\varphi_i, \theta_j, \lambda_k)$ podem ser estimados pelo método de máxima verossimilhança (MV).

Neste contexto, a estimação do modelo ARMAX (Equação 3) requer que o termo de erro ε_t seja homocedástico (ausência de volatilidade estocástica). No entanto, nas situações onde a distribuição condicional difere da distribuição incondicional, a suposição de variância do erro constante pode não ser verificada. Dessa forma, os modelos autorregressivos de heterocedasticidade condicional (ARCH) surgiram no início da década de 1980 (ver Engle, 1982), com o intuito de modelar, temporalmente, a variância condicional. Estes modelos foram generalizados por Bollerslev (1986), dando origem aos modelos autorregressivos de heterocedasticidade condicional generalizados (GARCH).

Para representar o modelo GARCH, toma-se ε_t com um processo estocástico real em tempo discreto. Neste estudo, ε_t são as inovações do processo ARMAX (Equação 3). Engle (1982) definiu um processo ARCH onde todos os ε_t são da forma,

$$\varepsilon_t = z_t \sigma_t, \tag{4}$$

em que z_t é um processo distribuído independente e identicamente com média zero e variância unitária. Embora ε_t seja serialmente não correlacionado por definição, sua variância condicional σ_t^2 pode ser autocorrelacionada e, portanto, mudar ao longo do tempo.

A equação da variância do GARCH(r,l) pode ser representada por (Bollerslev, 1986; Brockwell e Davis, 2002):

$$z_t = D_\theta(0,1), \tag{5}$$

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{r} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{l} \beta_j \sigma_{t-j}^2, \tag{6}$$

em que $D_\theta(0,1)$ é a função de densidade de probabilidade das inovações ou resíduos com média zero, variância unitária; e, $r \geq 0, l \geq 0;\ \alpha_0 > 0, \alpha_i \geq 0;\ i = 1,2,\cdots,r;\ \beta_j \geq 0,\quad j = 1,2,\cdots,l.$

Nota-se que, caso $l = 0$, o processo reduz-se para um ARCH(r). No mais, para $r = l = 0$, a variância condicional é constante, como em um modelo ARMA, e as inovações ε_t são reduzidas a ruídos brancos.

Bollerslev (1986) demonstrou que o processo GARCH(1,1) é estacionário com $E(\varepsilon_t) = 0$, $\text{var}(\varepsilon_t) = \alpha_0/(1-\alpha_1-\beta_1)$ e $\text{cov}(\varepsilon_t, \varepsilon_s) = 0$ para $t \neq s$ se, e somente se, $\alpha_1 + \beta_1 < 1$

(para mais detalhes, consultar Bollerslev (1986)). A estimação do modelo GARCH pode ser feita pelo método de máxima verossimilhança (MV).

3. RESULTADOS E DISCUSSÕES

3.1. Estimativas por ARMAX-GARCH

O primeiro passo na análise de séries temporais é verificar se as mesmas são estacionárias[6]. Se elas não forem estacionárias em nível deve-se realizar algum procedimento para estacionarizá-las (em geral, aplica-se a primeira diferença nas mesmas, dado que a maioria das séries é I(1), ou seja, integradas de primeira ordem). Os resultados dos testes *Augmented Dickey-Fuller – ADF* (Dickey e Fuller, 1981), *Phillips-Perron – PP* (Phillips e Perron, 1988) e *Kwiatkowski-Phillips-Schmidt-Shin – KPSS* (Kwiatkowski et al., 1992)[7] revelaram que todas as séries foram estacionárias em nível.

Para determinação do modelo ARMAX ideal, para cada estação de monitoramento, utilizou-se a FAC, a FACP e os critérios de informação de AIC, de SC e HQ. Foram estimados diversos modelos, sendo que o modelo com o melhor ajuste de cada estação encontra-se na Tabela 3. Importante mencionar que também foram estimados modelos ARMA, porém, pelos critérios de seleção, os mesmos foram desconsiderados.

No mais, todos os modelos estimados, constantes na Tabela 3, demonstraram resíduos (FAC) com características semelhantes à de um processo ruído branco, ou seja, não autocorrelacionados. Entretanto, quando se observou a FAC dos resíduos ao quadrado, verificou-se, para as três estações, um grande número de valores fora do intervalo de confiança. Logo, os resíduos ao quadrado não obedeceram à suposição de ruído branco, exibindo correlação na variância, ou seja, há heterocedasticidade condicional no processo. Para confirmar tal problema, realizou-se o teste de heterocedasticidade ARCH-LM[8] no modelo de cada estação (Tabela 3), e verificou-se que a variância condicional dos erros é autocorrelacionada. Vale lembrar que o teste foi realizado para diversos números de

[6] Uma série temporal (processo estocástico) é considerada estacionária quando apresentar média, variância e covariância constantes ao longo do tempo.
[7] Também foram analisados os gráficos e as funções de autocorrelação das séries.
[8] O teste ARCH-LM é utilizado para verificar se os resíduos (erros estimados) apresentam ou não volatilidade. A hipótese nula do teste é de ausência de volatilidade. Caso a hipótese nula não seja rejeitada, ocorre ausência de volatilidade e o modelo está adequado.

defasagens (a Tabela 3 apresenta o resultado para uma defasagem, ou seja, para defasagem de uma hora), sendo que, para todos, rejeitou-se a hipótese de ausência de volatilidade.

Tabela 3: Estimativas das equações das médias condicionais

	Laranjeiras		**Enseada do Suá**		**Cariacica**	
	Coef.	**Ep.**	**Coef.**	**Ep.**	**Coef.**	**Ep.**
Constante	31,45461*	3,678915	54,17850*	5,273380	14,14836*	4,277938
T	0,518867*	0,086809	0,240567**	0,122042	0,714503*	0,087138
UR	-0,041923*	0,017540	-0,166264*	0,030050	-0,047402*	0,018303
VV	1,104344*	0,152707	1,225437*	0,222257	1,396750*	0,145102
NO2	-0,675974*	0,007614	-1,000944*	0,014113	-0,459971*	0,008701
AR(1)	0,899618*	0,006371	0,824099*	0,012169	0,838127*	0,010965
AR(4)	-	-	0,029425*	0,009614	-0,031447*	0,008524
AR(12)	-	-	-	-	0,034332*	0,007915
AR(14)	-	-	-	-	-0,027510*	0,007968
AR(21)	0,047190*	0,010212	-	-	0,054735*	0,010831
AR(22)	0,029695**	0,014589	-	-	0,081962*	0,012935
AR(23)	0,060843*	0,013730	0,113795*	0,012975	-	-
AR(24)	-	-	0,092272*	0,016514	0,124145*	0,013372
AR(25)	-	-	-0,045601*	0,015871	-0,094428*	0,012322
AR(26)	-0,061539*	0,007720	-0,062426*	0,012344	-	-
MA(1)	0,173674*	0,016713	0,179530*	0,017111	0,161128*	0,015967
Teste de heterocedasticidade condicional						
ARCH-LM	529,010		231,581		578,191	
P-valor	0,00000		0,00000		0,00000	

Nota: * Significativo a 1%; ** Significativo a 5%; Coef.: coeficiente; e, Ep.: erro-padrão.

Dessa forma, adotou-se a técnica GARCH para modelar a volatilidade (variância condicional) da concentração de ozônio. Os modelos foram estimados considerando a suposição de que os erros do ARMAX seguem distribuição normal[9]. Foram testados vários modelos para o número de defasagens do GARCH, sendo que o melhor modelo para cada estação encontra-se na Tabela 4. Pelo teste ARCH-LM observou-se que não se rejeitou a hipótese nula de ausência de volatilidade, em todas as estações, eliminando-se o problema da heterocedasticidade condicional. Novamente, o teste foi realizado para diversos números de defasagens (a Tabela 4 apresenta o resultado para uma defasagem), sendo que, para todos, não se rejeitou a hipótese de ausência de volatilidade.

Para corroborar a utilização do modelo ARMAX-GARCH, realizou-se estimativas para os modelos ARMA, ARMAX e ARMAX-GARCH e verificou-se qual o melhor método para realizar previsões. Para isto, adotou-se o período de 0:30 horas de 01/01/2011 até as 23:30 horas de 30/12/2011 para fazer as estimações e reservou-se as 24 horas do dia

[9] No que tange ao teste de normalidade dos resíduos, o teste de Jarque-Bera rejeitou a hipótese nula de que os resíduos são normais, para todas as estações de monitoramento. Entretanto, assumiu-se, pela teoria assintótica sobre a média das distribuições de probabilidade, a suposição de que os resíduos são normais, dando seguimento à análise do modelo. Esta suposição é importante para realização dos testes de hipóteses do modelo estimado.

31.12.2011 para se fazer as previsões. A Tabela 5 demonstra os valores estimados para o Erro Absoluto Médio (MAE), para a Raiz Quadrada do Erro Quadrático Médio (RMSE) e para o Erro Percentual Absoluto Médio (MAPE), parar os modelos das estações de Laranjeiras, Enseada do Suá e Cariacica. Conforme observado, em todos os casos, o método ARMAX-GARCH apresentou os menores valores das estatísticas MAE, RSME e MAPE, sendo o mais adequado para realizar as previsões.

Tabela 4: Estimativas dos modelos de volatilidade condicional

	Laranjeiras		**Enseada do Suá**		**Cariacica**	
	Coef.	**Ep.**	**Coef.**	**Ep.**	**Coef.**	**Ep.**
Constante	2,713916*	0,304517	2,217166*	0,366801	1,111629*	0,161478
ARCH(1)	0,289816*	0,025081	0,297134*	0,024871	0,223160*	0,017258
GARCH(1)	0,522876*	0,032864	0,315804*	0,065489	0,736404*	0,016601
GARCH(2)	-	-	0,336062*	0,053538	-	-
Teste de heterocedasticidade condicional						
ARCH-LM	0,04678		0,81524		0,19865	
P-valor	0,82880		0,36360		0,65580	

Nota: * Significativo a 1%; Coef.: coeficiente; e, Ep.: erro-padrão.

Tabela 5: Critérios de seleção de modelo para previsão

Estação	**Modelo**	**MAE ($\mu g/m^3$)**	**RSME ($\mu g/m^3$)**	**MAPE (%)**
Laranjeiras	ARMA	3,556287	5,053671	9,675806
	ARMAX	3,035573	3,903782	8,249631
	ARMAX-GARCH	2,999764	3,899851	8,119810
Enseada do Suá	ARMA	4,847291	5,763318	13,39339
	ARMAX	3,408568	3,894462	9,290900
	ARMAX-GARCH	3,269677	3,794599	8,861550
Cariacica	ARMA	4,335348	5,779029	12,09641
	ARMAX	3,188361	3,853879	9,525701
	ARMAX-GARCH	3,055559	3,763219	8,921103

Ainda, para fins de comparação entre o pior modelo (ARMA) e o melhor modelo (ARMAX-GARCH) para realização de previsões, nas Figuras 2 e 3 encontram-se os valores observados e previstos para o dia 31/12/2011, para as estações de Laranjeiras, Enseada do Suá e Cariacica. Verifica-se a melhor adequação do modelo ARMAX-GARCH. O comportamento dos valores previstos é semelhante ao dos valores observados, inclusive no que diz respeito aos períodos de máxima concentração, que ocorrem durante as maiores temperaturas. Vale ressaltar que existem alguns pontos de subestimação e/ou superestimação, mas, as predições seguiram a trajetória observada da concentração de ozônio durante o dia 31/12/2011.

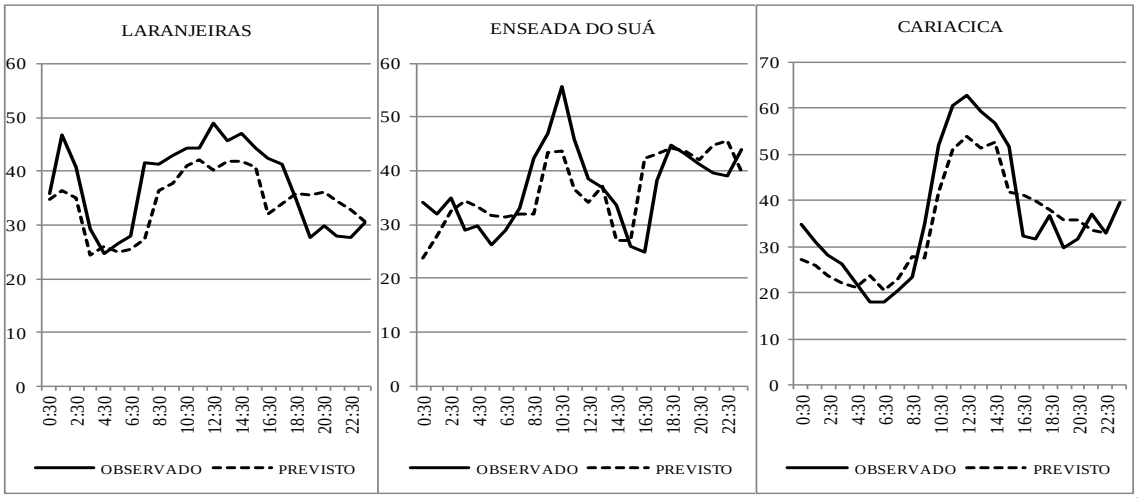


Figura 2: Valores observados e previstos para a concentração horária de ozônio (em $\mu g/m^3$), para o dia 31/12/2011, utilizando o modelo ARMA

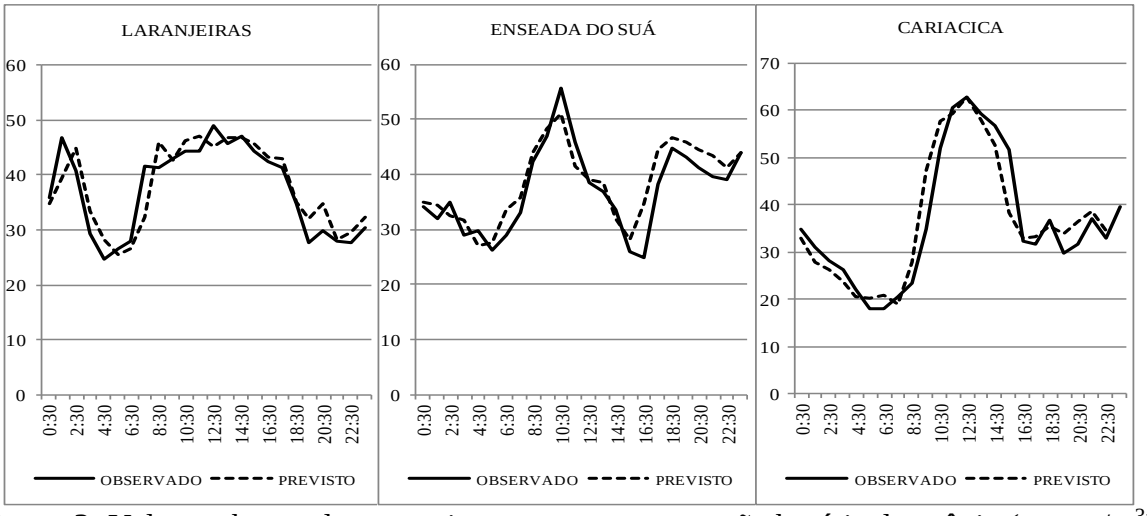


Figura 3: Valores observados e previstos para a concentração horária de ozônio (em $\mu g/m^3$), para o dia 31/12/2011, utilizando o modelo ARMAX-GARCH

Por fim, após estabelecer o melhor método para previsão, cabe dizer que os sinais estimados para as variáveis temperatura, umidade, velocidade do vento e dióxido de nitrogênio, estão de acordo com o esperado. Em relação à temperatura, o sinal positivo deve-se a correlação direta entre a radiação solar e a temperatura, uma vez que uma elevação da radiação solar implica em aumento da temperatura e, consequentemente, das reações fotoquímicas que contribuem para formação do ozônio.

No que se refere à correlação negativa entre a concentração de ozônio e a umidade relativa, conforme Seinfeld e Pandis (2006), os diversos mecanismos cinéticos podem ocorrer segundo a presença ou não de compostos orgânicos voláteis (COV) na atmosfera. Um destes mecanismos seria redução da formação de ozônio, quando da ocorrência de alta umidade

relativa. Conforme Carvalho et al. (2004), a relação negativa entre a concentração de ozônio e a umidade pode ser decorrência da radiação solar, uma vez que menores índice de umidade relativa podem estar relacionados a horários do dia com maior radiação solar e, assim, com maior formação de ozônio. Carvalho et al. (2004) salienta que "outro mecanismo cinético associado à presença dos COVs pode possibilitar a amplificação da produção do ozônio, devido à reação do radical peroxila com o óxido de nitrogênio para formar dióxido de nitrogênio, que numa segunda etapa, pode formar o poluente ozônio".

O sinal estimado da variável velocidade do vento foi positivo, demonstrando que aumentos na velocidade do vento elevam a concentração de ozônio. Esta relação pode ser influenciada por outros fatores meteorológicos, visto que o aumento da velocidade do vento na região pode surgir devido à atuação de sistemas sinóticos ou devido à circulação local que poderá ser gerada.

Independente da origem do aumento da velocidade do vento, existe um efeito direto entre o aumento da velocidade do vento e o aumento das concentrações de ozônio, devido à intensificação do transporte de gases precursores de outras regiões para as estações de Laranjeiras, Enseada do Suá e Cariacica, assim como o favorecimento do transporte do próprio ozônio já formado em outras regiões para as estações analisadas. Porém, em dias com atuação de sistemas sinóticos (por exemplo, frentes frias) existirá um aumento da velocidade do vento, mas a formação de ozônio ficará comprometida pela diminuição da incidência de radiação solar, consequentemente de temperatura do ar, aumento da umidade relativa do ar e nebulosidade e, a probabilidade de ocorrência de precipitação que contribuirá para a diluição dos poluentes precursores na atmosfera.

Em contrapartida, o aumento da velocidade do vento pode estar relacionado com a formação e intensificação de sistemas de circulação local, como a brisa marítima. Nestas situações, tem-se um aumento não somente da velocidade do vento, mas também da incidência de radiação solar, aumento da temperatura do ar, como consequência deste aumento, existirá uma intensificação do gradiente de pressão que é formado entre o mar e o continente, devido ao aquecimento diferenciado (mar-terra), surgindo com isso a circulação de brisa marítima. A existência da circulação de brisa marítima na RGV favorece o aumento da velocidade do vento local, logo o transporte de poluentes precursores é intensificado, assim como as condições meteorológicas são favoráveis para a ocorrência de reações fotoquímicas para a formação do ozônio.

Sob qualquer condição meteorológica que influencie a RGV, as horas do dia em que ocorrem as maiores concentrações de ozônio são próximas dos horários de máxima incidência de radiação solar e temperatura.

Já o dióxido de nitrogênio apresentou relação negativa com a concentração de ozônio. Por ser um poluente secundário, o ozônio é formado por reações fotoquímicas sobre os óxidos de nitrogênio e os compostos orgânicos voláteis. Assim, para formação do ozônio deve ocorrer o consumo de dióxido de nitrogênio, o que determina a relação negativa entre estes poluentes.

3.2. Desempenho estatístico do modelo ARMAX-GARCH

Além dos testes já realizados, baseando-se na metodologia de Ryan (1995) e Liu e Johnson (2003), calculou-se algumas estatísticas (Tabela 6), para comparar a capacidade de predição de episódios de concentração de ozônio do modelo ARMAX-GARCH, com os modelos ARMA e ARMAX. As estatísticas[10] mensuradas foram: a taxa de alarme falso (FAR), que mede a tendência de a predição superestimar os episódios de concentração de ozônio; a probabilidade de detecção (POD), que mensura a probabilidade de o modelo estimar corretamente os episódios horários da concentração de ozônio, ou seja, de predizer os episódios quando eles realmente ocorreram; e, a taxa de perda (MISS), que se refere à taxa na qual episódios de ozônio ocorreram, porém, não foram previstos.

Tabela 6: Desempenho estatístico dos modelos ARMA, ARMAX e ARMAX-GARCH, quando a predição da concentração de ozônio excede 80 $\mu g/m^3$ por hora

Estação	Modelo	FAR	POD	MISS
Laranjeiras	ARMA	0,16	0,38	0,62
	ARMAX	0,10	0,64	0,36
	ARMAX-GARCH	0,07	0,68	0,32
Enseada do Suá	ARMA	0,28	0,39	0,61
	ARMAX	0,21	0,49	0,51
	ARMAX-GARCH	0,18	0,53	0,47
Cariacica	ARMA	0,28	0,54	0,46
	ARMAX	0,25	0,63	0,38
	ARMAX-GARCH	0,21	0,65	0,35

Nota: FAR = Taxa de falso alarme; POD = Probabilidade de detecção; e, MISS = Taxa de perda.

Conforme descrito na introdução, pela avaliação dos dados das estações de Laranjeiras, Enseada do Suá e Cariacica, em diversas horas do ano de 2011, a concentração de

[10] A descrição das estatísticas encontra-se no apêndice.

ozônio esteve acima de 80 $\mu g/m^3$ e, até mesmo, superior a 100 $\mu g/m^3$, o que pode ser considerada como uma qualidade do ar regular (entre 80 e 160 $\mu g/m^3$), em termos de efeitos prejudiciais sobre a saúde, segundo a CETESB (2013). Logo, as concentrações acima de 80 $\mu g/m^3$ foram escolhidas como episódios neste estudo. Vale ressaltar, que níveis iguais ou inferiores a 80 $\mu g/m^3$ não estão isentos de causar riscos à saúde.

Para fins de exemplificação das medidas de desempenho, verifica-se que, para a estação de Laranjeiras, o modelo ARMAX-GARCH apresentou uma FAR de 0,07, o que significa que ele prevê 7% de falsos alarmes, enquanto que para o modelo ARMAX a taxa foi de 10% e para o ARMA de 16%. No caso da POD, nota-se que a probabilidade de detectar episódios corretamente é maior para o modelo ARMAX-GARCH do que para os outros modelos. O modelo ARMAX-GARCH obteve probabilidade de 68% no que diz respeito a estimar corretamente concentrações acima de 80 $\mu g/m^3$ por hora. Além disso, a taxa (MISS) referente à ocorrência de episódios que não foram detectados foi menor para o ARMAX-GARCH.

Para as demais estações, o comportamento da FAR, da POD e da MISS foi semelhante ao encontrado para a estação de Laranjeiras, corroborando, novamente, que o desempenho estatístico do modelo ARMAX-GARCH foi superior aos modelos ARMA e ARMAX, no que diz respeito à predição de episódios de poluição de ozônio. Por fim, vale dizer que, em geral, os melhores desempenhos das estatísticas FAR, POD e MISS foram para as estações de Laranjeiras, Cariacica e Enseada do Suá, respectivamente. Isto pode ser decorrência de a estação da Enseada do Suá ter apresentado um número muito maior de picos de concentração observados, quando comparado com as outras duas estações.

