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

HIGOR HENRIQUE ARANDA COTTA

ANÁLISE DE COMPONENTES PRINCIPAIS ROBUSTA EM DADOS DE POLUIÇÃO DO AR: APLICAÇÃO À OTIMIZAÇÃO DE UMA REDE DE MONITORAMENTO

VITÓRIA
2014

HIGOR HENRIQUE ARANDA COTTA

ANÁLISE DE COMPONENTES PRINCIPAIS ROBUSTA EM DADOS DE POLUIÇÃO DO AR: APLICAÇÃO À OTIMIZAÇÃO DE UMA REDE DE MONITORAMENTO

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

VITÓRIA
2014

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

Cotta, Higor Henrique Aranda, 1987-
C846a Análise de componentes principais robusta em dados de poluição do ar: aplicação à otimização de uma rede de monitoramento / Higor Henrique Aranda Cotta. – 2014.
75 f. : il.

Orientador: Valdério Anselmo Reisen.
Dissertação (Mestrado em Engenharia Ambiental) – Universidade Federal do Espírito Santo, Centro Tecnológico.

1. Análise de componentes principais. 2. Ar - Poluição. 3. Análise de séries temporais. I. Reisen, Valdério Anselmo. II. Universidade Federal do Espírito Santo. Centro Tecnológico. III. Título.

CDU: 628

AGRADECIMENTOS

- Ao meu orientador Valdério A. Reisen, ponto central neste trabalho. Muito obrigado pelos ensinamentos, paciência, convivência e, principalmente por toda amizade. Agradeço-o por ter me ensinado durante todo esse período, o que é e como se faz ciência.
- Aos professores do PPGEA e do DEST-UFES por todo ensinamento repassado.
- Aos membros da banca de defesa, Prof. Fábio Molinares, Prof. Márton Íspany e Prof. Neyval Reis pelas valorosas sugestões.
- Aos amigos de mestrado, do NuMEs e da graduação em estatística. Agradeço especialmente a Adriano Sgrâncio, Alessandro Sarnaglia, Bartolomeu Zamprogno, Edson Zambom, Emersom Matos, Gustavo Três, Fátima Leite, Juliana Bottoni, Mariana Figueira, Milena Machado, Pedro Berger, Wanderson Pinto, Wesley Corrêa e Wharley Borges pela ajuda direta ou indireta, conselhos dados e pelos momentos de descontração.
- A secretária do PPGEA Rose Mary Nunes Leão. Muito obrigado pelas diversas dúvidas respondidas.
- A secretária do Colegiado de Estatística da UFES Tânia Baldotto Ribeiro. Obrigado por tudo.
- A todos que de alguma forma contribuíram para que este trabalho fosse concluído.

"We have a large reservoir of engineers (and scientists) with a vast background of engineering know-how. They need to learn statistical methods that can tap into the knowledge. Statistics used as a catalyst to engineering creation will, I believe, always result in the fastest and most economical progress"
George Box

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E/OU SIGLAS

ACP	Análise de componentes principais
AC	Análise de clusters
ARMA	Autorregressivo médias móveis
B	Operador defasagem
BQM	Balanço químico de massa
CO	Carbono Orgânico
ES	Espírito Santo
EQM	Erro Quadrático Médio
FMP	Fatoração de matriz positiva
IBGE	Instituto Brasileiro de Geografia e Estatística
IEMA	Instituto Estadual do Meio Ambiente e Recursos Hídricos
ISJ	Instituto Jones do Santos Neves
MAD	Median absolute desviation
MAG	Modelo de regressão não linear aditivo generalizado
MAG-ACP	Modelo MAG com uso de ACP
PM_{10}	Particulate Matter (Material Particulado) menor que $10\mu m$ de diâmetro aerodinâmico
PIB	Produto Interno Bruto
PTS	Partículas Totais em Suspenção
RGV	Região da Grande Vitória
RR	Risco relativo
SO_2	Dióxidos de enxofre
NO_2	Óxidos de nítrico
OLS	Ordinary least squares
O_3	Ozônio
PC	Principal component
PCA	Principal components analysis
tr	traço de uma matriz
μm	micrômetro
$\Gamma_X(0)$	Matriz de covariância do modelo X na defasagem zero
Φ	Matriz com coeficientes do VARMA
Σ	Matriz de covariância do processo ruído branco
$\Psi(B)$	Operador defasagem de um processo linear
VAR	Modelo Vetorial autorregressivo
VARMA	Modelo Vetorial autorregressivo de médias móveis
WHO	*World Health Organization*

RESUMO

Os estudos de dados de Poluição do ar originados de uma rede de monitoramento envolvem um número considerável de variáveis e observações. Do ponto de vista de técnicas estatísticas, é possível analisar separadamente cada variável de interesse. Entretanto, esse tipo de análise pode não contemplar as diversas dinâmicas de relacionamento existentes entre essas variáveis. Devido a isso, faz-se necessário o uso de técnicas estatísticas capazes de lidar, medir e analisar conjuntamente esses dados gerados. Esse ramo da Estatística é conhecido como estatística multivariada. Na área da poluição do ar destaca-se a análise de componentes principais (ACP), que constrói combinações lineares das variáveis para explicar a estrutura de variância-covariância dos dados originais. Na poluição do ar, a análise de componentes principais é utilizada para: Criação de Índices de Qualidade do Ar, Identificação de fontes de poluição, Redimensionamento de uma Rede de Monitoramento, Pré-processador de variáveis para Modelos Aditivos Generalizados , alem de outras aplicações. Neste trabalho a Análise de Componentes Principais (ACP) é utilizada no estudo do redimensionamento da Rede de Monitoramento da Qualidade do Ar da Região da Grande Vitória (RAMQAr) para o poluente PM_{10}. A ACP assume que os dados sejam não correlacionados no tempo, característica não observada nos dados de poluição do ar. As componentes obtidas de séries temporais mantém a propriedade de ortogonalidade, entretanto, essas componentes são autocorrelacionadas e correlacionadas temporalmente. Esse resultado é demonstrado teórica e empiricamente. A segunda contribuição deste trabalho é estudar a ACP no contexto de séries temporais com *outliers* aditivos por meio de metodologia robusta. Como já explorado na literatura, os *outliers* aditivos destroem a estrutura de correlação dos dados e, como as componentes são calculadas da matriz de covariância, os *outliers* também afetam as propriedades das componentes.

Palavras-chave: análise de componentes principais, poluição do ar, análise de séries temporais, *outliers*, robustez

ABSTRACT

Studies of data from air pollution originating from a network of air monitoring involve a large number of variables and observations. From the standpoint of statistical techniques, it is possible to analyze separately each variable of interest. However, this type of analysis can not contemplate the relationship dynamics between these variables. Because of this, it is necessary to use statistical techniques to handle, measure and analyze these data generated jointly. This branch of statistics known as Multivariate Statistics. One important multivariate technique in the area of air pollution is the Principal Component Analysis (PCA), which builds linear combinations of variables to explain the variance-covariance structure of the original data. Air pollution in the Principal Component Analysis is used for: creating indexes of air quality, identification of pollution sources, management of air quality monitoring network, preprocessor variables for generalized additive models, besides other applications. In this work PCA is used to study the management and scaling of the Network for Monitoring Air Quality in the Greater Vitoria Region. This work deals with the use of Principal Component Analysis (PCA) in time series with additive outliers. The PCA is one of the most important multivariate techniques which are linear combinations constructed to explain the variance-covariance structure of the original data. Although PCA assumes that the data are serially independent, this assumption is not found in practice situation in time series, e.g. Air Pollution data. PCs calculated from time series observations maintains their orthogonality property, but the components are found to be auto and cross-correlated, which depends on the correlation structure of the original series. These properties and their impact in the use of PCA are one of main objective of this work. Another contribution is related to the study of PCA time series under the presence of additive outliers by proposing a Robust PCA (RPCA) method. It is well known that additive outliers in time series destroys the correlation structure of the data. Since the PCs are computed by using the covariance matrix, the outliers also affect the properties of PCs. Therefore the Robust PCA should be used in this context. The Robust PCA method proposed here is justified empirical and theoretically, and a real data set based on Air Pollution time serie is used to show the usefulness of the Robust PCA method in a real application.

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

SUMÁRIO

1 INTRODUÇÃO

A preocupação do homem com a poluição atmosférica aumentou consideravelmente nos últimos 50 anos. Em diversas regiões do mundo, principalmente, em países em desenvolvimento, a qualidade do ar tem sido degradada como consequência da industrialização, do crescimento populacional, das altas taxas de urbanização e além de políticas de controle da qualidade do ar inadequadas ou inexistentes. As adversidades causadas pela poluição do ar produzem impactos locais, regionais e globais.

Diversos problemas de natureza social e econômica podem ser atribuídos à poluição do ar. A confirmação sobre os efeitos adversos provocados pelos poluentes na saúde pode ser obtida a partir de estudos populacionais, como os quais referenciados pelo documento "Diretrizes para a Qualidade do ar"da Organização Mundial de Saúde (OMS). A exposição à um nível elevado de poluentes pode ocasionar desde tosses, irritação dos olhos, nariz e da garganta, diminuição das funções de deglutição, bronquite e pneumonia, até doenças respiratórias crônicas, câncer de pulmão, doenças cardiovasculares (WHO, 2005).

Por exemplo, a exposição a um nível elevado de ozônio (O_3) pode ocasionar efeitos na pele e em canais lacrimais (ROJAS et al., 2000), complicações em pacientes com asma (MORTIMER et al., 2000), aumento na mortalidade e internações hospitalares em pacientes com doenças pulmonares crônicas (BURNETT et al., 1997b), etc. Para o material particulado com diâmetro aerodinâmico inferior à $10\mu m$ (PM_{10}), a exposição pode ocasionar aumento na mortalidade e internações hospitalares em pacientes com doenças pulmonares crônicas (MACNEE; DONALDSON, 2003), piora nos sintomas de pacientes com asma (DONALDSON; GILMOUR; MACNEE, 2000), aumento na mortalidade e internações hospitalares em pacientes com doenças cardíacas (DONALDSON et al., 2005), dentre outros efeitos adversos.

Já a exposição ao dióxido de nitrogênio (NO_2) pode levar a alteração do metabolismo pulmonar (USA-EPA, 1993) e inflamação das vias respiratórias (BLOMBERG et al., 1997; BLOMBERG et al., 1999). O dióxido de enxofre (SO_2) pode causar redução da função pulmonar (LAWTHER et al., 1975), influenciar o sistema nervoso autônomo (TUNNICLIFFE et al., 2001), etc. Já a exposição ao monóxido de carbono (CO) pode aumentar o número de admissões hospitalares devido à doenças cardiovasculares (BURNETT et al., 1997a), mortalidade de cardiopatas (GOLDBERG et al., 2003), etc. Devido a esses efeitos, faz-se necessário legislar e controlar as emissões e concentrações de poluentes para assim planejar, prever e prevenir impactos no meio ambiente e na população.

O principal propósito da gestão da qualidade do ar é a proteção da saúde pública e do ambiente dos efeitos adversos da poluição do ar. Um controle adequado da qualidade do ar envolve uma série de atividades como gestão de riscos, definições de padrões para emissões e de qualidade do ar, monitoramento dos poluentes, implantação de medidas de controle e comunicação de risco (WHO, 2005). Um importante instrumento para controle da poluição atmosférica é o monitoramento da qualidade do ar. Esse monitoramento é realizado através de estações, sendo que cada estação mede diferentes poluentes de acordo com a necessidade da região onde está instalada. Para o monitoramento da qualidade do ar na Região da Grande Vitória (RGV) foi implantada a Rede de Monitoramento Automática da Qualidade do ar da Região da Grande Vitória (RAMQAr), constituída por 8 estações que monitoram diversos poluentes de acordo com a necessidade da região. Os dados de PM_{10} utilizados nesta dissertação foram obtidos da RAMQAr.

Pires et al. (2008a), Pires et al. (2008b) apontam que o número de estações que constituem uma rede de monitoramento deve ser otimizado com o objetivo de se reduzir custos e despesas. Idealmente, apenas uma estação de monitoramento deve operar em uma área caracterizada por um padrão específico de poluição do ar.

Os estudos de dados de Poluição do ar originados de uma rede de monitoramento do ar envolvem um número considerável de variáveis e observações. Do ponto de vista de técnicas estatísticas, é possível analisar separadamente cada variável de interesse. Entretanto, esse tipo de análise pode não contemplar as diversas dinâmicas de relacionamento existentes entre essas variáveis. Devido a isso, faz-se necessário o uso de técnicas estatísticas capazes de lidar, medir e analisar conjuntamente esses dados gerados. Esse ramo da estatística é conhecido como estatística multivariada.

A análise de componentes principais (ACP) é uma das principais técnicas de estatística multivariada. O objetivo da ACP é explicar a estrutura de covariância dos dados através de variáveis auxiliares chamadas de componentes. Essas componentes são construídas através de combinações lineares das variáveis originais e são não-correlacionadas. Sucintamente, a ACP consiste em calcular os autovalores e os autovetores da matriz de covariância ou de correlação. Na Poluição do ar a ACP é utilizada para: Criação de Índices de Qualidade do Ar (KHAMIS; ABDULLAH, 2004); Identificação de fontes de poluição (ALMEIDA et al., 2006) e (SONG et al., 2006); Redimensionamento de uma Rede de Monitoramento (LU; HE; DONG, 2011); Pré-processador de variáveis para Modelos Aditivos Generalizados (WANG; PHAM, 2011); alem de outras aplicações. Pires et al. (2008a), Pires et al. (2008b) empregaram a ACP como ferramenta não-paramétrica para identificação e classificação de estações que possuíam o mesmo comportamento para determinado poluente.

Não obstante as dificuldades e a complexidade da própria análise estatística multivariada, os dados de poluição do ar possuem problemas devido à observações faltantes, *missing data*, e observações discrepantes, *outliers*. Fox (1972) define como *outliers* as observações discordantes face às restantes. Já Rousseeuw e Zomeren (1990) definiram *outliers* como observações que se afastam das estimativas por um modelo estatístico sugerido pela maior parte do conjunto de dados. Essas observações discrepantes podem ser provenientes de diversos fatores:

- Erro de medição;
- Erro de digitação;
- Influência de intervenções;
- Alterações inesperadas de condições físicas.

Neste contexto é importante salientar que a ACP é sensível à *outliers*. Isso ocorre devido ao fato de que a própria estimação da matriz de médias, da matriz de covariância e da matriz de correlação é fortemente influenciada por *outliers*. Como consequência, a estimação dos autovalores e dos autovetores também sofrerá influências dos *outliers* presentes nos dados (FILZMOSER, 1999). Croux e Haesbroeck (2000) apontam que conclusões obtidas através das componentes principais calculadas de dados com *outliers* podem ser enganosas.

Existem diversos métodos estatísticos para definir se uma observação é ou não um *outlier*. A mais popular é definir como *outlier* observações afastadas da média por dois ou três desvios padrões (BARNETT; LEWIS, 1984). Métodos atuais para detecção de *outliers*, como o proposto por Filzmoser, Maronna e Werner (2008), são capazes de trabalhar com dados de grandes dimensões (MINGOTI, 2005).

Uma das formas de se lidar com *outliers* é utilizar estatísticas robustas. Huber e Ronchetti (2009) definem estatística robusta como insensibilidade a pequenos desvios em relação às suposições originais. Maronna, Martin e Yohai (2006) definem estatística robusta como uma ferramenta para aumentar a confiabilidade e a acurácia do modelo e da análise estatística. Já Tukey (1975) afirma ser possível utilizar a estatística clássica, não robusta e a estatística robusta concomitantemente, devendo-se atentar quando os dois métodos apresentarem valores muito distintos. Assim, pode-se considerar estatística robusta como uma estatística insensível à *outliers*.

Jackson (2004) aponta que existem 3 maneiras de se obter a robustez para a ACP:

- Obter uma estimativa robusta da matriz de covariância ou correlação. A partir daí, calcular as componentes através da ACP usual;

- Obter uma estimativa robusta dos autovalores e autovetores e, utilizá-las para calcular as estimativas robustas da matriz de covariância ou correlação;
- Aplicar algum tipo de análise ou alteração nos dados iniciais e só então calcular a ACP usual.

Neste interim, para cada estimador robusto da matriz de covariância ou correlação, obtém-se uma Análise de Componentes Principais Robusta (ACP Robusta). Dentre os métodos de estimação robusta da matriz de covariância ou correlação, destacam-se os estimadores propostos por Ma e Genton (2000), Maronna (1976), Rousseeuw (1984) e (ROUSSEEUW; YOHAI, 1984). Diversos estudos de simulação foram realizados afim de determinar qual o melhor estimador robusto. Nesta linha, pode-se citar os trabalhos de (DEVLIN; GNANADESIKAN; KETTENRING, 1981) e (CROUX; HAESBROECK, 2000). Hubert, Rousseeuw e Aelst (2008) apresentam uma visão geral dos recentes métodos robustos para dados multivariados.

A análise de componentes principais é amplamente utilizada na poluição do ar. Entretanto, pouca atenção é dada à análise de componentes principais robusta utilizando dados da poluição do ar. Neste contexto, o objetivo deste trabalho é propor uma análise de componentes principais robusta, aplicada à dados da poluição do ar obtidos através da RAMQAr. A ACP será aplicada como ferramenta de identificação e classificação de estações que possuem o mesmo padrão de comportamento para os poluentes monitorados.

As contribuições ressaltadas acima servem de base para a motivação deste trabalho e do artigo apresentado nesta dissertação. O artigo resultado é o ponto central desta dissertação, trazendo avaliações por meio de simulação computacional e posterior aplicação no gerenciamento das estações RAMQAr, e a utilização da ACP no contexto de séries temporais, considerando a presença *outliers* aditivos. Esta problemática não é encontrada na literatura e a desconsideração dessa característica dos dados pode levar a resultados equivocados.

Esta dissertação é dividida da seguinte forma: além da introdução, a Seção 2 apresenta os problemas, objetivos, gerais e específicos, que motivaram esta pesquisa. A revisão das principais referências é descrita na Seção 3. O ferramental estatístico utilizado nesta dissertação é sumarizado na Seção 4. A metodologia é apresentada na Seção 5. O Artigo resultado do estudo desta dissertação está anexado na Seção 6. A Seção 7 apresenta a discussão geral e as conclusões com as recomendações para pesquisas futuras.

2 PROBLEMAS E OBJETIVOS

2.1 PROBLEMAS

A Análise de Componentes Principais é extensamente utilizada na área de poluição do ar. Zamprogno (2013) mostrou que a correlação serial leva a um aumento da variabilidade total dos dados. Além disso, sabe-se que a ACP usual é extremamente sensível à um conjunto de dados com observações discrepantes, *outliers* aditivos, e que os resultados obtidos desta análise podem levar a conclusões equivocadas. De modo a mitigar esse problema, foram desenvolvidos diversos estimadores robustos dos parâmetros de locação e escala. Esses estimadores, substituindo os estimadores usuais da ACP, levam à Análise de Componentes Principais Robusta. Entretanto, pouca atenção foi dada a Análise de Componentes Principais Robusta utilizando dados de poluição do ar.

As estações de uma rede de estações de monitoramento da qualidade do ar podem apresentar padrões de comportamento semelhantes para um determinado poluente monitorado. Assim, caso detectado que duas ou mais estações possuem o mesmo padrão de comportamento para um poluente em comum, é possível realocar os equipamentos de medições desse poluente para uma outra estação. Esse tipo de avaliação ainda não foi realizada para a RAMQAr considerando a metodologia robusta.

2.2 OBJETIVO GERAL

Estudar o modelo multivariado ACP na presença de dados correlacionados com diferentes estruturas de dependência considerando *outliers* aditivos, e aplicar a metodologia em problemas da área da poluição do ar.

2.3 OBJETIVOS ESPECÍFICOS

1. Estudar empírica e analiticamente as propriedades estatísticas da ACP quando os vetores aleatórios de dados amostrais apresentam autocorrelacionados e e correlação cruzada no tempo;
2. Verificar o impacto de *outliers* aditivos na ACP;

3. Aplicar a ACP como metodologia para gerenciar RAMQAr, por meio da identificação das estações que possuem padrões de comportamento semelhantes.

3 REVISÃO BIBLIOGRÁFICA

A aplicação principal da metodologia proposta nesta dissertação é no gerenciamento de uma rede de monitoramento da qualidade do Ar. A ACP tem sido utilizada na poluição do ar para gerenciamento das estações de uma rede de monitoramento. Nesta linha destacam-se os trabalhos de Dominick et al. (2012), Lu, He e Dong (2011), Lau, Hung e Cheung (2009), Pires et al. (2009) e Pires et al. (2008a), Pires et al. (2008b). Utilizações semelhantes da ACP são encontradas na hidrologia para o gerenciamento de uma rede de monitoramento da qualidade da água. Nesta linha citam-se os trabalhos de Wu et al. (2014), Fang, Chang e Yu (2014) e Gomes et al. (2014).

Em recente estudo, Dominick et al. (2012) utilizaram ACP e análise de cluster (AC) para verificar o a contribuição de fontes e o padrão de comportamento dos poluentes CO, O_3, PM_{10}, SO_2, NO, NO_2 e O_3 em cinco diferentes estações na Malásia. As componentes principais extraídas da concentração de poluentes permitiram associar como fonte poluidoras veículos automotores, aeronaves, industrias e áreas densamente povoadas. Neste artigo, não foi considerada a correlação temporal existentes nos dados.

Lu, He e Dong (2011) empregaram a ACP no gerenciamento da rede de monitoramento da qualidade do ar de Hong Kong para os poluentes de SO_2, NO_2 e Partículas Respiráveis em Suspensão (PRS). Os resultados obtidos pelos autores demostram que para cada poluente as estações de monitoramento são agrupadas em diferentes padrões. Os autores atribuem esse comportamento devido a variabilidade na direção do vento. Além disso, os autores evidenciaram que estações muito próximas são caracterizadas pelo mesmo padrão de poluentes, o que permite realocar os equipamentos para outra região.

Kim et al. (2010) aplicaram a metodologia ACP no arranjo e desenvolvimento de uma redes de monitoramento da qualidade do ar em um sistema de metrô. A metodologia de monitoramento proposta permitiu a redução do número de alarmes falsos e outras falhas.

