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

ADRIANO MARCIO SGRANCIO

ANÁLISE FATORIAL EM SERIES TEMPORAIS COM LONG-MEMORY, OUTLIERS E SAZONALIDADE: APLICAÇÃO EM POLUIÇÃO DO AR NA REGIÃO DA GRANDE VITÓRIA, ES

VITÓRIA
2015

ADRIANO MARCIO SGRANCIO

ANÁLISE FATORIAL EM SERIES TEMPORAIS COM LONG-MEMORY, OUTLIERS E SAZONALIDADE: APLICAÇÃO EM POLUIÇÃO DO AR NA REGIÃO DA GRANDE VITÓRIA, ES

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

VITÓRIA
2015

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

Sgrancio, Adriano Marcio, 1970-

S523a Análise fatorial em series temporais com long-memory, outliers e sazonalidade : aplicação em poluição do ar na região da Grande Vitória-ES / Adriano Marcio Sgrancio. – 2015.

89 f. : il.

Orientador: Valdério Anselmo Reisen.

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

1. Análise fatorial. 2. Ar – Poluição. 3. Análise de séries temporais. 4. Valores estranhos (Estatística). 5. Estatística robusta. 6. Dióxido de enxofre. 7. Material particulado. I. Reisen, Valdério Anselmo. II. Universidade Federal do Espírito Santo. Centro Tecnológico. III. Título.

CDU: 628

AGRADECIMENTOS

- Agradeço a Deus, acima de tudo, por ter me dado a vida, o amor, a inteligência, a saúde, a minha esposa querida, os meus filhos amados, e por ter me conduzido até aqui. E agradeço a todos que contribuíram para a realização deste trabalho, e principalmente:
- à minha esposa, Adriana, pelo grande amor, companheirismo, compreensão, dedicação e por cobrir todas as necessidades da nossa família na minha ausência;
- aos meus filhos Mateus e Vitor pelo grande amor e por terem compreendido a minha ausência e acreditado na minha vitória;
- aos meus pais e irmãos, especialmente à minha mãe, Maria da Conceição, pelos cuidados com meus filhos;
- à minha sogra Zenir, meus cunhados Cristiane e Gustavo pelos cuidados com minha família;
- ao meu orientador, Valdério, pela dedicação, orientação precisa, conhecimentos compartilhados, envolvimento na pesquisa, confiança e amizade;
- ao prof Flavio pelas orientações e direcionamentos da pesquisa;
- aos professores Neyval, Taciana e Jane, pelas orientações e pelos conhecimentos transmitidos;
- aos professores da banca examinadora Bovas e Aerambamoorthy pelas sugestões e críticas a este trabalho;
- aos demais professores, coordenadores e funcionários do PPGEA, pelo apoio;
- aos amigos Wanderson, Bartolomeu, Nátaly e Edilson pelo apoio, e em especial ao Edson, Higor, Alessandro e Fabio pela incessante ajuda;
- aos alunos do NUMES e do PPGEA;
- aos professores e demais servidores do IFES pela viabilidade do doutorado, especialmente ao grande amigo José Geraldo;
- ao grande amigo Carlos Marcelo pelo apoio, compreensão e confiança;
- aos meus alunos pelo apoio e compreensão;
- ao IFES, CAPES, CNPQ, IEMA, FAPES pelo apoio a esta e outras pesquisas do nosso grupo de trabalho.

“ Dá, pois, ao teu servo um coração cheio de discernimento para governar o teu povo e capaz de distinguir entre o bem e o mal “
Reis 3:5-15 (Salomão)

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

Os estudos de poluição atmosférica geralmente envolvem medições e análises de dados de concentrações de poluentes, como é o caso do MP_{10} (material particulado), de SO_2 (dióxido de enxofre) e de outros poluentes. Estes dados normalmente possuem características importantes como autocorrelação, longa dependência, sazonalidade e observações atípicas, que necessitam de ferramentas de análise de séries temporais multivariadas para avaliar o seu comportamento na atmosfera. Neste contexto, propomos um estimador fracionário robusto da matriz de autocovariância robusta de longa dependência e frequência sazonal, para o modelo SARFIMA. O interesse prático em poluição do ar é avaliar o comportamento das séries de concentrações de SO_2 e fazer as previsões, mais acuradas, deste poluente. As previsões, do modelo SARFIMA estimado, são comparadas às previsões do modelo SARMA, através do erro quadrático médio.

Existe outra dificuldade na investigação dos poluentes atmosféricos, por modelos de séries temporais: os dados de SO_2, de MP_{10} e de outros poluentes possuem alta dimensionalidade. Este fato dificulta o tratamento dos dados através de modelos vetoriais autorregressivos, pelo excessivo número de parâmetros estimados. Na literatura, a abordagem do problema para séries temporais de grandes dimensões é feita através da redução da dimensionalidade dos dados, utilizando, principalmente, o modelo fatorial e o método de componentes principais. Porém, as características de longa dependência e de observações atípicas das séries de poluição atmosférica, normalmente, não são envolvidas na teoria de análise fatorial. Neste contexto, propomos aqui uma contribuição teórica para o modelo fatorial de séries temporais de grandes dimensões, envolvendo longa dependência e robustez na estimação dos fatores. O modelo sugerido é aplicado em séries de MP_{10} da rede de monitoramento da qualidade do ar da Grande Vitória - ES.

Palavras-chave: análise fatorial, poluição do ar, análise de séries temporais, *outliers*, robustez, longa dependência, MP_{10} e SO_2.