4. CONCLUSÕES

Este trabalho teve como objetivo estimar e prever a concentração horária de ozônio na Região da Grande Vitória, Espírito Santo, utilizando um modelo ARMAX-GARCH. Devido às limitações dos dados, as estimativas foram realizadas para as estações de Laranjeiras, Enseada do Suá e Cariacica. O modelo ARMAX apresentou volatilidade da variância dos resíduos, como já era esperado. Assim, estimou-se um ARMAX-GARCH para captar o efeito desta volatilidade. Ressalta-se que todas as séries utilizadas foram estacionárias em nível.

Neste contexto, destaca-se que, mesmo sendo um poluente de difícil modelagem estatística, em função da sua formação secundária, o modelo ARMAX-GARCH apresentou boas estimativas para as três estações de monitoramento. Os principais resultados foram:

i) As variáveis exógenas, temperatura, umidade, velocidade do vento e dióxido de nitrogênio, foram significativas e melhoram o ajuste do modelo final estimado;

ii) As estatísticas MAE, RMSE e MAPE revelaram que o modelo ARMAX-GARCH é mais adequado para realização de previsões do que os modelos ARMA e ARMAX;

iii) As previsões horárias do modelo ARMAX-GARCH, para o dia 31/12/2011, revelaram-se muito próximas dos valores observados, sendo que as estimativas, em geral, seguiram a trajetória diária da concentração de ozônio;

iv) Em comparação com os modelos ARMA e ARMAX, o modelo ARMAX-GARCH revelou-se mais eficaz na predição de episódios de poluição de ozônio (concentração horária superior a 80 $\mu g/m^3$), reduziu o número de falsos alarmes e apresentou menor taxa de ocorrência de episódios não detectados.

Ressalta-se que, estudos desta natureza são de grande importância, uma vez as preocupações quanto aos efeitos adversos que a poluição atmosférica pode causar na saúde humana tem aumento a cada dia. Sendo assim, estabelecer, por exemplo, como as condições meteorológicas afetam a concentração de ozônio e tentar prever os picos (episódios) de concentração deste poluente é fundamental, dado que pode auxiliar na tomada de decisões dos agentes públicos no que diz respeito ao combate à poluição, à prevenção de altas concentrações e a formulação de legislações para este fim.

Vale mencionar que este é um estudo preliminar. Para trabalhos futuros outras técnicas estatísticas podem ser utilizadas e comparadas com as adotadas nesta pesquisa, como os modelos de redes neurais, os modelos não lineares, e até mesmo os modelos vetoriais autorregressivos (VAR). Também podem ser utilizados outros métodos para modelar a variância condicional e tentar realizar estimativas separadas para as distintas estações do ano.

5. AGRADECIMENTOS

Os autores agradecem aos pareceristas anônimos por seus comentários construtivos. Ao IEMA pelo fornecimento das concentrações dos poluentes atmosféricos e das variáveis meteorológicas. Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)

e a Fundação de Amparo à Pesquisa do Espírito Santo (FAPES) pelo suporte financeiro parcial.

6. APÊNDICE A

a) Probabilidade de detecção (POD): mensura a ocorrência de horários de alta concentração de ozônio, iguais ou superiores a 80 μg/m^3, que foram preditas corretamente de acordo com os valores observados:

$$POD = \frac{A}{A+B}; \quad \text{(A1)}$$

b) Taxa de alarme falso (FAR): mede a tendência da predição de ozônio superestimar o valor observado da concentração de ozônio:

$$FAR = \frac{C}{C+A}; \quad \text{(A2)}$$

c) Taxa de perda (MISS): mensura a taxa a qual os episódios de ozônio observados não são preditos:

$$MISS = 1 - POD = \frac{B}{A+B}; \quad \text{(A3)}$$

em que

	Estimado – Sim	Estimado – Não
Observado – Sim	A	B
Observado – Não	C	D

Nota: as descrições, observado e estimado, referem-se aos valores característicos dos episódios.

7. REFERÊNCIAS BIBLIOGRÁFICAS

BOLLERSLEV, T. Generalized autoregressive conditional heteroskedasticity. **Journal of Econometrics**, v. 31, p. 307-327, 1986.

BROCKWEL, P, J; DAVIS, R. A. **Introduction to time series and forecasting**. 2ª ed. New York: Springer, 2002, 437 p.

CARVALHO, V. S. B.; CAVALCANTI, P. M. P. S.; CATALDI, M.; PIMENTEL, L. C. G. Avaliação da Concentração do Ozônio e de seus precursores na RMRJ e correlação deste com variáveis meteorológicas durante o ano de 2002. In: Congresso Brasileiro de Meteorologia, XIII, 2004, Fortaleza. **Anais eletrônicos...** Disponível em: <http://www.cbmet.com/cbm-files/22-106a74513a8169304ab1ec402bddd658.doc>. Acesso em: 05 nov. 2013.

CARVALHO, V. S. B. **Meteorologia da qualidade do ar no que tange as concentrações de ozônio e dos óxidos de nitrogênio na região Metropolitana do Rio de Janeiro**. Rio de janeiro, RJ. 2006. 134 f. Dissertação (Mestrado em Engenharia Mecânica). Programa de Pós-Graduação em Engenharia Mecânica, Universidade Federal do Rio de Janeiro, RJ.

CETESB. **Relatório da qualidade do ar do estado de São Paulo 2012**. São Paulo: CETESB, 2013.

CONSELHO NACIONAL DE MEIO AMBIENTE – CONAMA (Brasil). Resolução nº 08, de 6 de dezembro de 1990. **Diário Oficial [da] República Federativa do Brasil**, Brasília, 28 dez. 1990. Seção 1, p. 25539.

DICKEY, D. A.; FULLER, W. A. Likelihood ratio statistics for autoregressive time series with a unit root. **Econometrica**, v. 49, n. 4, p. 1057-1073, 1981.

ENGLE, R. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingtom inflation. **Econometrica**, v. 20, n.3, p. 339-350. 1982.

GOOGLE EARTH. **Informações geográficas**. 2014. Disponível em: <http://www.google.com.br/intl/pt-PT/earth/. Acessado em: 20 de mar. de 2014.

HAMILTON, J. D. Time series analysis, New Jersey: Princeton University Press, 1994, 820 p.

JORQUERA, H.; PÉREZ, R.; CIPRIANO, A.; ESPEJO, A.; LETELIER, M. V.; ACUÑA, G. Forecasting ozone daily maximum levels at Santiago, Chile. **Atmospheric Environment**. v. 32, n. 20, p. 3425-3424, 1998.

JUNGER, W.; LEON, A. P. **mtsdi: Multivariate time series data imputation**. R package version 0.3.3. 2012. Disponível em: http://CRAN.R-project.org/package=mtsdi. Acessado em 25.03.2014.

KUMAR, U.; RIDDER K. GARCH modeling in association with FFT-ARIMA to forecast ozone episodes. **Atmospheric Environment**, v. 44, p. 4252-4265, 2010.

KWIATKOWSKI, D.; PHILLIPS, P. C. B.; SCHMIDT, P.; SHIN, Y. Testing the null hypothesis of stationarity against the alternative of unit root. **Journal of Econometrics**, v. 54, n. 1, p. 159-178, 1992.

LIU, P. W. G.; JOHNSON, R. Forecasting peak daily ozone levels-I. A regression with time series errors model having a principal component trigger to fit 1991 ozone levels. **Journal of the Air & Waste Management Association**. v. 52, n. 9, p1064-1074, 2002.

LIU, P. W. G.; JOHNSON, R. Forecasting peak daily ozone levels: part 2. A regression with time series errors model having a principal component trigger to forecast 1999 and 2002 ozone levels. **Journal of the Air & Waste Management Association**. v. 53, n. 12, p. 1472-1489, 2003.

LIU, P. W. G.; TSAI, J. H.; LAI, H. C.; TSAI, D. M.; LI, L. W. Establishing multiple regression models for ozone sensitivity analysis to temperature variation in Taiwan. **Atmospheric Environment**, v. 79, p. 225-235, 2013.

MATTESON, D. S.; TSAY, R. S. Dynamic orthogonal components for multivariate time series. Journal of the American Statistical Association, v. 106, n. 496, p. 1450-1463, 2011.

MOREIRA, D. M.; TIRABASSI, T.; MORAES, M. R. Meteorologia e poluição atmosférica. **Ambiente & Sociedade**. v. 11, n. 1, p. 1-13, 2008.

PHILLIPS, P. C. B.; PERRON, P. Testing for unit roots in time series regression. **Biometrika**, v. 75, n. 3, p. 335-346, 1988.

REISEN, V. A.; SARNAGLIA, A, J. Q.; REIS JUNIOR, N. C; SANTOS, J. M. **Modeling and forecasting daily average PM_{10} concentrations by a seasonal long-memory model with volatility**. Environmental Modelling & Software, v. 51, p. 286-295, 2014.

RYAN, W. F. Forecasting severe ozone episodes in the Baltimore metropolitan area. **Atmospheric Environment**, v. 29, n. 17, p. 2387-2398, 1995.

RYAN, W. F.; PIETY, C. A; LUEBEHUSEN, E. D. Air quality forecasts in the Mid-Atlantic Region: current practice and benchmark skill. **Weather and Forecasting**, v. 15, n. 1, p. 46-60, 1999.

SEINFELD, J. H.; PANDIS, S. N. **Atmospheric chemistry and physics: from air pollution to climate change**. J. Wiley, New York, 2006.

Impactos das variáveis meteorológicas na qualidade do ar da Região da Grande Vitória, Espírito Santo, Brasil[1]

Impacts of meteorological variables on air quality in the Region of Grande Vitória, Espírito Santo, Brazil

Edson Zambon Monte[a], Taciana Toledo de Almeida Albuquerque[b], Valdério Anselmo Reisen[c]

[a] Universidade Federal do Espírito Santo (UFES), Departamento de Economia, Vitória, ES, Brasil e Doutorando do Programa de Pós-Graduação em Engenharia Ambiental, UFES, e-mail: edsonzambon@yahoo.com.br.

[b] Universidade Federal de Minas Gerais (UFMG), Departamento de Engenharia Sanitária e Ambiental, Belo Horizonte, MG, Brasil e Programa de Pós-Graduação em Engenharia Ambiental, UFES, e-mail: tacianatoledo26@gmail.com.

[c] Universidade Federal do Espírito Santo (UFES), Departamento de Estatística, Vitória, ES, Brasil e Programa de Pós-Graduação em Engenharia Ambiental, UFES, e-mail: valderioanselmoreisen@gmail.com.

Resumo: este trabalho objetivou verificar os impactos das variáveis meteorológicas temperatura, umidade relativa, velocidade do vento e precipitação sobre a qualidade do ar, na Região da Grande Vitória, Espírito Santo, Brasil, considerando o poluente material particulado inalável (MP_{10}), por meio do modelo Logit. O período de estudo foi de janeiro de 2005 a dezembro de 2010, onde a qualidade do ar foi classificada como "não boa" e "boa". Também foram estimados os efeitos dos dias da semana e das estações do ano sobre a probabilidade de ocorrência de qualidade do ar "não boa". Os resultados demonstraram que os fatores meteorológicos precipitação pluviométrica e velocidade do vento contribuíram significantemente para a redução da probabilidade de ocorrência de qualidade do ar "não boa". Além disso, os resultados simulados mostraram que, durante os finais de semana, as chances de ocorrer qualidade do ar "não boa" foram fortemente reduzidas e, nas estações do

[1] Artigo aceito para publicação na Revista Brasileira de Meteorologia.

outono e do inverno, a probabilidade de se verificar qualidade do ar “não boa” caiu de maneira relevante.

Palavras-chave: variáveis meteorológicas; poluição do ar; MP_{10}; modelo Logit.

Abstract: the objective of this study was to determine the impacts of meteorological variables, such as temperature, relative humidity, wind speed and precipitation, on the air quality in the Region of Grande Vitória, Espírito Santo, Brazil, considering the pollutant PM_{10} and using the Logit model. The period of study was from January 2005 to December 2010 and, in this study, the air quality was classified as “not good” and “good”. The day of the week and season effects over the air quality “not good” were also estimated. The results showed that the precipitation and the wind speed contributed significantly to the reduction of the probability of air quality “not good”. In addition, during the weekends the probability of air quality “not good” was greatly reduced and in the autumn and winter seasons the probability of air quality “not good” falls significantly.

Key-words: meteorological variables; air pollution; PM_{10}; Logit model.

1. Introdução

As questões relativas à qualidade do ar têm se tornado cada vez mais importantes, uma vez que vários problemas de saúde decorrem da poluição atmosférica, dentre eles: asmas, renites, ardor nos olhos, cansaço, tosse seca, doenças cardiovasculares e pulmonares, insuficiência cardíaca, etc. Autores como Brunekreef e Holgate (2002), Maynard (2004), World Health Organization (WHO, 2005), Curtis *et al*. (2006), entre outros, demonstraram a relação entre os poluentes clássicos (partículas inaláveis com diâmetro menor que 10 *microns* (MP_{10}), monóxido de carbono (CO), dióxido de enxofre (SO_2), óxidos de nitrogênio (NO_x) e ozônio (O_3)) e os problemas de saúde. No ano de 2012, por exemplo, a morte de 4,3 milhões de pessoas foi atribuída à poluição atmosférica (WHO, 2014). Além disso, a poluição do ar contribui para a degradação do meio ambiente, ajudando na propagação do efeito estufa.

Ressalta-se que, a intensificação do processo de industrialização ocorrida no século XIX, aliado ao crescimento populacional, especialmente, o crescimento da população urbana em detrimento da rural, vem aumentando as preocupações dos governos, sejam locais ou centrais, relacionadas à proteção do meio ambiente. Em relação à poluição do ar, de acordo com Vingarzan (2004) e Oltmans *et al*. (2006), em diversas partes do mundo esta vem

crescendo em função, principalmente, da industrialização, da urbanização e da queima de combustíveis fósseis. Conforme Gramsch et al. (2006), dado que a poluição atmosférica é mais concentrada em áreas urbanas e industriais, os esforços de monitoramento da qualidade do ar são maiores nestas áreas ou regiões.

Nesse contexto, vale ressaltar que, a economia do estado do Espírito Santo, onde se encontra a Região da Grande Vitória (RGV) [2], foco deste estudo, vem crescendo fortemente no decorrer dos últimos anos, especialmente, a partir de 2003, inclusive com taxas de crescimento do Produto Interno Bruto (PIB) superiores à média nacional. Com isso, diversas indústrias e empresas se instalaram ou ampliaram suas instalações no estado, principalmente, na RGV, o que tende, consequentemente, a elevar o nível de poluição atmosférica, mesmo existindo diversas regulamentações impostas pelos órgãos de controle ambiental a essas indústrias e empresas. Além disso, o crescimento da frota de veículos, o maior consumo de energia e etc., também contribuíram para a maior emissão de poluentes na região.

Destaca-se aqui que, no ano de 2010, a população do Espírito Santo era de 3.514.952 (INSTITUTO BRASILERO DE GEOGRAFIA E ESTATÍSTICA – IBGE, 2014). Desse total, 1.687.704 estava residindo na Região Metropolitana da Grande Vitória (RMGV), que é composta pelos municípios da RGV, mais Fundão e Guarapari. Tomando-se somente a RGV, a população chegou a 1.565.393, o que representou cerca de 44,54% da população capixaba. Logo, uma vez que Gramsch (2006) descreve que a maior concentração de poluentes está nas áreas urbanas e industriais, aproximadamente 45% da polução capixaba foi fortemente afetada pelas emissões de poluentes atmosféricos.

Importante destacar que, dentre os principais poluentes atmosféricos estão: CO, chumbo (Pb), material particulado (MP), SO_2, NO_x e O_3 (US EPA, 2014). Para Liu *et al*. (2013), os dois contaminantes atmosféricos que mais preocupam em relação à saúde humana são o ozônio e o material particulado, além de causarem vários danos ao meio ambiente. Segundo o último relatório de qualidade do ar da Secretaria Estadual do Ambiente do Espírito Santo (SEAMA), referente ao ano de 2013, a RGV vem apresentando problemas ambientais e sociais devido às concentrações de MP_{10} observadas nas estações locais de monitoramento. Devido aos problemas recorrentes da região com partículas, este trabalho teve como foco principal avaliar se as concentrações de material particulado inalável sofrem influência das condições meteorológicas observadas na região.

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

De acordo com Moreira, Tirabassi e Moraes (2008), as condições meteorológicas desempenham um papel importantíssimo na dispersão ou acumulação de poluentes. Liu e Johnson (2002) descreveram que a poluição do ar está associada, geralmente, à fatores como temperatura, umidade relativa, velocidade e direção do vento, entre outros. Como exemplo, tem-se que a baixa umidade relativa e a reduzida velocidade do vento tendem a elevar os níveis de poluentes. Já a ocorrência de precipitação pluviométrica e o aumento da velocidade do vento contribuem para a dispersão e diluição dos poluentes e, consequentemente, para a redução da concentração dos mesmos. Nesse contexto, destaca-se que, como na estação do verão ocorrem maiores volumes de chuvas do que no inverno, por exemplo, a tendência é que para poluentes como o MP_{10}, a concentração seja menor no período do verão.

Dessa forma, este trabalho objetivou avaliar os impactos das variáveis meteorológicas temperatura, umidade relativa, velocidade do vento e precipitação na qualidade do ar da RGV[3], considerando o poluente MP_{10}, por meio do modelo Logit simples. Além disso, foram estimados os efeitos de cada dia da semana e de cada estação do ano na probabilidade de acontecer incrementos de concentração de MP_{10}.

Uma das vantagens do modelo Logit é realizar estimativas de probabilidades de ocorrências em variáveis dependentes do tipo binário (*dummy*). Vale dizer que, apesar do modelo Logit ser amplamente utilizado para verificar os efeitos dos poluentes atmosféricos sobre os problemas de saúde, ver, por exemplo, Gent (2003) e Gehring et al. (2013), esse modelo, ainda, é pouco adotado nas análises que envolvem os impactos das variáveis meteorológicas na qualidade do ar.

No mais, nota-se, na literatura, um grande número de publicações que avaliam os impactos das variáveis meteorológicas sobre as concentrações de alguns poluentes, usando os modelos de regressão usuais (que consideram a variável dependente como contínua), por exemplo: 1) Liu e Johnson (2002), que realizaram previsões para picos diários de concentração de ozônio, em Milwaukee, Estados Unidos, adotando o modelo de regressão com erros de séries temporais (RTSE); 2) Liu et al. (2013), que investigaram a sensibilidade das concentrações de ozônio troposférico à variação da temperatura, em Taiwan, na China; 3) Lyra et al. (2011), que ajustaram e estimaram a concentração de MP_{10}, em função das condições meteorológicas, utilizando modelos de regressão linear múltipla, na cidade do Rio de Janeiro, Brasil; entre outros.

[3] Isto foi realizado por meio da classificação da qualidade do ar, no que tange ao poluente MP_{10}, em "não boa" e "boa", conforme detalhado no item "2.1. Região de estudo e apresentação das variáveis".

Entretanto, não foram encontrados muitos estudos que utilizaram o modelo Logit considerando a variável dependente como dicotômica (*dummy*) ou binária[4], conforme discutido no presente artigo. Com objetivo semelhante pode-se citar Kuchenho e Thamerus (1996), que utilizaram a técnica de regressão logística para estudar os eventos extremos de poluição do ozônio e do dióxido de enxofre, em Munique, na Alemanha, encontrando efeitos significativos das variáveis meteorológicas, tais como temperatura, velocidade do vento e umidade, sobre estes eventos extremos. Também, destaca-se Leite *et al*. (2011), que utilizaram a regressão logística simples para verificar a qualidade do ar atmosférico (considerando o poluente MP_{10}) na cidade de Uberlândia, Minas Gerais. Os autores verificaram que a umidade relativa e a precipitação influenciam significantemente a concentração de MP_{10}.

Importante dizer que, em seus resultados, Leite *et al*. (2011) focaram no cálculo da probabilidade de ocorrência de níveis de MP_{10} inferiores ou iguais a 50 $\mu g/m^3$, o que os autores denominaram de qualidade do ar "boa". No entanto, a pesquisa desenvolvida no presente artigo verificou o efeito marginal de cada variável preditora sobre a probabilidade de ocorrência de qualidade do ar "não boa", relativa ao MP_{10}, o que a difere da proposta de Leite *et al*. (2011).

Este artigo está estruturado da seguinte forma. Além desta introdução, a seção 2 traz uma descrição da região de estudo, as variáveis utilizadas e o modelo estatístico adotado. Na seção 3 apresentam-se as estimativas do modelo Logit. Por fim, as conclusões são apresentadas na seção 4.

2. Materiais e métodos

2.1. Região de estudo e apresentação das variáveis

A área de estudo compreendeu a Região da Grande Vitória, Espírito Santo, Brasil. Por estar situada na região litorânea, a RGV apresenta clima tropical quente (Aw), possuindo inverno ameno e seco, e verão chuvoso e quente. As temperaturas médias variam entre 24° C (Celsius) e 30° C e os ventos predominantes são de Norte/Nordeste na primavera/verão, sofrendo alterações durante outono e inverno devido ao posicionamento do sistema de alta pressão (Alta Pressão Subtropical do Atlântico Sul – ASAS) mais próximo do continente, possibilitando alterações na direção predominante do vento, a qual passa a variar entre as

[4] Variáveis dicotômicas (*dummies*) ou binárias são variáveis discretas que assumem valor igual a um se o evento ocorre e, zero, caso contrário (evento não ocorre).

direções Sul/Oeste (para mais detalhes, ver Instituto Estadual de Meio Ambiente e Recursos Hídricos (IEMA, 2014)). A RGV possui oito estações de monitoramento de qualidade do ar, que fazem parte da Rede Automática de Monitoramento da Qualidade do Ar (RAMQAr), a saber: Laranjeiras; Carapina; Jardim Camburi; Enseada do Suá; Vitória – Centro; Vila Velha – Ibes; Vila Velha – Centro; e, Cariacica (figura 1).

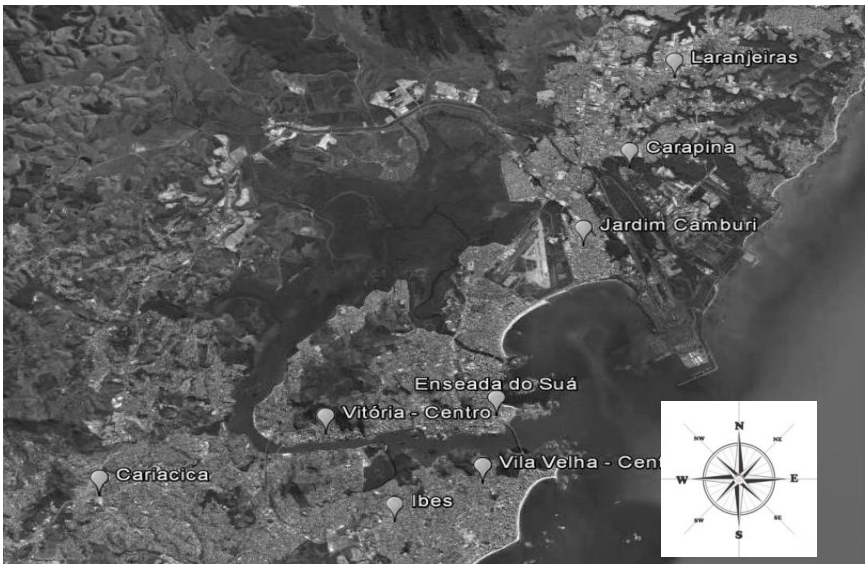


Figura 1 – Estações de monitoramento da qualidade do ar na Grande Vitória.
Fonte: Google Earth (2014).