Pires et al. (2009), Pires et al. (2008a), Pires et al. (2008b) utilizaram a ACP para identificar os locais de monitoramento com comportamento semelhante das concentrações dos poluentes PM_{10}, SO_2, CO, NO_2 e O_3 na região metropolitana da cidade do Porto (Portugal). No primeiro artigo os autores utilizaram dados mensais no período compreendido de janeiro de 2003 a dezembro de 2005. No segundo artigo foram considerados dados trimestrais no período de 2003 a 2004, totalizando oito trimestres. Em ambos os artigos os autores identificaram a necessidade de realocar os equipamentos para outros locais de modo a reduzir custos operacionais e expandir a área de atuação da rede de monitoramento. Em ambos artigos a ACP

foi aplicada desconsiderando a correlação existente nos dados.

Nas ultimas décadas diversos esforços foram empregados no desenvolvimento de metodologias robustas. O *M-estimador* provê estimativas robustas, equivariantes e de baixo custo computacional, entretanto possui baixo ponto de quebra, $\frac{1}{(p+1)}$. Ou seja, o ponto de quebra diminui a medida que se aumenta a dimensão dos dados. Estudos de simulações de Monte Carlo que comparam os diversos *M-estimadores* podem ser encontrados em (DEVLIN; GNANADESIKAN; KETTENRING, 1975) e (DEVLIN; GNANDESIKAN; KETTENRING, 1981).

Stahel (1981) e Donoho e Huber (1983) foram os primeiros a propor um estimador robusto equivariante afim com elevado ponto de quebra, 1/2, para os parâmetros de locação e escala, para qualquer dimensão p. Os parâmetros de locação e escala são, respectivamente, a média ponderada e a dispersão ponderada, onde, a ponderação é realizada de acordo com a medida de *outliers* nos dados. Tyler (1994) discutiu o ponto de quebra para amostras finitas desse estimador.

Outro estimador é o *S-estimator* proposto por (ROUSSEEUW; YOHAI, 1984) cujas a propriedades assintóticas foram obtidas por (DAVIES, 1987). O *S-estimator* é um estimador equivariante afim e, possui elevado ponto de quebra. O *S-estimator* pode ser escrito como um *M-estimator* (ROUSSEEUW; YOHAI, 1984), (LOPUHAA, 1989) e (ROCKE, 1996). Desta forma, possuem distribuição assintóticas semelhantes.

Ma e Genton (2000) propuseram um estimador para a função de autocorrelação de uma série temporal, estendido para o caso multivariado em (MA; GENTON, 2001). A ideia chave desse estimador é a eliminação do parâmetro de locação na estimação do parâmetro escala. De acordo com os autores, esse estimador é uma das melhores escolhas de estimador robusto, uma vez que combina baixa variabilidade e elevado ponto de quebra.

3.1 PPGEA-UFES

Os trabalhos apresentados nesta subseção foram realizados no Programa de Pós-graduação em Engenharia Ambiental da UFES. Esses trabalhos utilizaram, citaram ou definiram como trabalho futuro a utilização de Análise de Componentes Principais.

Zamprogno (2013) mostrou os efeitos da dependência temporal na estimação e inferência na ACP. Os resultados indicaram que cada componente é temporalmente correlacionada, uma propriedade não desejada na ACP.

Souza (2013) utilizou a ACP em um Modelo Aditivo Generalizado para estudo do Risco Relativo (RR) de admissão hospitalar. A remoção da correlação cruzada entre variáveis de poluentes permitiu verificar um aumento no risco relativo.

Soares (2011) utilizou as técnicas de fatoração de matriz positiva(FMP) e balanço químico de massa (BQM) para identificação de fontes poluidoras na região da grande Vitória. O autor apresentou como proposta para trabalho futuro, utilizar a ACP para determinar os elementos mais importantes para ajuste dos modelos, dessa forma, reduzindo a possibilidade de erros na determinação do número de fatores.

Apesar de não utilizar diretamente ACP, Freire (2009) aplicou a correlação canônica em seu trabalho. A correlação canônica é uma técnica estatística que utiliza a matriz de covariância. Assim, é possível robustificar essa matriz e obter resultados mais confiáveis para esse estudo.

Cruz (2005) utilizou Análise Fatorial para análise estatística de desempenho de uma Lagoa de Polimento de Efluente. Neste caso, a ACP foi utilizada como método de estimação das matrizes L e Ψ desse método.

4 CONCEITOS ESTATÍSTICOS

Nesta seção são apresentados os conceitos estatísticos e as notações matriciais utilizadas no decorrer desta dissertação. Para mais detalhes sobre os conceitos básicos e Análise de Componentes Principais consultar Mingoti (2005) e Johnson e Wichern (2007)

4.1 CONCEITOS BÁSICOS

Definição 1. *Um vetor aleatório é um vetor cujos elementos são variáveis aleatórias. Se existirem $X_1, ..., X_p$ variáveis aleatórias em um vetor X, então X é um vetor aleatório denotado por:*

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix}. \tag{1}$$

O vetor transposto do vetor aleatório X é denotador por:

$$X' = [X_1, ..., X_p]. \tag{2}$$

Definição 2. *Seja X um vetor aleatório. O vetor $\mu = E(X)$ é chamado de vetor de médias do vetor $X' = [X_1, ..., X_p]$ definido por:*

$$E(X) = \begin{bmatrix} E(X_1) \\ \vdots \\ E(X_p) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_p \end{bmatrix} = \mu, \tag{3}$$

onde $\mu_i = E(X_i)$ significa a esperança da variável aleatória $X_i, i = 1, 2, ..., p$.

Definição 3. *A variância do i-ésimo componente do vetor X é denotada por $VAR(X_i) =$*

$\sigma_i^2 = \sigma_{ii}, i = 1, ..., p$ *e sua representação matricial é:*

$$\begin{aligned}
\Sigma &= E(X-\mu)(X-\mu)' \\
&= E\left(\begin{bmatrix} (X_1-\mu_1) \\ \vdots \\ (X_p-\mu_p) \end{bmatrix} \begin{bmatrix} X_1-\mu_1, \ldots, X_p-\mu_p \end{bmatrix}\right) \\
&= \begin{bmatrix} E(X_1-\mu_1)^2 & E(X_1-\mu_1)(X_2-\mu_2) & \ldots & E(X_1-\mu_1)(X_p-\mu_p) \\ \vdots & \vdots & \ddots & \vdots \\ E(X_p-\mu_p)(X_1-\mu_1) & E(X_p-\mu_p)(X_2-\mu_2) & \ldots & E(X_p-\mu_p)^2 \end{bmatrix}
\end{aligned} \tag{4}$$

Definição 4. *A covariância entre o i-ésimo e o j-ésimo componente do vetor X é denotada por:*

$$COV(X_i, X_j) = \sigma_{ij} = E[(X_i-\mu_i)(X_j-\mu_j)] \tag{5}$$

Definição 5. *A matriz de variância e covariância do vetor aleatório X é definida como:*

$$\Sigma = COV(X) = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \ldots & \sigma_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \ldots & \sigma_{pp} \end{bmatrix}. \tag{6}$$

Definição 6. *O coeficiente de correlação entre a i-ésima e j-ésima variável do vetor X é:*

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}. \tag{7}$$

Assim, tem-se a matriz de correlação:

$$P_{p\times p} = \begin{bmatrix} 1 & \rho_{12} & \ldots & \rho_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1P} & \rho_{2p} & \ldots & 1 \end{bmatrix} \tag{8}$$

4.2 ESTIMAÇÃO DOS PARÂMETROS POPULACIONAIS

Em situações práticas, é necessário estimar o vetor de médias e as matrizes de covariância e de correlação através de dados amostrais. Então, seja X um vetor aleatório composto por p variáveis com n observações, denotam-se:

Média amostral:

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}, \quad j = 1, 2, ..., p. \tag{9}$$

Variância amostral:

$$s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2, \quad j = 1, 2, ..., p. \tag{10}$$

Covariância amostral:

$$s_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k), \quad j, k = 1, 2, ..., p. \tag{11}$$

Coeficiente de correlação amostral:

$$r_{jk} = \frac{s_{jk}}{\sqrt{jj}\sqrt{s_{kk}}} = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{\sqrt{\sum_{i=1}^{n}(x_{ij} - \bar{x}_i)^2}\sqrt{\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2}}. \tag{12}$$

Para se obter a forma matricial, basta substituir os valores populacionais pelo seu respectivo estimador.

4.3 DETECÇÃO DE *OUTLIERS*

4.3.1 Regra dos três Sigmas

A regra mais popular para determinar se uma observação é um *outlier* é definir como *outlier* observações afastadas da média dos dados por dois ou três desvios padrões (BARNETT; LEWIS, 1984):

$$t_i = \frac{x_i - \bar{x}}{s}, \tag{13}$$

onde $\bar{x}$ é a média amostral e s a variância amostral.

Maronna, Martin e Yohai (2006) apontam as consequências negativas de se utilizar esta regra. Se a amostra for de boa qualidade e de grande quantidade, esperar-se-á declarar algumas variáveis como *outliers* erroneamente. Para amostras de tamanhos menores que 10, pode-se provar que $|t_i| < 3$ sempre. Além disso, outros *outliers* podem mascarar *outliers* menores mas que não foram detectados como observações discrepantes.

4.3.2 Distância de Mahalanobis Populacional

A distância de Mahalanobis (MAHALANOBIS, 1936) é sugerida por muitos textos como um método para detectar *outliers* em dados multivariados. Para um vetor multivariado $x' = [x_1, ..., x_p]$ essa distancia é calculada como:

$$(d_i(x, \mu, \Sigma))^2 = (x_i - \mu_i)'\Sigma^{-1}(x_i - \mu_i), \tag{14}$$

onde, μ_i é o vetor de médias e Σ a matriz de variância e covariança entre os elementos de x.

Se $x \sim N_p(\mu, \Sigma)$, então d_i^2 terá distribuição χ_p^2 (KRZANOWSKI, 1988).

4.3.3 Distancia de Mahalanobis Amostral

A distancia de Mahalanobis amostral é obtida substituindo a matriz Σ por seu estimador S:

$$(d_i(x, \mu, \Sigma))^2 = (x_i - \mu_i)'S^{-1}(x_i - \mu_i). \tag{15}$$

Mardia, Kent e Bibby (1979) apontam que $d_i^2 \sim \frac{p(n-1)}{(n-p)}F_{p,n-p}$. Entretanto, Penny (1996) afirma que como x não é independente de S, esse resultado é equivocado. O correto é $d_i^2 \sim \frac{p(n-1^2F_{p,n-p-1})}{n(n-p-1+pF_{p,n-p-1})}$.

Como a distribuição de d_i^2 é conhecida, é possível analisar graficamente através de um gráfico quantil-quantil. Adicionalmente, pode-se substituir o estimador S por uma versão robusta desse estimador, obtendo-se, assim, um estimador robusto para a distância de Mahalanobis.

4.3.4 Procedimento de Johnson-Wichern

Johnson e Wichern (2007) apresentam um procedimento para se detectar *outliers* multivariados:

1. Plotar um diagrama de pontos para cada variável;
2. Plotar um diagrama de dispersão para todas as combinações de pares entre as variáveis;
3. Calcular o valor padronizado $z_{jk} = \frac{(x_{jk} - \bar{x}_k)}{\sqrt{s_{kk}}}$ para $j = 1, ..., n$ e cada coluna $k = 1, ..., p$;
4. Calcular a distância de Mahalanobis e observar a existência de valores discrepantes.

4.4 ANÁLISE DE COMPONENTES PRINCIPAIS

A ACP é uma importante ferramenta estatística multivariada. A ACP para dados de população e de amostra são apresentadas abaixo. Um texto mais completo pode ser encontrado no capítulo 8 de Johnson e Wichern (2007).

4.4.1 Componentes principais populacional

Seja o vetor aleatório $X' = [X_1, X_2, ..., X_p]$ com matriz de covariância Σ e seus autovalores $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$. Considere a seguinte combinação linear:

$$\begin{aligned} Y_1 &= a_1'X = a_{11}X_1 + a_{12}X_2 + ... + a_{1p}X_p \\ Y_2 &= a_2'X = a_{21}X_1 + a_{22}X_2 + ... + a_{2p}X_p \\ &\vdots \\ Y_p &= a_p'X = a_{p1}X_1 + a_{p2}X_2 + ... + a_{pp}X_p \end{aligned} \tag{16}$$

escrevendo de outra maneira:

$$Y_i = a_iX. \tag{17}$$

Essa combinação linear tem variância e matriz de covariância definidas por:

$$\begin{aligned} VAR(Y_i) &= a_i'\Sigma a_i \\ COV(Y_i, Y_k) &= a_i'\Sigma a_k \end{aligned} \tag{18}$$

As componentes principais são as combinações lineares tais quais o valor $VAR(Y_i)$ seja o maior possível com a restrição de $a_i' a_i = 1$. Ou seja, todos os vetores da combinação linear são ortonormais (norma igual a 1) e o produto interno entre cada par de vetor é igual a 0.

Seja Σ a matriz de covariância do vetor aleatório $\mathrm{X}^{'} = [X_1, X_2, \ldots, X_p]$ com pares de autovalores-autovetores $(\lambda_1, e_1), (\lambda_2, e_2), \ldots, (\lambda_p, e_p)$, onde $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$. Então, a i-ésima componente principal é dada por

$$Y_i = e_i^{'} X = e_{i1} X_1 + e_{i2} X_2 + \cdots + e_{ip} X_p, \quad i = 1, 2, \ldots, p. \tag{19}$$

Tem-se que:

$$\begin{aligned} VAR(Y_i) = e_i^{'} \Sigma e_i = \lambda_i \quad i = 1, 2, \ldots, p \\ COV(Y_i, Y_k) = e_i^{'} \Sigma e_k = 0 \quad i \neq k. \end{aligned} \tag{20}$$

Se alguns λ_i são iguais, a escolha do correspondente vetor de coeficientes e_i e de Y_i, não será única.

Seja Σ a matriz de covariância associada ao vetor aleatório $X^{'} = [X_1, X_2, \ldots, X_p]$ com pares de autovalores-autovetores $(\lambda_1, e_1), (\lambda_2, e_2), ..., (\lambda_p, e_p)$, onde $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$ e $Y_1 = e_1^{'} X, Y_2 = e_2^{'} X, \ldots, Y_p = e_p^{'} X$ as componentes principais. Então:

$$\sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} = \sum_{i=1}^{p} VAR(X_i) = \lambda_1 + \lambda_2 + \cdots + \lambda_p = \sum_{i=1}^{p} VAR(Y_i). \tag{21}$$

Se $Y_1 = e_1^{'} X, Y_2 = e_2^{'} X, \ldots, Y_p = e_p^{'} X$ são as componentes principais obtidas da matriz de covariâncias Σ, então

$$\rho_{Y_i, X_k} = \frac{e_{ik} \sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}} \quad i, k = 1, 2, \ldots, p, \tag{22}$$

são os coeficientes de correlação entre a componente Y_i e a variável X_k. $(\lambda_1, e_1), (\lambda_2, e_2), ..., (\lambda_p, e_p)$ são os pares de autovalores e autovetores de Σ.

4.4.2 Componentes principais amostral

Suponha que os dados $x_1, x_2, ..., x_n$ representem n amostras independentes de uma população p-dimensional com vetor de médias μ e matriz de covariância Σ. Sejam $\bar{x}$ e S, respectivamente, os valores estimados para o vetor de médias e matriz de covariância. Seja $S = \{s_{ik}\}$ a matriz $p \times p$ de covariância amostral com pares de autovalores e autovetores $(\hat{\lambda}_1, \hat{e}_1), (\hat{\lambda}_2, \hat{e}_2), \ldots, (\hat{\lambda}_p, \hat{e}_p)$, a i-ésima componente principal amostral é dada por:

$$\hat{y}_i = e_i^{'} x_i = \hat{e}_{i1} x_1 + \hat{e}_{i2} x_2 + \cdots + \hat{e}_{ip} x_p, \quad i = 1, 2, \ldots, p, \tag{23}$$

onde $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_p \geq 0$ e x_{ij}, $j = 1, ..., p$, é algum valor observado das variáveis $X_1, X_2, \ldots, X_p$.

Além disso:

$$\text{variância amostral}(\hat{y}_k) = \hat{\lambda}_k \quad k = 1, 2, \ldots, p \tag{24}$$

$$\text{covariância amostral}(\hat{y}_i, \hat{y}_k) = 0 \quad i \neq k. \tag{25}$$

$$\tag{26}$$

Adicionalmente:

$$\text{variância amostral total} = \sum_{i=1}^{p} s_{ii} = \hat{\lambda}_1 + \hat{\lambda}_2 + \cdots + \hat{\lambda}_p \tag{27}$$

e

$$r_{\hat{y}_i, \hat{x}_k} = \frac{\hat{e}_{ik} \sqrt{\hat{\lambda}_i}}{\sqrt{s_{kk}}} \quad i, k = 1, 2, \ldots, p. \tag{28}$$

4.5 PONTO DE QUEBRA

Uma das métricas para desempenho de estimadores robustos é o ponto de quebra. O ponto de quebra de um estimador é definido como a porcentagem máxima de observações discrepantes que o conjunto de dados pode possuir, sem que o estimador apresente resultados não confiáveis.

4.6 ESTIMAÇÃO ROBUSTA DA MATRIZ DE COVARIÂNCIA - ESTIMADOR DE MA E GENTON

Seja X um vetor de p variáveis com n observações. O estimador de Ma e Genton (MG) é definido como:

$$\hat{\gamma}(X_i, X_j) = \frac{\alpha\beta}{4}\left[Q_n^2\left(\frac{X_{ik}}{\alpha} + \frac{X_{jk}}{\beta}\right) - Q_n^2\left(\frac{X_{ik}}{\alpha} - \frac{X_{jk}}{\beta}\right)\right], \tag{29}$$

onde $i, j = 1, ..., p$, $k = 1, ..., n$, $\alpha = Q_n(X)$ e $\beta = Q_n(Y)$. A função Q_n é definida como:

$$Q_n(Z) = d|Z_i - Z_j|; i < j, i, j = 1, 2, ..., n, \tag{30}$$

onde $Z = (Z_1, ..., Z_n)^T$ é um vetor aleatório da variável aleatória Z.

Para se obter a matriz de covariância robusta do vetor aleatório X, basta calcular a combinação de todos os pares das p variáveis e dispô-las em forma de matriz.

4.7 ERRO QUADRÁTICO MÉDIO

Uma importante estatística é o erro quadrático médio (EQM).

Definição 7. *O erro quadrático médio de um estimador $\hat{\theta}$ do parâmetro θ é dado por:*

$$EQM[\hat{\theta}] = E[(\hat{\theta} - \theta)^2]. \tag{31}$$

Definição 8. *Um estimador $\hat{\theta}$ é dito ser não viciado para θ se:*

$$E[\hat{\theta}] = \theta. \tag{32}$$

Então, tem-se que:

$$EQM[\hat{\theta}] = VAR[\hat{\theta}]. \tag{33}$$

Seja o vetor aleatório $\mathbf{X}^{'} = [X_1, X_2, \ldots, X_p]$ com matriz de covariância Σ e autovalores $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$ cujo os autovalores estimados são $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_p \geq 0$. O EQM

para os estimadores dos autovalores da matriz Σ é:

$$EQM_{est} = \sum_{i=1}^{p} [\hat{\lambda}_i - \lambda_i], \tag{34}$$

onde *est* denota o nome do estimador utilizado.

O erro quadrático médio é comumente empregado na comparação de estimadores. Diz-se, então, que o melhor estimador é aquele que possuí o menor EQM dentre todos os estimadores utilizados.

5 MATERIAIS E MÉTODOS

5.1 REGIÃO DE ESTUDO

A Região da Grande Vitória (RGV) é constituída pelos municípios de Vitória, Vila Velha, Cariacica, Serra e Viana. A RMGV é localizada na região sudeste do estado do Espírito Santo. Sua área é de 146 Km^2 e possui uma população de aproximadamente 1,7 milhão habitantes, o que representa cerca de 46% da população total do estado (IBGE, 2011).

A região é o principal polo econômico do estado, representando aproximadamente 63,13% do Produto Interno Bruto (PIB) do Espírito Santo. Nessa região encontram-se atividades de siderurgia, pelotização, mineração (pedreiras), cimenteiras, indústria alimentícia, usina de asfalto, etc. No ano de 2007 a RGV possuía 49,18% da frota total do estado (IJSN, 2008).

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 (IEMA, 2005).

5.2 REDE AUTOMÁTICA DE MONITORAMENTO DA QUALIDADE DO AR DA GRANDE VITÓRIA - RAMQAR

A RAMQAr entrou em funcionamento no ano de 2000, sendo de propriedade e responsabilidade do Instituto Estadual de Meio Ambiente e Recursos Hídricos (IEMA). Para o período de estudo, a rede é composta 8 estações de monitoramento distribuídas nos municípios da RMGV seguinte forma: 3 estações no município de Serra, Laranjeiras e Carapina. O município de Vitória possui 3 estações, Jardim Camburi, Enseada do Suá e Centro. Vila Velha 2 estações, Ibes e Centro. O município de Cariacica possui 1 estação, localizada na Ceasa. A localização espacial das estações de monitoramento da RAMQAr encontra-se na Figura 1.