ABSTRACT

Studies about air pollution typically involve measurements and analysis of pollutants, such as PM_{10} (particulate matter), SO_2 (sulfur dioxide) and others. These data typically have important features like serial correlation, long dependency, seasonality and occurence of atypical observations, and many others, which may be analyzed by means of multivariate time series. In this context, a robust estimator of fractional robust autocovariance matrix of long dependence and seasonal frequency for SARFIMA model is proposed. The model is compared to SARMA model and is applied to SO_2 concentrations.

In addition of the mentioned features the data present high dimensionality in relation to sample size and number of variables. This fact complicates the analisys of the data using vector time series models. In the literature, the approach to mitigate this problem for high dimensional time series is to reduce the dimensionality using the factor analysis and principal component analysis. However, the long dependence characteristics and atypical observations, very common in air pollution series, is not considered by the standard factor analysis method. In this context, the standard factor model is extended to consider time series data presenting long dependence and outliers. The proposed method is applied to PM_{10} series of air quality monitoring network of the Greater Vitória Region - ES.

Keywords: factor analysis, air pollution, time series analysis, MP_{10}, SO_2, outliers, robustez and long memory.

SUMÁRIO

1 INTRODUÇÃO

A qualidade do ar de uma região é caracterizada pelos níveis de emissão dos poluentes, pela capacidade com que a atmosfera da região consegue absorver, dispersar e remover estes contaminantes. A qualidade do ar também é influenciada pela ocorrência de diversos fenômenos meteorológicos e pela topografia da região, que podem ampliar ou reduzir a capacidade do transporte e de dispersão atmosférica dos poluentes.

As emissões de poluentes atmosféricos podem ser classificadas em antropogênicas e naturais (GODISH, 1997). As emissões antropogênicas são aquelas provocadas pela ação do homem, geralmente nas indústrias, nos transportes e em processos de geração de energia. As emissões naturais são causadas por processos naturais, tais como as emissões vulcânicas, os processos microbiológicos, etc. Quanto à origem, os poluentes são classificados em níveis primários ou secundários. Os poluentes primários são aqueles lançados diretamente na atmosfera, como resultado dos processos industriais, dos gases de exaustão dos motores de combustão interna, construção civil e outros. Os poluentes secundários são aqueles formados a partir de reações químicas que ocorrem entre os poluentes primários, na atmosfera.

O material particulado (MP) é considerado um poluente atmosférico pelo potencial danoso à saúde das pessoas, dos animais e da vegetação; pela interferência nas mudanças climáticas regionais e globais e pelo incômodo de sua deposição nas superfícies dos materiais e edificações (WHO, 2005; JACOBSON, 2002). A gravidade dos danos causados pelo MP levou a uma maior abordagem sobre o assunto na literatura e ao crescimento do monitoramento de fontes de emissões e de dados da qualidade do ar em várias regiões do mundo.

O material particulado é composto de partículas capazes de permanecer em suspensão na atmosfera devido às suas pequenas dimensões. Como exemplos podem ser citados a poeira, a fuligem e as partículas de óleo (BRAGA et al., 2005). A forma e a composição química do MP podem ser bastante diversificadas. Normalmente a classificação é feita de acordo com o tamanho da partícula: nos casos de diâmetros aerodinâmicos inferiores a $2.5\mu m$ e $10\mu m$ são denominados $MP_{2.5}$ e MP_{10}, respectivamente. Na literatura, o MP_{10} também é definido como partículas inaláveis. Segundo Baird (2001), as partículas são classificação como partículas grossas (diâmetro maior que $2.5\mu m$) e finas (diâmetro menor que $2.5\mu m$). A importância do tamanho das partículas está relacionada aos danos que elas podem causar à saúde. Holgate et al. (1999) afirma que as partículas finas são as principais responsáveis por esses danos, uma vez que podem atingir e prejudicar o sistema respiratório inferior.

Tendo em vista os efeitos do material particulado na saúde e no meio ambiente, surgiram

as legislações ambientais para regulamentar os níveis de emissões e de qualidade do ar. A legislação brasileira, através da resolução CONAMA nº 3 de 1990, estabeleceu os seguintes padrões primário e secundário de concentração de partículas inaláveis, MP_{10}): (1) a concentração média aritmética anual deve ser no máximo de $50\mu g/m^3$ e (2) a concentração média de 24 horas deve ser no máximo de $150\mu g/m^3$ e não podendo ser ultrapassada mais de uma vez por ano. A Organização Mundial de Saúde (OMS) estabeleceu como diretriz $50\mu g/m^3$ para a concentração média de 24 horas e $20\mu g/m^3$ para a média artimética anual de MP_{10} (WHO, 2005).

Outro poluente atmosférico de grande importância é o dióxido de enxofre (SO_2). Este gás faz parte de um grupo de gases altamente reativos conhecidos como "óxidos de enxofre". A preocupação se deve ao grande número de fontes de emissão do gás para a atmosfera e aos efeitos causados ao meio ambiente e à saúde da população.

O SO_2 é gerado principalmente pela utilização de combustíveis fósseis (carvão e produtos petrolíferos). As fontes de emissão do gás para a atmosfera incluem processos industriais, tais como siderurgia e mineração; queima de combustíveis que contêm alto teor de enxofre, veículos de transporte, tais como locomotivas e navios de grande porte; indústria de celulose e fontes naturais, como as emissões vulcânicas (MALLIK; VENKATARAMANI; LAL, 2012; ANDERSSON et al., 2013). Os altos níveis de emissão de SO_2 aliado às condições meteorológicas e topográficas podem resultar em picos de concentração de SO_2 na atmosfera.