Os dados utilizados nessa pesquisa foram medidos pelas estações da RAMQAr, compreendendo o período de janeiro de 2005 a dezembro de 2010. A RAMQAr fornece dados horários de concentração de MP_{10}. De posse das concentrações horárias foi realizada a média de 24h para cada estação. Após o cálculo das médias de 24h, identificou-se a maior média de 24h entre as oito estações, essa maior concentração média foi selecionada para representar a concentração média de um determinado dia para toda região.

Para atingir o objetivo proposto, este trabalho tomou como base o relatório de qualidade do ar do Estado de São Paulo (CETESB, 2014). Nesse caso, foi considerada a classificação do índice de qualidade do ar (IQA) para o MP_{10}, índice este que está associado a efeitos à saúde e, portanto, independe do padrão de qualidade em vigor. A classificação da qualidade do ar em relação ao índice de MP_{10} é feita da seguinte forma: boa (0 – 50 $\mu g/m^3$), moderada (>50 – 100 $\mu g/m^3$), ruim (>100 – 150 $\mu g/m^3$), muito ruim (>150 – 250 $\mu g/m^3$) e péssima (> 250 $\mu g/m^3$). Na presente pesquisa, quando a maior média de 24h entre as oito estações foi igual ou inferior a 50 $\mu g/m^3$, a classificação foi considerada como “boa” em

termos de efeitos à saúde. Caso contrário, a classificação foi definida como “não boa” (moderada, ruim, muito ruim e péssima).

Como a regressão logística simples considera a variável dependente como dicotômica (*dummy*), a concentração de MP_{10} (maior média de 24h entre as oito estações) foi transformada em uma variável *dummy*, apresentando a seguinte classificação: um (1), para classificação “não boa” (qualidade do ar “não boa” em termos de efeitos à saúde) e, zero (0), para “boa”.

Para verificar os efeitos sobre a qualidade do ar em relação à concentração de MP_{10}, na RGV, foram consideradas, primeiramente, as variáveis meteorológicas temperatura (TEMP), umidade relativa (UMID), velocidade do vento (VELVENT) e precipitação pluviométrica (PREC). Em complemento, verificou-se os efeitos de cada dia da semana e de cada estação do ano na probabilidade de qualidade do ar “não boa”. Para representar os dias da semana foram criadas sete variáveis binárias (*dummies*) e para as estações do ano quatro variáveis binárias.

2.2. Modelo Logit

Para verificar a influência das variáveis preditoras na probabilidade de ocorrência de qualidade do ar “não boa”, foi utilizado o modelo Logit simples (GUJARATI; PORTER, 2008), que admite valores discretos, zero e um (variável binária), para a variável dependente. Um dos principais objetivos dos modelos de resposta binária é calcular a probabilidade de um dado evento, com determinado conjunto de atributos, de fato acontecer.

No modelo Logit utliza-se uma função de distribuição acumulada logística, dada por:

$$L(X_t\beta) = \frac{1}{1+e^{-X_t\beta}}, \qquad (1)$$

em que L representa a função de distribuição logística; X_t, vetor de variáveis independentes; β, vetor de parâmetros; e, e, base do logaritmo natural.

A ocorrência ou não de uma classificação “não boa” da qualidade do ar depende de vários fatores. Como os parâmetros dessa ocorrência não são observáveis para cada ponto do tempo t, pode-se definir uma variável latente ou não observada, Y_t^*, como

$$Y_t^* = X_t\beta + \mu_t, \qquad (2)$$

em que Y_t^* é variável dependente; β, parâmetros; X_t, conjunto de variáveis explicativas; μ_t, erro aleatório; e, $t = 1,..., n$.

A ocorrência de uma determinada classificação pode ser descrita pela variável binária, Y_t, tal que $Y_t = 1$, se a classificação é "não boa" e, $Y_t = 0$, se é "boa". Esses valores observados de Y_t estão relacionados com Y_t^*, como segue:

$Y_t = 1$, se $Y_t^* > 0$; e, $Y_t = 0$, se $Y_t^* = 0$,

$$\text{Prob}\left(Y_t = 1\right) = \text{Prob}\left(Y_t^* > 0\right) = \text{Prob}\left(\mu_t > -X_t\beta\right), \tag{3}$$

$$\text{Prob}\left(Y_t = 0\right) = \text{Prob}\left(Y_t^* = 0\right) = \text{Prob}\left(\mu_t \leq -X_t\beta\right). \tag{4}$$

O modelo é estimado pelo Método de Máxima Verossimilhança. A probabilidade de ocorrência da classificação "não boa" (a) e a probabilidade de ocorrência da classificação "boa" (b) podem ser calculadas pelas seguintes expressões:

$$\text{(a) } P_t = \frac{1}{1+e^{-X_t\beta}} \qquad \text{e} \qquad \text{(b) } 1-P_t = \frac{e^{-X_t\beta}}{1+e^{-X_t\beta}}, \tag{5}$$

sendo P_t igual a probabilidade de ocorrência da classificação "não boa"; $1-P_t$, probabilidade de ocorrência da classificação "boa"; X_t, variáveis explicativas do modelo; e β, coeficientes das variáveis explicativas.

Para determinar o efeito marginal de cada variável preditora, sobre a probabilidade de ocorrência da classificação "não boa", é necessário usar os valores médios das variáveis explicativas. O efeito marginal da variável X_t sobre a variável dependente é descrito pela expressão:

$$\frac{\partial P_t}{\partial X_t} = \beta \times \frac{1}{1+e^{-X_t\beta}} \times \frac{e^{-X_t\beta}}{1+e^{-X_t\beta}}, \tag{6}$$

considerando-se $P_t = \frac{1}{1+e^{-X_t\beta}}$ e $1-P_t = \frac{e^{-X_t\beta}}{1+e^{-X_t\beta}}$.

Como já mencionado, X_t representa o conjunto de variáveis explicativas. Assim, em função dos objetivos dessa pesquisa, pode-se dividir X_t em três grupos:

1) X_t igual as variáveis meteorológicas: aqui, X_t foi representado pelas variáveis contínuas temperatura, umidade relativa, velocidade do vento e precipitação pluviométrica. Estimou-e uma equação para captar o efeito dessas variáveis sobre a probabilidade de ocorrência de qualidade do ar "não boa";

2) X_t igual aos dias da semana (domingo, segunda, terça, quarta, quinta, sexta e sábado): nesse caso, foram criadas sete variáveis binárias (*dummies*), sendo uma para cada dia da semana. Para exemplificar, considere que X_t refere-se ao dia de domingo. Assim, $X_t = 1$,

se domingo e, $X_t = 0$, caso contrário. Lembrando que foi estimada uma regressão logística para cada dia da semana, perfazendo um total de sete regressões;

3) X_t igual às estações do ano (primavera, verão, outono e inverno): aqui, adotou-se quatro variáveis binárias (*dummies*), sendo uma para cada estação do ano. Exemplificando, suponha que X_t é a estação do inverno. Logo, $X_t = 1$, se inverno e, $X_t = 0$, caso contrário. Ressalta-se que foi estimada uma regressão logística para cada estação do ano, totalizando quatro regressões.

3. Resultados e discussões

3.1. Aspectos gerais das variáveis

Na tabela 1 são apresentadas as estatísticas descritivas das variáveis: concentrações de MP_{10}, temperatura (TEMP), umidade relativa (UMID), velocidade do vento (VELVENT) e precipitação pluviométrica (PREC). Em geral, observando-se os desvios-padrão e as diferenças entre os máximos e mínimos, nota-se que as variáveis apresentaram grande dispersão em termos estatísticos, exceção feita à temperatura, que climatologicamente não varia muito na RGV (IEMA, 2014). Especificamente em relação às concentrações de MP_{10} (maior média de 24h entre as oito estações), observa-se que, em média, as concentrações não ultrapassaram o valor de 50 $\mu g/m^3$. No entanto, nota-se um desvio-se padrão relativamente alto, o que demonstra que, um intervalo de confiança de um desvio padrão em relação à média teria um limite superior maior que 50 $\mu g/m^3$, gerando uma classificação do ar "não boa". Além disso, os resultados revelam que o valor máximo foi mais do que o dobro do valor médio, demonstrando a grande variabilidade das concentrações máximas de MP_{10} na RGV.

Tabela 1 – Estatísticas descritivas das variáveis contínuas

Estatística	Variáveis				
	PM_{10}	TEMP	UMID	VELVENT	PREC
Média	46,62	24,36	77,79	2,05	0,17
Mediana	44,83	24,35	77,33	1,99	0,00
Desvio-padrão	14,05	2,54	6,61	0,56	0,51
Mínimo	12,17	17,00	59,49	0,86	0,00
Máximo	117,33	30,80	97,65	5,49	6,74
Observações	2191	2191	2191	2191	2191

Fonte: elaborado a partir dos dados da pesquisa.

A figura 2 demonstra a evolução das concentrações de MP_{10} (maior média de 24h entre as oito estações) no período de análise deste estudo, que foi de janeiro de 2005 a

dezembro de 2010, perfazendo um total de 2.191 observações (dias). Desse total, cerca de 34% dos dias apresentou concentrações superiores a 50 $\mu g/m^3$, o que dá origem à uma classificação "não boa" da qualidade do ar. Além disso, constata-se, novamente, a grande variabilidade das concentrações de MP_{10} na RGV. Vale mencionar que, levando em consideração as estações de forma individual, o maior número de concentrações superiores a 50 $\mu g/m^3$ ocorreu, em ordem decrescente, nas estações de Cariacica, Laranjeiras, Vila Velha – Centro, Vila Velha – Ibes, Enseada do Suá, Jardim Camburi, Carapina e Vitória – Centro. Logo, como a concentração diária de MP_{10} utilizada para as estimativas desse estudo foi a maior média de 24h entre as oito estações, a estação de Cariacica foi a que mais contribuiu na formação da mesma. Esse resultado corrobora com a análise de tendência da concentração do MP_{10} apresentada no último relatório de qualidade do ar local. Os dados monitorados mostraram que de 2003 a 2013 a estação de Cariacica foi aquela que obteve as maiores concentrações de MP_{10} de toda a rede, apresentando uma tendência de aumento nos últimos anos (desde 2011). Todas as estações apresentaram tendência de diminuição da média de concentração anual de MP_{10}, exceto nas estações Jardim Camburi e Cariacica (IEMA, 2014).

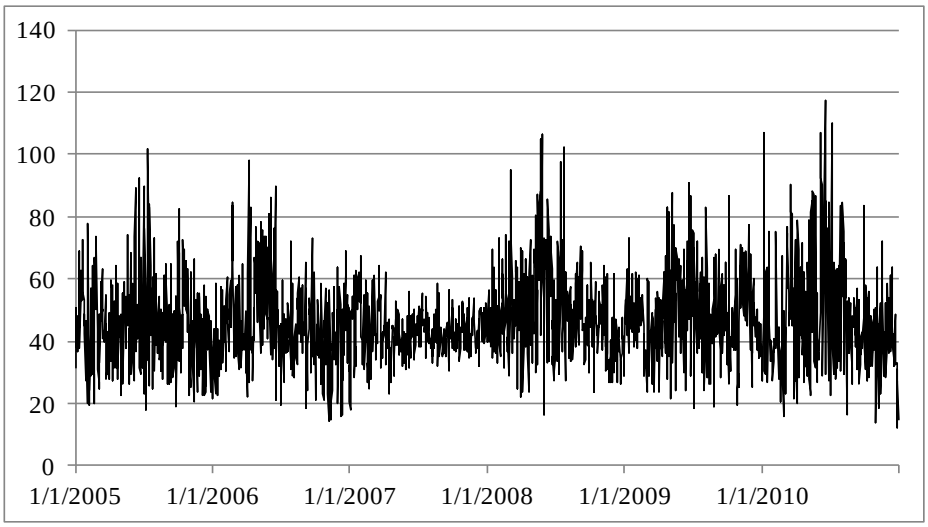


Figura 2 – Concentrações máximas diárias de PM_{10} ($\mu g/m^3$) na RGV, de 01/01/2005 a 31/12/2010.
Fonte: elaborado a partir dos dados da pesquisa.

3.2. Estimativas da regressão logística

Os resultados da tabela 2 representam a equação logística ajustada, considerando como variáveis exógenas a umidade relativa, a velocidade do vento e a precipitação. Também são apresentados os respectivos efeitos marginais de cada variável sobre a probabilidade de ocorrência de qualidade do ar "não boa". Verifica-se que a regressão como um todo foi estatisticamente significativa (Prob/LR estat.). Uma vez que a variável temperatura não foi

estatisticamente significativa, a mesma não se encontra na tabela 2. Vale dizer que, historicamente, a temperatura média da RGV não apresenta grandes variações ao longo do ano, o que pode justificar a não significância de tal variável. Já a umidade relativa, a velocidade do vento e a precipitação foram individualmente significantes na determinação de qualidade do ar "não boa" e os seus sinais coerentes com o esperado.

Destaca-se que, os coeficientes das variáveis explicativas, estimados pelo modelo Logit, não refletem seu efeito marginal sobre a probabilidade de ocorrência de um dado evento. Assim, para a determinação do efeito marginal foram usados os valores médios das variáveis preditoras, de acordo com a equação 6. Nota-se que, o maior efeito marginal ocorreu para variável precipitação pluviométrica, seguido, respectivamente, por velocidade do vento e umidade relativa. Ressalta-se, ainda, que o efeito marginal da umidade foi muito pequeno.

Tabela 2 – Equação logística considerando as variáveis umidade relativa, velocidade do vento e precipitação e seus efeitos marginais

Variáveis	Coeficientes	Erro-padrão	Valor de Z	Valor-p	Efeito marginal
C	6,2674***	0,9145	6,8533	0,0000	-
UMID	-0,0658***	0,0103	-6,3960	0,0000	-0,0141
VELVENT	-0,8526***	0,1052	-8,1016	0,0000	-0,1833
PREC	-1,1015***	0,2500	-4,4054	0,0000	-0,2368
Obs. com variável dependente = 0		1454	Total obs. = 2.191		
Obs. com variável dependente = 1		737			
Prob. (LR estat.)		0,0000			

Fonte: elaborado a partir dos dados da pesquisa.
Nota: 1) *** Significativo a 1%; e, 2) As estimativas foram realizadas utilizando o método de covariância robusta GLM (Modelo Linear Generalizado).

Para a variável precipitação, por exemplo, o efeito marginal, igual a -0,2368, significa que, o aumento de um milímetro (1 mm) na precipitação média diária na região ocasionou uma redução na probabilidade de qualidade do ar "não boa" de 23,68 pontos percentuais. Importante mencionar os elevados efeitos marginais negativos das variáveis velocidade do vento e precipitação. Isso indica que maiores velocidades do vento e altos volumes de chuvas contribuíram fortemente para redução da probabilidade de ocorrência de qualidade do ar classificada como "não boa", na RGV, uma vez que contribuem para a maior dispersão e diluição de poluentes.

Destaca-se que, conforme o Inventário de Fontes de Emissões Atmosféricas da Região da Grande Vitória, a principal fonte emissora de partículas na RGV são veículos automotores, especialmente as emissões ligadas à ressuspensão de partículas já depositadas nas vias (ECOSOFT, 2011). Dessa forma, a grande importância da variável precipitação na redução

das concentrações de MP_{10} pode estar relacionada ao processo de deposição úmida e a redução da ressuspensão de poeira do solo.

Vale ressaltar que, em relação ao coeficiente negativo da velocidade do vento, de acordo com Kukkonen et al. (2005), mesmo que o esperado seja que ventos fortes dissipem a poluição do ar gerada localmente, também pode haver aumento dos níveis de MP_{10}, sob certas condições meteorológicas, em função da ressuspensão de poeira do solo e de estradas. Segundo Vardoulakis e Kassomenos (2008), esse efeito ocorre com maior frequência em dias quentes e secos.

Referente ao coeficiente da variável umidade, embora pequeno, ele foi estatisticamente significativo, indicando que tal variável contribuiu para a redução das concentrações de MP_{10} na RGV. Como mencionado por Vardoulakis e Kassomenos (2008) e Lyra et al. (2011), em dias em que a umidade é alta, a tendência é de diminuição da ressuspensão do solo para o poluente MP_{10}, especialmente quando as velocidades dos ventos são baixas. Lembrando que a ressuspensão de poeira do solo é a principal fonte de emissão de partículas da RGV, segundo o inventário oficial do órgão ambiental.

A fim de enriquecer o trabalho, também foram estimadas regressões logísticas para cada dia da semana e para cada estação do ano (primavera, verão, outono e inverno). A partir dessas equações foram calculados os respectivos efeitos marginais. Na tabela 3 são demonstradas as estimativas das equações logísticas quando considerados os dias da semana (foi estimada uma equação para cada dia), assim como o efeito marginal relativo à cada dia. Como o dia de segunda-feira não foi significativo estatisticamente, o mesmo não foi apresentado. Observa-se que, nos dias relativos à terça, quarta, quinta e sexta-feira, a probabilidade de ocorrência de qualidade do ar "não boa" foi muito maior do que nos fins de semana (domingo e sábado). No domingo, por exemplo, a chance de qualidade do ar "não boa" reduziu-se em 15,44 pontos percentuais. Já na sexta-feira, a probabilidade de ocorrência de uma qualidade do ar "não boa" aumentou em 9,61 pontos percentuais.

Conforme o Inventário de Fontes de Emissões Atmosféricas da Região da Grande Vitória (ECOSOFT, 2011), em segundo lugar na lista dos grupos mais importantes para a emissão de partículas na RGV está o setor industrial minero-siderúrgico. Em terceiro lugar encontra-se o setor logístico, que inclui portos e aeroportos[5]. Logo, pode-se dizer que a menor probabilidade de ocorrência de qualidade do ar "não boa" nos fins de semana já era esperada,

[5] O inventário não contempla as atividades da construção civil.

uma vez que aos sábados e domingos ocorre redução da produção industrial, dos serviços de logística e diminuição da circulação de veículos.

Tabela 3 – Equações logísticas e efeitos marginais para cada dia da semana

Variáveis	Coeficientes	Erro-padrão	Valor de Z	Valor-p	Efeito marginal
Domingo					
C	-0,5342***	0,0478	-11,173	0,0000	-0,1544
DOM	-1,3030***	0,1710	-7,6196	0,0000	
Terça					
C	-0,9786***	0,0518	-18,897	0,0000	0,0588
TER	0,2663**	0,1310	2,0330	0,0421	
Quarta					
C	-0,9786***	0,0518	-18,897	0,0000	0,0588
QUAR	0,2663**	0,1310	2,0330	0,0421	
Quinta					
C	-0,9894***	0,0519	-19,058	0,0000	0,0752
QUIN	0,3343**	0,1300	2,5722	0,0101	
Sexta					
C	-1,0029***	0,0521	-19,258	0,0000	0,0961
SEX	0,4181***	0,1289	3,2436	0,0012	
Sábado					
C	-0,9022***	0,0509	-17,715	0,0000	-0,0488
SAB	-0,2702*	0,1424	-1,8971	0,0578	

Fonte: elaborado a partir dos dados da pesquisa.

Nota: 1) *** Significativo a 1%, ** Significativo a 5%, * Significativo a 10%; e, 2) As estimativas foram realizadas utilizando o método de covariância robusta GLM.

Por fim, a tabela 4 traz os resultados das estimativas das equações logísticas quando consideradas as estações do ano (foi estimada uma equação para cada estação), assim como o efeito marginal relativo à cada estação. Todas as estações do ano foram estatisticamente significativas. Como era esperado, os coeficientes da primavera e do verão foram negativos e os do outono e do inverno positivos, indicando que nos períodos de temperaturas mais baixas, com menores volumes de chuva, a chance de ocorrência de qualidade do ar "não boa" aumentou. No inverno, por exemplo, o efeito marginal igual a 0,1588 demonstra que nesta estação a probabilidade de uma qualidade do ar "não boa" elevou-se em 15,88 pontos percentuais, ao passo que no verão houve uma redução de 7,68 pontos percentuais nessa probabilidade.

4. Conclusões

Esse trabalho teve como objetivo verificar os impactos das variáveis meteorológicas temperatura, umidade relativa, velocidade do vento e precipitação sobre a qualidade do ar, na

RGV, considerando o poluente MP_{10}, por meio do modelo Logit. Para isto, a qualidade do ar, no que se refere ao MP_{10}, foi classificada como "não boa" e "boa". Também foram estimados os efeitos dos dias da semana e das estações do ano sobre a probabilidade de ocorrência de qualidade do ar "não boa".

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

Por fim, vale destacar que, as preocupações referentes à poluição do ar vêm aumento ao longo do tempo, dado que, cada vez mais, a poluição tem afetado a saúde humana, a fauna e a flora. Dessa forma, este estudo visa contribuir na tomada de decisões dos agentes públicos no que diz respeito ao combate à poluição, à prevenção de altas concentrações e à formulação de legislações para esse fim.

Tabela 4 – Equações logísticas e efeitos marginais para cada estação do ano

Variáveis	Coeficientes	Erro-padrão	Valor de Z	Valor-p	Efeito marginal
			Primavera		
C	-0,8413***	0,0536	-15,693	0,0000	-0,0733
PRIM	-0,4283***	0,1172	-3,6554	0,0003	
			Verão		
C	-0,8370***	0,0535	-15,650	0,0000	-0,0768
VER	-0,4536***	0,1181	-3,8405	0,0001	
			Outono		
C	-0,8238***	0,0538	-15,326	0,0000	0,1310
OUT	0,5351***	0,1010	5,2958	0,0000	
			Inverno		
C	-0,6238***	0,0438	-14,258	0,0000	0,1588
INV	0,6351***	0,0501	12,6750	0,0001	

Fonte: elaborado a partir dos dados da pesquisa.

Nota: 1) *** Significativo a 1%; e, 2) As estimativas foram realizadas utilizando o método de covariância robusta GLM.

5. Referências bibliográficas

BRUNEKREEF B.; HOLGATE, S. T. Air pollution and health. **Lancet**, v. 360, n. 9341, p. 1233-1242, 2002.