3333

A RAMQAr monitora os seguintes poluentes: dióxido de enxofre,partículas totais em sus-

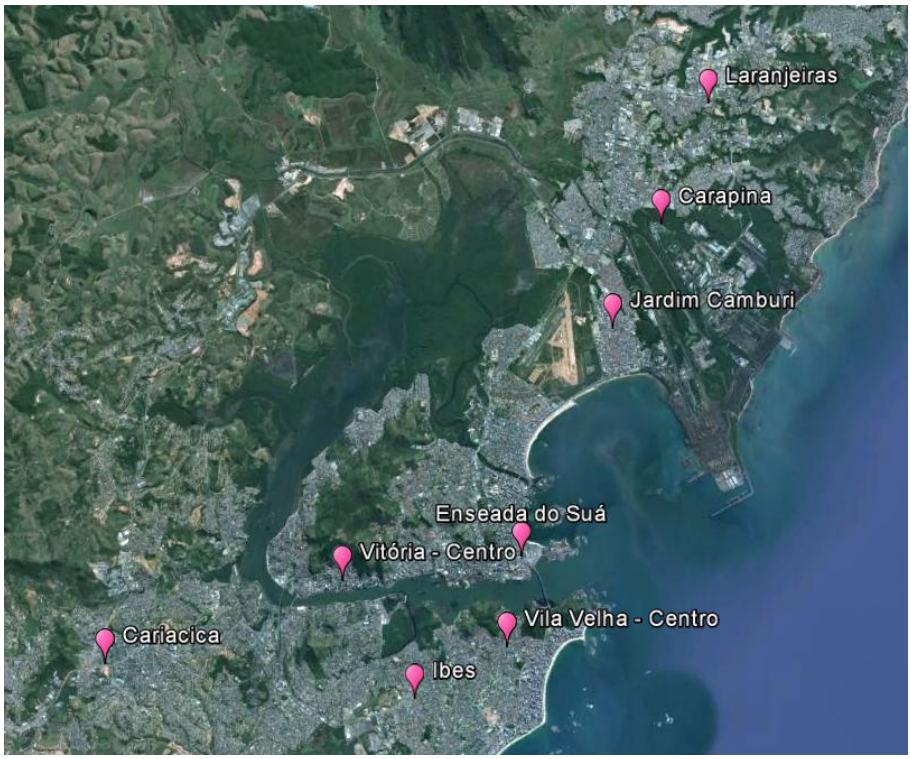


Figura 1: Localização espacial das estações da RAMQAr.

pensão,partículas inaláveis, ozônio, óxidos de nitrogênio, monóxido de carbono e hidrocarbonetos (hc). Além desses poluentes, alguns parâmetros meteorológicos são monitorados: direção dos ventos (DV), velocidade dos ventos (VV), precipitação pluviométrica (PP), umidade relativa do ar (UR), temperatura (T), pressão atmosférica (P) e radiação solar (I). Nem todos os poluentes e parâmetros meteorológicos são monitorados por todas estações. A lista dos poluentes e parâmetros monitorados por cada estação é apresentada na Tabela 1.

Tabela 1: Poluentes e parâmetros meteorológicos em cada estação da RAMQAr.

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

De todas estações, a da Enseada do Suá e a estação do Íbes são as únicas que registram as concentrações de todos os poluentes. A estação de Carapina monitora todos os parâmetros meteorológicos.

5.3 DADOS

Este trabalho foi realizado com os dados de PM_{10} obtidos pela RAMQAr. A concentrações referem-se ao período de janeiro de 2005 a dezembro de 2009, sendo a periodicidade diária e medidas em $\mu g/m^3$.

5.4 *SOFTWARE* ESTATÍSTICO

A metodologia proposta e toda análise efetuada foi realizada por meio do *Software* R. 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. Tem uma grande quantidade de funções para o desenvolvimento de ambiente gráfico e criação de diversos tipos de apresentação de dados (REISEN; SILVA, 2011).

6 RESULTADOS

Anexado nesta seção encontra-se o artigo resultado desta dissertação.

Robust principal component analysis with air pollution data: an application to the clustering of RAMQAr

Higor Henrique Aranda Cotta, Valdério Anselmo Reisen
DEST-CCE, PPGEA-CT – Universidade Federal do Espírito Santo

Abstract

This paper deals with the use of principal component analysis (PCA) in multivariate time series with additive outliers. The PCA is a set of linear combinations constructed to explain the variance-covariance structure of the original data which is one of the most important multivariate techniques. Although PCA assumes that the data are time independent, this assumption is not found in a practice situation in time series, e.g. air pollution data. PCs calculated from time series observations maintain their orthogonality property, but the components are found to be auto and cross-correlated, which depends on the correlation structure of the original series. These properties and their impact in the use of PCA are one of main objective of this work. Another contribution is related to the study of PCA time series under the presence of additive outliers by proposing a Robust PCA method. It is well known that additive outliers in time series destroy the correlation structure of the data. Since the PCs are computed by using the covariance matrix, the outliers also affect the properties of PCs. Therefore the Robust PCA should be used in this context. The Robust PCA method proposed here is justified empirically and theoretically, and a real data set based on Air Pollution time series is used to show the usefulness of the Robust PCA method in a real application.

Keywords: air pollution, principal component analysis, autocorrelation, robustness, eigenvalues.

1 Introduction

The concern about air pollution problems has increased considerably in the last 50 years. Especially in developing countries, the air quality has been degraded as a result of industrialization, population growth, high rates of urbanization, and inadequate or nonexistent policies to control air pollution. The problems caused by air pollution produce local, regional and global impacts.

The main purpose for air quality management is to protect the public health and the environment from the adverse effects of air pollution. An adequate control of air quality involves a number of activities such as risk management, setting standards for emissions and air quality, implementation of control measures and risk communication (WHO 2005). The realization of an efficient management of air quality is essential for to identifying and quantifying the pollutants found in the region and their sources. The monitoring of air quality is an important tool for air pollution control. This is accomplished using stations to monitor different pollutants according to the need of the regions where the stations are installed.

It is desirable that only one monitoring station operates in an area characterized by a specific pattern of air pollution. Pires et al. (2008*a*) indicate that number of stations that constitute a monitoring network must be optimized in order to reduce costs and expenses. If there are stations with similar patterns of pollution for a specific pollutant, the monitoring

equipment could be properly relocated to another area of interest. In this context, the principal component analysis (PCA) has been used in air pollution area for managing a network of monitoring stations in several studies such as Dominick et al. (2012), Lu et al. (2011), Lau et al. (2009), Pires et al. (2009) and Pires et al. (2008*a*). In a recent study, Dominick et al. (2012) used PCA and Cluster Analysis (CA) to check the pattern of behavior of the pollutants carbon monoxide (CO), ozone (O_3), particulate matter of diameter $< 10\mu m$ (PM_{10}), sulphur dioxide (SO_2), nitric oxide (NO) and nitrogen dioxide (NO_2) in five different stations in Malaysia.

Pires et al. (2009, 2008*a*,*b*) applied PCA to identify monitoring sites with similar concentrations of pollutants for PM_{10}, SO_2, CO, NO_2 and O_3 in the metropolitan area of Porto (Portugal). Lu et al. (2011) employed the PCA network management monitoring of air quality in Hong Kong for the pollutants of SO_2, NO_2 and Respirable Suspended Particulate (RSP). The authors found that the monitoring stations located in nearby areas are characterized by the same specific air pollution characteristics and suggested that redundant equipment should be transferred to other monitoring stations allowing for further enlargement of the monitored area.

The principal component analysis is one of the main multivariate statistical techniques. The goal of PCA is to explain the covariance structure of the data by means of auxiliary variables called components. These components are constructed from linear combinations of the original variables and are uncorrelated. Briefly, PCA calculates the eigenvalues and eigenvectors of the covariance or correlation matrix. The main application of PCA is to reduce the dimensionality of a correlated data matrix of n dimension to a m dimension, where $m < n$. The reduction is performed such that the new set of variables captures the same quantity of variability contained in the original data.

The classical theory of PCA is based on data set that is not time dependent. The use of PCA in multivariate time series requires some caution in its application, especially, if more than very weak dependence is present in the series (Jollife 2002). Since the PCs are linear combination of the original variables, the temporal correlation of these series will translate to the PCs. Therefore, neglecting the temporal structure of the observations can lead to totally erroneous inferencial interpretations or even derail the calculation when the components of the multivariate time series are not stationary in the mean.

Furthermore, the PCA is sensitive to outliers due to the fact the very estimation of the mean vector, the covariance matrix and the correlation matrix are directly influenced by outliers. As a consequence the estimation of the eigenvalues and eigenvectors of the covariance or correlation matrix will also suffer influences of outliers present in the data (Filzmoser 1999). Croux & Haesbroeck (2000) indicate that conclusions obtained from principal component analysis calculated from a data set with outliers may be misleading.

Among the methods of robust estimation of the covariance or correlation matrix for time independent data sets, there is an estimator proposed by Ma & Genton (2001). This estimator uses the estimator proposed by $Q_n(.)$ Rousseeuw & Croux (1993) which is independent of the position measurement data set. The robustness and efficiency properties of the estimators have also been investigated through analysis of numerical experiments and real data analysis for the univariate time series. For further details on these theoretical and numerical studies, see Lévy-Leduc et al. (2011*b*).

In this work, the PCA was employed in order to identify pollution behavior of PM_{10} pollutant in the metropolitan area of Vitória (RGV), Brazil, to enable better management of the local monitoring network. The effect of different time correlation structures and additive outliers on a vector linear process are considered and their implications in the analysis and interpretation of the principal components calculated from the correlation matrix of this

process.

Besides the introduction, this paper is organized as follows. In Section 2, the effect of additive outliers in the covariance and correlation matrix function of a multivariate time series is shown. From these results, the impact of multivariate time series and additive outliers in the principal component analysis is derived. In Section 3, the robust estimators of the correlation matrix function (ACF) and covariance matrix function (ACOVF) are suggested. Section 4 presents some Monte Carlo experiments so as to support our theoretical claims. The data obtained from RAMQAr stations are studied as an example of application in Section 5. Some concluding remarks are provided in Section 6.

2 Impact of outliers in multivariate time series and in PCA

2.1 Multivariate time series model

Let $\boldsymbol{X}_t = [X_{1t}, X_{2t}, \ldots, X_{kt}]'$, $t = 0, \pm 1, ...,$ $t \in \mathbb{Z}$ be a *k-dimensional* linear vector process of random variables defined as follows

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

where $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_k]'$ is the mean vector of the process. $\Psi(0)$ is identity matrix of $k \times k$ dimension, $\Psi(j)$, $j = 1, ..., \infty$ are $k \times k$ matrices of coefficients satisfying $\sum_{j=0}^{\infty} ||\Psi(j)||^2 < \infty$, where $||A||$ denotes a norm for the matrix A such as $||A||^2 = tr(A'A)$. $\boldsymbol{\varepsilon}_t = [\varepsilon_{1t}, \ldots, \varepsilon_{kt}]'$ is a vector white noise process such that $\mathbb{E}(\boldsymbol{\varepsilon}_t) = 0$ and covariance matrix

$$\Gamma_{\boldsymbol{\varepsilon}}(h) = \mathbb{C}ov(\boldsymbol{\varepsilon}_t, \boldsymbol{\varepsilon}_{t+h}) = \mathbb{E}(\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_{t+h}^{'}) = \begin{cases} \Sigma_{\boldsymbol{\varepsilon}}, & h = 0, \\ 0, & h \neq 0, \end{cases} \tag{2}$$

where $\Sigma_{\boldsymbol{\varepsilon}}$ is assumed to be positive-definite, see for example, Reinsel (2003). Thus, although the elements of $\boldsymbol{\varepsilon}_j$ at different times are uncorrelated, they may be contemporaneously correlated. By the definition stated above, the process $\{\boldsymbol{X}_t\}$ is second order stationary and has covariance matrix given by

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

where the (i,j)th element of matrix $\Gamma_{\boldsymbol{X}}(h)$ is $\gamma_{ij}(h) = \mathbb{E}[(X_{it} - \mu_i)(X_{j(t+h)} - \mu_j)]$, $i, j = 1, \ldots, k$.

The lag-h correlation matrix function for the vector process $\{\boldsymbol{X}_t\}$ is defined by

$$\boldsymbol{\rho}_{\boldsymbol{X}}(h) = \boldsymbol{D}^{-1/2}\boldsymbol{\Gamma}_{\boldsymbol{X}}(h)\boldsymbol{D}^{-1/2} = [\rho_{ij}(h)], h \geq 0, \tag{3}$$

for $i, j = 1, ..., k$, where $\boldsymbol{D}$ is the diagonal matrix in which the ith diagonal element is the variance of the ith process , i.e, $diag[\gamma_{11}(0), ..., \gamma_{kk}(0)]$. The ith diagonal element of $\boldsymbol{\rho}_{\boldsymbol{X}}(h)$,

$\boldsymbol{\rho}_{ii}(h)$, is the autocorrelation function of ith process, and the (i,j)th off-diagonal element of $\rho_{ij}(h)$ given by

$$\rho_{ij}(h) = \frac{\mathbb{C}ov(X_i, X_j)}{\sqrt{\mathbb{V}ar(X_{ii})\mathbb{V}ar(X_{jj})}} = \frac{\gamma_{ij}(h)}{[\gamma_{ii}(0)\gamma_{jj}(0)]^{1/2}}, \tag{4}$$

represents the cross-correlation function between X_{it} and X_{jt} . For more details, see, for example, Wei (1994).

A parametric class of models belonging to the linear process described by Equation 1 is the Vector Autoregressive Moving Average of p and q orders, VARMA(p,q), stated below

$$\boldsymbol{\Phi}_p(B)(\boldsymbol{X}_t - \boldsymbol{\mu}) = \boldsymbol{\Theta}_q(B)\boldsymbol{\varepsilon}_t, \tag{5}$$

where B is the backshift operator, such as $B\boldsymbol{X}_t = \boldsymbol{X}_{t-1}$. $\boldsymbol{\Phi}_p = I - \sum_{i=1}^{p} \boldsymbol{\Phi}_i B^i$ and $\boldsymbol{\Theta}_q = I - \sum_{i=1}^{q} \boldsymbol{\Theta}_i B^i$ are $k \times k$ of orders p and q, referring to autoregressive and moving average, respectively. $\{\boldsymbol{\varepsilon}_t\}$ is a vector white noite process as stated before. The VARMA(p,q) process is said to be stationary and invertible if the roots of $|\boldsymbol{\Phi}(B)| = 0$ and the roots of $|\boldsymbol{\Theta}(B)| = 0$ all are greater than one in absolute value, respectively.

2.2 Impact of additive outliers in multivariate time series

Additive outliers can affect the dependence structure of a multivariate time series. In this section, some results related to the effects of additive outliers on the covariance and correlation structure of a correlated process are derived. The following definition introduces the model for stationary vector process with additive outliers.

Let $\{\mathbf{Z}_t\}, t = 1, ..., t \in \mathbb{Z}$ be a vector process contaminated by additive outliers defined as follow:

$$\underbrace{\mathbf{Z}_t}_{(k\times 1)} = \underbrace{\boldsymbol{X}_t}_{(k\times 1)} + \underbrace{\boldsymbol{\omega} \circ \boldsymbol{\delta}_t}_{(k\times 1)}, \tag{6}$$

where "$\circ$" is the Hadamard product (Johnson 1989). $\boldsymbol{\omega} = [\omega_1, ..., \omega_k]'$ is a magnitude vector of additive outliers, for simplify, is assumed that $\omega_1 = \omega_2 = ... = \omega_k = \omega$. $\boldsymbol{\delta}_t = [\delta_{1t}, ..., \delta_{kt}]'$ is a random vector indicating the occurrence of an outlier at time t, in variable k, such as $\mathbb{P}(\delta_{kt} = -1) = \mathbb{P}(\delta_{kt} = 1) = p/2$ and $\mathbb{P}(\delta_{kt} = 0) = 1 - p$, where $\mathbb{E}[\delta_{kt}] = 0$ and $\mathbb{E}[\delta_{kt}^2] = \mathbb{V}ar(\delta_{kt}) = p$. 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, $\mathbb{E}(\boldsymbol{\delta}_t\boldsymbol{\delta}_t^{'}) = \Sigma_{\boldsymbol{\delta}} = diag(p, ..., p)$ and $\mathbb{E}(\boldsymbol{\delta}_t\boldsymbol{\delta}_{t+h}^{'}) = 0$ for $h \neq 0$.

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.*

The following proposition elucides the impact of additive outliers on the temporal covariance matrix.

Proposition 1. *Let $\boldsymbol{Z}_t = [Z_{1t}, Z_{2t}, \ldots, Z_{kt}]'$ be a linear vector process defined in Equation 6. Without any loss of generality $\mathbb{E}(\boldsymbol{X}_t) = 0$ is assumed. Then,*

$$\mathbb{E}[\boldsymbol{Z}_t] = \boldsymbol{\mu} = 0, \tag{7}$$

$$\Gamma_{\boldsymbol{Z}}(h) = \mathbb{C}ov(\boldsymbol{Z}_t, \boldsymbol{Z}_{t+h}) = \begin{cases} \Gamma_{\boldsymbol{X}}(0) + \Gamma_{\boldsymbol{\omega}\circ\boldsymbol{\delta}}(0), & h = 0, \\ \Gamma_{\boldsymbol{X}}(h), & h \neq 0, \end{cases} \tag{8}$$

where $\Gamma_{\boldsymbol{\omega}\circ\boldsymbol{\delta}}(0) = \omega\Sigma_{\boldsymbol{\delta}}\omega'$.

Proof.

$$\mathbb{E}[\boldsymbol{Z}_t] = \mathbb{E}[\boldsymbol{X}_t + \boldsymbol{\omega} \circ \boldsymbol{\delta}_t] = \mathbb{E}[\boldsymbol{X}_t] + \mathbb{E}[\boldsymbol{\omega} \circ \boldsymbol{\delta}_t] = \boldsymbol{\mu} + \boldsymbol{\omega} \circ E[\boldsymbol{\delta}_t] = \boldsymbol{\mu} = 0. \tag{9}$$

$$\begin{aligned} \Gamma_{\boldsymbol{Z}}(h) &= \mathbb{C}ov(\mathbf{Z}_t, \mathbf{Z}_{t+h}) \\ &= \mathbb{C}ov(\boldsymbol{X}_t + \boldsymbol{\omega} \circ \boldsymbol{\delta}_t, (\boldsymbol{X}_{t+h} + \boldsymbol{\omega} \circ \boldsymbol{\delta}_{t+h})) \\ &= \mathbb{C}ov(\boldsymbol{X}_t, \boldsymbol{X}_{t+h}) + \mathbb{C}ov(\boldsymbol{X}_t, (\boldsymbol{\omega} \circ \boldsymbol{\delta}_{t+h})) + \mathbb{C}ov(\boldsymbol{\omega} \circ \boldsymbol{\delta}_t, \boldsymbol{X}_{t+h}) \\ &+ \mathbb{C}ov(\boldsymbol{\omega} \circ \boldsymbol{\delta}_t, (\boldsymbol{\omega} \circ \boldsymbol{\delta}_{t+h})) \\ &= \begin{cases} \Gamma_{\boldsymbol{X}}(0) + \Gamma_{\boldsymbol{\omega}\circ\boldsymbol{\delta}}(0), & h = 0 \\ \Gamma_{\boldsymbol{X}}(h), & h \neq 0 \end{cases}, \end{aligned} \tag{10}$$

$\mathbb{C}ov(\boldsymbol{X}_t, (\boldsymbol{\omega}_{t+h} \circ \boldsymbol{\delta}_{t+h})) = 0$ and $\mathbb{C}ov(\boldsymbol{\omega}_t \circ \boldsymbol{\delta}_t, \boldsymbol{X}_{t+h}) = 0$ according to model suppositions. □

Note that, according to Definition 6, the occurrence of additive outliers is random, which is a plausible assumption in a practical situation.

Remark 2. *Proposition 1 states that the variability increases due to the presence of additive outliers in the vector process, which is an expected result. Now, consider the lag-h correlation matrix function for the vector process contaminated by additive outliers $\{\boldsymbol{Z}_t\}$. Since the correlation matrix function given by Equation 4 depends on $\mathbb{V}ar(X_{ii})$ and $\mathbb{V}ar(X_{jj})$, Therefore, the increase of the variance will lead to a decreasing of the correlation. The recent works of Molinares et al. (2009), Lévy-Leduc et al. (2011*b*) and Lévy-Leduc et al. (2011*a*) discuss these in univariate time series with short and long memory properties.*

2.3 PCA in multivariate time series with additive outliers

As described in the introduction, additive outliers can affect the variability structure of a vector time serie and, as a consequence, the PCA will also be affected, since it is a linear combination of the original vector. Returning to the model given by Equation 6, the next proposition shows the effects of time correlation and additive outliers in the PCA.