A preocupação com as altas concentrações de SO_2 na atmosfera se agrava com os vários efeitos sobre a saúde da população, como no caso da bronquite crônica (KANAROGLOU et al., 2013). A exposição a SO_2, mesmo por curtos períodos de tempo (variando desde 5 minutos até 24 horas), pode causar problemas respiratórios, incluindo o aumento dos sintomas de asma e, consequentemente, maior número de internações hospitalares, principalmente em idosos, crianças e pessoas com pré-disposição a doenças respiratórias (WHO, 2005). Além disto, o SO_2 é oxidado na troposfera, formando ácido sulfúrico (H_2SO_4) que é depositado na superfície da terra, através da chuva ácida (GEORGOULIAS et al., 2009).

No caso do SO_2, o padrão primário estabelecido pelo CONAMA é $80\mu g/m^3$ para a concentração média anual, e $365\mu g/m^3$ para a concentração média de 24h, que não deve ser excedido mais de uma vez ao ano. E o padrão secundário é de $40\mu g/m^3$ para a concentração média anual, e de $100\mu g/m^3$ para a concentração média de 24h, que não deve ser excedido mais de uma vez ao ano. A Organização Mundial de Saúde estabeleceu como diretriz $500\mu g/m^3$ para a média de 10 minutos, e $20\mu g/m^3$ para a média artimética anual de SO_2 (WHO, 2005).

Para manter os níveis de concentrações dentro dos padrões da legislação, é necessário fazer o

controle dos poluentes, através do monitoramento da qualidade do ar. A região metropolitana da Grande Vitória (RGV) possui uma Rede Automática de Monitoramento da Qualidade do Ar (RAMQAR), que pertence ao Instituto Estadual de Meio Ambiente e Recursos Hídricos (IEMA). A rede foi inaugurada em julho de 2000 e fornece dados horários de concentração de MP e de SO_2.

Esta disponibilidade de dados monitorados em intervalos de tempo igualmente espaçados, aliada a necessidade de estimar a poluição do ar e fazer previsões destes poluentes, justificam a abordagem de séries temporais de poluição atmosférica feita para a RGV. Outra justificativa para a escolha da RGV se deve aos seguintes fatos: a RGV possui fontes de emissão de poluentes do ar nas indústrias e no crescente desenvolvimento urbano. E acrescentando, no que se refere ao efeito dos poluentes na saúde da população, a região apresentou aumento do número de atendimentos hospitalares por doenças respiratórias e cardiovasculares em função do crescimento da região (SOUZA et al,2014).

Na literatura atual, as séries de concentrações de SO_2 têm sido abordadas através de modelos que avaliam os impactos do SO_2 no meio ambiente, e investigam a influência das fontes de emissão e de condições meteorológicas nas concentrações de SO_2. Destacam-se os estudos envolvendo análise de componentes principais e de cluster (Cheng e Lam, 2000), análise de regressão múltipla (LUVSAN et al., 2012) e, especialmente, pesquisas utilizando modelos de séries temporais, tais como autorregressivos integrados de médias móveis (ARIMA) (PRYBUTOK; YI; MITCHELL, 2000; HASSANZADEH; HOSSEINIBALAM; ALIZADEH, 2009).

Contribuições teóricas ao modelo ARIMA foram feitas por Granger e Joyeux (1980) e Hosking (1981) que acrescentaram as propriedades de longa dependência e propuseram o modelo ARFIMA. Reisen et al. (2014a) estenderam o modelo ARFIMA e estimaram os parâmetros fracionários e sazonais do modelo SARFIMA e, posteriormente, Reisen et al. (2014b) estudaram as propriedades do modelo SARFIMA para um e dois períodos sazonai, e componentes de memória longa e curta, permitindo a aplicação mais apurada desta classe de modelos em séries de poluição do ar.

Embora o estudo de modelos de memória longa, na presença de outliers, seja um assunto de interesse recente para os investigadores, especialmente nas áreas de economia e finanças (Tolvi, 2003), poucos trabalhos têm sido dedicados a este tema na área ambiental.

Neste contexto, existe uma grande preocupação em estimar os modelos de séries temporais utilizando a autocovariância na sua forma clássica, e posterior utilização em dados de poluentes atmosféricos. Com isto, propomos um estimador robusto para os parâmetros não-sazonais

e sazonais do modelo SARFIMA, baseado na autocovariância robusta. O estimador robusto dos parâmetros do modelo SARFIMA está apresentado no artigo 1 desta tese, onde é feita a combinação do modelo SARFIMA com o estimador robusto da covariância, e sugerido um estimador fracional robusto para caracterizar a longa dependência na presença de sazonalidade. O efeito da ocorrência de observações atípicas também é incluído no processo de estimação da matriz de covariância clássica e de covariância robusta. A contribuição prática deste artigo na área de poluição atmosférica é feita através da utilização do modelo SARFIMA, em previsões de séries de concentrações de SO_2 da RAMQAR. É feita também uma comparação do modelo SARFIMA com o modelo SARMA no processo de previsão da série de SO_2.

Em pesquisas recentes sobre a poluição atmosférica, muita ênfase é dada aos modelos matemáticos denominados modelos receptores. Os modelos receptores são importantes por identificarem as fontes emissoras de poluentes no receptor (VIANA et al., 2006). Esses modelos têm sido amplamente estudados nos últimos anos, recebendo contribuições de autores de diversas áreas. Segundo Seinfeld e Pandis (2006) os modelos receptores identificam as fontes de emissão de poluentes a partir das características químicas das partículas no receptor, e das características das fontes de emissão dos poluentes. O artigo publicado por Belis et al. (2013) mostra a utilização dos modelos receptores no mundo. O autor faz uma avaliação crítica, comparando os diversos modelos receptores que identificam as fontes de emissão de MP nos países da Europa.