COMPANHIA AMBIENTAL DO ESTADO DE SÃO PAULO (CETESB). **Relatório de Qualidade do ar no Estado de São Paulo 2014**. São Paulo, 2014. Disponível em: http://ar.cetesb.sp.gov.br/publicacoes-relatorios/. Acessado em: 25 de ago. de 2014.

CURTIS, L.; REA, W.; SMITH-WILLIS, P.; FENYVES, E.; PAN, YAQIN. Adverse health effects of outdoor air pollutants. **Environment International**, v. 32, n. 6, p. 815-830, 2006.

ECOSOFT CONSULTORIA E SOFTWARES AMBIENTAIS. **Inventário de emissões atmosféricas da Região da Grande Vitória**. Vitória, 2011. Disponível em: http://www.es.gov.br/Banco%20de%20Documentos/PDF/Maio/100511/RTC10131-R1.pdf. Acessado em: 20 de mar. de 2014.

GEHRING, ULRIKE ET AL. Air pollution exposure and lung function in children: the escape project. **Environmental Health Perspectives**, v. 121, n. 11-12, p. 1357-1364, 2013.

GENT, J. F.; TRICHE, E. W.; HOLFORD, T. R.; BELANGER, K.; BRACKEN, M. B.; BECKETT, W. S.; LEADERER, B. P. Association of low-level ozone and fine particles with respiratory symptoms in children with asthma. **American Medical Association**, v. 8, n. 14, p. 1859-1867, 2003.

GOOGLE EARTH. **Informações geográficas**. 2014. Disponível em: <http://www.google.com.br/intl/pt-PT/earth/. Acessado em: 20 de mar. de 2014.

GRAMSCH, E.; CERECEDA-BALIC, F.; OYOLA, P.; VON BAER, D. Examination of pollution trends in Santiago de Chile with cluster analysis of PM_{10} and ozone data. **Atmospheric Environment**, v. 40, n. 28, p. 5464-5475, 2006.

GUJARATI, D. N.; PORTER, D. C. **Basic Econometrics**. 5 ed. New York: McGraw-Hill/Irwin, 2008. 944 p.

INSTITUTO BRASILERO DE GEOGREFIA E ESTATÍSTICA (IBGE). 2014. **Banco de dados. Cidades@**. Disponível em: http://www.cidades.ibge.gov.br/xtras/home.php. Acessado em: 20 de mar. de 2014.

INSTITUTO ESTADUAL DE MEIO AMBIENTE E RECURSOS HÍDRICOS DO ESTADO DO ESPÍRITO SANTO. Relatório da qualidade do ar da Região da Grande Vitória. Vitória, 2013. Disponível em: <http://www.meioambiente.es.gov.br>. Acessado em: 27 de jun. de 2015.

KUCHENHO, H.; THAMERUS, M. Extreme value analysis of Munich air pollution data. **Environmental and Ecological Statistics**, v. 3, n. 2, p. 127-141, 1996.

KUKKONEN, J.; POHJOLA, M.; SOKHI, R. S.; LUHANA, L.; KITWIROON, N.; FRAGKOU, L.; RANTAMAKI, M.; BERGE, E.; ØDEGAARD, V.; HAVARD SLØRDAL, L.; DENBY, B.; FINARDI, S. Analysis and evaluation of selected local-scale PM_{10} air pollution episodes in four European cities: Helsinki, London, Milan and Oslo. **Atmospheric Environment**, v. 39, p. 2759-2773, 2005.

LEITE, R. C. M.; GUIMARÃES, E. C.; LIMA, E. A. R. L.; BARROZO, M. A. S. B.; TAVARES, M. Utilização de regressão logística simples na verificação da qualidade do ar atmosférico de Uberlândia. **Engenharia Sanitária Ambiental**, v. 16, n. 1, p. 175-180, 2011.

LIU, P. W. G.; JOHNSON, R. Forecasting peak daily ozone levels-I. A regression with time series errors model having a principal component trigger to fit 1991 ozone levels. **Journal of the Air & Waste Management Association**. v. 52, n. 9, p1064-1074, 2002.

LIU, P. W. G.; TSAI, J. H.; LAI, H. C.; TSAI, D. M.; LI, L. W. Establishing multiple regression models for ozone sensitivity analysis to temperature variation in Taiwan. **Atmospheric Environmen**t, v. 79, p. 225-235, 2013.

LYRA, G. B.; ODA-SOUZA, M.; VIOLA, D. N. Modelos lineares aplicados à estimativa da concentração do material particulado (MP_{10}) na cidade do Rio de Janeiro, RJ. **Revista Brasileira de Meteorologia**, v. 26, n. 3, p. 392-400, 2011.

MAYNARD, R. Key airborne pollutants: the impact on health. **Science of the Total Environment**, v. 334-335, p. 9-13, 2004.

MOREIRA, D. M.; TIRABASSI, T.; MORAES, M. R. Meteorologia e poluição atmosférica. **Ambiente e Sociedade**. v. 11, n. 1, p. 1-13, 2008.

OLTMANS, S. J.; LEFOHN, A. S.; HARRIS, J. M.; GALBALLY, I.; SCHEEL, H. E.; BODEKER, G.; BRUNKE, E.; CLAUDE, H.; TARASICK, D.; JOHNSON, B. J.; SIMMONDS, P.; SHADWICK, D.; ANLAUF, K.; HAYDEN, K.; SCHMIDLIN, F.; FUJIMOTO, T.; AKAGI, K.; MEYER, C.; NICHOL, S.; DAVIES, J.; REDONDAS, A.; CUEVAS, E. Long-term changes in tropospheric ozone. **Atmospheric Environment**, v. 40, n. 17, p. 3156-3173, 2006.

UNITED STATES ENVIRONMENTAL PROTECTION AGENCY (US EPA**). Air and Radiation – Air Pollutants**. 2014. Disponível em: <http://www.epa.gov/air/>. Acessado em: 24 mar. 2014.

VARDOULAKIS, S.; KASSOMENOS, P. Sources and factors affecting PM_{10} levels in two European cities: implications for local air quality management. **Atmospheric Environment**, v. 42, n. 17, p. 3949–3963, 2008.

VINGARZAN, R. A review of surface ozone background levels and trends. **Atmospheric Environment**, v. 38, n. 21, p. 3431-3442, 2004.

WORLD HEALTH ORGANIZATION (WHO). **Air pollution estimates**. 2014. Disponível em: <http://www.who.int/phe/health_topics/outdoorair/databases/FINAL_HAP_AAP_BoD_24March2014.pdf?ua=1>. Acessado em: 26 de mar. 2014.

_____. **WHO air quality guidelines global update 2005**. Report on a working group meeting, Bonn/Germany, 18-20 October 2005, 2005. Disponível em: <http://www.euro.who.int/__data/assets/pdf_file/0008/147851/E87950.pdf>. Acessado em: 20 de mar. de 2014.

Inter-relações entre as concentrações de ozônio e de dióxido de nitrogênio na Região da Grande Vitória, Espírito Santo, Brasil[1]

Interrelationships between the ozone and nitrogen dioxide concentrations in the Region of Grande Vitória, Espírito Santo, Brazil

Edson Zambon Monte
Doutorando em Engenharia Ambiental pela Universidade Federal do Espírito Santo (UFES). Professor do Departamento de Economia da UFES.

Taciana Toledo de Almeida Albuquerque
Doutora em Meteorologia pela Universidade Federal de São Paulo (USP). Professora do Departamento de Engenharia Sanitária e Ambiental da Universidade Federal de Minas Gerais (UFMG) e do Programa de Pós-graduação em Engenharia Ambiental, da UFES.

Valdério Anselmo Reisen
Doutor em Estatística pela University of Manchester Institute of Science And Technology. Professor do Departamento de Estatística e do Programa de Pós-graduação em Engenharia Ambiental, da UFES.

Resumo

Este trabalho objetivou verificar as inter-relações entre as concentrações de ozônio (O_3) e de dióxido de nitrogênio (NO_2), na Região da Grande Vitória (RGV), Espírito Santo, Brasil. Adotou-se a metodologia vetorial auto-regressiva (VAR) e o teste de causalidade de Granger. O modelo VAR captura as interdependências lineares entre várias séries temporais, sendo que, cada variável possui uma equação estimada que representa sua evolução em termos de suas próprias defasagens e das defasagens das outras variáveis. Já o teste de causalidade de Granger baseia-se em um sistema de equações bivariado, para verificar se uma variável é capaz de prever a outra. Os resultados revelaram que as concentrações de O_3 e de NO_2 da região (estação) de Laranjeiras foram as menos afetadas por concentrações de outras estações. Dada à localização, as concentrações de O_3 e NO_2 da Enseadá do Suá tiveram significativa influência de outras

[1] Artigo aceito para publicação na revista Engenharia Sanitária e Ambiental.

regiões, especialmente de Jardim Camburi, Ibes e Vitória – Centro. A concentração de ozônio na região do Ibes foi fortemente influenciada pelas concentrações de O_3 e de NO_2 da Enseada do Suá. Além disso, as concentrações de Cariacica sofreram impactos relevantes das concentrações da Enseada do Suá, provavelmente devido à direção do vento Norte/Nordeste, predominante na RGV.

Palavras-chaves: ozônio; dióxido de nitrogênio; poluição do ar; séries temporais; vetores auto-regressivos.

Abstract

The objective of this paper was to determine the interrelationships between the ozone (O_3) and nitrogen dioxide (NO_2) concentrations, in the Grande Vitória Region (RGV), Espírito Santo, Brazil, using the methodology VAR and the Granger causality test. The VAR model captures the linear interdependencies between multiple time series. In this context, each variable has an estimated equation that represents its evolution in terms of its own lags and the lags of other variables. Granger causality test is based on a system of equations bivariate to check whether a variable is able to forecast the other. The results showed that the O_3 and NO_2 concentrations at Laranjeiras station were less affected by concentrations of other stations. The concentrations of Enseada do Suá were significantly affected by other regions, especially Jardim Camburi, IBES and Vitória – Centro. The Ibes ozone concentrations were strongly influenced by the O_3 and NO_2 concentrations from Enseada do Suá. Furthermore, the O_3 and NO_2 concentrations of Cariacica had significant impacts of concentrations of the Enseada do Suá, probably due to the prevailing North/Northeast wind direction in the RGV.

Key-words: ozone; nitrogen dioxide; air pollution; time series; autoregressive vectors.

INTRODUÇÃO

O processo de industrialização, aliado ao grande crescimento populacional ocorrido nas últimas décadas, vem aumentando as preocupações relacionadas à proteção do meio ambiente. Nesse contexto, cada vez mais tem se dado atenção para os efeitos adversos que a poluição atmosférica pode causar à fauna, à flora e, também, à saúde humana (tais como: irritação dos olhos, problemas pulmonares, alergias, etc.). Liu *et al*. (2013) consideram que os dois poluentes do ar que mais preocupam em relação à saúde humana são o ozônio (O_3) e o material particulado.

Vale ressaltar que os males causados pelo ozônio ocorrem na faixa de ar perto da superfície terrestre, onde o gás é muito tóxico. Conforme Seinfeld e Pandis (2006), o ozônio apresenta um duplo paradoxo na atmosfera, pois, esse poluente tem um papel benéfico na estratosfera e maléfico na troposfera.

O ozônio troposférico é um oxidante fotoquímico formado a partir de reações químicas na atmosfera originadas pela presença de dióxido de nitrogênio (NO_2) e radiação proveniente do sol. Além do dióxido de nitrogênio, os hidrocarbonetos (HC), também conhecidos como compostos orgânicos voláteis (COV), são importantes precursores do ozônio na baixa troposfera (Instituto Estadual de Meio Ambiente e Recursos Hídricos – IEMA, 2014). De acordo com Holgate *et al*. (1999), o processo de conversão do óxido nítrico (NO) em dióxido de nitrogênio é um fator fundamental na formação fotoquímica do ozônio, em regiões urbanas poluídas. De modo simplificado, as reações de formação do ozônio troposférico se iniciam pela fotodissociação do dióxido de nitrogênio em monóxido de nitrogênio e oxigênio atômico. Então, o oxigênio atômico reage com o oxigênio molecular presente no ar e forma o ozônio. Posteriormente, o ozônio reage com o óxido de nitrogênio para formar o dióxido de nitrogênio e oxigênio molecular, fechando o ciclo fotoquímico. Assim, o ozônio é formado por uma reação fotoquímica na atmosfera e requer a presença de precursores e também de radiação solar.

De acordo com o inventário de fontes de emissão oficial do IEMA (Ecosoft Consultoria e Softwares Ambientais, 2011), ano base 2009, a principal fonte de NO_X, na Região da Grande Vitória (RGV), foco deste estudo, são as indústrias minero-siderúrgicas (47,6%), com concentração das emissões na Ponta de Tubarão. Em segundo lugar aparecem as emissões veiculares (33,4%) e, em terceiro, atividade logística (17,2%).

Segundo Seinfeld e Pandis (2006), a formação do ozônio depende de diversos fatores químicos e físicos, que variam no espaço e no tempo de forma não linear. É importante ressaltar que, somente as reações entre NO, NO_2 e O_3 não explicam totalmente os altos níveis de ozônio formados na baixa atmosfera, pois não há produção líquida de O_3. Reações adicionais envolvendo os COV na atmosfera consomem NO e o transformam em NO_2, gerando mais O_3. Dessa forma, a presença de COV na atmosfera aumenta significativamente os níveis de O_3. Os COV, por sua vez, são emitidos na atmosfera em ambientes urbanos e industriais por diversas fontes como combustão incompleta de combustíveis fósseis, plásticos e outros compostos de carbono e evaporação de reservatórios, entre outras. A relação entre COV e óxidos de nitrogênio, variando entre 4 e 10, favorece a formação de ozônio. Assim, a concentração de

ozônio na troposfera depende da presença de outros poluentes e das condições meteorológicas (Seinfeld e Pandis, 2006).

Portanto, devido as suas características, os fatores meteorológicos têm importante papel na formação do ozônio. Ryan, Piety e Luebehusen (1999) destacaram que a radiação ultravioleta tem papel fundamental na formação de O_3. Moreira, Tirabassi e Moraes (2008) descreveram que as condições meteorológicas desempenham papel relevante na dispersão ou acumulação de poluentes. Já Liu e Johnson (2002) salientaram que a poluição do ar, particularmente a concentração de ozônio, é altamente correlacionada no tempo, estando associada, geralmente, a fatores como temperatura, umidade relativa, velocidade e direção do vento, entre outros.

De acordo com Carvalho (2006), a formação secundária do ozônio faz com que, para esse poluente, as modelagens estatísticas sejam mais complexas do que para outros poluentes. No entanto, a natureza da concentração de ozônio tem sido abordada por vários estudos estatísticos, tais como: Ryan (1995); Jorquera *et al*. (1998); Liu e Johnson (2002); Liu e Johnson (2003); entre outros. Porém, nota-se que não é comum à adoção da abordagem vetorial auto-regressiva (VAR) e de testes de causalidade, para a modelagem da concentração de poluentes, especialmente de ozônio, sendo que essas metodologias, por serem multivariadas, podem verificar de melhor forma as inter-relações entre as variáveis que dão origem à determinado poluente.

No que tange à adoção da modelo VAR, pode-se citar o estudo de Hsu (1992), que verificou a interdependência entre os poluentes O_3, NO e NO_2, na cidade de Taipei, Taiwan. Cai (2008), que analisou a associação entre a concentração mensal de monóxido de carbono e as variáveis, precipitação, temperatura, radiação solar e tráfego de veículos, na costa sul da Califórnia, Estados Unidos. E, Wang e Niu (2009), que aplicaram a técnica VAR para avaliar a associação entre a concentração mensal de material particulado fino ($PM_{2,5}$) e as variáveis tráfego de veículos, velocidade do vento, temperatura, temperatura do solo e pontos de orvalho, na região de Los Angeles/Long Bech, Estados Unidos. Quanto à utilização de testes de causalidade, tem-se o trabalho de Sfetsos e Vlachogiannis (2013), que buscaram, por meio do teste de causalidade de Granger, verificar as relações de causa entre a concentração de ozônio e variáveis como temperatura e óxido de nitrogênio, na região de Atenas, Grécia.

Segundo Hsu (1992), o *smog* fotoquímico tornou-se um fenômeno comum em quase todas as grandes cidades, em diversas regiões do mundo. Reações químicas entre os poluentes primários, tais como hidrocarbonetos e óxidos de nitrogênio, produzem oxidantes como o

ozônio e o nitrato de peroxiacetila (PAN), que são responsáveis pela irritação dos olhos e pulmão em seres humanos e por danos aos animais e vegetações (Haagen-Smit, 1952). Conforme Liu *et al*. (1980), dependendo da estação do ano e da latitude, o tempo de vida do ozônio pode variar desde alguns dias até um mês.

Nesse contexto, vale dizer que a Região da Grande Vitória vem crescendo muito no decorrer dos últimos anos. Diversas indústrias e empresas se instalaram ou ampliaram suas instalações na região, o que tende, consequentemente, a elevar o nível de poluição atmosférica, mesmo existindo diversas regulamentações impostas pelos órgãos de controle ambiental à essas indústrias e empresas. Além disso, o crescimento da frota de veículos, o maior consumo de energia, e, etc., também contribuem para a maior emissão de poluentes na RGV. Assim, este trabalho objetivou analisar as inter-relações entre as concentrações de ozônio e de dióxido de nitrogênio, por meio da abordagem vetorial auto-regressiva, em complemento com o teste de causalidade de Granger, na RGV, Espírito Santo, Brasil, no período de janeiro a dezembro de 2010.

Ressalta-se que, mesmo não tendo ultrapassado os padrões primário e secundário (160 $\mu g/m^3$) estabelecidos pela Resolução CONAMA (CONAMA, 1990), no período de estudo, em diversos momentos as concentrações de ozônio e de dióxido de nitrogênio ultrapassaram o valor de 80 $\mu g/m^3$ e, até mesmo, o de 100 $\mu g/m^3$. Conforme a Companhia Ambiental do Estado de São Paulo (CETESB, 2013), esses níveis de concentração já podem gerar efeitos prejudiciais à saúde humana, principalmente para população mais sensível, como idosos e crianças. Logo, esta pesquisa torna-se importante no que diz respeito, especialmente, ao levantamento de dados que podem subsidiar a formulação de medidas preventivas por parte dos órgãos competentes, uma vez que a concentração de ozônio, na RGV, embora não tenha atingido níveis alarmantes, tem-se elevado nos últimos anos.

MATERIAIS E MÉTODOS

A área de estudo compreendeu a RGV, Espírito Santo, Brasil. Por estar situada na região litorânea, a RGV apresenta clima tropical quente (Aw), possuindo inverno ameno e seco, e verão chuvoso e quente. As temperaturas médias variam entre 24° C (Celsius) e 30° C, e os ventos predominantes de Norte/Nordeste na primavera – verão, sofrendo alterações durante outono e inverno devido ao posicionamento do sistema de alta pressão (Alta Pressão Subtropical do Atlântico Sul – ASAS) mais próximo do continente, possibilitando alterações na direção predominante do vento, a qual passa a variar entre as direções Sul/Oeste.

A RGV possui oito estações de monitoramento de qualidade do ar, a saber: Laranjeiras; Carapina; Jardim Camburi; Enseada do Suá; Vitória – Centro; Vila Velha – Ibes; Vila Velha – Centro; e, Cariacica (Figura 1). Essas estações fazem parte da Rede Automática de Monitoramento da Qualidade do Ar (RAMQAr), e pertencem ao Instituto Estadual de Meio Ambiente e Recursos Hídricos (IEMA). Vale lembrar que a RAMQAR monitora os seguintes poluentes: partículas totais em suspensão (PTS); PM_{10}; SO_2; CO; NO_x; hidrocarbonetos (HC); e, O_3. Existe, também, o monitoramento de alguns parâmetros meteorológicos, a saber: direção do vento (DV); velocidade do vento (VV); umidade relativa (UR); precipitação pluviométrica (PP); pressão atmosférica (P); temperatura (T); e, radiação solar (I). Um resumo dos poluentes e parâmetros meteorológicos que são medidos em cada estação pode ser visto na Tabela 1.

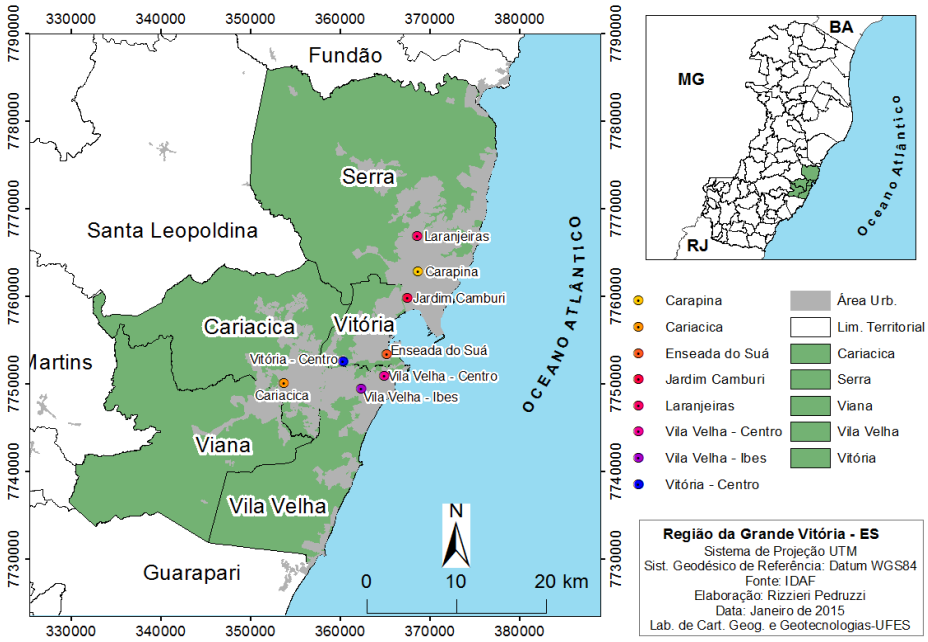


Figura 1 – Estações de monitoramento da qualidade do ar na RGV.

Fonte: Pedruzzi (2016).

Especificamente, em relação à este artigo, as análises estatísticas foram realizadas para o período de janeiro a dezembro de 2010, sendo os dados tomados de forma horária. As variáveis foram relativas às concentrações de ozônio ($\mu g/m^3$) e de dióxido de nitrogênio ($\mu g/m^3$) e, algumas variáveis meteorológicas (temperatura, umidade relativa e velocidade do vento), para dar maior robustez à identificação das inter-relações entre as concentrações de ozônio e de dióxido de nitrogênio.

Tabela 1 – Poluentes e parâmetros meteorológicos monitorados nas estações da RAMQAr

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

Fonte: elaborado com base nas informações do IEMA (2014).