Proposition 2. *Let $\boldsymbol{Z}_t = [Z_{1t}, Z_{2t}, \ldots, Z_{kt}]'$ be a vector process contaminated by outliers defined in Equation 6, with covariance matrix $\Gamma_{\boldsymbol{Z}}(h), h = 1, ..., h < n$, given by Proposition 1. Let $\Gamma_{\boldsymbol{Z}}(0)$ have the eigenvalue-eigenvector pairs $(\lambda_1,\boldsymbol{\varepsilon}_1), ..., (\lambda_k,\boldsymbol{\varepsilon}_k)$ where $\lambda_1 \geq, ..., \geq \lambda_k$ Then the ith principal component is given by*

$$Y_{it} = \boldsymbol{e}_i^{'}\boldsymbol{Z}_t = \boldsymbol{e}_{i1}Z_{1t} + \boldsymbol{e}_{i2}Z_{2t} + \cdots + \boldsymbol{e}_{ik}Z_{kt}, \quad i = 1, 2, \ldots, k, \quad t \in \mathbb{Z}, \tag{11}$$

and, for $h = 0$,

$$a)\mathbb{V}ar(Y_{it}) = \boldsymbol{e}_i^{'}\Gamma_{\boldsymbol{Z}}(0)\boldsymbol{e}_i = \boldsymbol{e}_i^{'}(\Gamma_{\boldsymbol{X}}(0) + \Gamma_{\boldsymbol{\omega}\circ\boldsymbol{\delta}}(0))\boldsymbol{e}_i = \lambda_i, \qquad i = 1, 2, \ldots, k \quad (12)$$

$$b)\mathbb{C}ov(Y_{it}, Y_{jt}) = \boldsymbol{e}_i^{'}\Gamma_{\boldsymbol{Z}}(0)\boldsymbol{e}_j = \boldsymbol{e}_i^{'}(\Gamma_{\boldsymbol{X}}(0) + \Gamma_{\boldsymbol{\omega}\circ\boldsymbol{\delta}}(0))\boldsymbol{e}_j = 0, \qquad i \neq j. \quad (13)$$

for $h \neq 0$,

$$c)\mathbb{C}ov(Y_{it}, Y_{i(t+h)}) = \boldsymbol{e}_i^{'}\Gamma_{\boldsymbol{Z}}(h)\boldsymbol{e}_i = \boldsymbol{e}_i^{'}\Gamma_{\boldsymbol{X}}(h)\boldsymbol{e}_i, \qquad i = 1, 2, \ldots, k. \quad (14)$$

$$d)\mathbb{C}ov(Y_{it}, Y_{j(t+h)}) = \boldsymbol{e}_i^{'}\Gamma_{\boldsymbol{Z}}(h)\boldsymbol{e}_j = \boldsymbol{e}_i^{'}\Gamma_{\boldsymbol{X}}(h)\boldsymbol{e}_j, \qquad i \neq j. \quad (15)$$

If some λ_i *are equal, the choices of the corresponding coefficient vectors,* $\boldsymbol{\varepsilon}_i$*, and hence* Y_{it} *are not unique.*

Proof. For a) and b), let $h = 0$.

a)Consider the following maximization of a quadratic form

$$\max_{\boldsymbol{e}_1 \neq 0} \frac{\boldsymbol{e}_1'\Gamma_{\boldsymbol{Z}}(0)\boldsymbol{e}_1}{\boldsymbol{e}_1'\boldsymbol{e}_1} = \lambda_1 \; and \; \max_{\boldsymbol{e}_1 \perp, \ldots, \boldsymbol{e}_k} \frac{\boldsymbol{e}_{k+1}'\Gamma_{\boldsymbol{Z}}(0)\boldsymbol{e}_{k+1}}{\boldsymbol{e}_{k+1}'\boldsymbol{e}_{k+1}} = \lambda_{k+1}, \quad (16)$$

where, the symbol $\perp$ means "is perpendicular to". The proof of Equation 16 is given in Johnson & Wichern (2013, p. 80). But, $\boldsymbol{e}_i'\boldsymbol{e}_i = 1$, $i = 1, ..., k$, since the eigenvectors are normalized. Thus,

$$\max_{\boldsymbol{e}_i \neq 0} \frac{\boldsymbol{e}_i'\Gamma_{\boldsymbol{Z}}(0)\boldsymbol{e}_i}{\boldsymbol{e}_i'\boldsymbol{e}_i} = \boldsymbol{e}_i'\Gamma_{\boldsymbol{Z}}(0)\boldsymbol{e}_i = \lambda_i = \mathbb{V}ar(Y_{it}) \; i = 1, ..., k. \quad (17)$$

b)It is known that $\mathbb{C}ov(Y_i, Y_j) = \boldsymbol{e}_i'\Gamma_{\boldsymbol{Z}}(0)\boldsymbol{e}_j$ and $\boldsymbol{e}_i'\boldsymbol{e}_j = 0$, for any $i \neq j$. Then,

$$\mathbb{C}ov(Y_{it}, Y_{jt}) = \boldsymbol{e}_i'\Gamma_{\boldsymbol{Z}}(0)\boldsymbol{e}_j = \boldsymbol{e}_i'\lambda_j\boldsymbol{e}_j = \lambda_j\boldsymbol{e}_i'\boldsymbol{e}_j = 0. \quad (18)$$

For c) and d), let $h \neq 0$.

c)Let in Equation 18 $i = j$, then

$$\mathbb{C}ov(Y_{it}, Y_{i(t+h)}) = \boldsymbol{e}_i'\Gamma_{\boldsymbol{Z}}(h)\boldsymbol{e}_i = \boldsymbol{e}_i^{'}\Gamma_{\boldsymbol{X}}(h)\boldsymbol{e}_i, \quad i = 1, 2, \ldots, k. \quad (19)$$

d)Follows from Equation 19 when $i \neq j$.

□

In the above proposition, the item *a*) and *b*) show that the properties of the time independent case are preserved when computing the PCs from time series data, whereas the item *c*) and *d*) show that the PCs are time correlated, which is, indeed, an undesirable property.

The theoretical results presented above are empirically studied in Section 4.

3 Robust estimation of the covariance and correlation matrix functions

The estimation of the covariance matrix is the central point in the PCA tool. Rousseeuw & Croux (1993) proposed a robust estimator of the dispersion for one random variable X. The method is summarized as follows

Let $x_1, ..., x_n$ be a sample of X. The $Q_n(.)$ estimator is defined by

$$Q_n(X) = c\ \{|x_i - x_j|; i < j\}_{\{k\}}, \quad i, j = 1, .., n, \tag{20}$$

where c is a constant to guarantee consistency (c = 2.2191 for the gaussian distribution) and $k = \lfloor (\binom{n}{r}+2)/4 \rfloor + 1$, which corresponds the kth order statistic of $\binom{n}{2}$ distances $\{|x_i - x_j|, i < j\}_{\{k\}}\}$. $\lfloor . \rfloor$ denotes the integer part. Croux & Rousseeuw (1992) proposed a computationally efficient algorithm to calculate the $Q_n(.)$ function. Rousseeuw & Croux (1993) showed that the asymptotic breakdown point of $Q_n(.)$ is 50% when X_i, $i = 1, ...n$, are independent random variables.

The covariance between two random variables X and Y may be obtained from the following identity

$$\mathbb{C}ov(X, Y) = \frac{\alpha\beta}{4}\left[\mathbb{V}ar\left(\frac{X}{\alpha} + \frac{Y}{\beta}\right) - \mathbb{V}ar\left(\frac{X}{\alpha} - \frac{Y}{\beta}\right)\right], \tag{21}$$

where $\alpha = \frac{1}{\sqrt{\mathbb{V}ar(X)}}$ and $\beta = \frac{1}{\sqrt{\mathbb{V}ar(Y)}}$ (Huber 2004).

Now, let X_t, $t \in \mathbb{Z}$, a time series with $\mathbb{V}ar(X_t) = \gamma_{X_t}(0)$. Let $\alpha = \beta = 1/\sqrt{\gamma_{X_t}(0)}$, based on Equation 21 and replacing the $\mathbb{V}ar(.)$ by $Q_n^2(.)$, Ma & Genton (2000) proposed the following highly robust ACOVF for single time series

$$\hat{\gamma}_{Q_n}(h, X_t) = \frac{1}{4}\left[Q_{n-h}^2(U + V) - Q_{n-h}^2(U - V)\right], \tag{22}$$

where U and V are vectors containing the initial $n - h$ and the final $n - h$ observations of the single time series X_t, respectively. Then, the autocorrelation function can be obtained from

$$\hat{\rho}_{Q_n}(h, X_t) = \frac{Q_{n-h}^2(U + V) - Q_{n-h}^2(U - V)}{Q_{n-h}^2(U + V) + Q_{n-h}^2(U - V)}, \tag{23}$$

where U and V are also vectors containing the initial $n - h$ and the final $n - h$ observations of X_t.

The asymptotical results of the above robust autocovariance in time series with short and long memory properties were the motivation for the papers of Lévy-Leduc et al. (2011*b*) and Lévy-Leduc et al. (2011*a*). Theorem 4 in Lévy-Leduc et al. (2011*b*) presents the central limit theorem for the autocorrelation given by 23. The theoretical results given in Lévy-Leduc et al. (2011*b*) can be extended to a multivariate time series context. However, this is still an open problem. This work is mainly centered around to introducing the multivariate robust ACF and ACOVF, to measure its robustness against additive outliers and to use those suggested functions to compute robust principal components in multivariate time models.

In this work, the estimators of the robust covariance and correlation matrices proposed by Ma & Genton (2001) are extended to the multivariate time series, i.e, the covariance and correlation matrix functions, respectively.

Based from the univariate case given by Equation 22 and $\hat{\gamma}_{Q_n}(.)$ matrix from Ma & Genton (2001), the following robust estimator of covariance matrix function of a vector process $\boldsymbol{X}_t$, $t = 1, ..., n$, is suggested

$$\hat{\Gamma}_{\boldsymbol{X}_t, Q_n}(h) = \begin{bmatrix} \hat{\gamma}_{Q_{n-h}}(X_{1t}, X_{1t}) & \hat{\gamma}_{Q_{n-h}}(X_{1t}, X_{2t}) & \dots & \hat{\gamma}_{Q_{n-h}}(X_{1t}, X_{kt}) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\gamma}_{Q_{n-h}}(X_{kt}, X_{1t}) & \hat{\gamma}_{Q_{n-h}}(X_{kt}, X_{2t}) & \dots & \hat{\gamma}_{Q_{n-h}}(X_{kt}, X_{kt}) \end{bmatrix}, \tag{24}$$

where $\hat{\gamma}_{Q_{n-h}}(X_{it}, X_{jt})$ is estimated from

$$\hat{\gamma}_{Q_{n-h}}(X_{it}, X_{jt}) = \frac{\alpha\beta}{4}\left[Q^2_{n-h}\left(\frac{U}{\alpha}+\frac{V}{\beta}\right) - Q^2_{n-h}\left(\frac{U}{\alpha}-\frac{V}{\beta}\right)\right], i, j = 1, ..., k. \tag{25}$$

In the above, U corresponds to the first $n-h$ observations from X_{it} and V corresponds to the last $n-h$ from X_{jt}. $\alpha = Q_n(X_{it})$ and $\beta = Q_n(X_{jt})$. For $\hat{\gamma}_{Q_{n-h}}(X_{it}, X_{jt})$, when $i = j$, the robust cross-covariance becomes the robust ACOVF given by Equation 22.

The robust correlation matrix function is suggested below

$$\hat{\boldsymbol{\rho}}_{\boldsymbol{X}_t, Q_n}(h) = \begin{bmatrix} \hat{\rho}_{Q_{n-h}}(X_{1t}, X_{1t}) & \hat{\rho}_{Q_{n-h}}(X_{1t}, X_{2t}) & \dots & \hat{\rho}_{Q_{n-h}}(X_{1t}, X_{kt}) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\rho}_{Q_{n-h}}(X_{kt}, X_{1t}) & \hat{\rho}_{Q_{n-h}}(X_{kt}, X_{2t}) & \dots & \hat{\rho}_{Q_{n-h}}(X_{kt}, X_{kt}) \end{bmatrix}, \tag{26}$$

where $\hat{\rho}_{Q_{n-h}}(X_{it}, X_{jt})$ is estimated from

$$\hat{\rho}_{Q_{n-h}}(X_{it}, X_{jt}) = \frac{Q^2_{n-h}\left(\frac{U}{\alpha}+\frac{V}{\beta}\right) - Q^2_{n-h}\left(\frac{U}{\alpha}-\frac{V}{\beta}\right)}{Q^2_{n-h}\left(\frac{U}{\alpha}+\frac{V}{\beta}\right) + Q^2_{n-h}\left(\frac{U}{\alpha}-\frac{V}{\beta}\right)}, i, j = 1, ..., k. \tag{27}$$

U and V are obtained in similar way as in Equation 25. For $\hat{\rho}_{Q_{n-h}}(X_{it}, X_{jt})$, when $i = j$, the robust cross-correlation becomes the robust ACF given by Equation 23.

Molinares et al. (2009) considered $\hat{\Gamma}_{\boldsymbol{X}_t, Q_n}(h)$ to estimate the robust ACOVF of X_t when the series has a long-memory property. In the case of periodic process, Sarnaglia et al. (2010) suggests a robust estimation procedure for the parameters of the periodic AR (PAR) models when the data contains additive outliers.

4 Monte Carlo Studies

This section reports the results of several Monte Carlo experiments to analyze the effect of additive outliers and of time correlated observations on the PCs and theirs corresponding eigenvalues and eigenvectors. In addition, the performance of the robust ACOVF and ACF matrices, discussed previously, are investigated under different scenarios.

Now, let the vector $\boldsymbol{X}_t$ generated by Equation 5 with $p = 1$, $q = 0$. That is, $\boldsymbol{X}_t$ follows a VAR(1) process. $\boldsymbol{\varepsilon}_t$ is a vector white noite process with mean 0 and covariance matrix given by

$$\Gamma_{\boldsymbol{\varepsilon}}(0) = E(\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t^{'}) = \begin{bmatrix} 127.4089 & 30.5878 & 47.4390 \\ 30.5878 & 58.7881 & 33.8929 \\ 47.4390 & 33.8929 & 64.1786 \end{bmatrix}.$$

In the simulations, the VAR(1) models are: Model 1 is a vector white noise process and for Models 2, 3 and 4, the Φ coefficients of the $\boldsymbol{X}_t$ processes are presented in Table 1. Note that Model 2 corresponds a process with no temporal correlation off the diagonal, Models 3 and 4 present a more complex correlation structure, with values outside the diagonal and strong correlation.

The $\boldsymbol{Z}_t$ models were simulated in accord to Equation 6 with $\boldsymbol{\delta}_t = [p_1, p_2, p_3]'$ and $\boldsymbol{\omega} = [4\sigma_1, 4\sigma_2, 4\sigma_3]'$, where $\sigma_i = \sqrt{\mathbb{V}ar(X_{it})}, i = 1, 2, 3$. For each model, 1000 $\boldsymbol{Z}_t$ processes of size $n = 500$ were generated. The p_i and σ_i, for $i = 1, 2, 3$, values are given in the tables. Note that $\boldsymbol{Z}_t \equiv \boldsymbol{X}_t$, when $p_i = 0$, $\forall\, i$. The covariance matrices $\Gamma_{\boldsymbol{Z}}(h)$, $h < n$, were estimated using the proposed robust estimator $\widehat{\Gamma}_{\boldsymbol{Z},Q_n}(h)$ and the classical estimator of ACOVF matrix, denoted here by $\widehat{\Gamma}_{\boldsymbol{Z}}(h)$.

Table 1: Φ matrices for VAR(1) process.

Φ_1 (**Model 2**)			Φ_1 (**Model 3**)			Φ_1 (**Model 4**)		
0.5	0.0	0.0	0.2	0.0	0.6	0.6	0.3	0.0
0.0	0.5	0.0	0.0	0.3	0.0	0.1	0.2	0.0
0.0	0.0	0.5	0.2	0.0	0.5	0.1	0.8	0.4

Table 2 shows the empirical mean of $\widehat{\Gamma}_{\boldsymbol{Z}}(0)$ when $\boldsymbol{\delta}_t = [p_1, 0, 0]'$, $p_1 = 0.01, 0.05$ and 0.15, and $\boldsymbol{\omega} = [4\sigma_1, 0, 0]'$. In the rows of $p = 0$, it can be seen that the presence of time correlation in the vector inflate the variability of $\widehat{\Gamma}_{\boldsymbol{Z}}(0)$. When $p \neq 0$, the variability also inflates. In this situation the expected number of outliers are 5, 25 and 75, for 1% ,5% and 15%, respectively. These empirical evidences are in accordance with the results discussed in Proposition 1. Note that, the increasing of the variability is more expressive for a stronger weak dependence vector process, for example, Models 3 and 4.

The simulation results of the robust covariance matrix estimator $\widehat{\Gamma}_{\boldsymbol{Z},Q_n}(0)$ are displayed in Table 3. $\widehat{\Gamma}_{\boldsymbol{Z},Q_n}(0)$ presented similar estimates of $\widehat{\Gamma}_{\boldsymbol{Z}}(0)$ when the vector is outlier free ($p_1 = 0$), which is also in accordance with the asymptotical behavior of the robust autocovariance when dealing with a univariate outlier free time series given in Lévy-Leduc et al. (2011*b*). When the serie has outliers, the robust estimator performs well, especially, when $p_1 = 0.01$. The percentage of outliers in only one vector seems to be, in general, not strong enough to destroy the robustness of properties of $\widehat{\Gamma}_{\boldsymbol{Z},Q_n}(0)$, which is an important empirical evidence.

Now, let $\boldsymbol{\delta}_t = [p_1, p_2, p_3]'$, $p_1 = p_2 = p_3 = p$, and $\boldsymbol{\omega} = [4\sigma_1, 4\sigma_2, 4\sigma_3]'$. The expected number of outliers are 15, 75 and 225 for $p = 0.01, 0.05, 0.15$, respectively, which is in such way a large amount of atypical observations in the matrix. The simulation results of $\widehat{\Gamma}_{\boldsymbol{Z},Q_n}(0)$ are given in Table 4. The performance of $\widehat{\Gamma}_{\boldsymbol{Z},Q_n}(0)$ slightly changes when more than one vector contains outliers. Even with this large amount of outliers, the robust ACOVF estimates seems to be, in general, quite robust.

Figure 1 displays the plots of ACF and robust ACF for the Model 4, where, Figures 1(a) and 1(b) and Figure 1(c) give the classical and the robust covariance functions, respectively. Figure 1(a) shows the results of $\widehat{\boldsymbol{\rho}}_{\boldsymbol{Z}}(h)$, when $\boldsymbol{Z}_t \equiv \boldsymbol{X}_t$. The effect of outliers in $\widehat{\boldsymbol{\rho}}_{\boldsymbol{X}}(h)$ is virtually observed in Figure 1(b), where in this, the values of $\widehat{\boldsymbol{\rho}}_{\boldsymbol{Z}}(h)$ are reduced compared to the ones in Figure 1(a). Note that, in the Figure 1(b), the scale of $y - axis$ is much smaller compared with Figure 1(a). The reduction of the ACF is also observed in $\widehat{\rho}_{1j}(h)$, for $j = 2, 3$. This is an also expected result. Note that, since the outliers occurs only in Z_{1t}, $\widehat{\rho}_{ij}(h)$ for $i = j = 2, 3$, maintains similar performance of the case without outliers.

Figure 1(c), where the results are related to $\widehat{\boldsymbol{\rho}}_{\boldsymbol{Z},Q_n}(h)$, shows higher accuracy compared to Figure 1(b). The behaviour of the $\widehat{\boldsymbol{\rho}}_{\boldsymbol{Z},Q_n}(h)$ in Figure 1(c) is similar to the Figure 1(a). In addition, the values of the ACF and the cross-correlation of Z_{2t} and Z_{3t} are not far from those plots in Figure 1(a), since they are outlier free processes. Therefore, the percentage of outliers used in this empirical exercise was not strong enough to destroy the robustness properties of the proposed multivariate robust ACF matrix, which corroborates with the results given in Table 3.

Table 2: $\widehat{\Gamma}_{\boldsymbol{Z}}(0)$ matrices for VAR(1) process, $\boldsymbol{\delta}_t = [p_1, 0, 0]'$ and $\boldsymbol{\omega} = [4\sigma_1, 0, 0]'$.

	$\widehat{\Gamma}_{\boldsymbol{Z}}(0)$ (**Model 2**)			$\widehat{\Gamma}_{\boldsymbol{Z}}(0)$ (**Model 3**)			$\widehat{\Gamma}_{\boldsymbol{Z}}(0)$ (**Model 4**)		
$p_1 = 0$	167.41	40.17	62.24	152.92	36.07	83.83	238.38	57.40	149.96
	40.17	77.91	44.44	36.07	64.34	41.88	57.40	66.29	63.41
	62.24	44.44	84.67	83.83	41.88	115.61	149.96	63.41	202.71
	$\widehat{\Gamma}_{\boldsymbol{Z}}(0)$ (**Model 2**)			$\widehat{\Gamma}_{\boldsymbol{Z}}(0)$ (**Model 3**)			$\widehat{\Gamma}_{\boldsymbol{Z}}(0)$ (**Model 4**)		
$p_1 = 0.01$	187.69	40.06	62.07	172.86	36.08	83.84	259.10	57.40	150.02
	40.06	77.91	44.44	36.08	64.34	41.88	57.40	66.29	63.41
	62.07	44.44	84.67	83.84	41.88	115.61	150.02	63.41	202.71
	$\widehat{\Gamma}_{\boldsymbol{Z}}(0)$ (**Model 2**)			$\widehat{\Gamma}_{\boldsymbol{Z}}(0)$ (**Model 3**)			$\widehat{\Gamma}_{\boldsymbol{Z}}(0)$ (**Model 4**)		
$p = 0.05$	267.92	40.18	62.04	253.57	36.09	83.78	340.96	57.47	150.12
	40.18	77.91	44.44	36.09	64.34	41.88	57.47	66.29	63.41
	62.04	44.44	84.67	83.78	41.88	115.61	150.12	63.41	202.71
	$\widehat{\Gamma}_{\boldsymbol{Z}}(0)$ (**Model 2**)			$\widehat{\Gamma}_{\boldsymbol{Z}}(0)$ (**Model 3**)			$\widehat{\Gamma}_{\boldsymbol{Z}}(0)$ (**Model 4**)		
$p_1 = 0.15$	475.42	39.89	62.31	459.75	35.82	83.36	543.32	57.44	150.20
	39.89	77.91	44.44	35.82	64.34	41.88	57.44	66.29	63.41
	62.31	44.44	84.67	83.36	41.88	115.61	150.20	63.41	202.71

Table 3: $\widehat{\Gamma}_{\boldsymbol{Z},Q_n}(0)$ matrices for VAR(1) process, $\boldsymbol{\delta}_t = [p_1, 0, 0]'$ and $\boldsymbol{\omega} = [4\sigma_1, 0, 0]'$.

	$\widehat{\Gamma}_{\boldsymbol{Z},Q_n}(0)$ (**Model 2**)			$\widehat{\Gamma}_{\boldsymbol{Z},Q_n}(0)$ (**Model 3**)			$\widehat{\Gamma}_{\boldsymbol{Z},Q_n}(0)$ (**Model 4**)		
$p_1 = 0.0$	167.53	40.26	62.24	153.68	36.00	84.22	238.68	57.37	150.06
	40.26	78.11	44.49	36.00	64.26	41.83	57.37	66.43	63.49
	62.24	44.49	84.57	84.22	41.83	115.88	150.06	63.49	203.00
	$\widehat{\Gamma}_{\boldsymbol{Z},Q_n}(0)$ (**Model 2**)			$\widehat{\Gamma}_{\boldsymbol{Z},Q_n}(0)$ (**Model 3**)			$\widehat{\Gamma}_{\boldsymbol{Z},Q_n}(0)$ (**Model 4**)		
$p_1 = 0.01$	174.32	40.92	63.41	159.69	36.63	85.79	247.52	58.06	152.03
	40.92	78.11	44.49	36.63	64.26	41.83	58.06	66.43	63.49
	63.41	44.49	84.57	85.79	41.83	115.88	152.03	63.49	203.00
	$\widehat{\Gamma}_{\boldsymbol{Z},Q_n}(0)$ (**Model 2**)			$\widehat{\Gamma}_{\boldsymbol{Z},Q_n}(0)$ (**Model 3**)			$\widehat{\Gamma}_{\boldsymbol{Z},Q_n}(0)$ (**Model 4**)		
$p_1 = 0.05$	204.33	43.52	67.48	187.93	39.42	92.73	287.48	60.61	159.86
	43.52	78.11	44.49	39.42	64.26	41.83	60.61	66.43	63.49
	67.48	44.49	84.57	92.73	41.83	115.88	159.86	63.49	203.00
	$\widehat{\Gamma}_{\boldsymbol{Z},Q_n}(0)$ (**Model 2**)			$\widehat{\Gamma}_{\boldsymbol{Z},Q_n}(0)$ (**Model 3**)			$\widehat{\Gamma}_{\boldsymbol{Z},Q_n}(0)$ (**Model 4**)		
$p_1 = 0.15$	310.16	46.38	72.78	286.90	41.81	101.19	414.34	62.43	167.38
	46.38	78.11	44.49	41.81	64.26	41.83	62.43	66.43	63.49
	72.78	44.49	84.57	101.19	41.83	115.88	167.38	63.49	203.00

Table 4: $\widehat{\Gamma}_{\boldsymbol{Z},Q_n}(0)$ matrices for VAR(1) process, $\boldsymbol{\delta}_t = [p_1, p_2, p_3]', p_1 = p_2 = p_3 = p$, and $\boldsymbol{\omega} = [4\sigma_1, 4\sigma_2, 4\sigma_3]'$.