Na literatura, os modelos receptores mais estudados são: balanço químico de massa (BQM), técnicas de análise multivariada, técnicas de análise de componentes principais (ACP), modelo fatorial (AF), regressão linear múltipla, análise de cluster, fatoração de matriz positiva (FMP), entre outros (WATSON et al., 2002). Segundo Trindade (2009) os modelos receptores mais comumente empregados são o BQM, a FMP e o modelo fatorial. Informações referentes à utilização dos modelos receptores são disponibilizadas pela United States Environmental Protection Agency (COULTER, 2004).

Os estudos referentes aos modelos receptores também são relevantes em pesquisas desenvolvidas na área de poluição atmosférica do PPGEA/UFES. Alguns estudos utilizaram os modelos receptores no processo de identificação das fontes de material particulado: (1) Trindade (2009) identificou as fontes de emissão de $MP_{2.5}$ e avaliou as informações fornecidas pelos modelos BQM e FMP para relacionar as fontes ao receptor, na cidade de Brighton, região rural do Colorado, EUA. A autora verificou que os dois modelos reproduziram os dados do receptor com ajustes aceitáveis; (2) Maioli (2011)desenvolveu um trabalho experimental coletando e analisando amostras de $MP_{2.5}$ na região metropolitana da Grande Vitória, ES. A autora utilizou o modelo receptor BQM para identificar as contribuições e a relevância dos

grupos de fontes de emissão responsáveis pelo $MP_{2.5}$ no receptor; (3) Soares (2011) estudou a contribuição das fontes de partículas totais em suspensão (PTS) e de partículas sedimentáveis (PS) na região metropolitana da Grande Vitória, ES utilizando os modelos BQM e FMP. A análise fatorial está dentro do contexto da poluição do ar. Os principais estudos nesta área envolvem a identificação das fontes de emissão de poluentes, o gerenciamento das redes de monitoramento, a análise de regressão, a análise de cluster, dentre outros (PIRES et al., 2008a; PIRES et al., 2008b; VIANA et al., 2006).

Acrescenta-se que os parâmetros do modelo fatorial são normalmente estimados através da análise de componentes principais (ACP) que, assim como o modelo fatorial, é afetada pela a autocorrelação dos dados. Consequentemente o problema é transportado das componentes principais para o modelo fatorial que, neste caso, sofre as consequências do problema de utilizar dados correlacionados no tempo. Neste contexto, a tese de Zamprogno (2013), desenvolvida no PPGEA, explora este fato e revela que negligenciar a estrutura de autocorrelação temporal das observações pode acarretar em análises e interpretações equivocadas e podem inviabilizar a estimação das componentes principais quando os dados não forem estacionários na média. Zamprogno (2013) também aplica os conceitos de ACP no processo de identificação de fontes de emissão e no gerenciamento da RAMQAR através da redução da dimensão dos dados de poluição do ar.

Atualmente, trabalhos desenvolvidos no PPGEA/UFES abordam o tema em questão: (1) Souza (2013) aplica a técnica de ACP em dados de poluição da rede de monitoramento da qualidade do ar da região da Grande Vitória, objetivando avaliar a associação entre: o número de atendimentos hospitalares por causas respiratórias e a qualidade do ar na região; (2) Pinto (2013) trata do processo de estimação das autocorrelações das séries, utilizando dados faltantes; (3) Cotta (2014) analisa as componentes principais pela técnica de estimação robusta, considerando os outliers nas séries de dados de poluentes na Grande Vitória; (4) Monroy (2013) estuda a dependência entre observações monitoradas em cada região da RAMQAR e a dependência sequencial das séries temporais de poluição do ar, e estima uma classe de modelos espaço-temporais, considerando os fenômenos de longa dependência que ocorrem nas séries de poluição do ar.

O desenvolvimento teórico e a aplicação prática do modelo fatorial em séries temporais de grandes dimensões já é uma linha de pesquisa na área de econometria. O modelo fatorial é amplamente utilizado na área econômica, principalmente em séries temporais financeiras. Segundo Bai (2003) o modelo fatorial é um método útil para tratar grandes conjuntos de dados econômicos. O autor desenvolve a teoria inferencial para modelos fatoriais aplicados a grandes dimensões de séries temporais.

O uso do modelo fatorial na área de poluição do ar é um dos fatores que motivam o desenvolvimento desta pesquisa. Os trabalhos citados enfatizam o processo de redução da dimensionalidade de séries temporais de poluição do ar, e do gerenciamento das redes de monitoramento da qualidade do ar. Entretanto, a preocupação com a estrutura de autocorrelação, presente nestes dados, motivaram as investigações apresentadas aqui.

Outra motivação, na abordagem teórica, feita aqui, é proveniente do seguinte fato: de acordo com Johnson, Wichern et al. (1992), Harman (1960) e Anderson et al. (1958), o modelo fatorial, para ser aplicado na sua forma clássica, exige a hipótese de independência dos dados. Essa é uma grande limitação para utilizá-lo em séries temporais de poluentes atmosféricos. O monitoramento das séries de poluentes consiste em medições de concentrações feitas no tempo e, consequentemente, as séries de dados apresentam a autocorrelação temporal, o que não atende à hipótese de independência dos dados.

O ajuste do modelo fatorial em séries temporais de poluição do ar, sem considerar a autocorrelação dos dados, pode levar à estimação da matriz de covariância de forma viesada, ou seja, ocorre um erro na estimação da matriz, que pode gerar resultados inconsistentes no modelo fatorial e comprometer a avalição das contribuições das fontes responsáveis pela poluição atmosférica (ZAMPROGNO, 2013). Este problema não foi devidamente explorado em trabalhos desenvolvidos na área de poluição atmosférica, justificando as investigações propostas neste estudo.