Como pode ser verificado na Tabela 1, a concentração de ozônio é medida nas estações de Laranjeiras (LAR), Enseada do Suá (SUA), Vila Velha – Ibes (IBES) e Cariacica (CAR). Já a concentração de dióxido de nitrogênio é monitorada nas estações de Laranjeiras (LAR); Jardim Camburi (CAMB); Enseada do Suá (SUA); Vitória – Centro (VIX); Vila Velha – Ibes (IBES); e, Cariacica (CAR). No caso da temperatura e da umidade relativa, adotou-se como referência a estação de Carapina, que possui dados mais confiáveis desses parâmetros. Quanto à velocidade do vento, calculou-se uma média entre as estações de Carapina, Enseada do Suá, Vila Velha – Ibes e Cariacica.

Para atingir ao objetivo proposto, adotou-se a metodologia VAR (ver, Lütkepoh (2007) e Bueno (2011)), inicialmente proposta por Sims (1980), em complemento com o teste de causalidade de Granger (para detalhes, consultar Gujarati (2008)). O modelo VAR captura as interdependências lineares entre várias séries temporais, sendo que, cada variável possui uma equação estimada que representa sua evolução em termos de suas próprias defasagens e das defasagens das outras variáveis. O teste de causalidade de Granger objetiva verificar a relação de causalidade temporal existente entre duas variáveis. A adoção dessas metodologias justifica-se pelo fato de que existe correlação temporal cruzada entre as variáveis em estudo. Nesse caso, os modelos univariados (por exemplo, o modelo auto-regressivo (AR)) não são capazes de capturar essas correlações cruzadas, gerando resultados viesados (ou enganosos). Tal problema é corrido utilizando-se de metodologias multivariadas como o modelo VAR e o teste de causalidade de Granger. Pode-se expressar um modelo VAR de ordem p em função de um

vetor com n variáveis endógenas, X_t, sendo que essas se conectam por meio de uma matriz A, da seguinte forma (Equação 1):

$$AX_t = B_0 + \sum_{i=1}^{p} B_i X_{t-i} + B\varepsilon_t, \quad (1)$$

em que: A é uma matriz $n \times n$ que define as restrições contemporâneas entre as variáveis que constituem o vetor $n \times 1$, X_t; B_0, vetor de constantes $n \times 1$; B_i, matrizes $n \times n$, com $i = 0,1,2,...,p$, sendo p o número de defasagens (informação disponível no passado); B, matriz diagonal $n \times n$ de desvios-padrão; e, ε_t, vetor $n \times 1$ de perturbações aleatórias não correlacionadas entre si contemporânea ou temporalmente, isto é, $\varepsilon_t \sim i.i.d(0; I_n)$.

A Equação 1 expressa as relações entre as variáveis endógenas e é denominada de forma estrutural. No entanto, devido à endogeneidade das variáveis do VAR, o modelo é normalmente estimado em sua forma reduzida, dada pela Equação 2:

$$X_t = A^{-1}B_0 + \sum_{i=1}^{p} A^{-1}B_i X_{t-i} + A^{-1}B\varepsilon_t = \Phi_0 + \sum_{i=1}^{p} \Phi_i X_{t-i} + e_t, \quad (2)$$

em que: $\Phi_i = A^{1}B_i; i = 0,1,2,...,p$; e, $B\varepsilon_t = Ae_t$.

A metodologia VAR pode ser estimada por meio do método de Mínimos Quadrados Ordinários (MQO), levando-se em conta, principalmente, a interação entre as variáveis do sistema considerado. Dentre as suas principais vantagens na análise econométrica estão a obtenção das funções de impulso-resposta (FRI) e a decomposição da variância (DV). Em relação às funções de impulso-reposta, o interesse está na variação das variáveis em torno de suas médias. O que se quer, por exemplo, é verificar com um choque de um desvio-padrão na variável "X_1" afeta a variação da variável "X_2", nos períodos subsequentes ao choque. Esses períodos são mensurados na mesma escala de medida dos dados que estão sendo analisados. Para exemplificar, este estudo considerou dados horários. Assim, o choque inicial ocorre no tempo zero ($t = 0$). O tempo $t = 1$ representa uma hora após o choque e, assim, sucessivamente.

RESULTADOS E DISCUSSÕES

O primeiro passo na análise de séries temporais é verificar se as variáveis são estacionárias. Uma série temporal (processo estocástico) é considerada estacionária quando apresentar média, variância e covariância constantes ao longo do tempo. Em geral, os testes estatísticos verificam se essa pressuposição é válida. Se as séries não forem estacionárias em nível (sem transformações) deve-se realizar algum procedimento para estacionarizá-las (em geral, aplica-se a primeira diferença nas mesmas, dado que a maioria das séries é I(1), ou seja, integradas de primeira ordem). Os resultados dos testes *Augmented Dickey-Fuller* – ADF (Dickey e Fuller, 1981), *Phillips-Perron* – PP (Phillips e Perron, 1988) e, *Kwiatkowski-Phillips-Schmidt-Shin* – KPSS (Kwiatkowski *et. al.*, 1992) demonstraram que todas as séries utilizadas nesta pesquisa foram estacionárias em nível. Nos testes ADF e PP a hipótese nula equivale à existência de uma raiz unitária na série de dados. Já no teste KPSS, a hipótese nula refere-se à estacionariedade da série. Destaca-se que o teste KPSS é um teste assintótico, e que o mesmo deve ser utilizado em complemento aos demais testes de raiz unitária (Bueno, 2011). Também foram analisados os gráficos e os correlogramas (funções de autocorrelação) das séries.

Os critérios do Erro de Previsão Final (FPE), de Akaike (AIC), de Schwarz (SC) e de Hannan-Quinn (HQ) revelaram um modelo VAR ideal com 26 defasagens. Esse apresentou todas as raízes do polinômio dentro do círculo unitário, satisfazendo a condição de estabilidade do VAR. Os resultados foram satisfatórios para não autocorrelação e também para ausência de heteroscedasticidade. No que tange ao teste de normalidade dos resíduos, o teste de Jarque-Bera rejeitou a hipótese nula de que os resíduos são normais. Isso já era esperado, devido à assimetria das variáveis. Entretanto, assumiu-se, pela teoria assintótica sobre a média das distribuições de probabilidade, a hipótese de que os resíduos são normais, dando seguimento a análise do modelo.

Finalizada a etapa de identificação do modelo, foram analisadas as funções de impulso-resposta, em conjunto com os testes de causalidade de Granger, quando necessário. Ressalta-se que, antes de estimar as funções de impulso-resposta, é fundamental identificar o ordenamento de Cholesky do modelo VAR, um dos métodos mais populares para tal finalidade. Isto porque as funções de impulso-resposta são sensíveis à ordenação das variáveis.

Esta pesquisa adotou, como método de ordenação das variáveis, o teste de causalidade de Granger (*Block Exogeneity Wald Tests*), complementado pelo que se pode denominar de método de informação *a priori* [teorias de engenharia ambiental e meteorologia (Seinfeld e Pandis (2006), Holgate et al. (1999), artigos (Hsu (1992), Caio (1998), etc.), dentre outros]. A

ordenação adotada foi: TEMP, VELVENT, UMID, NO2LAR, NO2CAMB, O3SUA, O3LAR, NO2CAR, NO2SUA, O3CAR, NO2VIX, NO2IBES e O3IBES. Devido à importância da ordenação de Cholesky para a correta estimação das funções de impulso-resposta, testou-se outros ordenamentos (o que pode ser considerado um teste de robustez), que não alteraram significantemente tais funções.

Na Figura 2 são apresentadas as funções de impulso-resposta, considerando os efeitos de choques nas variáveis O3LAR, O3SUA, O3IBES, O3CAR, NO2LAR, NO2OCAMB, NO2SUA, NO2VIX, NO2IBES e NO2CAR, sobre as variáveis O3LAR e NO2LAR. As linhas contínuas equivalem às funções impulso-resposta e as linhas tracejadas equivalem à intervalos de confiança correspondentes a dois erros-padrão. Inicialmente, para exemplificar a análise de uma função de impulso-resposta, toma-se o caso do efeito da concentração de dióxido de nitrogênio em Laranjeiras (NO2LAR) sobre a concentração de ozônio em Laranjeiras (O3LAR), Figura 2e.

Em geral, os choques nas concentrações das outras regiões não afetaram ou afetaram muito pouco as concentrações de O3LAR e de NO2LAR, seja de forma positiva ou negativa, o que pode ser função da localização da estação de Laranjeiras, que fica ao Norte das demais regiões, quando se tem como ponto de partida a direção Norte/Nordeste (ver Figura 1). No mais, algumas relações apresentaram-se contrárias ao esperado, possivelmente devido ao grande número de variáveis no modelo estimado, ou por algum fator não captado pelo modelo ajustado. Por exemplo, os impactos positivos de NO2CAMB (Figura 2p), NO2SUA (Figura 2q), NO2VIX (Figura 2r) e NO2CAR (Figura 2t) em NO2LAR. Neste último caso, os resultados demonstraram que NO2CAR não causou Granger NO2LAR. Para as demais situações, é importante frisar que, apesar do vento Norte/Nordeste ser predominante na região da Grande Vitória, com a chegada de Frentes Frias, por exemplo, ocorrem alterações na direção do vento para Sul, Sul/Sudeste e Sul/Sudoeste, o que, dependendo da intensidade, poderia explicar os pequenos efeitos positivos de NO2CAMB, NO2SUA e NO2VIX em NO2LAR.

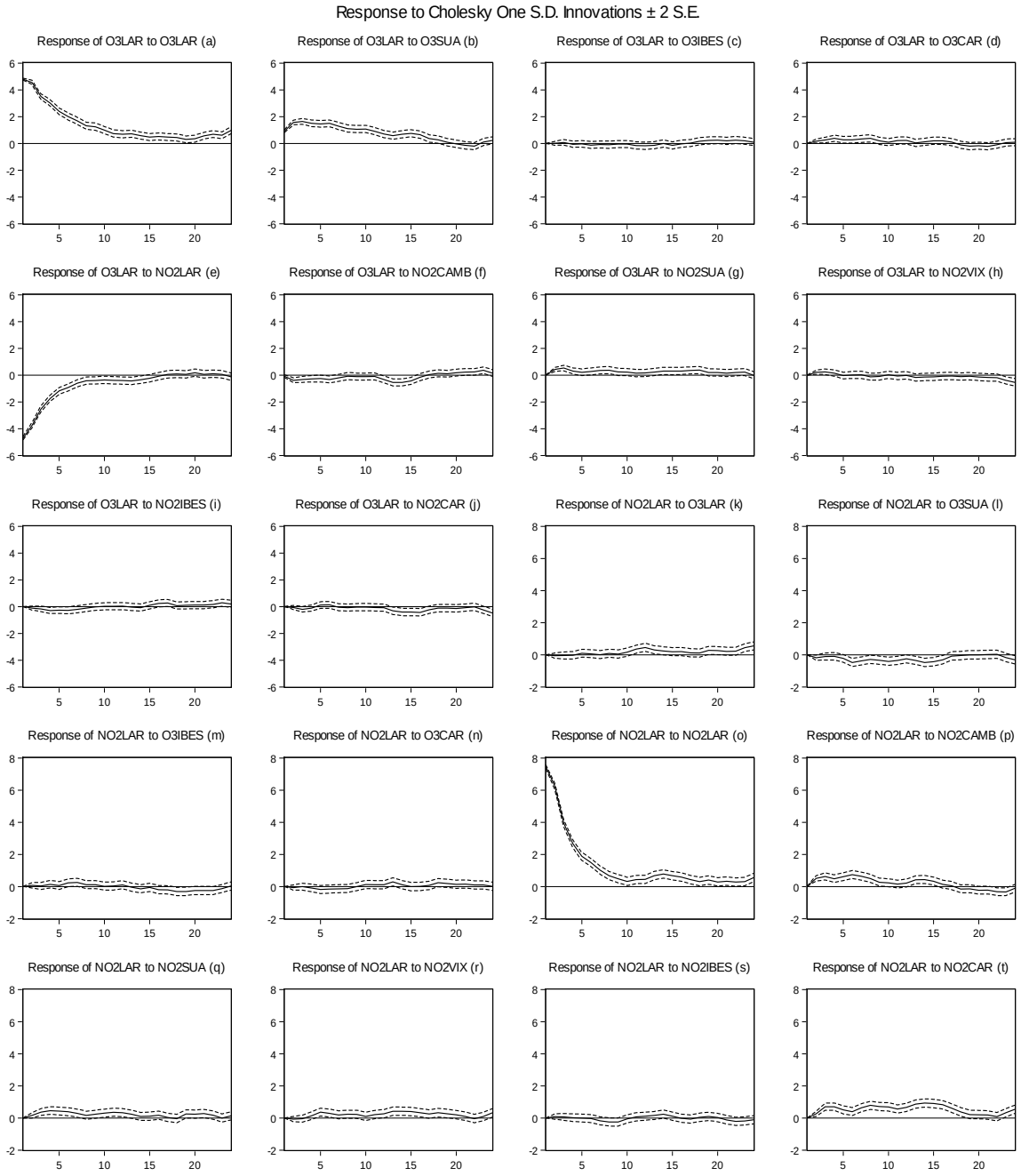


Figura 2 – Funções de impulso-resposta para as concentrações de ozônio e de dióxido de nitrogênio de Laranjeiras.

Nota: as linhas contínuas equivalem às funções impulso-resposta e as linhas tracejadas equivalem à intervalos de confiança correspondentes a dois erros-padrão.

Os resultados da Figura 3 são relativos às funções de impulso-resposta, quando considerados os impactos de choques nas variáveis O3LAR, O3SUA, O3IBES, O3CAR, NO2LAR, NO2OCAMB, NO2SUA, NO2VIX, NO2IBES e NO2CAR, sobre as variáveis

O3SUA e NO2SUA. No que se refere aos efeitos de O3SUA em NO2SUA (Figura 3l), nota-se reduções de NO2SUA até a sexta hora, quando para os períodos seguintes a tendência é de estabilidade. Conforme observado, a variável NO2SUA não apresentou efeitos sobre O3SUA na primeira hora após o choque (Figura 3g). A partir da primeira hora ocorreram pequenas elevações na concentração de NO2SUA. O teste de causalidade de Granger demonstrou uma causalidade bidirecional entre O3SUA e NO2SUA. Isto revela que pode haver produção e destruição local do ozônio na estação da Enseada do Suá, devido, principalmente, às reações fotoquímicas. Importante destacar que a formação do ozônio na troposfera se dá pela reação de fotólise do dióxido de nitrogênio. Entretanto, também ocorrem reações em que o óxido nítrico reage com o ozônio dando origem ao dióxido de nitrogênio (Orlando, 2008).

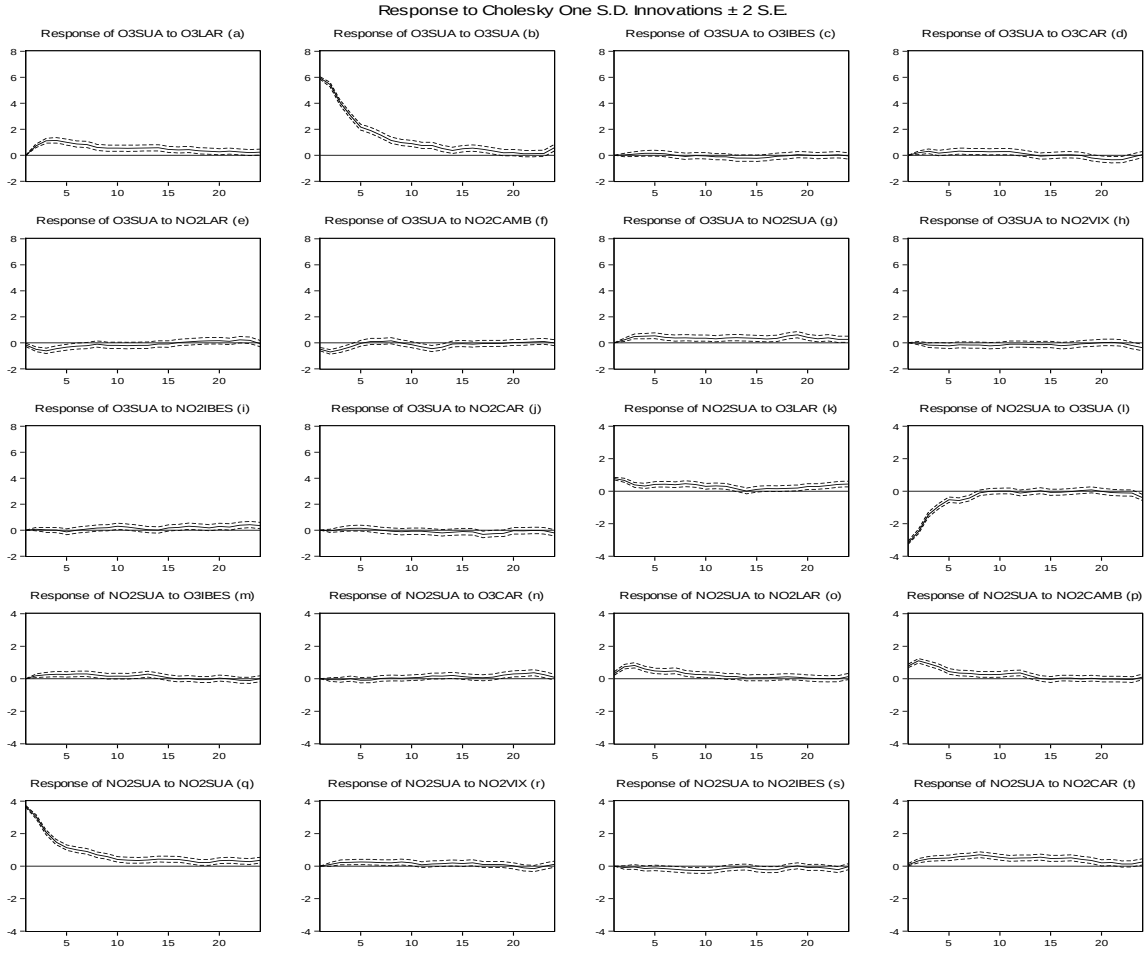


Figura 3 – Funções de impulso-resposta para as concentrações de ozônio e de dióxido de nitrogênio da Enseada do Suá.

Nota: as linhas contínuas equivalem às funções impulso-resposta e as linhas tracejadas equivalem à intervalos de confiança correspondentes a dois erros-padrão.

Uma vez que, a região da Enseada do Suá está localizada em um local que pode ser considerado como central às outras regiões (ver Figura 1), as concentrações de ozônio e de dióxido de nitrogênio de outras estações afetaram consideravelmente as concentrações da Enseada do Suá. Destaca-se que tal região apresenta grande fluxo de veículos, principalmente as 6 horas e as 18 horas. Pode-se observar, por exemplo, que NO2CAMB teve efeitos positivos sobre NO2SUA nas primeiras quatro horas após o choque inicial (Figura 3p). Isso mostra que existe um deslocamento regional de NO2 entre estas estações que pode ser resultado da direção predominante do vento na Grande Vitória (Norte/Nordeste na maioria dos meses).

Ainda, algumas variáveis não causaram efeitos sobre O3SUA e NO2SUA e certas relações podem ser consideradas como contrárias ao esperado, como a existente entre O3SUA e O3LAR (Figura 3a). Neste particular, o teste de causalidade de Granger revelou uma causalidade entre as duas variáveis, que pode ser advinda de uma falha do modelo estimado, ou de recorrentes mudanças na direção padrão do vento na região da Grande Vitória. Não é objetivo deste trabalho analisar esses efeitos, o que pode ser aprofundado em estudos futuros. Outro resultado não esperado foi o choque positivo de NO2CAR em NO2SUA (Figura 3t). Porém, o teste de causalidade demonstrou que NO2CAR não causou Granger NO2SUA.

Em relação à Figura 4, essa demonstra as funções de impulso-resposta, analisando-se os efeitos de choques nas variáveis O3LAR, O3SUA, O3IBES, O3CAR, NO2LAR, NO2OCAMB, NO2SUA, NO2VIX, NO2IBES e NO2CAR, sobre as variáveis O3IBES e NO2IBES. Devido à proximidade da estação da Enseada do Suá e à direção do vento em certas horas do dia, nota-se que O3SUA teve grande impacto sobre O3IBES (Figura 4b), com pico na terceira hora, revelando a ocorrência de transporte de ozônio entre as regiões. Verifica-se, também, que choques em NO2IBES reduziram a concentração de O3IBES (Figura 4i), principalmente até a quinta hora, demonstrando uma destruição local do ozônio quando da formação de NO2IBES.

Ainda, em relação ao transporte de poluentes entre as regiões, observa um impacto negativo relevante de O3SUA sobre NO2IBES (Figura 4l), especialmente na segunda e terceira hora após o choque inicial em O3SUA. Isto indica que a fotólise de NO2IBES pode ter contribuído para formação de O3SUA. No mais, NO2CAMB (Figura 4p), NO2SUA (Figura 4q) e NO2VIX (Figura 4r) contribuíram para formação da concentração de NO2IBES. Como a concentração de ozônio e de dióxido de nitrogênio de outras estações afetou a concentração de O_3 e NO_2 da estação do Ibes, a formação de ozônio na região da estação Ibes não necessariamente reduz a concentração de NO_2 no local.