	$\widehat{\Gamma}_{\boldsymbol{Z},Q_n}(0)$ (**Model 2**)			$\widehat{\Gamma}_{\boldsymbol{Z},Q_n}(0)$ (**Model 3**)			$\widehat{\Gamma}_{\boldsymbol{Z},Q_n}(0)$ (**Model 4**)		
	167.53	40.26	62.24	153.68	36.00	84.22	238.68	57.37	150.06
$p = 0.0$	40.26	78.11	44.49	36.00	64.26	41.83	57.37	66.43	63.49
	62.24	44.49	84.57	84.22	41.83	115.88	150.06	63.49	203.00
	$\widehat{\Gamma}_{\boldsymbol{Z},Q_n}(0)$ (**Model 2**)			$\widehat{\Gamma}_{\boldsymbol{Z},Q_n}(0)$ (**Model 3**)			$\widehat{\Gamma}_{\boldsymbol{Z},Q_n}(0)$ (**Model 4**)		
	174.32	41.69	64.70	159.69	37.43	86.98	247.52	59.34	153.23
$p = 0.01$	41.69	81.26	46.25	37.43	67.00	43.31	59.34	69.17	65.40
	64.70	46.25	88.03	86.98	43.31	120.18	153.23	65.40	209.01
	$\widehat{\Gamma}_{\boldsymbol{Z},Q_n}(0)$ (**Model 2**)			$\widehat{\Gamma}_{\boldsymbol{Z},Q_n}(0)$ (**Model 3**)			$\widehat{\Gamma}_{\boldsymbol{Z},Q_n}(0)$ (**Model 4**)		
	204.33	47.51	73.81	187.93	43.94	99.10	287.48	67.17	165.23
$p = 0.05$	47.51	95.52	53.10	43.94	79.08	48.91	67.17	81.36	72.29
	73.81	53.10	103.34	99.10	48.91	139.35	165.23	72.29	234.72
	$\widehat{\Gamma}_{\boldsymbol{Z},Q_n}(0)$ (**Model 2**)			$\widehat{\Gamma}_{\boldsymbol{Z},Q_n}(0)$ (**Model 3**)			$\widehat{\Gamma}_{\boldsymbol{Z},Q_n}(0)$ (**Model 4**)		
	310.16	58.15	93.82	286.90	56.42	123.01	414.34	80.18	181.81
$p = 0.15$	58.15	143.54	67.41	56.42	120.43	60.20	80.18	124.43	82.92
	93.82	67.41	155.48	123.01	60.20	202.73	181.81	82.92	314.39

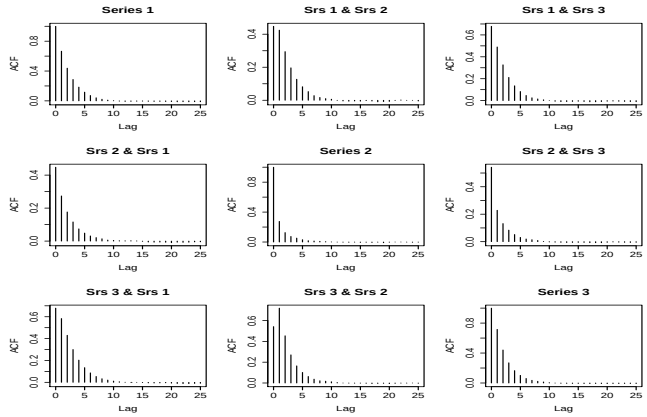
Related to the results of Figure 1, a more clear comparison between $\widehat{\boldsymbol{\rho}}_{\boldsymbol{Z}}(h)$ and $\widehat{\boldsymbol{\rho}}_{\boldsymbol{Z},Q_n}(h)$ is given in Figure 2, which displays the boxplot of the difference between $\widehat{\boldsymbol{\rho}}_{\boldsymbol{Z}}(h)$ and $\widehat{\boldsymbol{\rho}}_{\boldsymbol{Z},Q_n}(h)$. These corroborates to the empirical evidence discussed from Tables 2 and 3. As can be seen, for each lag h, the results are very close. This result indicates that the robust ACF presents similar behavior of the standard ACF even when the series has no outliers. The results for the case when all series are contaminated with outliers are shown in Figure 3. As one can see, for small h, $\widehat{\boldsymbol{\rho}}_{\boldsymbol{Z},Q_n}(h)$ presents a slightly change, but it still maintains its robustness properties.

In the context of PCA, this study evaluates the behavior of the eigenvalues of Models 2,3 and 4, comparing to the eigenvalues of Model 1. Tables 5 and 6 present the eigenvalues of covariance matrix for Models 1, 2, 3 and 4, along with the percentages of variability, for the estimators $\widehat{\Gamma}_{\boldsymbol{Z}}(0)$ and $\widehat{\Gamma}_{\boldsymbol{Z},Q_n}(0)$, respectively. It is observed that the effects of outliers are translated to the eigenvalues, in the sense of the increasing of variability. This phenomena provokes, in general, the increasing of the percentual of variance explained by the first eigenvalue. Therefore, the percentage of explained variance for the first component also increases. Thus, if the time correlation and the outliers observations are not taken in account, these may lead to an erroneous interpretations of the PCA technique in practical situation.

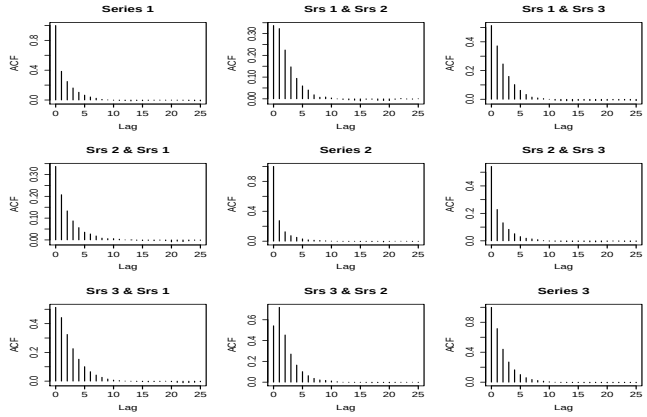
When the eigenvalues are computed based on $\widehat{\Gamma}_{\boldsymbol{Z},Q_n}(0)$, one can see that the quantities are not much affected by the outliers. Therefore, these empirical evidence show that robust $\widehat{\Gamma}_{\boldsymbol{Z},Q_n}(0)$ is an alternative ACOVF function to be used when there is a suspicion of an existence of outliers in the vector series.

5 Application to the real data: Clustering of RAMQAr

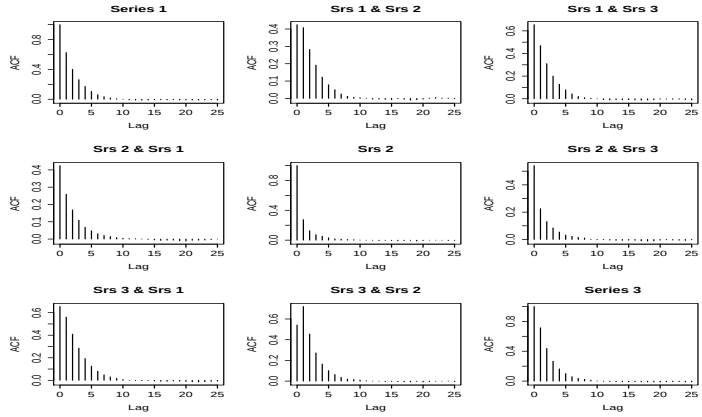
Following the idea adopted by Pires et al. (2008*a*,*b*) to use the PCA to reduce the number of monitoring stations for a given pollutant, this section presents the application of the methodology proposed in this paper for the management of PM_{10}'s monitoring equipment of the Automatic Monitoring Network Air Quality in Greater Vitória Region (RAMQAr), without neglecting the temporal correlation structure and the outliers in the data set.



(a) $\widehat{\boldsymbol{\rho}}_{\boldsymbol{Z}}(h)$ - $(p_1 = 0)$



(b) $\widehat{\boldsymbol{\rho}}_{\boldsymbol{Z}}(h)$ - $(p_1 = 0.05)$



(c) $\widehat{\boldsymbol{\rho}}_{\boldsymbol{Z},Q_n}(h)$ - $(p_1 = 0.05)$

Figure 1: Correlation matrix function, classical and robust ones, for the VAR(1) Model 4, $\boldsymbol{\delta}_t = [p_1, 0, 0]'$ and $\boldsymbol{\omega} = [4\sigma_1, 0, 0]'$.

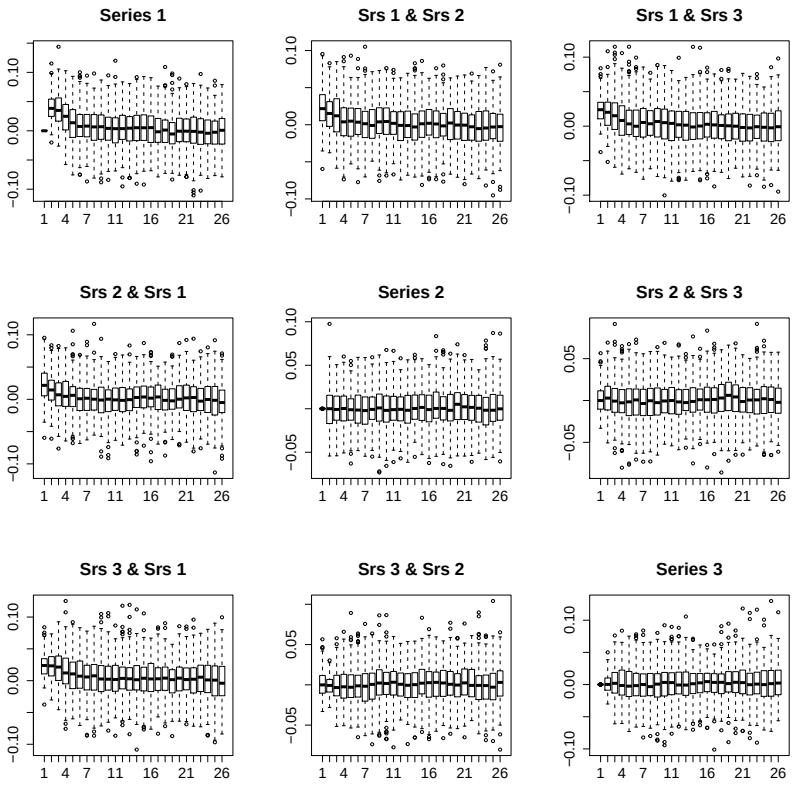


Figure 2: Boxplot of differences between $\widehat{\boldsymbol{\rho}}_{\boldsymbol{Z}}(h)$ and $\widehat{\boldsymbol{\rho}}_{\boldsymbol{Z},Q_n}(h)$ for Model 4, $\boldsymbol{\delta}_t = [0.05, 0, 0]'$ and $\boldsymbol{\omega} = [4\sigma_1, 0, 0]'$.

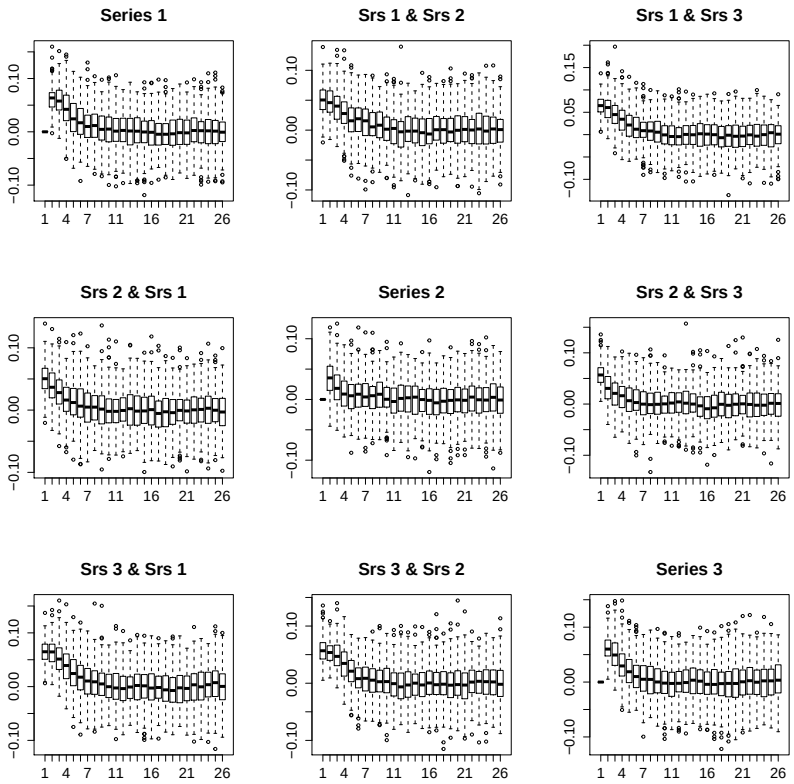


Figure 3: Boxplot of differences between $\widehat{\boldsymbol{\rho}}_{\boldsymbol{Z}}(h)$ and $\widehat{\boldsymbol{\rho}}_{\boldsymbol{Z},Q_n}(h)$ for Model 4, $\boldsymbol{\delta}_t = [0.05, 0.05, 0.05]'$ and $\boldsymbol{\omega} = [4\sigma_1, 4\sigma_2, 4\sigma_3]'$.

Table 5: Eigenvalues of $\widehat{\Gamma}_{\boldsymbol{Z}}(0)$ - VAR(1) Model 4, $\boldsymbol{\delta}_t = [p_1, 0, 0]'$ and $\boldsymbol{\omega} = [4\sigma_1, 0, 0]'$.

No outliers

Models	$\widehat{\lambda}_1$	$\widehat{\lambda}_2$	$\widehat{\lambda}_3$	% $\widehat{\lambda}_1$	% $\widehat{\lambda}_2$	% $\widehat{\lambda}_3$
1	169.72	54.82	25.83	67.79	21.90	10.32
2	223.06	72.42	34.51	67.60	21.95	10.46
3	237.27	57.45	38.14	71.28	17.26	11.46
4	393.66	72.39	41.34	77.59	14.27	8.15

($p_1 = 0.01$) **outliers**

Models	$\widehat{\lambda}_1$	$\widehat{\lambda}_2$	$\widehat{\lambda}_3$	% $\widehat{\lambda}_1$	% $\widehat{\lambda}_2$	% $\widehat{\lambda}_3$
2	236.39	78.89	34.98	67.49	22.52	9.99
3	248.56	65.14	39.11	70.45	18.46	11.09
4	404.76	82.00	41.35	76.64	15.53	7.83

($p_1 = 0.05$) **outliers**

Models	$\widehat{\lambda}_1$	$\widehat{\lambda}_2$	$\widehat{\lambda}_3$	% $\widehat{\lambda}_1$	% $\widehat{\lambda}_2$	% $\widehat{\lambda}_3$
2	299.13	95.57	35.79	69.49	22.20	8.31
3	303.86	89.47	40.18	70.09	20.64	9.27
4	454.73	113.88	41.36	74.55	18.67	6.78

($p_1 = 0.15$) **outliers**

Models	$\widehat{\lambda}_1$	$\widehat{\lambda}_2$	$\widehat{\lambda}_3$	% $\widehat{\lambda}_1$	% $\widehat{\lambda}_2$	% $\widehat{\lambda}_3$
2	490.35	111.38	36.27	76.86	17.46	5.68
3	483.59	115.50	40.61	75.60	18.06	6.35
4	610.81	160.15	41.37	75.19	19.72	5.09

Table 6: Eigenvalues of $\widehat{\Gamma}_{\boldsymbol{Z},Q_n}(0)$ - VAR(1) Model 4, $\boldsymbol{\delta}_t = [p_1, 0, 0]'$ and $\boldsymbol{\omega} = [4\sigma_1, 0, 0]'$.

No outliers

Models	$\widehat{\lambda}_1$	$\widehat{\lambda}_2$	$\widehat{\lambda}_3$	% $\widehat{\lambda}_1$	% $\widehat{\lambda}_2$	% $\widehat{\lambda}_3$
2	223.21	72.50	34.50	67.60	21.96	10.45
3	238.07	57.58	38.16	71.32	17.25	11.43
4	394.06	72.62	41.44	77.55	14.29	8.15

($p_1 = 0.01$) **outliers**

Models	$\widehat{\lambda}_1$	$\widehat{\lambda}_2$	$\widehat{\lambda}_3$	% $\widehat{\lambda}_1$	% $\widehat{\lambda}_2$	% $\widehat{\lambda}_3$
2	228.87	73.60	34.53	67.91	21.84	10.25
3	243.05	58.52	38.26	71.52	17.22	11.26
4	400.77	74.75	41.44	77.52	14.46	8.02

($p_1 = 0.05$) **outliers**

Models	$\widehat{\lambda}_1$	$\widehat{\lambda}_2$	$\widehat{\lambda}_3$	% $\widehat{\lambda}_1$	% $\widehat{\lambda}_2$	% $\widehat{\lambda}_3$
2	254.03	78.21	34.76	69.22	21.31	9.47
3	267.00	62.48	38.58	72.54	16.98	10.48
4	431.25	84.24	41.43	77.44	15.13	7.44

($p_1 = 0.15$) **outliers**

Models	$\widehat{\lambda}_1$	$\widehat{\lambda}_2$	$\widehat{\lambda}_3$	% $\widehat{\lambda}_1$	% $\widehat{\lambda}_2$	% $\widehat{\lambda}_3$
2	343.99	93.36	35.48	72.75	19.75	7.50
3	344.92	82.47	39.64	73.85	17.66	8.49
4	522.63	119.71	41.43	76.43	17.51	6.06

The RAMQAr is consisted by eight monitoring stations distributed in the cities of RGV as follows: two stations in Serra, Laranjeiras and Carapina. The city of Vitória has three stations, Jardim de Camburi, Suá and Centro (VixCentro). Vila Velha has two stations, Centro (VVCentro) and Ibes. The city of Cariacica has one station at Ceasa. The PM_{10} is monitored in all stations. Figure 4 presents the geographical location of each station. The PM_{10} series corresponds to the daily average observed at all stations from January 2005 to December 2009. The PM_{10} is measured in $\mu g/m^3$.

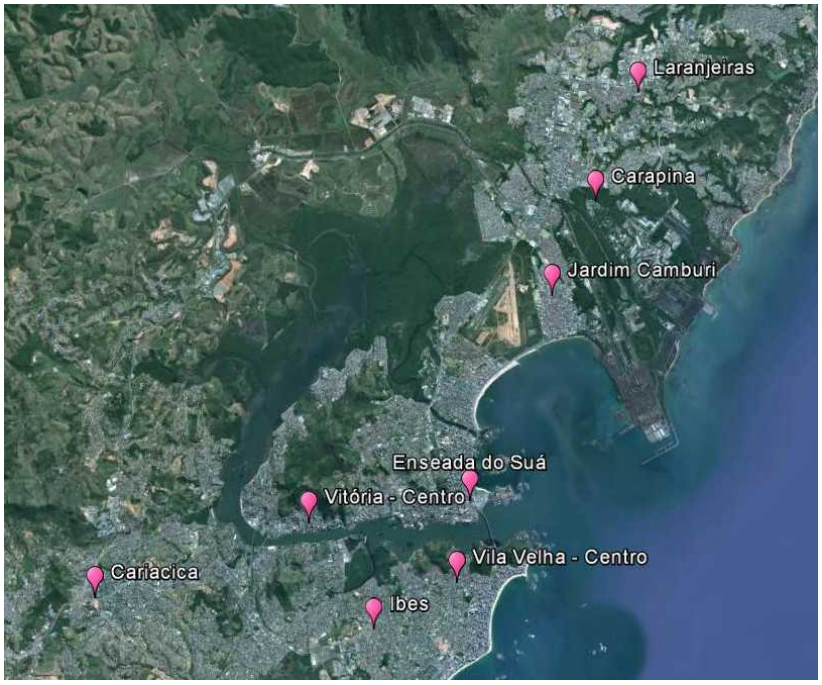


Figure 4: Geographical location of RAMQAr's stations.

Table 7: The descriptive statistics for all RAMQAr's stations

Laranjeiras	Carapina	Camburi	Sua
Min. : 6.08	Min. : 5.75	Min. : 8.67	Min. : 7.50
1st Qu.:24.50	1st Qu.:19.33	1st Qu.:23.64	1st Qu.:22.71
Median :31.27	Median :23.00	Median :28.33	Median :27.00
Mean :32.26	Mean :24.13	Mean :28.97	Mean :28.08
3rd Qu.:38.07	3rd Qu.:27.71	3rd Qu.:33.46	3rd Qu.:32.46
Max. :**86.46**	Max. :**88.25**	Max. :**78.08**	Max. :**74.58**

VixCentro	Ibes	VVCentro	Cariacica
Min. : 5.63	Min. : 7.00	Min. : 5.92	Min. : 8.92
1st Qu.:21.46	1st Qu.:22.01	1st Qu.:21.51	1st Qu.: 36.14
Median :25.25	Median :27.29	Median :27.21	Median : 43.33
Mean :26.01	Mean :28.13	Mean :28.94	Mean : 44.16
3rd Qu.:29.78	3rd Qu.:32.91	3rd Qu.:33.92	3rd Qu.: 50.79
Max. :**70.42**	Max. :**88.13**	Max. :**94.75**	Max. :**106.33**

The boxplot of the data, the series of the PM_{10} and their robust ACF are given in Figure 5, 6 and 7 , respectively.