O interesse teórico do problema da relação de dependência entre as séries temporais é de longa data. Pena e Box (1987) desenvolveram uma importante metodologia que utiliza a análise fatorial para identificar os fatores que representam as séries temporais multivariadas, considerando a correlação temporal e a autocorrelação. Lam, Yao et al. (2012) se basearam no modelo de Pena e Box e desenvolveram uma teoria assintótica, para estimar a dimensão de redução das séries temporais multivariadas não correlacionadas.

Entretanto, Lam, Yao et al. (2012) trabalharam somente com a correlação temporal em séries de memória curta. O artigo de Pena e Box (1987) utilizou a correlação temporal e a autocorrelação, para estimar o modelo. Porém, existem limitações para utilização das teorias desses autores em dados de poluição atmosférica, visto que as séries de poluição normalmente apresentam propriedades específicas em sua estrutura temporal. Dentre estas propriedades encontra-se a longa dependência, que influencia de forma significativa na estimação da matriz de covariância do modelo fatorial.

Portanto, uma contribuição, proposta nesta tese, é agregar as propriedades de longa dependência à análise fatorial proposta na teoria de Pena e Box (1987) e Lam, Yao et al.

(2012), permitindo uma melhoria significativa na estimação dos modelos fatoriais e, consequentemente, na aplicação do modelo fatorial em dados de poluição atmosférica. Estas contribuições propostas estão apresentadas no artigo 2.

Outro fato de grande relevância, considerado aqui, provém das seguintes limitações: os artigos de Pena e Box (1987) e Lam, Yao et al. (2012) não consideram o problema de observações atípicas (outliers) que ocorrem nas séries temporais. Estas observações atípicas são muito comuns em dados de concentrações de poluentes atmosféricos. Normalmente, estas altas concentrações são geradas por emissões excessivas de fontes de poluição em curtos intervalos de tempo ou por condições meteorológicas adversas. Neste sentido, propomos uma contribuição teórica, com o objetivo de agregar o efeito das observações atípicas ao modelo fatorial. Além disso, propomos uma contribuição de ordem prática através da utilização do modelo fatorial na área de poluição atmosférica. Estas contribuições estão apresentadas no artigo 2 desta tese.

As observações atípicas revelam um problema que é recorrente na área de poluição do ar: as estações de monitoramento da qualidade do ar normalmente medem picos de concentrações de poluentes, em curtos intervalos de tempo. Estes picos de concentração podem representar uma deteriorização da qualidade do ar de uma região. A influência desses outliers em séries de concentrações de MP_{10} da RAMQAR (IEMA, 2013) também está apresentada no artigo 2. A investigação é feita através da avaliação da influência dessas observações atípicas na estimação da matriz de correlação robusta e, consequentemente, no modelo fatorial. Estudos referentes a esta influência foram feitos por Cotta (2014).

Portanto, o modelo fatorial, proposto no artigo 2, é construído com base no processo de redução da dimensionalidade das séries temporais multivariadas que apresentam sazonalidade, longa dependência e observações atípicas. Inicialmente estes comportamentos das séries foram tratatados de forma univariada no modelo SARFIMA, proposto no artigo 1. As contribuições teóricas e práticas do artigo 1 foram utilizadas na construção do artigo 2. E, em termos de aplicação prática na área de poluição do ar, o artigo 1 avaliou uma série de SO_2 e, o artigo 2, as séries multivariadas de MP_{10}. Em ambos os casos, os modelos estimados foram utilizados para a previsão das concentrações destes poluentes na Grande Vitória.

O texto desta tese está dividido da seguinte forma: na subseção seguinte são apresentados os objetivos: geral e específicos; na seção 2 é apresentada uma revisão da literatura referente à estimação dos parâmetros do modelo fatorial para séries dependentes e independentes, longa dependência e outliers; na seção 3 são apresentados os materiais e métodos utilizados; na seção 4 são apresentados os resultados da tese na forma de dois artigos: o artigo 1 e o artigo

2, que incluem a teoria e os resultados; na seção 5 são apresentadas as conclusões e, em sequência, as referências utilizadas.

1.1 OBJETIVOS

1.1.1 Objetivo Geral

O objetivo geral desta tese é propor um modelo fatorial robusto para séries temporais correlacionadas de grandes dimensões de poluentes atmosféricos, e que apresentam as propriedades de longa dependência, sazonalidade e observações atípicas.

1.1.2 Objetivos Específicos

Os objetivos específicos desta tese são os seguintes:

- Estimar um modelo fatorial para séries temporais multivariadas que apresentem autocorrelação, longa dependência, observações atípicas e sazonalidade e, consequentemente, reduzir a dimensão das séries;

- Aplicar o modelo fatorial em séries de MP_{10} da rede de monitoramento da qualidade do ar da Grande Vitória, utilizado a matriz de autocovariância e de autocovariância robusta, no processo de estimação dos fatores e da previsão das concentrações do poluente;

- Desenvolver o modelo sazonal para séries temporais univariadas com longa dependência e observações atípicas (SARFIMA);

- Aplicar o modelo SARFIMA em dados de concentração de SO_2 da rede de monitoramento da qualidade do ar da Grande Vitória, e utilizar o modelo para fazer a previsão de concentrações do poluente.

2 REVISÃO BIBLIOGRÁFICA

A necessidade de conhecer a origem dos contaminantes atmosféricos tem levado os cientistas a inovarem as técnicas experimentais e as ferramentas matemáticas e estatísticas, que permitem identificar as fontes de emissão dos contaminantes, através da sua composição física e química no receptor, e também caracterizar a qualidade do ar de uma região.