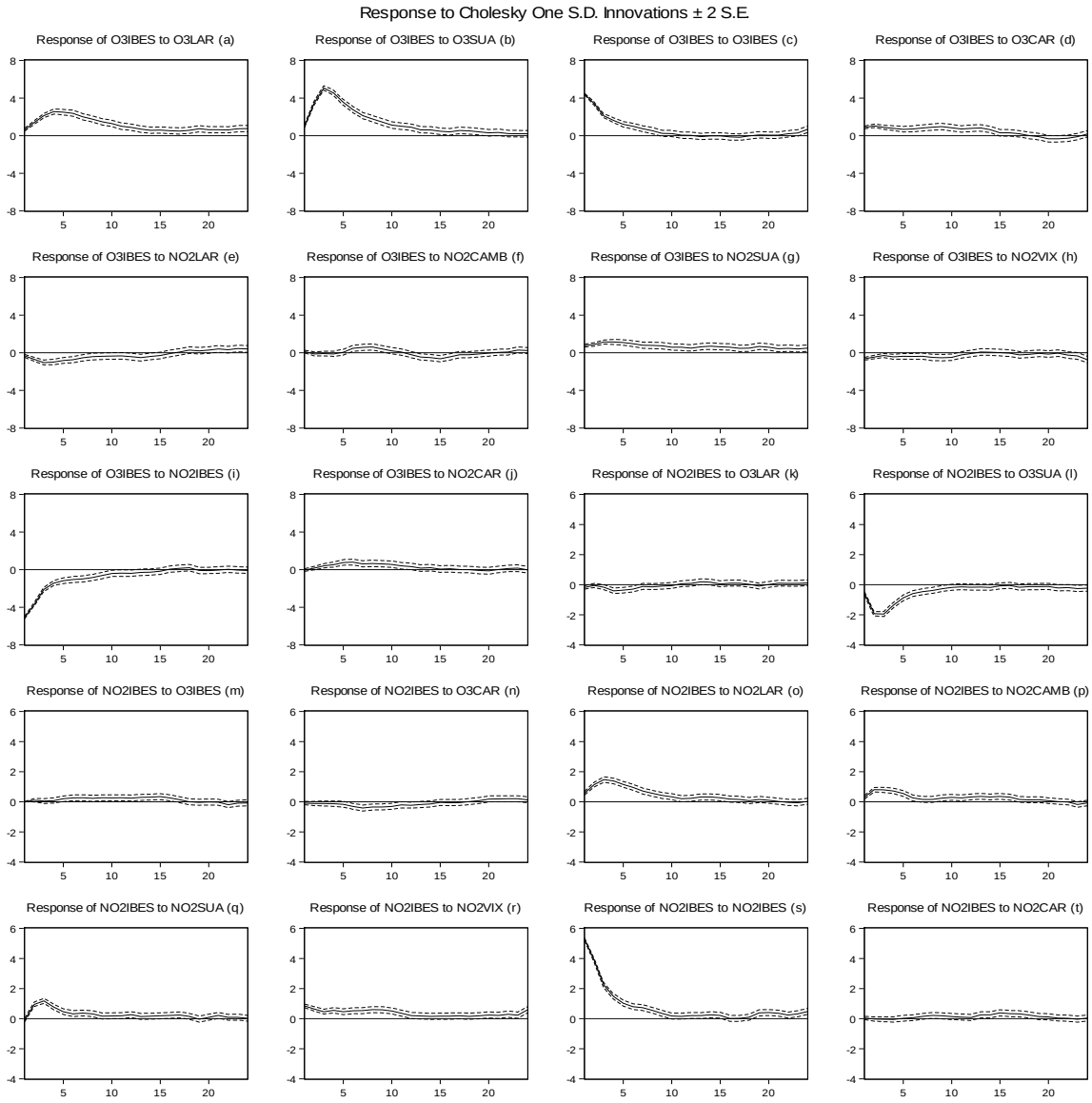


Figura 4 – Funções de impulso-resposta para as concentrações de ozônio e de dióxido de nitrogênio do Ibes.

Nota: as linhas contínuas equivalem às funções impulso-resposta e as linhas tracejadas equivalem à intervalos de confiança correspondentes a dois erros-padrão.

Algumas relações tiveram efeitos insignificantes e, mais uma vez, alguns resultados não eram esperados. Por exemplo: i) o impacto positivo de O3CAR em O3IBES (Figura 4d). O teste de causalidade de Granger revelou que O3CAR não causou O3IBES; ii) o efeito negativo (na maior parte do período de 24 horas) de NO2LAR em O3IBES (Figura 4e). O teste de causalidade foi realizado e constatou-se que NO2LAR não causou Granger O3IBES; e, iii)

efeitos positivos de NO2LAR em NO2IBES (Figura 4o). Verificou-se não causalidade de Granger de NO2LAR para NO2IBES.

Por fim, na Figura 5 são apresentados os resultados referentes às funções de impulso-resposta, levando-se em conta os impactos dos choques em O3LAR, O3SUA, O3IBES, O3CAR, NO2LAR, NO2OCAMB, NO2SUA, NO2VIX, NO2IBES e NO2CAR, sobre O3CAR e NO2CAR. Vale ressaltar que a região de Cariacica se localiza relativamente distante, ficando a Sudoeste das demais regiões, quando se tem como ponto de partida a direção Norte/Nordeste (ver Figura 1), sendo esta a direção predominante do vento na maioria dos meses na Região da Grande Vitória. Um primeiro ponto a destacar é que tanto O3SUA (Figura 5b) quanto NO2SUA (Figura 5g) afetaram positivamente as concentrações de ozônio (O3CAR) em Cariacica, o que parece bastante plausível, dada a direção do predominante do vento na RGV.

Destaca-se que o aumento da concentração de ozônio na estação da Enseada do Suá causou redução na concentração de dióxido de nitrogênio em Cariacica (Figura 5l). Isso significa que, com as reações fotoquímicas produzindo mais ozônio na Enseada do Suá, uma menor concentração de dióxido de nitrogênio da Enseada do Suá foi transportada pelo vento para a região de Cariacica. Além disso, pode-se observar que o aumento de NO2SUA provocou uma elevação da concentração de NO_2 em Cariacica (Figura 5q). Logo, as concentrações de ozônio e de dióxido de nitrogênio da Enseada do Suá parecem ter papel fundamental na formação de O_3 e NO_2, em Cariacica.

No mais, em função da direção do vento, constata-se, também, que NO2CAMB (Figura 5p) e NO2VIX (Figura 5r) tiveram efeitos positivos sobre o NO2CAR. Além disso, é possível notar que o crescimento da concentração de NO2CAR causou efeitos negativos em O3CAR até a quinta hora após o choque inicial em NO2CAR (Figura 5j). Por fim, algumas inter-relações mostraram-se insignificantes e poucos resultados foram incoerentes, como por exemplo, O3LAR afetando NO2CAR (Figura 5k), o que não foi rejeitado pelo teste de causalidade Granger.

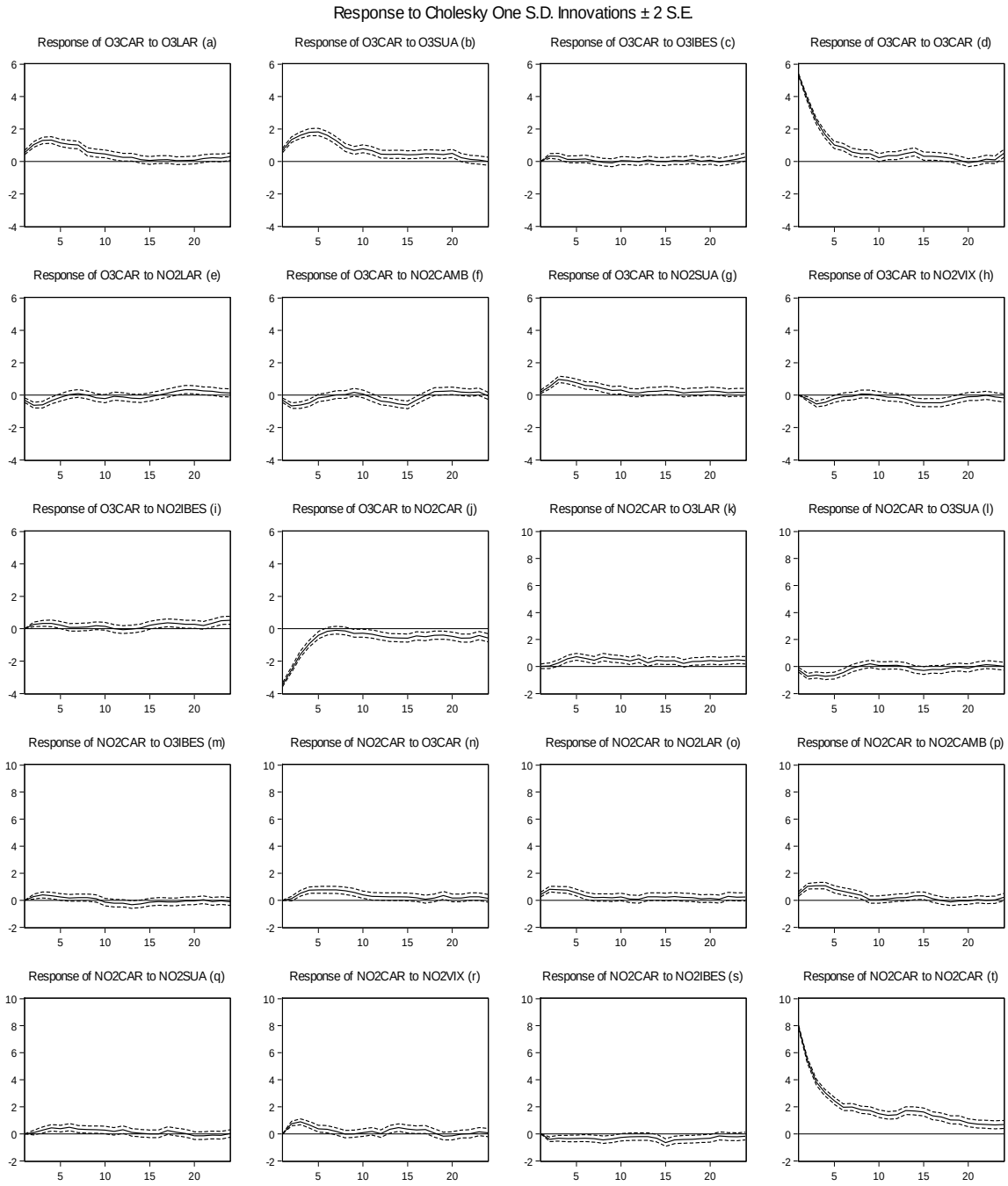


Figura 5 – Funções de impulso-resposta para as concentrações de ozônio e de dióxido de nitrogênio de Cariacica.

Nota: as linhas contínuas equivalem às funções impulso-resposta e as linhas tracejadas equivalem à intervalos de confiança correspondentes a dois erros-padrão.

CONCLUSÕES

Mesmo sendo um poluente de difícil modelagem estatística, em função da sua formação secundária, o modelo VAR estimado apresentou todas as raízes do polinômio dentro do círculo unitário, satisfazendo a condição de estabilidade do VAR, revelou resíduos não autocorrelacionados e ausência de heteroscedasticidade. Os resíduos não seguiram distribuição normal. Além disso, apesar de apresentar algumas inconsistências em termos de relações entre as variáveis, o modelo estimado conseguiu captar de forma satisfatória as inter-relações entre as concentrações de ozônio e dióxido de nitrogênio das diversas estações de monitoramento consideradas no trabalho.

As concentrações de ozônio e dióxido de nitrogênio da estação (região) de Laranjeiras foram as que receberam menos influência das concentrações de outras regiões e as que menos influenciaram as concentrações de outras regiões. Isto pode ser decorrência da localização da estação de Laranjeiras, que fica à montante e ao Norte das demais regiões, quando se tem como ponto de partida a direção Norte/Nordeste (ver Figura 1), e da direção predominante do vento na Região da Grande Vitória, que Norte/Nordeste boa parte do ano. No entanto, vale ressaltar que parece existir pequenos efeitos de NO2CAMB, NO2SUA e NO2VIX em NO2LAR. Nesse caso, isso pode reflexo das variações na direção do vento que ocorrem na Região da Grande Vitória ao longo do ano, especialmente devido ao posicionamento do sistema de alta pressão (ASAS) mais próximo ao continente nos meses de outono inverno, alterando-se de Nordeste para Sul, Sul/Oeste e Sul/Sudoeste. Entretanto, análises mais profundas são necessárias, como por exemplo, a comparação de modelos VAR para diferentes períodos do ano, como, por exemplo, aqueles em que a ASAS se encontra sobre o oceano Atlântico favorecendo os ventos de Norte/Nordeste e quando esse sistema se desloca para o continente alterando as direções predominantes para Oeste/Sul.

Pela localização mais centralizada da estação da Enseada do Suá em relação às outras estações de monitoramento, as concentrações de ozônio e dióxido de nitrogênio da mesma tiveram significativa influência de outras regiões (estações), especialmente, das estações de Jardim Camburi, Ibes e Vitória – Centro. Pela predominância do vento ser Norte/Nordeste, há forte transporte de NO_2 de Jardim Camburi para a região da Enseada do Suá. Ressalta-se que a região da Enseada do Suá apresenta grande fluxo de veículos durante alguns horários do dia. Para algumas inter-relações que *a priori* são incoerentes com o esperado, caberiam análises mais pontuais, o que não é objetivo deste estudo.

No que se refere à região da estação Ibes, os resultados revelaram que a concentração de ozônio nessa região foi fortemente influenciada pelas concentrações de ozônio e de dióxido de nitrogênio da estação Enseada do Suá. Observou-se, também, que a concentração de NO_2 na estação do Ibes foi significantemente afetada pela concentração de NO_2 das estações de Jardim Camburi e Enseada do Suá. Como são regiões próximas, a tendência é ocorrer um grande transporte de poluentes entre as mesmas, o que na prática foi cofirmado.

Outro ponto importante revelado pelos resultados foi a forte influência das concentrações de ozônio e de dióxido de nitrogênio da estação Enseada do Suá sobre as concentrações destes poluentes na região de Cariacica, o que comprova que a predominância do vento Norte/Nordeste na Região Grande Vitória contribui significantemente para o transporte dos mesmos. Até mesmo o NO_2 de Camburi contribuiu para a concentração de NO_2 em Cariacica. Mais uma vez, alguns resultados não esperados requerem maior atenção para trabalhos futuros.

Por fim, considera-se que esta pesquisa atingiu ao objetivo proposto, ao demonstrar que existe uma inter-relação entre as concentrações de ozônio e de dióxido de nitrogênio nas estações de qualidade do ar da Região da Grande Vitória. Vale mencionar, ainda, que este é um estudo preliminar. Para trabalhos futuros outras técnicas estatísticas podem ser utilizadas para aprimorar os resultados encontrados nesse trabalho, como os modelos multivariados com volatilidade estocástica, os modelos multivariados robustos e considerar a modelagem de memória longa (longa dependência) das séries estudadas.

REFERÊNCIAS

BUENO, R. D. L. S. *Econometria de séries temporais*. 2 ed. São Paulo: Cengage Learning, 2011, 338 p.

CAI, X. H. *Time Series Analysis of Air Pollution CO in California South Coast Area, with Seasonal ARIMA model and VAR model*. Los Angeles, California. 2008. 46 f. Thesis (Master of Science in Statistics). University of California, Los Angeles.

CARVALHO, V. S. B. *Meteorologia da qualidade do ar no que tange as concentrações de ozônio e dos óxidos de nitrogênio na região Metropolitana do Rio de Janeiro*. Rio de janeiro, RJ. 2006. 134 f. Dissertação (Mestrado em Engenharia Mecânica). Programa de Pós-Graduação em Engenharia Mecânica, Universidade Federal do Rio de Janeiro, RJ.

CETESB. *Relatório da qualidade do ar do estado de São Paulo 2012*. São Paulo: CETESB, 2013.

CONSELHO NACIONAL DE MEIO AMBIENTE – CONAMA (Brasil). Resolução nº 08, de 6 de dezembro de 1990. *Diário Oficial [da] República Federativa do Brasil*, Brasília, 28 dez. 1990. Seção 1, p. 25539.

DICKEY, D. A.; FULLER, W. A. Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica*, v. 49, n. 4, p. 1057-1073, 1981.

ECOSOFT CONSULTORIA E SOFTWARES AMBIENTAIS. *Inventário de emissões atmosféricas da Região da Grande Vitória*. Vitória, 2011. Disponível em: <http://www.es.gov.br/Banco%20de%20Documentos/PDF/Maio/100511/RTC10131-R1.pdf>. Acesso em: 20 de mar. de 2014.

GUJARATI, D. N.; PORTER, D. C. *Basic Econometrics*. 5 ed. New York: McGraw-Hill/Irwin, 2008. 944 p.

HAAGEN-SMIT A. J. Chemistry and physiology of Los Angeles smog. *Industrial & Engineering Chemistry*, v. 44, n. 2, p. 1342-1346, 1992.

HOLGATE, S. T.; SAMET, J. M.; KOREN, H. S.; MAYNARD, R. L. *Air Pollution and Health*. London: Academic Press, 1999, 1065 p.

HSU, K, J. Time series analysis of the interdependence among air pollutants. *Atmospheric Environment*, v. 26B, n. 4, p. 491-503, 1992.

INSTITUTO ESTADUAL DO MEIO AMBIENTE E RECURSOS HÍDRICOS (IEMA). *Relatório da qualidade do ar da Região da Grande Vitória: 2013*. Vitória, 2014. Disponível em: <http://www.meioambiente.es.gov.br>. Acesso em: 27 de jun. de 2015.

IEMA. *Rede automática de monitoramento da qualidade do ar da região da Grande Vitória (RAMQAR)*. 2014. Disponível em: <http://www.meioambiente.es.gov.br>. Acessado em: 20 de mar. de 2014.

JORQUERA, H.; PÉREZ, R.; CIPRIANO, A.; ESPEJO, A.; LETELIER, M. V.; ACUÑA, G. Forecasting ozone daily maximum levels at Santiago, Chile. *Atmospheric Environment*. v. 32, n. 20, p. 3425-3424, 1998.

KWIATKOWSKI, D.; PHILLIPS, P. C. B.; SCHMIDT, P.; SHIN, Y. Testing the null hypothesis of stationarity against the alternative of unit root. *Journal of Econometrics*, v. 54, n. 1, p. 159-178, 1992.

LIU, P. W. G.; JOHNSON, R. Forecasting peak daily ozone levels-I. A regression with time series errors model having a principal component trigger to fit 1991 ozone levels. *Journal of the Air & Waste Management Association*. v. 52, n. 9, p1064-1074, 2002.

LIU, P. W. G.; JOHNSON, R. Forecasting peak daily ozone levels: part 2. A regression with time series errors model having a principal component trigger to forecast 1999 and 2002 ozone levels. *Journal of the Air & Waste Management Association*. v. 53, n. 12, p. 1472-1489, 2003.

LIU, P. W. G.; TSAI, J. H.; LAI, H. C.; TSAI, D. M.; LI, L. W. Establishing multiple regression models for ozone sensitivity analysis to temperature variation in Taiwan. *Atmospheric Environment*, v. 79, p. 225-235, 2013.

LIU, S. C.; KLEY, D.; MCFARLAND, M.; MAHLMAN, J. D.; LEVY, H. On the origin of tropospheric ozone. *Journal of Geophysical Research*, v. 86, p. 7546-7552, 1980.

MOREIRA, D. M.; TIRABASSI, T.; MORAES, M. R. Meteorologia e poluição atmosférica. *Ambiente & Sociedade*. v. 11, n. 1, p. 1-13, 2008.

LÜTKEPOH, H. *New introduction to multiple time series analysis*. New York, Springer, 2007. 764 p.

ORLANDO, J. P. *Estudo dos precursores de ozônio da cidade de São Paulo através de simulação computacional*. São Paulo, SP. 116 f. Dissertação (Mestrado em Tecnologia Nuclear). Programa de Pós-Graduação em Tecnologia Nuclear. Instituto de Pesquisas Energéticas e Nucleares, SP.

PEDRUZZI, R. *Influência das condições de contorno nas simulações do modelo CMAQ para Região Metropolitana da Grande Vitória – ES*. Vitória, ES. 110 f. Dissertação (Mestrado em Engenharia Ambiental). Programa de Pós-Graduação em Engenharia Ambiental. Universidade Federal do Espírito Santo, ES.

PHILLIPS, P. C. B.; PERRON, P. Testing for unit roots in time series regression. *Biometrika*, v. 75, n. 3, p. 335-346, 1988.

RYAN, W. F. Forecasting severe ozone episodes in the Baltimore metropolitan area. *Atmospheric Environment*, v. 29, n. 17, p. 2387-2398, 1995.

RYAN, W. F.; PIETY, C. A; LUEBEHUSEN, E. D. Air quality forecasts in the Mid-Atlantic Region: current practice and benchmark skill. *Weather and Forecasting*, v. 15, n. 1, p. 46-60, 1999.

SEINFELD, J. H.; PANDIS, S. N. *Atmospheric chemistry and physics: from air pollution to climate change*. J. Wiley, New York, 2006.

SFETSOS, A.; SVLACHOGIANNIS, D. An analysis of ozone variation in the Greater Athens area using Granger Causality. *Atmospheric Pollution Research*. v. 4, n. 3, p. 290-297, 2013.

SIMS, C. "Macroeconomics and reality". *Econometrica*, v. 48, n. 1, p. 1-48, 1980.

WANG, W.; NIU, Z. VAR model of $PM_{2.5}$, weather and traffic in Los Angeles Long Beach. In: INTERNATIONAL CONFERENCE ON PRINT, 2009, China. *Anais eletrônicos...* Disponível em: <http://ieeexplore.ieee.org>. Acesso em: 20 de maio 2014.

Robust factor modeling for high-dimensional time series: An application to air pollution data

Adriano Marcio Sgrancio[a], Valdério Anselmo Reisen[a,b,*], Flávio Augusto Ziegelmann[c], Edson Zambon Monte[a], Higor Henrique Aranda Cotta[a], C. Lévy-Leduc[d]

[a] *Department of Statistics, Federal University of Espírito Santo, Espírito Santo, Brazil*
[b] *Graduate Program in Environmental Engineering, Federal University of Espírito Santo, Espŕito Santo, Brazil*
[c] *Department of Statistics, Ppge and Ppga, Federal University of Rio Grande do Sul, Rio Grande do Sul, Brazil*
[d] *AgroParisTech/UMR INRA MIA 518, France*

Abstract

This paper considers the factor modeling for high-dimensional time series in the presence of additive outliers by proposing a robust variant of the estimation methods given in Lam & Yao (2012). The estimator of the number of factors is obtained by an eigenanalysis of a robust non-negative definite covariance matrix. Some asymptotic properties of the robust eigenvalues are derived, in particular, it is shown that the eigenvalues of the robust estimator covariance matrix have the same rate of convergence as the eigenvalues of the standard covariance matrix estimator. Simulations are conducted to analyse the finite sample size performances of the robust estimator of the number of factors under the scenarios of contaminated and non-contaminated multivariate time series. As an example of application, the robust factor analysis is performed to identify pollution behavior for the pollutant PM_{10} of the Region of Greater Vitória, Brazil, aiming to reduce the dimensionality of the data and to produce good forecasts for the PM_{10} pollution levels.

Keywords:
Factor analysis; Time series; Robustness; Eigenvalues; Reduced rank; Air pollution.

In the last 50 years, issues related to air pollution problems have increased considerably, especially in developing countries, where the air quality has been degraded as a result of industrialization, population growth, high rates of urbanization, and inadequate or non-existent policies to control air pollution, among other reasons. The problems caused by air pollution produce local, regional and global impacts. Among different environmental problems, air pollution is reported to cause the greatest damage to health and loss of quality of life. The most common human health problems caused by air pollution are asthma, rhinitis, burning eyes , fatigue, dry cough, heart and lung diseases and heart failure. The works by Brunekreef & Holgate (2002), Maynard (2004), WHO (2005), Curtis et al. (2006) showed the relationship between the legislated pollutants (inhalable particles with smaller diameter than 10 micrometers (PM_{10}),

*Corresponding author. Department of Statistics, Federal University of Espírito Santo, 29075-910, 514, Vitoria, ES, Brazil. Tel.: +5502740092903.
E-mail address: valderioanselmoreisen@gmail.com (V. A. Reisen).

carbon monoxide (CO), sulfur dioxide (SO_2), nitrogen oxides (NO_x) and ozone (O_3)) and health problems. In 2012, for example, the death of 4.3 million people have been attributed to air pollution (WHO, 2014). In addition, air pollution contributes to the degradation of the environment, the greenhouse effect, among many others problems.