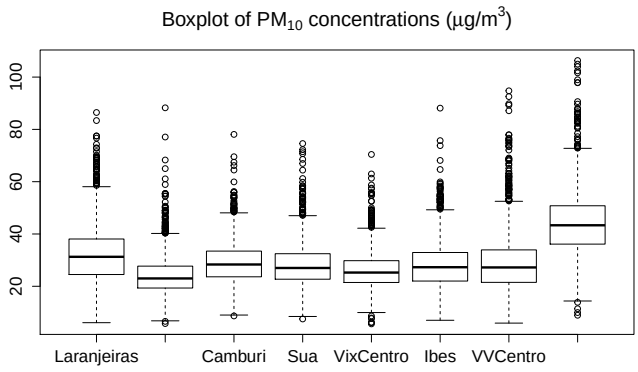


Figure 5: Boxplot of PM_{10}'s of RAMQAr's stations.

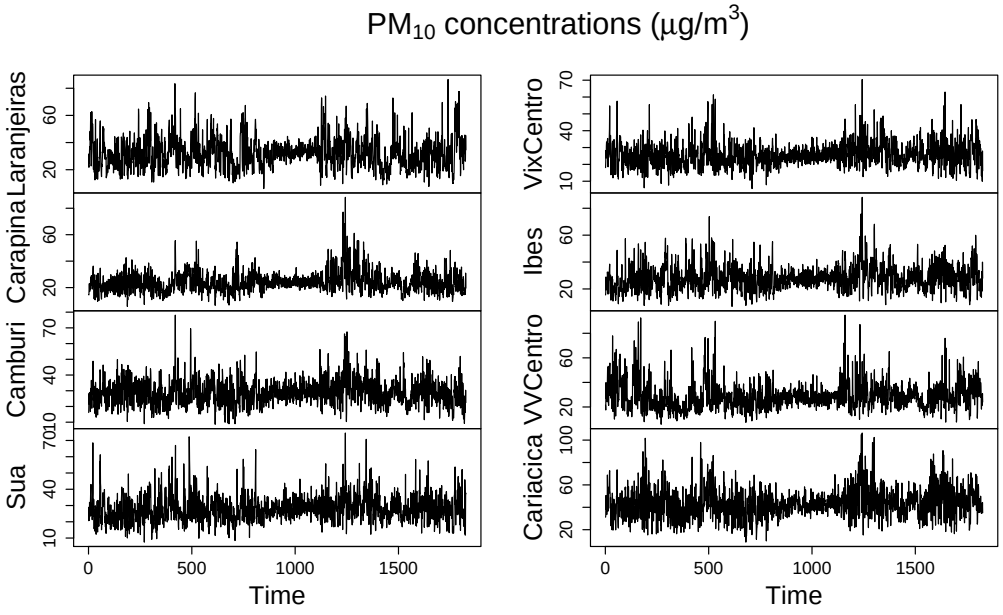


Figure 6: PM_{10}'s concentrations of RAMQAr's stations

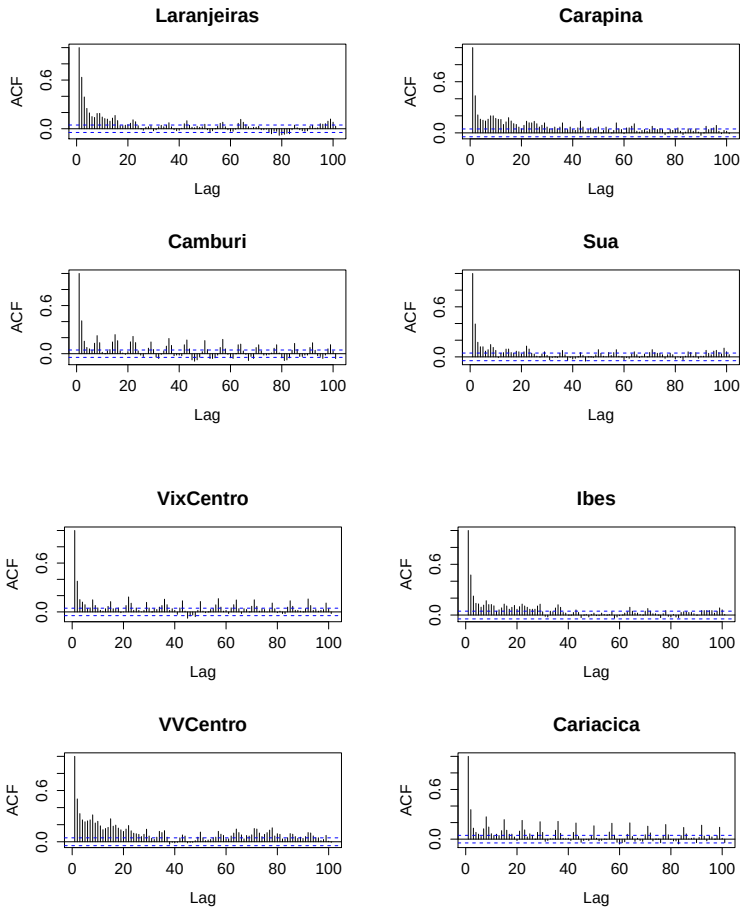


Figure 7: PM_{10} $\widehat{\rho}_{\boldsymbol{Z},Q_n}(h)$ function for RAMQAr's Stations

From the descriptive statistics, the boxplot and the plots of the series, one can observe high levels of PM_{10} pollutant in this period. Although, the high levels of PM_{10} are important information that should be considered in the context of the air pollution, these observations can be identified, from a statistical point of view, as being outliers. Therefore, the high levels of PM_{10} presented in the series, justifies the use of robust ACF.

As one example of the comparison between the classical and the robust ACFs, the classical ACF was computed for the Vila Velha Centro station. The minimum and maximum difference of $|\widehat{\Gamma}_{\boldsymbol{Z},Q_n}(h) - \widehat{\Gamma}_{\boldsymbol{Z}}(h)|, h = 1, ..., 6$, was 0.11 and 0.07, respectively. This indicates that the $\widehat{\Gamma}_{\boldsymbol{Z},Q_n}(h)$ performs quite differently compared to $\widehat{\Gamma}_{\boldsymbol{Z}}(h)$. Therefore, these high levels of PM_{10} may be a provoking the effect of additive outliers in the calculation of the ACF and may indicate the presence of some outliers among the observed concentrations of PM_{10}.

In the PCA tool, the estimates of the eigenvalues and theirs corresponding eigenvectors using $\widehat{\rho}_{\boldsymbol{Z}}(0)$ and $\widehat{\rho}_{\boldsymbol{Z},Q_n}(0)$ are given in Table 8.

For both estimators, four components could explain approximately 84 % of the total variability of data set leading to a dimension reduction of data. It is observed that PCA computed

Table 8: PCA results for PM_{10} in RAMQAr

Stations	PCA - $\widehat{\rho}_{\boldsymbol{Z}}(0)$				PCA - $\widehat{\rho}_{\boldsymbol{Z},Q_n}(0)$			
	1	2	3	4	1	2	3	4
Laranjeiras	-0.3002	**0.7193**	-0.1756	0.1460	-0.3123	**0.6998**	0.0533	0.0683
Carapina	-0.3554	**-0.4004**	0.2628	0.1750	-0.3488	**-0.4144**	-0.1961	0.2701
Camburi	-0.3472	0.1700	0.0502	**0.7019**	-0.3446	0.2356	-0.2115	**0.7037**
Suá	-0.3632	0.2163	0.0406	**-0.6118**	**-0.3722**	0.1519	-0.0045	**-0.5144**
VixCentro	**-0.3864**	-0.2265	-0.1026	-0.1629	**-0.3745**	-0.2867	-0.0211	-0.1276
Ibes	**-0.3869**	0.1787	0.2359	-0.2271	**-0.3863**	0.1902	-0.0881	-0.3395
VVCentro	-0.3055	-0.2942	**-0.8391**	0.0141	-0.3203	-0.1838	**0.8942**	0.1475
Cariacica	**-0.3721**	-0.2766	0.3542	0.0507	-0.3625	-0.3283	-0.3259	-0.0962
Eigenvalue	4.8971	0.7744	0.6282	0.4973	5.146	0.7568	0.5334	0.4612
Proportion	61.22	9.68	7.85	6.22	64.25	9.46	6.67	5.77
Cumulative	61.22	70.90	78.75	84.97	64.25	73.71	80.38	86.14

using robust $\widehat{\rho}_{\boldsymbol{Z},Q_n}(0)$ preserved a greater percentage of variability in the components.

Besides the use for dimensionality reduction, the PCA technique can be used for clustering of the variables of a data matrix. Cadima & Jolliffe (1995) discuss in their paper the clustering of variables by means of the eigenvectors of the PCA. The grouping of variables consists of choosing variables that have similar values for its eigenvectors in module and are higly correlated to the principal component. The correlation between the group and the PC is given by

$$r_m = \lambda_j^{1/2}(\mathbf{e}_j^{k'}\widehat{\rho}_{\boldsymbol{Z},Q_n}(0)_k^{-1}\mathbf{e}_j^k), \tag{28}$$

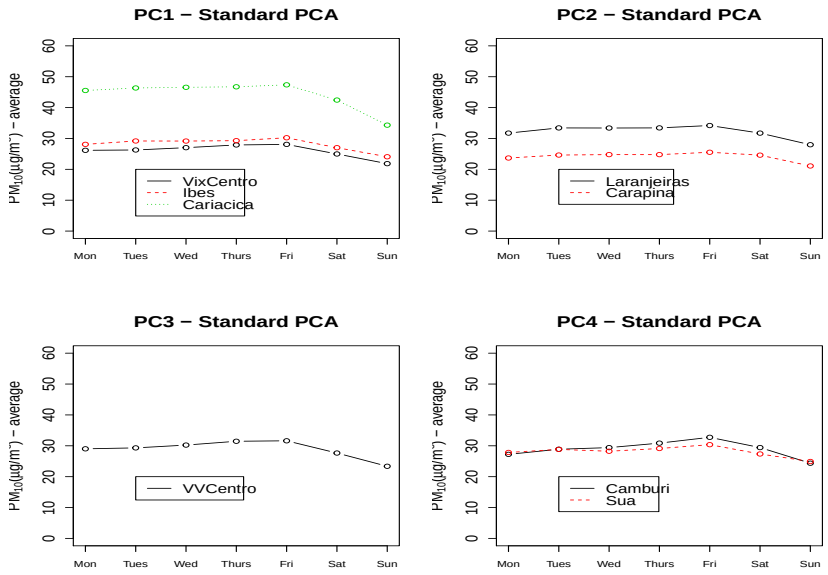
where λ_j is eigenvalue of $j$ith, $\mathbf{e}_j^k$ is the clustered vector of $\mathbf{e}$ containing k variables and $\widehat{\rho}_{\boldsymbol{Z},Q_n}(0)_k^{-1}$ is the submatrix of $\widehat{\rho}_{\boldsymbol{Z},Q_n}(0)$, which involves lines and columns corresponding to the k grouped variables.

In the application, the grouping of stations consists of choosing stations that have similar values for its eigenvectors in module. That is, stations having the same contribution component will have similar values for their eigenvectors. The cutoff point was arbitrarily chosen to be 0.37 in the module, leading to a correlation of 0.92.

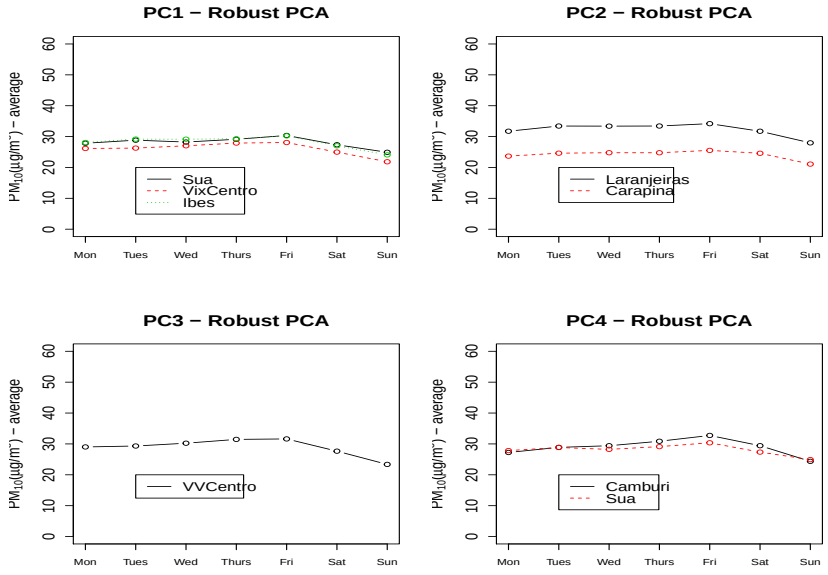
Thus, for the method of moments estimator for the first component it is possible to visualize the existence of a group of stations formed by Ibes, VVCentro and Cariacica. In the second component, the group is formed by Laranjeiras and Carapina. For the third component VVCentro forms a group. Finally, the fourth component is the group formed by Camburi and Suá.

For the grouping through robust PCA, in the first component Ibes, Suá and VixCentro can be grouped. For the second component, Laranjeiras and Carapina form a group. In third component, VVCentro is the only station in the group. For the fourth component the groups is formed by Suá and Camburi. Therefore, the proposed method allocated groups differently from $\widehat{\rho}_{\boldsymbol{Z}}(0)$. However, based on the propositions stated above and the simulations results, the grouping based on $\widehat{\rho}_{\boldsymbol{Z},Q_n}(0)$ is suggested here. Thus, any of the stations of a group formed by two or more stations may be reallocated to other areas of interest. For example, it is possible the move the equipment from Suá to create a new station in Cariacica providing a further expansion of the local network.

In order to confirm the grouping results for both estimators, the average daily profiles of PM_{10} daily averages for the groups are shown in Figure 8. It is seen that the grouping using $\widehat{\rho}_{\boldsymbol{Z},Q_n}(0)$ is superior, since for the first principal the grouped stations have similar concentrations.



(a) Standard PCA



(b) Robust PCA

Figure 8: Average daily profile of the PM_{10} daily average.

6 Conclusions

This article applied the Robust Principal Component Analysis to identify pollution behavior for the pollutant PM_{10} in the Metropolitan Region of Vitória to enable better management of the local monitoring network. The article considers the effects of different correlation structures and additive outliers on a vector linear process and its implication in the analysis and interpretation of principal components calculated from the correlation matrix of this process.

It was demonstrated that the principal components obtained from usual methodology are time correlated and may present significant cross-correlation at different lags. In addition, it was shown that the existence of outliers destroys cross-correlation. In this case the use of principal component regression correlated in Models may cause spurious results.

Also, it was observed that the structure of temporal correlation and the presence of additive outliers affect the eigenvalues of covariance matrix. As a direct consequence, the percentage of explanation of each component is changed and, as a result, the eigenvectors will also be modified. These results indicate that the results and conclusions obtained in the literature by the application of PCA in time series data and outliers may be wrong. Therefore, the use the Robust ACF in all steps to compute the robust PCA and applying this procedure to the techniques that make use of PCA to analyze multivariate time series data is strongly encouraged.

7 Acknowledgement

The authors would like to thank the support from CNPq, CAPES and FAPES.

References

Cadima, J. & Jolliffe, I. T. (1995), 'Loading and correlations in the interpretation of principle compenents', *Journal of Applied Statistics* **22**(2), 203–214.

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

Croux, C. & Rousseeuw, P. J. (1992), 'Time-efficient algorithms for two highly robust estimators of scale', *Computational Statistics* **1**, 1–18.

Dominick, D., Juahir, H., Latif, M. T., Zain, S. M. & Aris, A. Z. (2012), 'Spatial assessment of air quality patterns in malaysia using multivariate analysis', *Atmospheric Environment* **60**, 172–181.

Filzmoser, P. (1999), 'Robust principal component and factor analysis in the geostatistical treatment of environmental data', *Environmetrics* **10**, 363–375.

Huber, P. (2004), *Robust Statistics*, Wiley Series in Probability and Statistics - Applied Probability and Statistics Section Series, Wiley.

Johnson, C. (1989), *Matrix theory and applications*, American Mathematical Soc.

Johnson, R. & Wichern, D. (2013), *Applied Multivariate Statistical Analysis: Pearson New International Edition*, Pearson Education, Limited.

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

Lau, J., Hung, W. & Cheung, C. (2009), 'Interpretation of air quality in relation to monitoring station's surroundings', *Atmospheric Environment* **43**(4), 769 – 777.

Lévy-Leduc, C., Boistard, H., Moulines, E., Taqqu, M. S. & Reisen, V. A. (2011*a*), 'Large sample behavior of some well-known robust estimators under long-range', *Statistics* **45**(1), 59–71.

Lévy-Leduc, C., Boistard, H., Moulines, E., Taqqu, M. S. & Reisen, V. A. (2011*b*), 'Robust estimation of the scale and of the autocovariance function of gaussian short-and long-range dependent processes', *Journal of Time Series Analysis* **32**(2), 135–156.

Lu, W.-Z., He, H.-D. & yun Dong, L. (2011), 'Performance assessment of air quality monitoring networks using principal component analysis and cluster analysis', *Building and Environment* **46**(3), 577 – 583.

Ma, Y. & Genton, M. G. (2000), 'Highly robust estimation of the autocovariance function', *Journal of Time Series Analysis* **21**, 663–684.

Ma, Y. & Genton, M. G. (2001), 'Highly robust estimation of dispersion matrices', *Journal of Multivariate Analysis* **78**, 11–36.

Molinares, F. F., Reisen, V. A. & Cribari-Neto, F. (2009), 'Robust estimation in long-memory processes under additive outliers', *Journal of Statistical Planning and Inference* **139**(8), 2511–2525.

Pires, J., Pereira, M., Alvim-Ferraz, M. & Martins, F. (2009), 'Identification of redundant air quality measurements through the use of principal component analysis', *Atmospheric Environment* **43**(25), 3837 – 3842.

Pires, J., Sousa, S., Pereira, M., Alvim-Ferraz, M. & Martins, F. (2008*a*), 'Management of air quality monitoring using principal component and cluster analysis-part i: SO2 and PM10', *Atmospheric Environment* **42**(6), 1249 – 1260.

Pires, J., Sousa, S., Pereira, M., Alvim-Ferraz, M. & Martins, F. (2008*b*), 'Management of air quality monitoring using principal component and cluster analysis-part ii: CO, NO2 and O3', *Atmospheric Environment* **42**(6), 1261 – 1274.

Reinsel, G. (2003), *Elements of Multivariate Time Series Analysis*, Springer Series in Statistics, Springer New York.

Rousseeuw, P. J. & Croux, C. (1993), 'Alternatives to the median absolute deviation', *Journal of the American Statistical Association* **88**(424), 1273–1283.

Sarnaglia, A., Reisen, V. & Lévy-Leduc, C. (2010), 'Robust estimation of periodic autoregressive processes in the presence of additive outliers', *Journal of multivariate analysis* **101**(9), 2168–2183.

Wei, W. W.-S. (1994), *Time series analysis*, Addison-Wesley Redwood City, California.

WHO (2005), *WHO Air quality guidelines for particulate matterand ozoneand nitrogen dioxide and sulfur dioxide.*

7 CONCLUSÃO E TRABALHOS FUTUROS

Os estudos e resultados aqui apresentados foram motivados pela aplicação da análise de componente principal no contexto da poluição do ar, em especial, no uso da técnica no gerenciamento da RAMQAR. Os dados obtidos por meio da RAMQAr possuem *outliers* e são provenientes de uma série temporal. A não consideração dessa característica pode levar a resultados equivocados que não retratam de modo real a verdadeira natureza das informações estudadas. Além disso, a ACP baseia-se no pressuposto de variáveis independentes, geradas por uma distribuição normal multivariada, propriedades também não observadas nos dados da RAMQAr.

Os resultados deste trabalho corroboram e estendem os resultados obtidos por Zamprogno (2013) e mostram que as análises oriundas da técnica aplicada em processos multivariados podem levar em interpretações espúrias. Por exemplo, autocorrelações fortemente positivas das variáveis dos poluentes atmosféricos podem acarretar no subdimensionamento da rede de monitoramento da poluição do ar.

A contribuição científica desta dissertação é apresentada no artigo resultante do estudo. O artigo apresenta a fundamentação teórica e empírica do impacto da correlação temporal e de *outliers* aditivos na técnica ACP. Em termos teóricos e empíricos, por meio de simulações de Monte Carlo, foi demonstrado que as componentes principais obtidas do método usual de ACP são autocorrelacionadas e podem apresentar correlação cruzada significativa nas defasagens diferentes de zero. Também foi demonstrado que a existência de *outliers* mascara a correlação cruzada. Nesse caso, o uso de componentes principais correlacionadas em modelos de regressão pode provocar resultados espúrios.

Além disso, verificou-se que a estrutura de correlação e os *outliers* presentes nas variáveis alteram os autovalores da matriz de covariância. Como consequência direta, os percentuais de explicação de cada componente foi alterado. Além disso, os autovetores também serão modificados. Esses resultados indicam que os resultados e conclusões obtidas pela literatura por meio da aplicação da ACP em dados de séries temporais e *outliers* podem estar equivocados.

7.1 TRABALHOS FUTUROS

O filtro VAR utilizado para remoção da correlação temporal entre as varáveis também sofre influência de *outliers*. Como trabalho futuro, pode-se explorar a estimação robusta do modelo VAR e posterior aplicação da ACP.

O foco desta dissertação foi aplicação direta da ACP, isto é, nos resultados dos autovetores e autovalores. Não foi objetivo da dissertação aplicar testes estatísticos e outras inferências sobre os autovetores e autovalores. Entretanto, como trabalho futuro é possível empregar a técnica *bootstrap* para se obter a distribuição dos autovalores de modo a realizar testes de hipóteses.

A técnica ACP é uma das mais importantes ferramentas da estatística multivariada. Outras áreas do conhecimento cujos dados são provenientes de séries temporais e possuem *outliers* poderão utilizar os resultados aqui apresentados. Além disso, a ACP é utilizada como ponto inicial na análise fatorial, ferramenta empregada na identificação de fontes poluidoras. Deste modo, como trabalho futuro, pode-se considerar o emprego da metodologia aqui proposta para identificação robusta das fontes poluidoras na RGV.

REFERÊNCIAS

ALMEIDA, S. et al. Approaching pm2.5 and pm2.5 source apportionment by mass balance analysis, principal component analysis and particle size distribution. **Science of The Total Environment**, v. 368, p. 663–674, 2006.