Segundo (WATSON et al., 2002), os principais modelos receptores são: Balanço químico de massa (BQM), análise multivariada, regressão linear múltipla, análise de cluster, entre outros. A análise multivariada pode ser separada em: modelo de fatoração de matriz positiva (FMP), análise fatorial (AF) e análise de componentes principais (ACP). O modelo fatorial é desenvolvido aqui para aplicações em séries de grandes dimensões de poluentes atmosféricos.

Na litaratura, os trabalhos científicos procuraram avaliar e aprimorar os modelos receptores e utilizá-los na poluição do ar. Se destacam as pesquisas de Usero e Gracia (1986), Paatero et al. (2002), Anderson et al. (2002), Hopke (1991), Lee et al. (2008), dentre outros. Pesquisas recentes no Brasil, envolvendo modelos receptores, foram desenvolvidas por Albuquerque (2005), Maioli (2011), Trindade (2009), Soares (2011), Souza (2013), Zamprogno (2013) e Cotta (2014).

Alguns trabalhos aplicaram os modelos receptores BQM e FMP e merecem destaque: (1) Trindade (2009) que avaliou as informações fornecidas pelos modelos receptores BQM e FMP, para relacionar as fontes de $MP_{2.5}$ no receptor; (2) Maioli (2011) coletou amostras de $MP_{2.5}$ na região metropolitana da Grande Vitória, ES e utilizou o modelo receptor BQM para identificar as contribuições e a relevância dos grupos de fontes de emissão; (3) Soares (2011) utilizou o BQM e FMP para determinar a contribuição das fontes de partículas totais em suspensão (PTS) e de partículas sedimentáveis (PS) na região metropolitana da Grande Vitória, ES.

A importância do modelo fatorial na área de poluição atmosférica, está no fato de poder ser usado para estimar a composição das fontes de emissão, através de monitoramento de elementos químicos coletados em um determinado local (receptor), e contribuir no processo de identificação das fontes de emissão que influenciam no receptor (SEINFELD; PANDIS, 2006).

Segundo Seinfeld e Pandis (2006), as principais suposições de análise fatorial para aplicação do modelo em poluição atmosférica são: as espécies químicas usadas na modelagem não interagem entre si; os erros de medições são aleatórios e não correlacionados; a variabilidade

das concentrações entre as amostras ocorre, principalmente, por alterações de contribuições da fonte e não por medições incertas e mudanças na composição da fonte; o efeito de processos que afetam todas as fontes igualmente é muito menor que o efeito de processos que influenciam fontes individuais; e por último, existem muito mais amostras que os tipos de fontes, para cálculos estatísticos significantes.

O modelo fatorial começou a ser estudado na área da psicologia, com avaliações da habilidade mental. A análise era feita utilizando matrizes de correlações de um conjunto de testes cognitivos. Esta técnica foi estudada por Charles Spearman, em 1904, que apresentou a teoria do fator de inteligência geral. Os resultados mostraram que o grau de inteligência geral poderia ser descrito por um fator, e a inteligência específica poderia ser expressa por um fator relacionado ao assunto ou tipo do teste (SPEARMAN, 1904) e Johnson, Wichern et al. (1992).

Segundo Mingoti (2005), a análise fatorial tem como objetivo principal descrever a variabilidade do vetor aleatório original, em termos de um número menor de variáveis aleatórias, chamadas fatores comuns. Os fatores são relacionados ao vetor original por um modelo linear. Uma parte da variabilidade do vetor aleatório original é atribuída aos fatores comuns, e a outra parte é atribuída aos erros aleatórios, que são as variáveis que não foram incluídas no modelo. O modelo AF agrupa as variáveis originais em subconjuntos (fatores), de novas variáveis mutuamente não correlacionadas. No caso de um grande conjunto de variáveis medidas e correlacionadas entre si, o modelo fatorial permite identificar um número menor de variáveis alternativas, capazes de explicar as informações contidas nos dados originais. Os valores numéricos dos fatores identificados são denominados escores.

O problema a ser resolvido pela análise fatorial é estimar a relação linear entre o vetor aleatório (representando os dados), a matriz dos fatores e o respectivo vetor de erros aleatórios. Na literatura, os métodos de estimação mais utilizados são: O método de máxima verossimilhança, que é indicado quando o vetor aleatório tem distribuição normal multivariada; e o método de componentes principais, que tem a vantagem de não exigir suposições sobre a distribuição de probabilidade do vetor aleatório (ANDERSON et al., 1958).

Nos últimos anos, a análise multivariada, envolvendo o modelo fatorial e componentes principais, tem sido muito utilizada no processo de identificação de fontes de emissão de poluentes (BELIS et al., 2013; WATSON et al., 2002). Uma das vantagens destes modelos é a não exigência do conhecimento da fonte, utilizando apenas dados de composição química do contaminante, medidos no receptor. As limitações desses métodos são a dificuldade de interpretação dos resultados, e a necessidade de grande quantidade amostral das concentrações

do contaminante no receptor (HOPKE, 1991).

Pires et al. (2008a), Pires et al. (2008b) utilizaram ACP para identificar locais similares de poluição do ar e identificar as fontes emissoras na região metropolitana de Oporto, Portugal. Viana et al. (2006) identificaram as fontes de material particulado através de ACP e avaliaram a influência dos dados de direção do vento, em área de grande industrialização no norte da Espanha.

Além da utilização de ACP em dados de material particulado e em dados meteorológicos, a técnica pode ser empregada para avaliar e dimensionar uma rede de monitoramento ambiental: Pires et al. (2009) utilizou a ACP para identificar as medições desnecessárias, ou de pouca relevância, no gerenciamento da rede de monitoramento da qualidade do ar na região metropolitana de Oporto, Portugal. Os resultados mostraram que poderiam ser alterados os pontos de monitoramento de poluentes.