In the recent studies related to air pollution much attention has been paid to the mathematical receptor models, which attempt to measure and analyse concentrations at their sources from a given site without reconstructing the dispersion patterns of the pollutants, such as particular matter (PM), SO_2, among others. These methodologies have mathematical and statistical tools which are mainly used to provide the identification of the pollutant emission sources from chemical characteristics of the particles on the receiver and the pollutant emission sources (Seinfeld & Pandis, 2006). In the literature, the majority of receptor models studied are as following: chemical balance of mass (CBM), multivariate analysis, principal component analysis techniques (PCA), factor analysis model (FA), multiple linear regression, cluster analysis, factoring positive matrix (FPM), among others (Watson et al., 2002). Regarding the classical factor analysis, this technique has been widely used in the area of air pollution, especially for the identification of emission sources, the management of monitoring networks, regression analysis, cluster analysis and prediction, among others.

Many time series arising in practice are best considered as components of multivariate time series models, which accommodate the serial dependence of each component and, also, the interdependence between different components. However, it should be noted that, among the studies that adopted the classical PCA and techniques of factor analysis, especially in the area of air pollution, the time-dependence of the data is a common feature neglected, since the standard assumption of these multivariate statistical tools is the independence of the data (see, for example, Anderson (2003) and Johnson & Wichern (2007)). To deal with this problem, Peña & Box (1987), Stock & Watson (2002), Lam et al. (2011) and Lam & Yao (2012) studied the factor modeling for multivariate time series from a dimension-reduction point of view. Differently from the PCA and factor analysis for independent observations, these papers look for factors which drive the serial dependence of the original time series. Further discussions and additional references in this direction can be found in Lam & Yao (2012).

Apart from the purpose of dimension reduction, factor analysis has been widely used with the aim of forecasting in the sense that this technique can drastically reduce the order of the estimated model. According to Stock & Watson (2002), dimension reduction can be a central concern in forecasting investigation when the number of candidate predictor series (say, k) is very large. This issue can make impractical the forecast investigation, for example, in the use of vector autoregressive moving average (VARMA) models with a large number of parameters. This high-dimensional problem is simplified by modeling the common dynamics in terms of a relatively small number of unobserved latent factors. Forecasting can then be carried out in a two-step process: first, a time series of the factors is estimated from the predictors; second, the relationship between the variable to be forecast and the factors is estimated by a linear regression, for example.

Environmental time series are often of a high dimension due to the large number of measurements recorded across many different locations. These data may also present interesting phenomena to be considered from applied and theoretical statistical points of view, for example, pollution data may have observations which can be defined as outliers.

As well known (see, for example, Chang et al. (1988), Tsay (1988), Chen & Liu (1993) and the references therein), outliers can destroy the statistical properties of sample functions such as the standard mean and covariance. Since the estimation of time series models is connected with these sample functions, the final estimated model can be strongly affected by large peaks of concentrations. One way to deal with model estimation in case the series has additive outliers is to use the robust ACF function based on the robust scale function $Q_n(.)$ proposed by Rousseeuw & Croux (1993a) and studied recently by Lévy-Leduc et al. (2011a), Lévy-Leduc et al. (2011b) and Lévy-Leduc et al. (2011c).

This paper considers most of the above issues using factor analysis for dimension reduction and forecasting PM_{10} concentrations. In this context, it is proposed here a robust version of the dimension reduction estimator given in Lam & Yao (2012). In this direction, this paper makes the use of the robust scale estimator $Q_n(.)$, proposed by Rousseeuw & Croux (1993a), to identify the number of the factors of multivariate time series under additive outliers. Some theoretical results are discussed and the method performance is investigated through Monte Carlo simulations. The proposed methodology is applied to PM_{10} series measured at the Air Quality Automatic Monitoring Network (AQAMN), Region of Greater Vitória (RGV), Brazil.

This paper is divided as follows. In Section 1, the model and the estimation methods are presented. Section 2 gives asymptotic properties of the robust eigenvalues to support theoretically the robust approach proposed here. Section 3 presents some Monte Carlo experiments. In Section 4 the data obtained from AQAMN stations is modeled and forecasts are computed and compared with a vector AR (VAR) model. Some concluding remarks are provided in Section 5.

1. Factor model in time series

1.1. The factor model

Let $\boldsymbol{Z}_t$ be a k-dimensional zero-mean vector of an observed time series. Let also $\boldsymbol{X}_t$ be an unobserved r-dimensional vector of common factors ($r \leq k$). It is assumed that $\boldsymbol{Z}_t$ is generated by

$$\boldsymbol{Z}_t = \boldsymbol{P}\boldsymbol{X}_t + \boldsymbol{\varepsilon}_t, \tag{1}$$

where $\boldsymbol{P}$ is an unknown $k \times r$ matrix of parameters of rank r, denoted by the factor-loading matrix, and $\boldsymbol{\varepsilon}_t$ is a k-dimensional white-noise sequence with full-rank covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$. When r is small relative to k, the model presented in Equation (1) is most useful, since it will result in a multivariate time series with a reduced dimension and, consequently, can lead to a much simpler multivariate time series for forecasting.

In the sequel, it is made the following assumption

(**A1**) $\boldsymbol{X}_t$ is a multivariate stationary process. Moreover, $\boldsymbol{X}_t$ and $\boldsymbol{\varepsilon}_t$ are assumed to be independent processes and $\boldsymbol{P}'\boldsymbol{P} = \boldsymbol{I}_r$, where $\boldsymbol{I}_r$ denotes the $r \times r$ identity matrix.

Note that the assumption above is here to ensure identifiability in Equation (1); see Lam & Yao (2012) and Peña & Box (1987) for further details.

It follows from Equation (1) and under Assumption A1 that the covariance matrices of $\boldsymbol{Z}_t$ are given by

$$\boldsymbol{\Gamma}_Z(0) = \boldsymbol{P}\boldsymbol{\Gamma}_X(0)\boldsymbol{P}' + \boldsymbol{\Sigma}_\varepsilon, \tag{2}$$

$$\boldsymbol{\Gamma}_Z(h) = \boldsymbol{P}\boldsymbol{\Gamma}_X(h)\boldsymbol{P}', \quad h \geq 1, \tag{3}$$

where $\boldsymbol{\Gamma}_X(h) = E[\boldsymbol{X}_{t-h}\boldsymbol{X}_t']$ is the covariance matrix of $\boldsymbol{X}_t$.

The key to the inference for the model in Equation (1) is to determine the number of factors r and to estimate the $k \times r$ factor loading matrix $\boldsymbol{P}$. Once an estimator is obtained, say, $\widehat{\boldsymbol{P}}$, a natural estimator for the factor process is

$$\widehat{\boldsymbol{X}}_t = \widehat{\boldsymbol{P}}'\boldsymbol{Z}_t. \tag{4}$$

For further details on the estimation of $\boldsymbol{P}$, see Lam & Yao (2012).

Following the same lines as in Lam & Yao (2012), it is defined the following estimator for the number of factors r as follows:

$$\widehat{r} = \underset{1 \leq i \leq R}{\operatorname{argmin}}\, \widehat{\lambda}_{i+1}/\widehat{\lambda}_i, \tag{5}$$

where $r < R < k$ is a constant, $\widehat{\lambda}_1 \geq \ldots \geq \widehat{\lambda}_k$ are the eigenvalues of $\widehat{\boldsymbol{M}}$ defined by

$$\widehat{\boldsymbol{M}} = \sum_{h=1}^{h_0} \widehat{\boldsymbol{\Gamma}}_Z(h)\widehat{\boldsymbol{\Gamma}}_Z(h)', \tag{6}$$

where $\widehat{\boldsymbol{\Gamma}}_Z(h)$ denotes the sample covariance matrix of $\boldsymbol{Z}_t$ at lag h. Lam & Yao (2012) derive the asymptotical properties of the above results.

In this context, the aim of this paper is to propose robust estimators of $\boldsymbol{M}$ and r against additive outliers which are based on a robust covariance matrix estimator for $\boldsymbol{Z}_t$ and these are discussed in the following sections.

1.1.1. Robust estimator of M ($\hat{\boldsymbol{M}}_Q$)

Let $(Y_i)_{i\geq 1}$ be a stationary Gaussian process. Given the observations $Y_{1:n} = (Y_1, \ldots, Y_n)$, the estimator of scale proposed by Rousseeuw & Croux (1993b) is defined by

$$Q_n(Y_{1:n}) = c(\Phi)\left\{\left|Y_p - Y_q\right|;\ 1 \leq p, q \leq n\right\}_{(\lfloor n^2/4 \rfloor)}, \tag{7}$$

where $c(\Phi) = 1/(\sqrt{2}\Phi^{-1}(5/8) = 2.21914$.

Now, consider the following assumption on $\boldsymbol{X}_t$.

(**A2**) $\boldsymbol{X}_t = (X_{1,t}, X_{2,t}, \ldots, X_{r,t})'$ is a multivariate stationary zero-mean Gaussian process satisfying

$$\sum_{h\geq 1} |\gamma_{ij}^X(h)| < \infty, \text{ for all } i, j \in \{1, \ldots, r\},$$

where $\gamma_{ij}^X(h) = \text{Cov}(X_{i,t}, X_{j,t+h})$.

By using Equations (1) and (3), then $(\boldsymbol{Z}_t)$ is also a multivariate stationary zero-mean Gaussian process satisfying

$$\sum_{h\geq 1} |\gamma_{ij}(h)| < \infty, \text{ for all } i, j \in \{1, \ldots, k\}, \tag{8}$$

where $\gamma_{ij}(h) = \text{Cov}(Z_{i,t}, Z_{j,t+h})$.

From the estimator Q_n defined in (7) and from the observations $(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)$, it is proposed a robust estimator of $\gamma_{ij}(h) = \text{Cov}(Z_{i,t}, Z_{j,t+h})$ for all i, j in $\{1, \ldots, k\}$ defined as follows

$$\widehat{\gamma}_{i,j}^Q(h) = \frac{1}{4}\left[Q_{n-h}^2(Z_{i,1:n-h} + Z_{j,h+1:n}) - Q_{n-h}^2(Z_{i,1:n-h} - Z_{j,h+1:n})\right], \tag{9}$$

where $Z_{i,1:n-h} = (Z_{i,1}, \ldots, Z_{i,n-h})$ and $Z_{j,h+1:n} = (Z_{j,h+1}, \ldots, Z_{j,n})$, which is the multivariate extension of the one proposed by Ma & Genton (2000).

From this estimator, the following robust estimator of the covariance matrix of $\boldsymbol{Z}_t$ is given by

$$\hat{\boldsymbol{\Gamma}}_Q(h) = \begin{bmatrix} \hat{\gamma}_{1,1}^Q(h) & \hat{\gamma}_{1,2}^Q(h) & \ldots & \hat{\gamma}_{1,k}^Q(h) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\gamma}_{k,1}^Q(h) & \hat{\gamma}_{k,2}^Q(h) & \ldots & \hat{\gamma}_{k,k}^Q(h) \end{bmatrix}, \tag{10}$$

Based on Equation (6) and on the robust estimator of the covariance matrix, the robust version of the estimator $\widehat{\boldsymbol{M}}$ is suggested here as follows

$$\widehat{\boldsymbol{M}}_Q = \sum_{h=1}^{h_0} \widehat{\boldsymbol{\Gamma}}_Q(h)\widehat{\boldsymbol{\Gamma}}_Q(h)'. \tag{11}$$

where $\widehat{\boldsymbol{\Gamma}}_Q(h)$ denotes the robust covariance matrix estimator of $\boldsymbol{Z}_t$ at lag h.

Therefore, the robust estimator $\widehat{r}_Q$ of r is similarly obtained from Equation (5) by replacing $\widehat{\lambda}_{i+1}$ and $\widehat{\lambda}_i$ by $\widehat{\lambda}_{i+1}^Q$ and $\widehat{\lambda}_i^Q$, respectively, where $(\widehat{\lambda}_j^Q)_{1\leq j\leq k}$ are the eigenvalues of $\widehat{\boldsymbol{M}}_Q$.

2. Theoretical results

This section provides some analytical results to support theoretically the robust approach discussed in the previous section.

Theorem 1. Let h be a fixed positive integer and $\left(\widehat{\Gamma}_Q(h)\right)_{1\leq i,j\leq k} = \left(\widehat{\gamma}^Q_{i,j}(h)\right)_{1\leq i,j\leq k}$, where $\widehat{\gamma}^Q_{i,j}(h)$ is defined in Equation (9). Assume that (A2) holds, then

$$\sqrt{n}\sup_{1\leq j\leq k}\left|\widehat{\lambda}^Q_j - \lambda_j\right| = O_p(1), \text{ as } n\to\infty,$$

where $(\widehat{\lambda}^Q_j)_{1\leq j\leq k}$ and $(\lambda_j)_{1\leq j\leq k}$ denote the eigenvalues of $\left(\sum_{h=1}^{h_0}\widehat{\Gamma}_Q(h)\widehat{\Gamma}_Q(h)'\right)$ and $\left(\sum_{h=1}^{h_0}\Gamma(h)\Gamma(h)'\right)$, respectively, where $(\Gamma(h))_{1\leq i,j\leq k} = (\gamma_{i,j}(h))_{1\leq i,j\leq k}$ and h_0 is a fixed integer larger than 1.

The proof of this theorem directly follows from Lemmas 1, 2 and 3 given below and proved in Section 7.

Remark 1. By Theorem 1 and Lam & Yao (2012, Proposition 1), it can see that the eigenvalues of the robust estimator covariance matrix of $\boldsymbol{Z}_t$ have the same rate of convergence as the eigenvalues of the standard estimator of the covariance matrix of $\boldsymbol{Z}_t$.

Lemma 1. Let $\widehat{A}_n$ be a sequence of $k\times k$ symmetric matrices and A a $k\times k$ symmetric matrix such that $u_n(\widehat{A}_n - A) = O_p(1)$, where u_n is a sequence of positive numbers tending to infinity as n tends to infinity, then

$$u_n\sup_{1\leq j\leq p}|\lambda_j(\widehat{A}) - \lambda_j(A)| = O_p(1), \quad \text{as } n\to\infty,$$

where $(\lambda_j(\widehat{A}))_{1\leq j\leq k}$ and $(\lambda_j(A))_{1\leq j\leq k}$ are the eigenvalues of $\widehat{A}_n$ and A, respectively.

Lemma 2. Let $\widehat{A}_n(h)$ be a sequence of $k\times k$ symmetric matrices and $A(h)$ a $k\times k$ symmetric matrix such that $u_n(\widehat{A}_n(h) - A(h)) = O_p(1)$, for each fixed $h\in\{1,\ldots,h_{max}\}$, where u_n is a sequence of positive numbers tending to infinity as n tends to infinity, then

$$u_n\left(\sum_{h=1}^{h_{max}}\widehat{A}_n(h)\widehat{A}_n(h)' - \sum_{h=1}^{h_{max}}A(h)A(h)'\right) = O_p(1),$$

as n tends to infinity.

Lemma 3. Let h be a non negative integer and i and j two integers in $\{1,\ldots,k\}$. Assume that (A2) holds, then the robust autocovariance estimator $\widehat{\gamma}^Q_{i,j}(h)$ defined in (9) satisfies the following central limit theorem

$$\sqrt{n}(\widehat{\gamma}^Q_{i,j}(h) - \gamma_{ij}(h)) \xrightarrow{d} \mathcal{N}(0,\widetilde{\sigma}^2_{i,j}(h)), \text{ as } n\to\infty,$$

where

$$\widetilde{\sigma}^2_{i,j}(h) = \mathbb{E}[\psi(Z_{i,1},Z_{j,1+h})^2] + 2\sum_{\ell\geq 1}\mathbb{E}[\psi(Z_{i,1},Z_{j,1+h})\psi(Z_{i,\ell+1},Z_{j,\ell+1+h})],$$

where ψ is defined in Equation (13).

3. Monte Carlo studies

This section reports the results of several Monte Carlo experiments to analyze the effect of high-dimensional time series with additive outliers on the factor modeling. In this context, the empirical study considered the VAR(1) model presented for simulating $\boldsymbol{X}_t$, with $r = 3$. The VAR(1) model was generated with independent white noise vector from N($\mathbf{0}, \boldsymbol{I}$) and $\boldsymbol{\Phi}$ coefficients, which are displayed in Table 1. The sample size is $n = 50, 100, 200, 400, 800$ and 1600, and $k = 0.2n, 0.5n, 0.8n$. The factor model (Equation 1) was generated as follows: first, all $k \times r$ elements of matrix $\boldsymbol{P}$ were generated as independent observations from the uniform distribution on the interval $[-1, 1]$ (see, also, Lam & Yao (2012)). The process $\boldsymbol{\epsilon}_t$ in Equation (1) consists of independent N(0, 1) components and they are also independent across t.

Table 1: $\boldsymbol{\Phi}$ matrices for VAR(1) process

$\boldsymbol{\Phi}_1$ (Model 1)			$\boldsymbol{\Phi}_1$ (Model 2)		
0.60	0.00	0.00	0.60	0.35	0.10
0.00	-0.50	0.00	0.05	-0.50	0.65
0.00	0.00	0.30	0.80	0.00	0.30

Note that, in Table 1, each model has its particularities. Model 1 corresponds to a process with no temporal correlation outside the diagonal; that is, each $X_{i,t}$, $i = 1, 2, 3$, has serial dependence only, while in Model 2 , $X_{i,t}$ has not only serial dependence, but also the interdependence between different series $X_{i,t}$ and $X_{j,t}$.

The main interest in this empirical study is to verify the performance of the statistics $\widehat{r}$ and $\widehat{r}_Q$ in the context of VAR(1) models with and without outliers. For this, the relative frequencies of $\widehat{r} = r$, denoted here as $f_{rel.}(\widehat{r} = r)$, were computed, where $\hat{r}$ is the estimator of r. It was similarly computed for the $\widehat{r}_Q$ estimator. The statistical quantities were computed based on 1000 replications.

Now, let $\{\mathbf{Z}_t\}, t = 1, ..., t \in \mathbb{Z}$, be a vector process contaminated by additive outliers defined as follows:

$$\mathbf{Z}_t = \boldsymbol{X}_t + \boldsymbol{\omega} \circ \boldsymbol{\delta}_t, \tag{12}$$

where "$\circ$" is the Hadamard product (Johnson, 1989). $\boldsymbol{\omega} = [\omega_1, ..., \omega_k]'$ is a magnitude vector of additive outliers. $\boldsymbol{\delta}_t = [\delta_{1t}, ..., \delta_{kt}]'$ is a random vector indicating the occurrence of an outlier at time t, in variable k, such as $\mathbb{P}(\delta_{k,t} = -1) = \mathbb{P}(\delta_{k,t} = 1) = p/2$ and $\mathbb{P}(\delta_{k,t} = 0) = 1 - p$, where $\mathbb{E}[\delta_{k,t}] = 0$ and $\mathbb{E}[\delta_{k,t}^2] = \mathbb{V}ar(\delta_{k,t}) = p$. The model described above assumes that $\{\mathbf{Z}_t\}$ and $\{\boldsymbol{\delta}_t\}$ are independent processes. Also, it is assumed that the elements of $\boldsymbol{\delta}_t$ are not correlated and temporally uncorrelated, i.e., $\mathbb{E}(\boldsymbol{\delta}_t \boldsymbol{\delta}_t^{'}) = \Sigma_\delta = diag(p, ..., p)$ and $\mathbb{E}(\boldsymbol{\delta}_t \boldsymbol{\delta}_{t+h}^{'}) = 0$ for $h \neq 0$.

Remark 2. $\boldsymbol{\delta}_{kt}$ is the product of *Bernoulli*(p) random variable with *Rademacher* random variable, the latter equals 1 or -1, both with probability 1/2.

Here, in the empirical investigation, the probability of an outlier occurring at time t is $p = 0.05$ and, without loss of generality, it is also assumed that $\boldsymbol{\omega} = [\omega, 0, 0]'$; that is, $Z_{1,t}$, $t = 1, ..., n$, is the only process in $\mathbf{Z}_t = (Z_{1,t}, Z_{2,t}, Z_{3,t})'$ contaminated with outliers and $\omega = 15$.

Table 2 reports the relative frequency estimates $f_{rel.}(\widehat{r} = 3)$ for Model 1. Observe that the ratio-based estimator of r improves when n is very large. Similar performance of the ratio is observed when the dimension k increases.

Table 3 displays the empirical investigation when the VAR(1) is Model 2; that is, now the process is generated with a non-diagonal $\boldsymbol{\Phi}$ matrix. As expected the empirical convergence of the estimated relative frequencies to 1 is much slower compared to the results related to Model 1. The estimation for r is very accurate for $n \geq 800$. Therefore, the ratio-based estimator has difficulty addressing the number of factors correctly when there is inter-correlation among variables. This phenomenon is more evident in more complex models (the simulation is not presented here but is available upon request). These results indicate that the presence of a more complex structure of correlation leads to an incorrect estimation of the dimensional reduction.

Table 2: Relative frequency estimates for $f_{rel.}(\widehat{r} = 3)$ - Model 1

n	**50**	**100**	**200**	**400**	**800**	**1600**
$k = 0.2n$	0.170	0.585	0.870	0.995	1	1
$k = 0.5n$	0.395	0.710	0.975	1	1	1
$k = 0.8n$	0.435	0.740	0.960	1	1	1
$k = 1.2n$	0.470	0.785	0.960	1	1	1

Table 3: Relative frequency estimates for $f_{rel.}(\widehat{r} = 3)$ - Model 2

n	**50**	**100**	**200**	**400**	**800**	**1600**
$k = 0.2n$	0.080	0.095	0.145	0.360	0.815	1
$k = 0.5n$	0.180	0.155	0.205	0.450	0.875	1
$k = 0.8n$	0.155	0.165	0.250	0.455	0.830	1
$k = 1.2n$	0.180	0.160	0.285	0.465	0.915	1

Now, the investigation is directed to the case where the process $\boldsymbol{Z}_t$ contains additive outliers. Table 4 shows the relative frequency estimates for the dimensional reduction ($\widehat{r}$ and $\widehat{r}_Q$) when $r = 3$ for Model 1 considering the presence of outliers. The standard case ($\widehat{\boldsymbol{\Gamma}}_{\boldsymbol{Z}}$ and $p = 0$) is in accordance with the results given by Table 2. Fourth column gives the simulation results using $\widehat{\boldsymbol{\Gamma}}_Q$ when $p = 0$. As can be seen, the $\widehat{r}$ estimates using $\widehat{\boldsymbol{\Gamma}}_Q$ present similar results as $\widehat{\boldsymbol{\Gamma}}_{\boldsymbol{Z}}$ when $p = 0$, which is in accordance with the asymptotic results discussed previously (see Remark 1), that is, the eigenvalues of the robust covariance matrix estimator of $\boldsymbol{Z}_t$ have the same rate of convergence as the eigenvalues of the standard estimator of the covariance matrix of $\boldsymbol{Z}_t$.