BARNETT, V.; LEWIS, T. **Outliers in Multivariate Data**. [S.l.]: John Wiley, 1984.

BLOMBERG, A. et al. Persistent airway inflammation but accommodated antioxidant and lung function responses after repeated daily exposure to nitrogen dioxide. **American Journal of Respiratory and Critical Care Medicine**, v. 159, p. 536–543, 1999.

BLOMBERG, A. et al. The inflammatory effects of 2 ppm no2 on the airways of healthy subjects. **American Journal of Respiratory and Critical Care Medicine**, v. 156, p. 418–424, 1997.

BURNETT, R. et al. Association between ambient carbon monoxide levels and hospitalizations for congestive heart failure in the elderly in 10 canadian cities. **Epidemiology**, v. 8, p. 162–167, 1997.

BURNETT, R. T. et al. Association between ozone and hospitalization for respiratory diseases in 16 canadian cities. **Environmental Research**, v. 72, p. 24–31, 1997.

CROUX, C.; HAESBROECK, G. Principal component analysis based on robust estimators of the covariance or correlation matrix: Influence functons and efficiencies. **Biometrika**, v. 87, p. 603–618, 2000.

CRUZ, L. S. **Variação temporal das comunidades fitoplantônicas em uma lagoa de polimento de efluente de um reator anaeróbio compartimentado tratando esgoto sanitário**. 2005. 173 p. Dissertação (Mestrado em Engenharia Ambiental) — Programa de Pós-graduação em Engenharia Ambiental, Universidade Federal do Espírito Santo, Vitória, 2005.

DAVIES, P. L. Asymptotic behaviour of s-estimates of multivariate location parameters and dispersion matrices. **The Annals of Statistics**, v. 15, p. 1269–1292, 1987.

DEVLIN, S. J.; GNANADESIKAN, R.; KETTENRING, J. R. Robust estimation and outlier detection with correlation coefficients. **Biometrika**, v. 62, n. 3, p. 531–545, 1975.

DEVLIN, S. J.; GNANDESIKAN, R.; KETTENRING, J. R. Robust estimation of dispersion matrices and principal components. **Journal of the American Statistical Association**, v. 76, n. 374, p. 354–362, 1981.

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

DONALDSON, K.; GILMOUR, M.; MACNEE, W. Asthma and pm10. **Respiratory Research**, v. 1, p. 12–15, 2000.

DONALDSON, K. et al. Role of inflammation in cardiopulmonary health effects of pm. **Toxicology and Applied Pharmacology**, v. 207, p. 483–488, 2005.

DONOHO, D.; HUBER, P. The notion of breakdown point. **A Festschrift for Erich L. Lehmann**, p. 157–184, 1983.

FANG, S.-C.; CHANG, I.-C.; YU, T.-Y. Analysis of spatial features of coastal oil pollution using multivariate methods. **Journal of Coastal Research**, The Coastal Education and Research Foundation, 2014.

FILZMOSER, P. Robust principal component and factor analysis in the geostatistical treatment of environmental data. **Environmetrics**, v. 10, p. 363–375, 1999.

FILZMOSER, P.; MARONNA, R.; WERNER, M. Outlier identification in high dimensions. **Computational Statisticts and Data Analysis**, v. 52, p. 1694–1711, 2008.

FOX, A. J. Outliers in time series. **Journal of the Royal Statistical Society**, v. 34, p. 350–363, 1972.

FREIRE, A. P. **Correlação do uso do solo e qualidade de água utilizando ferramentas de geoprocessamento e técnica de análise estatística multivariada**. 2009. 171 p. Dissertação (Mestrado em Engenharia Ambiental) — Programa de Pós-graduação em Engenharia Ambiental, Universidade Federal do Espírito Santo, Vitória, 2009.

GOLDBERG, M. S. et al. Associations between ambient air pollution and daily mortality among persons with congestive heart failure. **Environmental Research**, v. 91, p. 8–20, 2003.

GOMES, A. I. et al. Optimization of river water quality surveys by multivariate analysis of physicochemical, bacteriological and ecotoxicological data. **Water Resources Management**, Springer, v. 28, n. 5, p. 1345–1361, 2014.

HUBER, P.; RONCHETTI, E. **Robust Statistics**. [S.l.]: Wiley, 2009.

HUBERT, M.; ROUSSEEUW, P. J.; AELST, S. V. High-breakdown robust multivariate methods. **Statistical Science**, v. 23, n. 1, p. 92–119, 2008.

IBGE. **Sinopse do Censo Demográfico 2010**. Rio de Janeiro, RJ, 2011.

IEMA. **Relatório da Qualidade do Ar da Região da Grande Vitória - 2005**. Espírito Santo, ES, 2005.

IJSN. **Perfil Regional - Região Metropolitana da Grande Vitória**. Espírito Santo, ES, 2008.

JACKSON, J. E. **A User's Guide to Principal Components**. [S.l.]: John Wiley & Sons, Inc., 2004.

JOHNSON, R. A.; WICHERN, D. W. **Applied Multivariate Statistical Analysis (6th Edition)**. [S.l.]: Prentice Hall, 2007.

KHAMIS, A.; ABDULLAH, M. On robust environmental quality indices. **Pertanika Journal of Science and Technology**, v. 12, p. 1–9, 2004.

KIM, Y. et al. Predictive monitoring and diagnosis of periodic air pollution in a subway station. **Journal of hazardous materials**, Elsevier, v. 183, n. 1, p. 448–459, 2010.

KRZANOWSKI, W. **Principles of multivariate analysis: a user's perspective**. [S.l.]: Clarendon Press, 1988.

LAU, J.; HUNG, W.; CHEUNG, C. Interpretation of air quality in relation to monitoring station's surroundings. **Atmospheric Environment**, v. 43, n. 4, p. 769 – 777, 2009.

LAWTHER, P. et al. Pulmonary function and sulphur dioxideand some preliminary findings. **Environmental Research**, v. 10, n. 3, p. 355–367, 1975.

LOPUHAA, H. On the relation between s-estimators and m-estimators of multivariate location and covariance. **The Annals of Statistics**, p. 1662–1683, 1989.

LU, W.-Z.; HE, H.-D.; DONG, L. yun. Performance assessment of air quality monitoring networks using principal component analysis and cluster analysis. **Building and Environment**, v. 46, n. 3, p. 577–583, 2011.

MA, Y.; GENTON, M. G. Highly robust estimation of the autocovariance function. **Journal of Time Series Analysis**, v. 21, p. 663–684, 2000.

MA, Y.; GENTON, M. G. Highly robust estimation of dispersion matrices. **Journal of Multivariate Analysis**, v. 78, p. 11–36, 2001.

MACNEE, W.; DONALDSON, K. Mechanism of lung injury caused by pm10 and ultrafine particles with special reference to copd. **European Respiratory Journal**, v. 40, p. 47–51, 2003.

MAHALANOBIS, P. On the generalized distance in statistics. In: NEW DELHI. **Proceedings of the National Institute of Sciences of India**. [S.l.], 1936. v. 2, n. 1, p. 49–55.

MARDIA, K.; KENT, J.; BIBBY, J. **Multivariate analysis**. [S.l.]: Academic Press, 1979.

MARONNA, R.; MARTIN, R. D.; YOHAI, V. **Robust Statistics**. [S.l.]: Wiley, 2006.

MARONNA, R. A. Robust m-estimators of multivariate location and scatter. **Annals of Statistics**, v. 4, p. 51–67, 1976.

MINGOTI, S. **Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada**. [S.l.]: Editora UFMG, 2005.

MORTIMER et al. The effect of ozone on inner-city children with asthma: identification of susceptible subgroups. **American Journal of Respiratory and Critical Care Medicine**, v. 162, p. 1838–1845, 2000.

PENNY, K. I. Appropriate critical values when testing for a single multivariate outlier by using the mahalanobis distance. **Journal of the Royal Statistical Society. Series C (Applied Statistics)**, v. 45, n. 1, p. 73–81, 1996.

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.

PIRES, J. et al. Management of air quality monitoring using principal component and cluster analysis - part i: So2 and pm10. **Atmospheric Environment**, v. 42, n. 6, p. 1249 – 1260, 2008.

PIRES, J. et al. Management of air quality monitoring using principal component and cluster analysis - part ii: Co, no2 and o3. **Atmospheric Environment**, v. 42, n. 6, p. 1261 – 1274, 2008.

REISEN, V. A.; SILVA, A. N. **O uso da linguagem R para cálculos de estatística básica.** [S.l.]: EDUFES, 2011.

ROCKE, D. Robustness properties of s-estimators of multivariate location and shape in high dimension. **The Annals of Statistics**, p. 1327–1345, 1996.

ROJAS, E. et al. Evaluation of DNA damage in exfoliated tear duct epithelial cells from individuals exposed to air pollution assessed by single cell gel electrophoresis assay. **Mutation Research**, v. 468, p. 11–17, 2000.

ROUSSEEUW, P.; ZOMEREN, B. V. Unmasking multivariate outliers and leverage points. **Journal of the American Statistical Association**, v. 85, p. 633–639, 1990.

ROUSSEEUW, P. J. Least median of squares regression. **Journal of the American Statistical Association**, v. 79, n. 388, p. 871–880, 1984.

ROUSSEEUW, P. J.; YOHAI, V. Robust regression by means of s-estimators. **in Robust and Nonlinear Time Series Analysis, edited by J.Franke, W. Hiirdle, and R. D. Martin, Lecture Notes in Statistics No. 26,**, p. 256–272, 1984.

SOARES, I. P. **Avaliação do uso de diferentes modelos receptores para determinação da contribuição das fontes de Partculas totais em suspensão**. 2011. 153 p. Dissertação (Mestrado em Engenharia Ambiental) — Programa de Pós-graduação em Engenharia Ambiental, Universidade Federal do Espírito Santo, Vitória, 2011.

SONG, Y. et al. Source apportionment of pm2.5 in beijing using principal component analysis/absolute principal component scores and unmix. **Science of the Total Environment**, v. 372, p. 278–286, 2006.

SOUZA, J. B. **Análise de componentes principais e a modelagem linear generalizada: uma associação entre o número de atendimentos hospitalares por causas respiratórias e a qualidade do ar, na região da grande Vitória, ES**. 2013. 66 p. Dissertação (Mestrado em Engenharia Ambiental) — Programa de Pós-graduação em Engenharia Ambiental, Universidade Federal do Espírito Santo, Vitória, 2013.

STAHEL, W. **Breakdown of covariance estimators**. [S.l.]: Fachgruppe für Statistik, Eidgenössische Techn. Hochsch., 1981.

TUKEY, J. Useable resistant/robust techniques of analysis. In: **Proceedings of the first ERDA Symposium**. [S.l.: s.n.], 1975.

TUNNICLIFFE, W. et al. The effect of sulphur dioxide exposure on indices of heart rate variability in normal and asthmatic adults. **European Respiratory Journal**, v. 17, p. 604–608, 2001.

TYLER, D. Finite sample breakdown points of projection based multivariate location and scatter statistics. **The Annals of Statistics**, p. 1024–1044, 1994.

USA-EPA - US Environmental Protection Agency. **Air quality criteria for oxides of nitrogen.** [S.l.], 1993.

WANG, Y.; PHAM, H. Analyzing the effects of air pollution and mortality by generalized additive models with robust principal components. **International Journal of Systems Assurance Engineering and Management**, v. 3, p. 253–259, 2011.

WHO - World Health Organization. **WHO Air quality guidelines for particulate matterand ozoneand nitrogen dioxide and sulfur dioxide**. [S.l.], 2005.

WU, E. M.-Y. et al. The application of water quality monitoring data in a reservoir watershed using amos confirmatory factor analyses. **Environmental Modeling & Assessment**, Springer, p. 1–9, 2014.

ZAMPROGNO, B. **O uso e interpretação de análise de componentes principais, em séries temporais, com enfoque no gerenciamento da qualidade do ar**. 2013. 117 p. Tese (Doutorado em Engenharia Ambiental) — Programa de Pós-graduação em Engenharia Ambiental, Universidade Federal do Espírito Santo, Vitória, 2013.

8 APÊNDICE A - FIGURAS EM MAIOR RESOLUÇÃO

Neste apêndice são encontradas as figuras utilizadas neste trabalho em maior resolução e definição, na ordem que aparecem no texto.

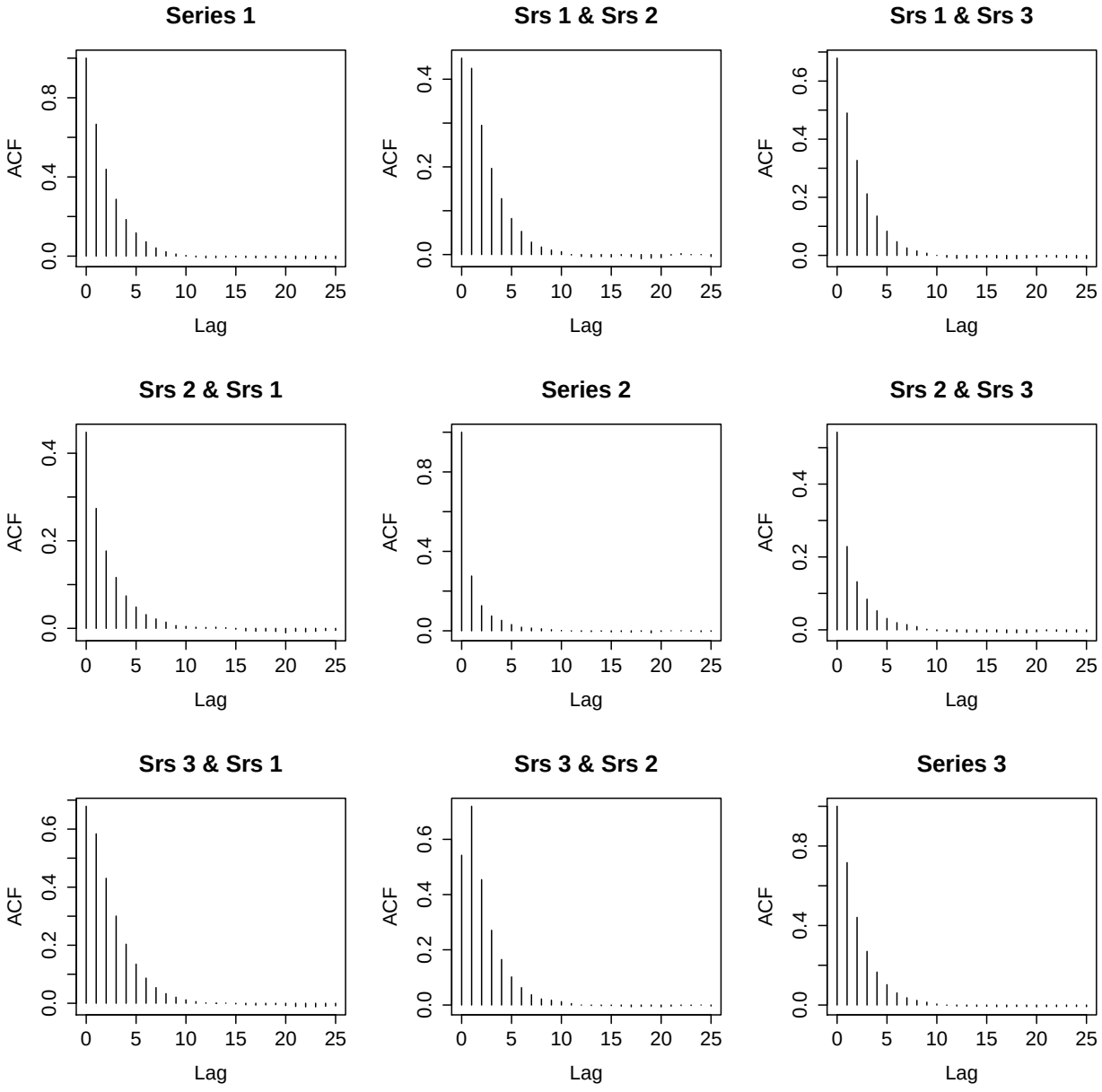


Figura 2: Refere-se à Figura 1(a) do artigo.

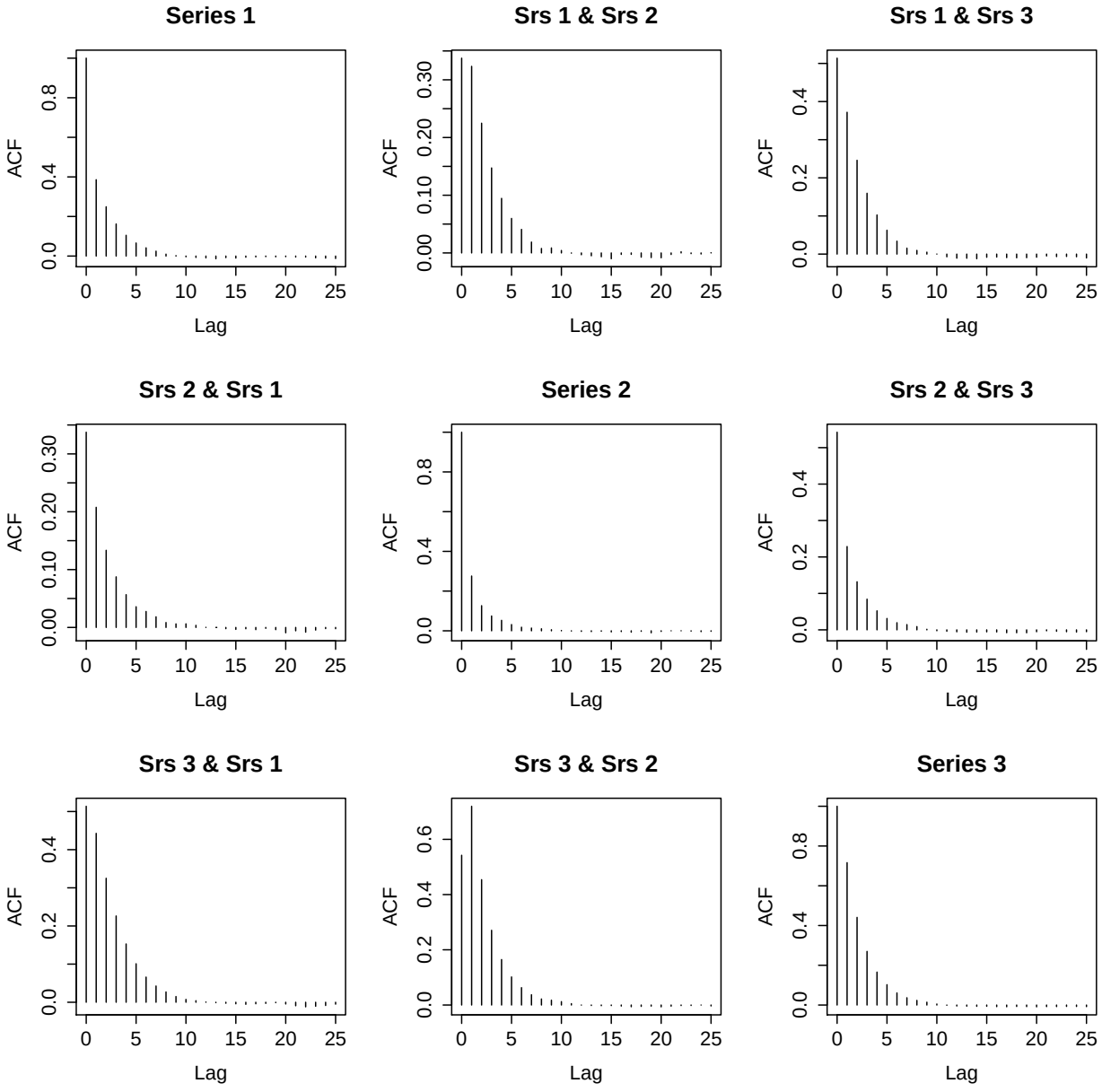


Figura 3: Refere-se à Figura 1(b) do artigo.

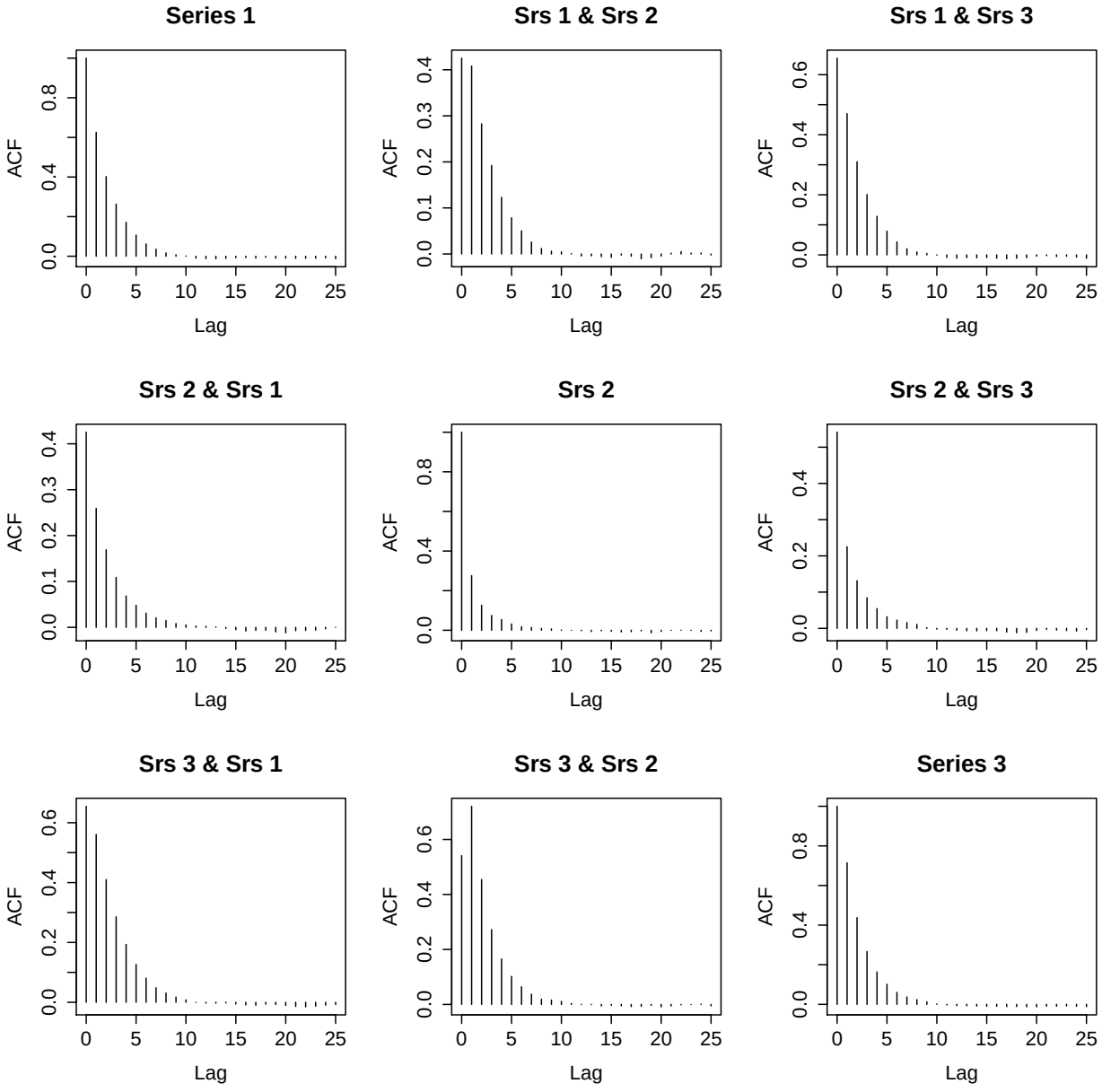


Figura 4: Refere-se à Figura 1(c) do artigo.