Viana et al. (2008b) sugeriram combinações entre os modelos AF, ACP, FMP com o BQM para interpretar as fontes e avaliar as limitações de cada modelo receptor. Os autores utilizaram a ACP como método de estimação dos loadings do modelo fatorial. Os autores Johnson, Wichern et al. (1992) também sugeriram utilizar a ACP como método de estimação dos loadings do modelo fatorial.

No modelo fatorial clássico, o número de fatores pode ser estimado pelo método das componentes principais (ACP). Neste caso, a escolha do número de fatores é determinada pelos autovalores não nulos da matriz de correlação. Escolhe-se o número de fatores que explicam certa proporção da variância amostral total (JOHNSON; WICHERN et al., 1992). O número de fatores comuns do modelo deve aumentar, até que uma proporção adequada da variação total amostral seja explicada. Johnson, Wichern et al. (1992) afirmam que o número de fatores deve ser menor que o número de variáveis, para que o sistema de equações do modelo fatorial seja linearmente dependente, e a variabilidade seja explicada por um número menor de fatores do que o número de variáveis.

A análise fatorial apresenta a dificuldade de interpretação dos fatores estimados no modelo. Para uma melhor interpretação dos fatores é comum fazer uma rotação ortogonal nos mesmos. Segundo Mingoti (2005), a rotação ortogonal mantém a orientação original entre os fatores, mantendo-os ortogonais após esta transformação. Após a estimação do modelo fatorial, é sempre possível encontrar uma nova solução através de uma matriz estimada rotacionada (JOHNSON; WICHERN et al., 1992). Isto permite concluir que a solução do modelo fatorial não é única. Esta possibilidade de diversas soluções se torna uma grande vantagem, por permitir a busca interativa da solução mais fácil de ser interpretada, e que melhor explique

o comportamento das variáveis analisadas.

Existem vários métodos de rotação fatorial, e dentre eles pode ser destacado o varimax, proposto por Kaiser (KAISER, 1958). O varimax se baseia na rotação dos fatores utilizando uma matriz ortogonal, e é comumente utilizado por apresentar soluções mais simples (MINGOTI, 2005). Segundo Johnson, Wichern et al. (1992). após a rotação dos fatores. a matriz de covariância estimada (ou correlação) não se altera. O mesmo acontece com os valores das comunalidades e com as variâncias específicas do modelo. Assim, afirma que trabalhar matematicamente com as matrizes originais ou rotacionadas é equivalente.

A estimação dos parâmetros do modelo fatorial pode ser feita se o modelo fatorial for considerado, por analogia, como um modelo de regressão linear múltipla. É possível estimar o vetor dos fatores pelo método dos mínimos quadrados ponderados (JOHNSON; WICHERN et al., 1992). Este método não faz exigências quanto à distribuição de probabilidades do vetor aleatório, consequentemente, é de grande abrangência em termos de aplicabilidade do modelo fatorial em dados que não têm, necessariamente, comportamento normal (MINGOTI, 2005).

Os estudos na área de séries temporais de poluentes atmosféricos envolvem o monitoramento de dados medidos no receptor e nas fontes de emissão. Em geral são feitas medições de variáveis que são autocorrelacionadas no tempo. O tratamento de séries temporais deve considerar as relações entre as variáveis (WEI, 1994). Cada série é uma componente de um vetor de séries temporais multivariadas, que especifica a dependência dentro de cada série, e a interdependência entre diferentes séries (HAMILTON, 1994). Por exemplo, é possível considerar o poluente MP_{10} medido no tempo como um vetor, e suas relações com outras séries de poluentes como o $MP_{2.5}$, em cada estação de monitoramento.

Além disso, as séries de poluentes atmosféricos são fornecidas em termos de médias horárias de concentrações dos poluentes e, portanto, apresentam grandes dimensões no tempo e com grande número de variáveis. Para tratar o problema da análise de séries temporais de grandes dimensões, utilizando a análise fatorial, alguns autores desenvolveram importantes teorias, apresentadas a seguir.

Pena e Box (1987) desenvolveram um método para identificar os fatores comuns num vetor de séries, na presença e na ausência de autocorrelação temporal. Neste caso, o procedimento é feito da seguinte forma: uma parte do vetor de séries temporais é escrito por uma combinação linear de fatores comuns não observáveis, e outra parte é representada pelos fatores aleatórios. Os autores ainda demonstraram que a redução da dimensionalidade ocorre através da estimação do número de autovalores e de autovetores da matriz de covariância do vetor de

séries temporais. Esta utilização de autovalores e autovetores para representar as matrizes de dados multivariados já havia sido estudada por (BOX; TIAO, 1977), dentre outros.

(BAI; NG, 2002) utilizaram a teoria de Pena e Box (1987) e desenvolveram novas abordagens do modelo fatorial, para tratar as séries de grandes dimensões. Os autores desenvolveram um procedimento estatístico formal, para um estimador consistente do número de fatores de séries temporais com alta dimensionalidade, que contribuiu significativamente com os modelos fatoriais na literatura. Inicialmente estabeleceram a taxa de convergência para os fatores estimados, que permitiu a estimativa consistente do número de fatores. Posteriormente mostraram que as simulações, utilizando o critério proposto, é convergente e com boas propriedades amostrais. Os estudos foram aplicados em dados econômicos.