This fact indicates that the robust methodology may be used when the presence of outliers in the series is uncertain. The impact of additive outliers in the number of estimated factors can be verified from the second column where the presence of atypical observations in the data leads to a reduction of the estimated frequencies when $\widehat{r} = 3$ for all values of k. This does not

occur when the robust estimator is used, and the results are quite close to the ones from the first column. The percentage of outliers in only one vector seems to be, in general, not strong enough to destroy the robustness of the proposed method. Other simulation cases presented similar conclusions and are available upon request.

Table 4: Relative frequency estimates for dimensional reduction, $n = 100$ - Model 1

	$p = 0$			$p = 0.05$ and $\omega = 15$			$p = 0$			$p = 0.05$ and $\omega = 15$		
	$\widehat{r} = 1$	$\widehat{r} = 2$	$\widehat{r} = 3$	$\widehat{r} = 1$	$\widehat{r} = 2$	$\widehat{r} = 3$	$\widehat{r}_Q = 1$	$\widehat{r}_Q = 2$	$\widehat{r}_Q = 3$	$\widehat{r}_Q = 1$	$\widehat{r}_Q = 2$	$\widehat{r}_Q = 3$
$k = 0.2n$	0.110	0.330	0.585	0.250	0.230	0.290	0.140	0.410	0.450	0.180	0.380	0.440
$k = 0.5n$	0.100	0.280	0.710	0.240	0.240	0.260	0.130	0.320	0.550	0.160	0.310	0.530
$k = 0.8n$	0.040	0.200	0.785	0.130	0.120	0.210	0.040	0.270	0.690	0.060	0.290	0.650

4. Application to the pollutant PM_{10}

This section presents an application of the methodology discussed previously for PM_{10} concentrations measured at the AQAMN (Air Quality Automatic Monitoring Network) of the Region of Greater Vitória (RGV), Espírito Santo, Brazil. RGV is comprised of seven cities with a population of approximately 1.9 million inhabitants in an area of 2,319 km^2. The AQAMN consists of eight monitoring stations distributed in the cities of RGV; Laranjeiras, Carapina, Camburi, Suá, VixCentro, Vila Velha (VVCentro), Ibes and Cariacica. The application was divided in two parts: 1) reduction of matrix dimensions, and 2) forecasting. PM_{10} is a daily average value expressed in $\mu g/m^3$, monitored in all stations ($k = 8$) and measured from January 2005 to December 2009 ($n = 1826$).

Figure 1 shows the plots of the PM_{10} concentrations. Based on this figure, the series indicated that they have high levels of concentrations which can be identified, from a statistical point of view, as outliers (additive), since they produce similar impact on the sample ACF to that caused by additive outliers, that is, they lead to a reduction of the sample autocovariance values. This empirical evidence justifies the use of both robust and non-robust methods to verify whether or not these high levels make any impact on the model estimation.

Figure 2 displays the robust ACFs of the series. The plots of the classical sample ACF are not presented here to save space, however, robust and classical sample autocorrelation functions presented similar behavior. This is an indication that the high levels of the pollutant were not large enough to destroy the sample structure correlation of the data set. The robust ACFs show possible seasonal pattern of period $s = 7$.

From the above discussion, it is expected that the FA estimated model and forecasting issues will show similar performance for both methodologies, that is, for the standard and robust ones. The estimates of the number of factors r were computed by performing eigenanalysis on $\widehat{\boldsymbol{M}}$ and on $\widehat{\boldsymbol{M}}_Q$ of Equations (6) and (11), respectively, with $h_0 = 14$. The eigenvalues obtained (in decreasing order) and their ratios obtained using $\widehat{\boldsymbol{\Gamma}}_{\boldsymbol{Z}}$ are shown in Figure 3 (first two panels, respectively). The corresponding robust version, i.e., using $\widehat{\boldsymbol{\Gamma}}_Q$, is shown in Figure 4. The plots show similar results which is, as previously stated, an expected result. The plots indicate that $\widehat{r} = \widehat{r}_Q = 1$. The reduction was not affected when varying the value of h_0.

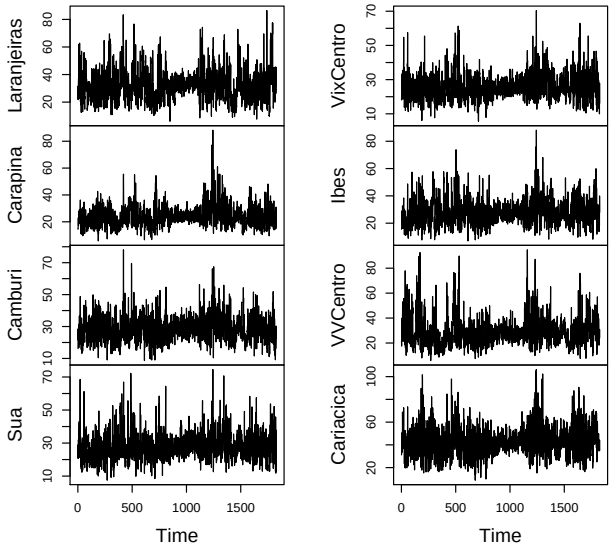


Figure 1: PM_{10} concentrations of the AQAMN stations.

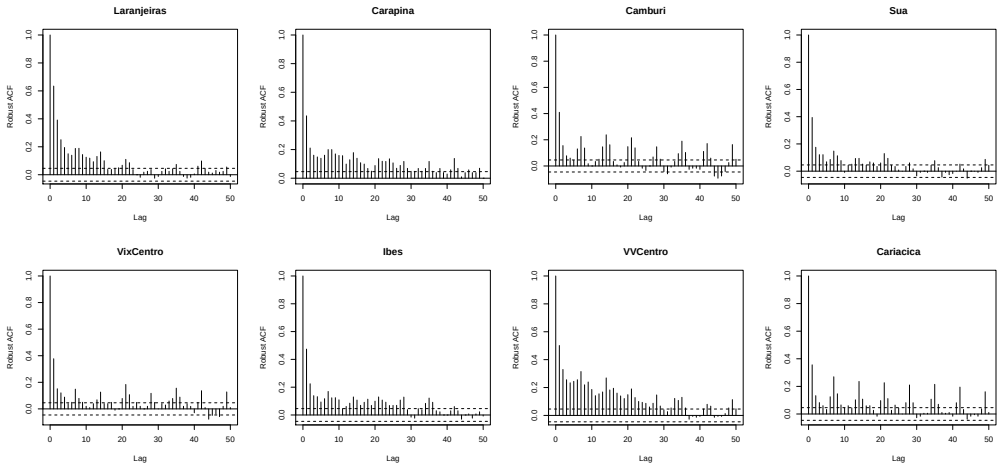


Figure 2: Robust ACF of PM_{10} concentrations.

The last panels in Figures 3 and 4 display the time series plots of the estimated $\widehat{X}_t$ defined in Equation 4, for $\widehat{\mathbf{\Gamma}}_{\mathbf{Z}}$ and $\widehat{\mathbf{\Gamma}}_{Q}$, respectively. The robust ACF of $\widehat{X}_t$ (factor) showed a stochastic seasonal behavior of period $s = 7$ remained from the original data set (result available upon request). Thus, the forecast $\hat{X}_{T+h}^{(h)}$ was obtained by means of a standard univariate SARMA model.

For forecasting purpose, the factor series $(\widehat{X}_t)$ was divided in two parts: learning and prediction sets. The 1626 observations from January 1st, 2005 to June 14th, 2009 are considered to be the learning set and the remaining 200 observations are considered for the forecasting study. Based on statistical analysis, the SARMA$(1,0)\times(1,0)_7$ model was chosen for the factor series. The robust ACFs of the residuals is presented in Figure 5, where it can be observed that the filter captured the seasonality of the factor series quite well. The Box-Pierce and Ljung-Box

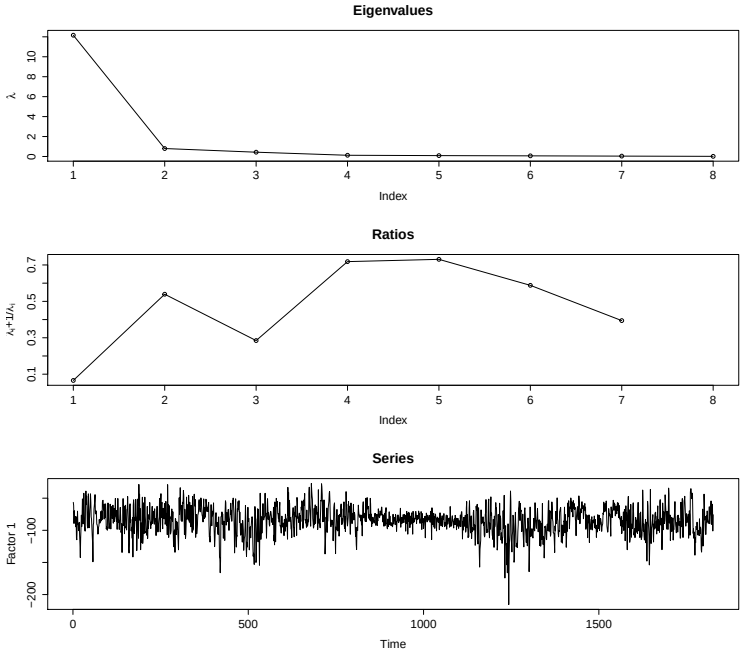


Figure 3: Plots of estimated eigenvalues, ratios of estimated eigenvalues of $\hat{\boldsymbol{M}}$ and estimated first factor.

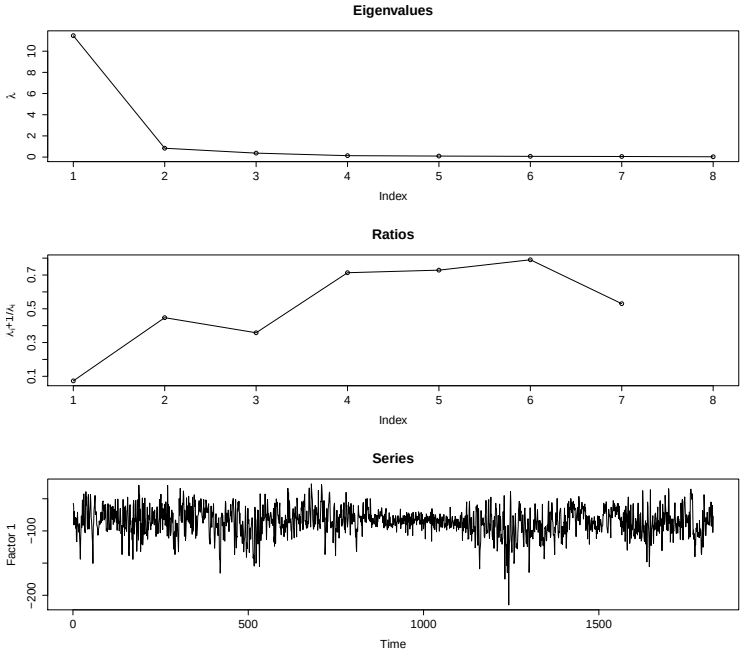


Figure 4: Plots of estimated eigenvalues, ratios of estimated eigenvalues of $\widehat{\boldsymbol{M}}_Q$ and estimated first factor.

statistics (robust tests) indicated that the sample residuals are not time-correlated (the results are available upon request).

The forecasts of the observed series ($\boldsymbol{Z}_t$) was computed by $\widehat{\boldsymbol{Z}}_{T+h}^{(h)} = \widehat{\boldsymbol{P}}\widehat{\boldsymbol{X}}_{T+h}^{(h)}$ (FA-SARMA model). Based on the one-step ahead forecast, the performance of the FA-SARMA model was compared with the standard VAR(1) model. The latter model was applied on the original series. The forecast comparison was made for Suá station.

To measure the accuracy of the forecasts, the criteria Mean Square Prediction Error (MSPE),

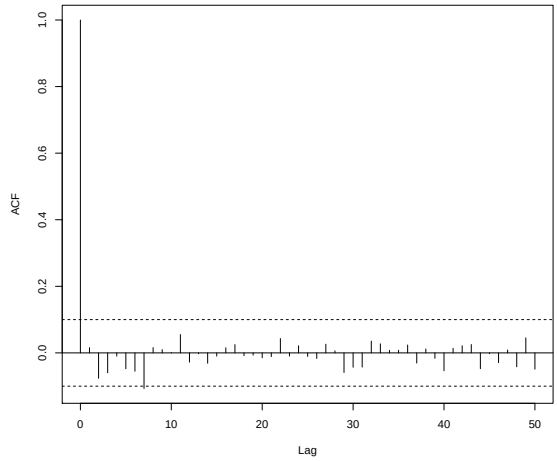


Figure 5: Robust ACFs for residuals of the SARMA model, for the resulting factor series

Mean Percent Prediction Error (MPPE) and Mean Absolute Percent Prediction Error (MAPPE) were used. The values are displayed in Table 5. From this table, it can be seen that the FA-SARMA model yield more accurate forecasts than the than the VAR(1) model.

Table 5: MSPE, MPPE and MAPPE of the fitted models, for Suá station

	MSPE	MPPE	MAPPE
FA-SARMA	8.22	6.52	23.16
VAR(1)	12.34	9.27	40.28

Figure 6 presents a visual analysis of the one-step-ahead forecast values of PM_{10} measured at Suá station using the FA-SARMA model, from June 15th 2009 to December 31st 2009. It indicates a reasonably good performance of the model proposed here.

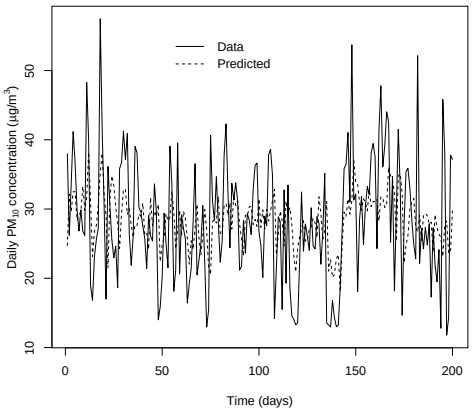


Figure 6: One-step-ahead forecasts of PM_{10} concentrations at Suá station using the FA-SARMA model, from June 15th 2009 to December 31st 2009,

5. Conclusions

In this paper a robust method for high-dimensional time series with additive outliers is proposed. Some theoretical results are discussed and these were empirically investigated by Monte Carlo experiments under different scenarios. They showed the effect of the additive outliers on the reduction of the factor dimension. The proposed robust estimator performed quite well and it can be very useful in practical applications where there is any evidence of atypical observations, such as, high levels of concentrations in the pollution area. In addition, the proposed methodology was used to identify pollution behavior of the pollutant PM_{10} in the Region of Greater Vitória (RGV), Espírito Santo, Brazil, and to forecast the observations, which can be very useful for the management of the air quality network. The results in this paper will hopefully stimulate further research on this theme.

6. Acknowledgments

The authors would like to thank CNPq, CAPES and FAPES for their financial support.

7. Proofs

Proof of Lemma 1. By Weyl's Theorem, see Horn & Johnson (1985, p. 184), for all $j \in \{1, \ldots, k\}$, it follows that

$$\lambda_j(\widehat{A}) - \lambda_j(A) \leq \lambda_k(\widehat{A} - A) \leq \sup_{1 \leq \ell \leq k} |\lambda_\ell(\widehat{A} - A)|.$$

By exchanging the role of $\widehat{A}$ and A, for all $j \in \{1, \ldots, k\}$, it follows that

$$\lambda_j(A) - \lambda_j(\widehat{A}) \leq \sup_{1 \leq \ell \leq k} |\lambda_\ell(\widehat{A} - A)|.$$

Hence,

$$\sup_{1 \leq j \leq k} |\lambda_j(\widehat{A}) - \lambda_j(A)| \leq \sup_{1 \leq \ell \leq k} |\lambda_\ell(\widehat{A} - A)| = \|\widehat{A} - A\|_2,$$

where $\|X\|_2$ denotes the largest absolute value of the eigenvalues of a matrix X. Since $u_n(\widehat{A}_n - A) = O_p(1)$, the result follows. □

Proof of Lemma 2. The proof of this lemma directly follows from the application of the continuous mapping theorem; see van der Vaart (1998, Theorem 2.3). □

Proof of Lemma 3. Observe that the autocovariance of the process $(Z_{i,t} + Z_{j,t+h})_{t \geq 1}$ at lag ℓ is equal to

$$\gamma_{i,j}^{(+)}(\ell) = \mathrm{Cov}(Z_{i,t} + Z_{j,t+h}, Z_{i,t+\ell} + Z_{j,t+h+\ell}) = \gamma_{i,i}(\ell) + \gamma_{i,j}(h + \ell) + \gamma_{j,i}(\ell - h) + \gamma_{j,j}(\ell),$$

and that the autocovariance of the process $(Z_{i,t} - Z_{j,t+h})_{t\geq 1}$ at lag ℓ is equal to

$$\gamma_{i,j}^{(-)}(\ell) = \mathrm{Cov}(Z_{i,t} - Z_{j,t+h}, Z_{i,t+\ell} - Z_{j,t+h+\ell}) = \gamma_{i,i}(\ell) - \gamma_{i,j}(h+\ell) - \gamma_{j,i}(\ell - h) + \gamma_{j,j}(\ell).$$

By A2 and Equation (8), $\sum_{\ell\geq 1} |\gamma_{i,j}^{(+)}(\ell)| < \infty$ and $\sum_{\ell\geq 1} |\gamma_{i,j}^{(-)}(\ell)| < \infty$. The proof of this lemma thus follows the same lines as the ones of Lévy-Leduc et al. (2011c, Theorem 2) by replacing X_i and X_{i+h} by $Z_{i,t}$ and $Z_{j,t+h}$, respectively and the summations on i by summations on t which leads to

$$\sqrt{n-h}\left(\widehat{\gamma}_{i,j}^{Q}(h) - \gamma_{ij}(h)\right) = \frac{1}{\sqrt{n-h}} \sum_{t=1}^{n-h} \psi(Z_{i,t}, Z_{j,t+h}) + o_P(1),$$

where

$$\begin{aligned}
&\psi(x,y) = \\
&\quad \frac{1}{2}\left(\gamma_{i,i}(0) + \gamma_{j,j}(0) + \gamma_{i,j}(h) + \gamma_{j,i}(-h)\right) \mathrm{IF}\left(\frac{x+y}{\sqrt{\gamma_{i,i}(0) + \gamma_{j,j}(0) + \gamma_{i,j}(h) + \gamma_{j,i}(-h)}}, Q, \Phi\right) \\
&- \frac{1}{2}\left(\gamma_{i,i}(0) + \gamma_{j,j}(0) - \gamma_{i,j}(h) - \gamma_{j,i}(-h)\right) \mathrm{IF}\left(\frac{x-y}{\sqrt{\gamma_{i,i}(0) + \gamma_{j,j}(0) - \gamma_{i,j}(h) - \gamma_{j,i}(-h)}}, Q, \Phi\right), \quad (13)
\end{aligned}$$

where IF is defined in Equation (20) of Lévy-Leduc et al. (2011c). By applying Arcones (1994, Theorem 4), the result is obtained. □

References

Anderson, T. W. (2003). *An introduction to multivariate statistical analysis*. (3rd ed.). New Jersey: John Wiley & Sons.

Arcones, M. A. (1994). Limit theorems for nonlinear functionals of a stationary gaussian sequence of vectors. *Ann. Probab.*, *22*, 2242–2274. doi:10.1214/aop/1176988503.

Brunekreef, B., & Holgate, S. T. (2002). Air pollution and health. *The Lancet*, *360*, 1233–1242.

Chang, I., Tiao, G. C., & Chen, C. (1988). Estimation of time series parameters in the presence of outliers. *Technometrics*, *30*, 193–204.

Chen, C., & Liu, L.-M. (1993). Joint estimation of model parameters and outlier effects in time series. *Journal of the American Statistical Association*, *88*, 284–297.

Curtis, L., Rea, W., Smith-Willis, P., Fenyves, E., & Pan, Y. (2006). Adverse health effects of outdoor air pollutants. *Environment International*, *32*, 815–830.

Horn, R. A., & Johnson, C. R. (1985). *Matrix Analysis*. Cambridge University Press. URL: http://dx.doi.org/10.1017/CBO9780511810817 cambridge Books Online.

Johnson, C. (1989). *Matrix theory and applications*. American Mathematical Soc.

Johnson, R., & Wichern, D. (2007). *Applied multivariate statistical analysis*. (6th ed.). New Jersey: Prentice Hall.

Lam, C., & Yao, Q. (2012). Factor modeling for high-dimensional time series: Inference for the number of factors. *Ann. Statist.*, *40*, 694–726. doi:10.1214/12-AOS970.

Lam, C., Yao, Q., & Bathia, N. (2011). *Estimation of latent factors for high-dimensional time series*. LSE Research Online Documents on Economics London School of Economics and Political Science, LSE Library. URL: http://EconPapers.repec.org/RePEc:ehl:lserod:31549.

Lévy-Leduc, C., Boistard, H., Moulines, E., Taqqu, M. S., & Reisen, V. A. (2011a). Asymptotic properties of U-processes under long-range dependence. *The Annals of Statistics*, *39*, 1399–1426.

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

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

Ma, Y., & Genton, M. G. (2000). Highly robust estimation of the autocovariance function. *Journal of Time Series Analysis*, *21*, 663–684.

Maynard, R. (2004). Key airborne pollutants: the impact on health. *Science of The Total Environment*, *334-335*, 9–13.

Peña, D., & Box, G. E. P. (1987). Identifying a simplifying structure in time series. *Journal of the American Statistical Association*, *82*, pp. 836–843.

Rousseeuw, P. J., & Croux, C. (1993a). Alternatives to the median absolute deviation. *Journal of the American Statistical association*, *88*, 1273–1283.

Rousseeuw, P. J., & Croux, C. (1993b). Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, *88*, 1273–1283.

Seinfeld, J. H., & Pandis, S. N. (2006). *Atmospheric chemistry and physics: from air pollution to climate change*. New York: J. Wiley.

Stock, J. H., & Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, *97*, 1167–1179.

Tsay, R. S. (1988). Outliers, level shifts, and variance changes in time series. *Journal of forecasting*, *7*, 1–20.

van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Watson, J. G., Zhu, T., Chow, J. C., Engelbrecht, J., Fujita, E. M., & Wilson, W. E. (2002). Receptor modeling application framework for particle source apportionment. *Chemosphere*, *49*, 1093–1136.

WHO (2005). *WHO air quality guidelines global update 2005. Report on a working group meeting, Bonn/Germany*. WHO - World Health Organization. URL: `http://www.euro.who.int/__data/assets/pdf_file/0008/147851/E87950.pdf`.

WHO (2014). *Air pollution estimates*. WHO - World Health Organization. URL: `http://www.who.int/phe/health_topics/outdoorair/databases/FINAL_HAP_AAP_BoD_24March2014.pdf?ua=1`.