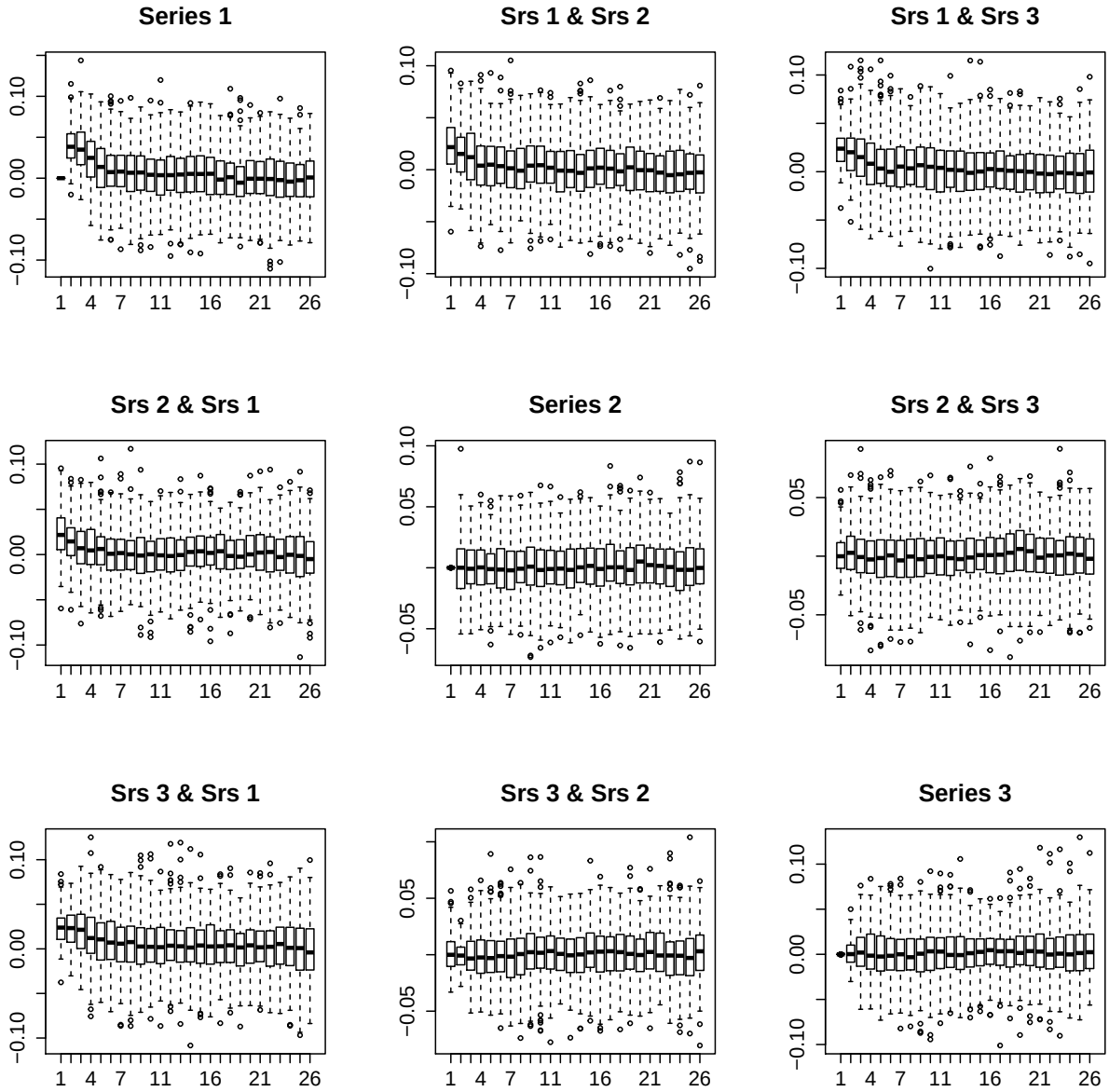


Figura 5: Refere-se à Figura 2 do artigo.

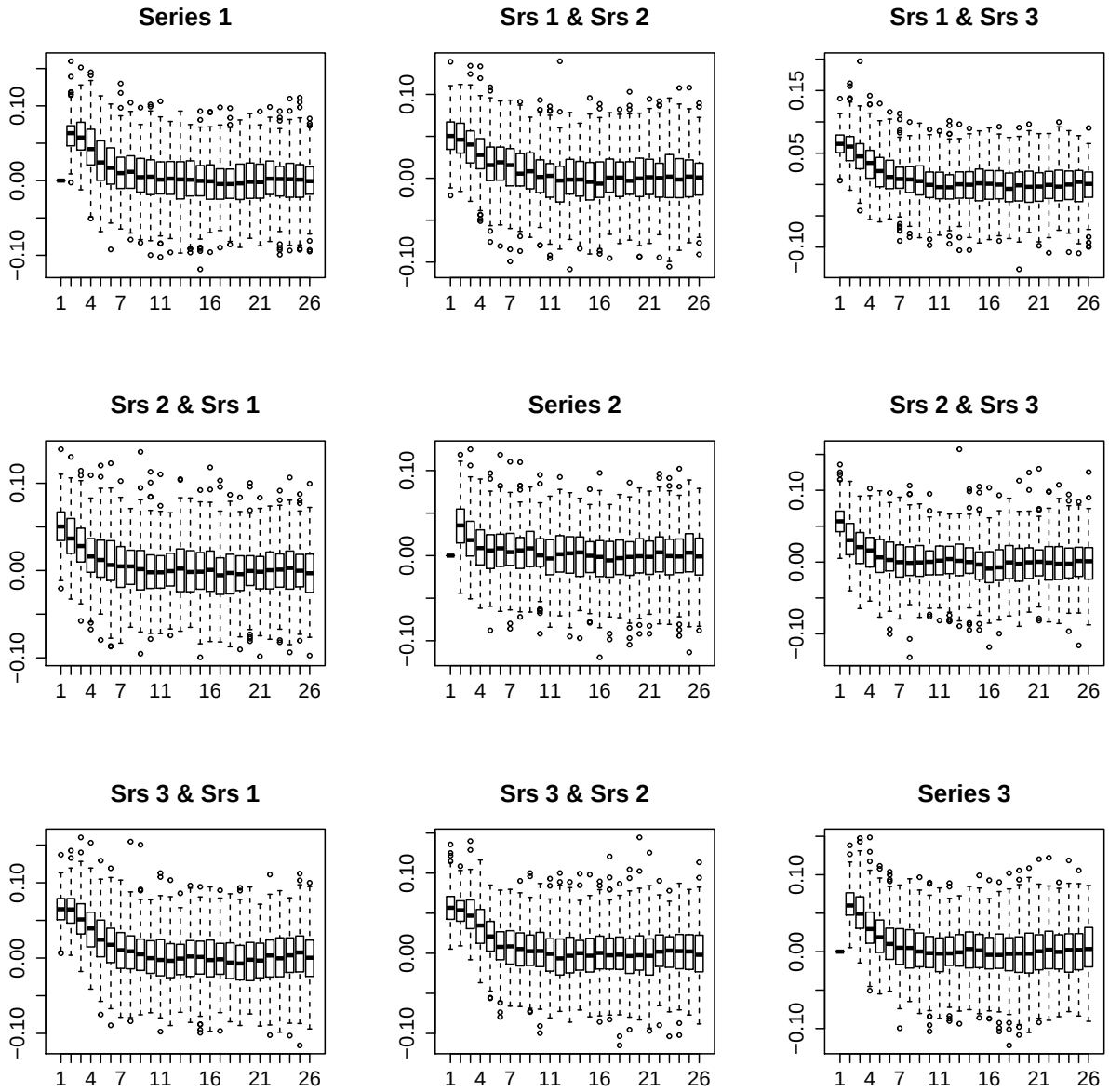


Figura 6: Refere-se à Figura 3 do artigo.

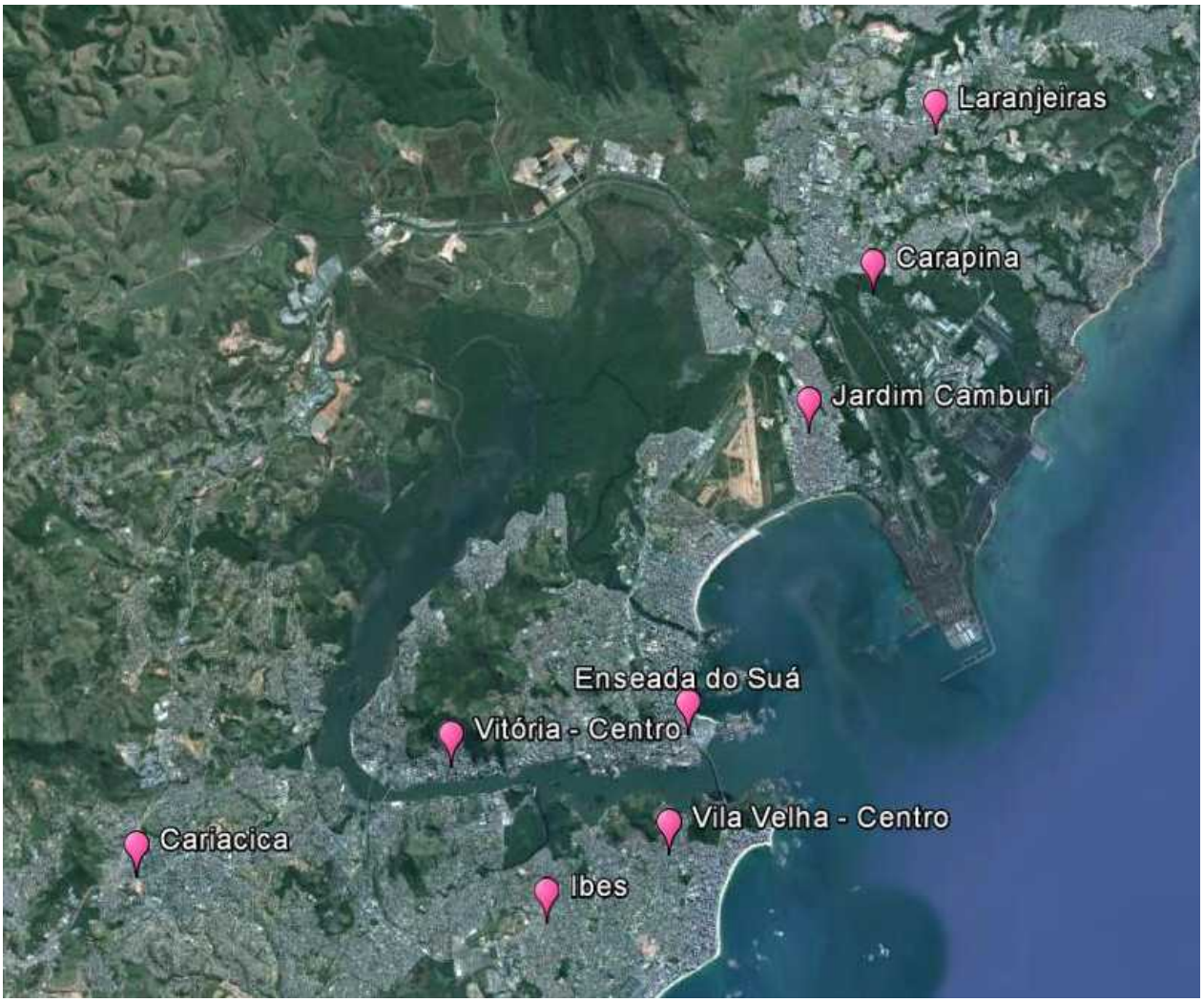


Figura 7: Refere-se à Figura 4 do artigo.

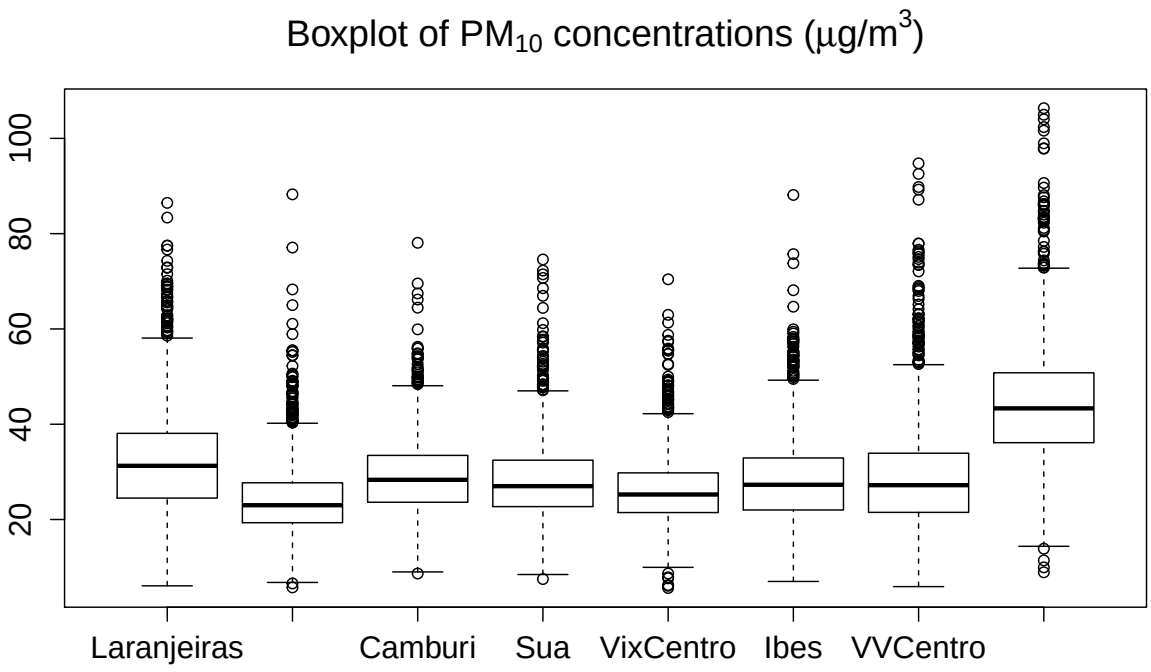


Figura 8: Refere-se à Figura 5 do artigo.

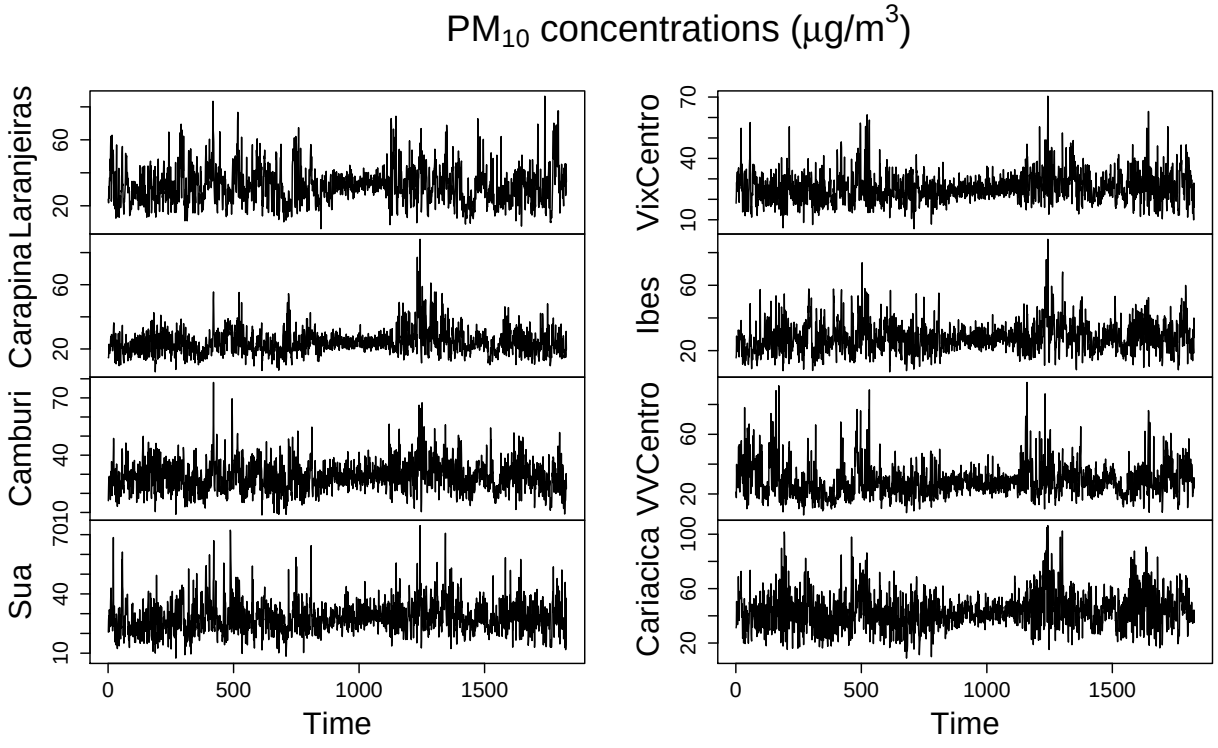


Figura 9: Refere-se à Figura 6 do artigo.

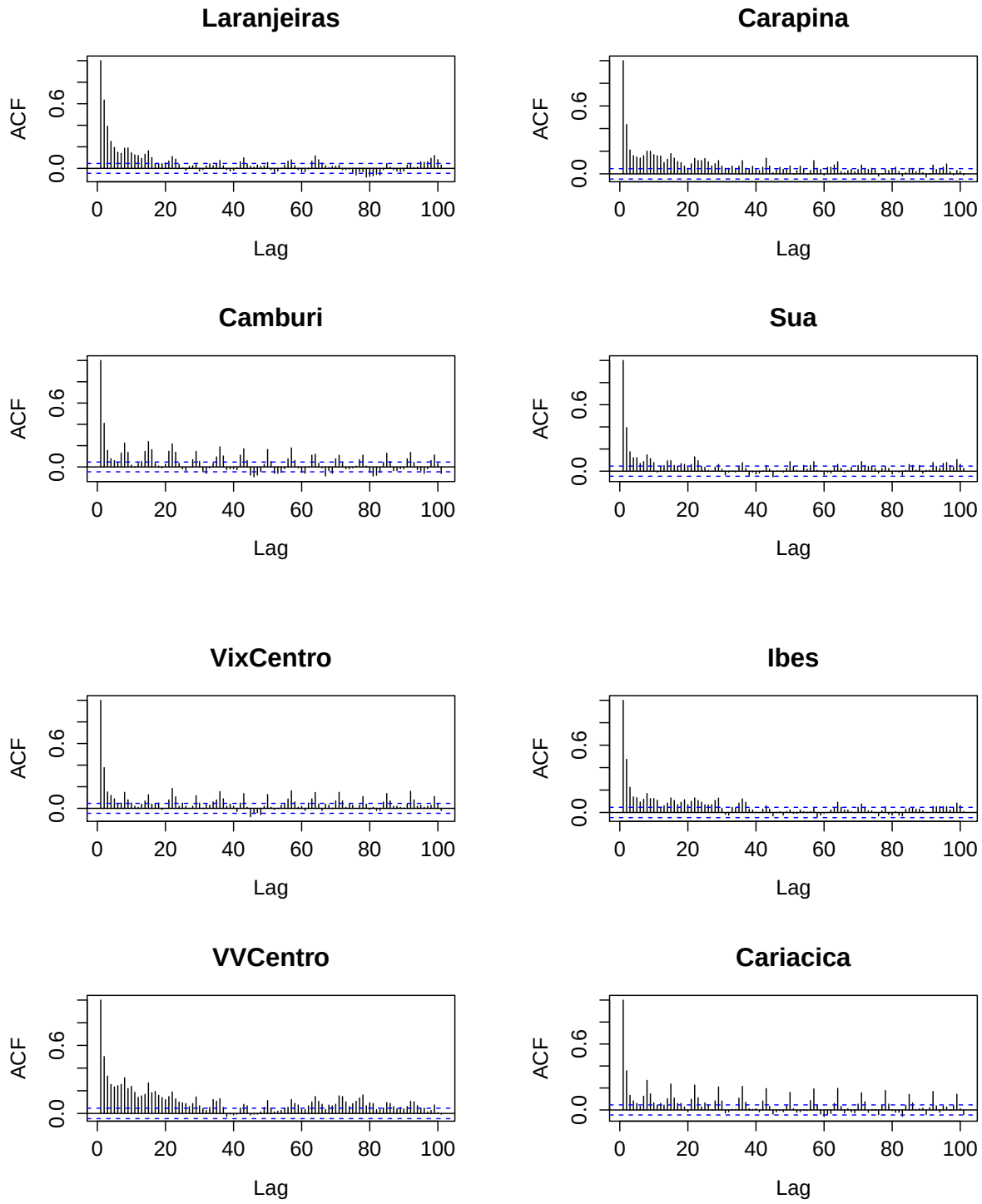


Figura 10: Refere-se à Figura 7 do artigo.