Outros autores como Stock e Watson (2002) e Stock e Watson (2011)também estudaram o modelo fatorial usando o método das componentes principais, em séries temporais macroeconômicas. Utilizaram o modelo fatorial para fazer previsões em séries de alta dimensionalidade, e mostraram que as previsões usando os fatores são assintoticamente eficientes, quando o tamanho da mostra e o número de variáveis crescem. Referências sobre o modelo de previsão utilizado por Stock e Watson podem ser encontradas em (FORNI et al., 2004) e (BAI; NG, 2002).

Peña e Poncela (2006) verificaram que o modelo fatorial é mais adequado para ser utilizado no tratamento de séries temporais de grandes dimensões, se comparado à abordagem tradicional, que utiliza modelos vetoriais autorregressivos de médias móveis (VARMA). No caso do modelo VARMA, há necessidade da estimação de muitos parâmetros, tornando o modelo parcimonioso. Alguns trabalhos relacionados a este tema podem ser encontrados em Priestley, Rao e Tong (1974), Box e Tiao (1977), entre outros.

Modificações nos estimadores de (BAI; NG, 2002) foram propostas por Amengual e Watson (2007). Os autores utilizaram os estimadores propostos por Stock e Watson (2006), sugeriram um estimador consistente do número de fatores dinâmicos, e aplicaram a técnica na estimação do número de fatores dinâmicos em séries temporais econômicas de grandes dimensões.

Portanto, a teoria de Pena e Box (1987) se difundiu e foi utilizada também em diversos outros estudos. Dentre eles, uma excelente abordagem é encontrada no artigo de Lam, Yao et al. (2012), que é utilizada como base de referência para os estudos teóricos e para as simulações feitas aqui nesta tese. Lam, Yao et al. (2012) se basearam na teoria de Pena e Box (1987) e desenvolveram uma teoria assintótica para estimar a dimensão de redução das séries temporais, através do modelo fatorial.

O modelo de Lam, Yao et al. (2012) pode ser utilizado em aplicações que apresentam a

correlação temporal, em séries temporais de memória curta. O artigo de Pena e Box (1987) pode ser utilizado em aplicações em que a correlação cruzada e a autocorrelação estão presentes. Entretanto, existem limitações para utilização dessas teorias em séries de poluição atmosférica, que comumente apresentam propriedades específicas em sua estrutura temporal. Dentre estas propriedades se destaca a longa dependência, que influencia, de forma significativa, na estimação da matriz de covariância do modelo fatorial. Esta influência não está devidamente explorada na literatura atual de análise fatorial.

Paralelamente aos estudos de análise fatorial, os modelos de séries temporais se desenvolveram de forma independente. No contexto de longa dependência, a literatura é bem vasta e conhecida. Os estudos nesta área foram iniciados por Hosking (1981). Reisen (1994) estudou o modelo de longa dependência (ARFIMA) e desenvolveu um estimador do parâmetro de diferenciação fracionário do modelo, utilizando a função periodograma suavizado.

Atualmente, o efeito da sazonalidade foi incorporado ao modelo de longa dependência, com aplicações em poluição atmosférica. Neste contexto, os seguintes trabalhos são destacados na literatura: Reisen et al. (2014), Zamprogno (2013) e Monroy (2013).

Apesar da relevância das contribuições de Pena e Box (1987) e Lam, Yao et al. (2012), seus artigos não abordam outro problema bem comum nas séries de poluentes atmosféricos: observações atípicas (outliers), que estão presentes na maioria dos dados de poluição do ar. Normalmente estes outliers são frequentes e ocorrem devido a emissões excessivas de fontes de poluição, em curtos intervalos de tempo, ou por condições meteorológicas adversas.

A importância dos efeitos causados por observações atípicas no modelo fatorial deve ser considerada, pois os outliers provocam grande alteração na matriz de covariância Cotta (2014), Lévy-Leduc et al. (2011). E consequentemente, alteram a dimensão do modelo fatorial estimado através desta matriz. O artigo 2 trata o problema da presença de outliers e avalia o seu efeito no modelo fatorial.

Portanto, no que se refere às características de longa dependência e de sazonalidade, na presença de outliers em séries temporais de grandes dimensões, a literatura referente à análise fatorial necessita de contribuições teóricas e de estudos com aplicações práticas, na área de poluição atmosférica, que estão sendo propostas nesta tese.

3 MATERIAIS E MÉTODOS

3.1 REGIÃO DE ESTUDO

A Região Metropolitana da Grande Vitória (RMGV) é constituída pelos municípios de Vitória, Vila Velha, Cariacica, Serra, Fundão, Guarapari e Viana. A RMGV é localizada na região sudeste do estado do Espírito Santo. Sua área é de 2331.03 Km^2 e possui uma população de aproximadamente 1,884 milhão de habitantes, representando cerca de 48% da população total do estado (IBGE, 2011).

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

No que se refere ao relevo, a RMGV é caracterizada 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, 2013).

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

O início do funcionamento da RAMQAr foi no ano de 2000. A rede é de propriedade e responsabilidade do Instituto Estadual de Meio Ambiente e Recursos Hídricos (IEMA). Para o período de estudo, a rede é composta 9 estações de monitoramento, sendo que foram utilizados de séries históricas de dados de 8 estações distribuídas nos municípios da RMGV, da seguinte forma: 2 estações no município de Serra (Laranjeiras e Carapina); 3 estações no município de Vitória (Jardim Camburi, Enseada do Suá e Centro); 2 estações no município de Vila Velha (Ibes e Centro) e 1 estação no município de Cariacica (Ceasa). A localização espacial das estações de monitoramento da RAMQAr encontra-se na Figura 1.

Os poluentes monitorados nas 8 estações utilizadas da RAMQAR são: dióxido de enxofre,

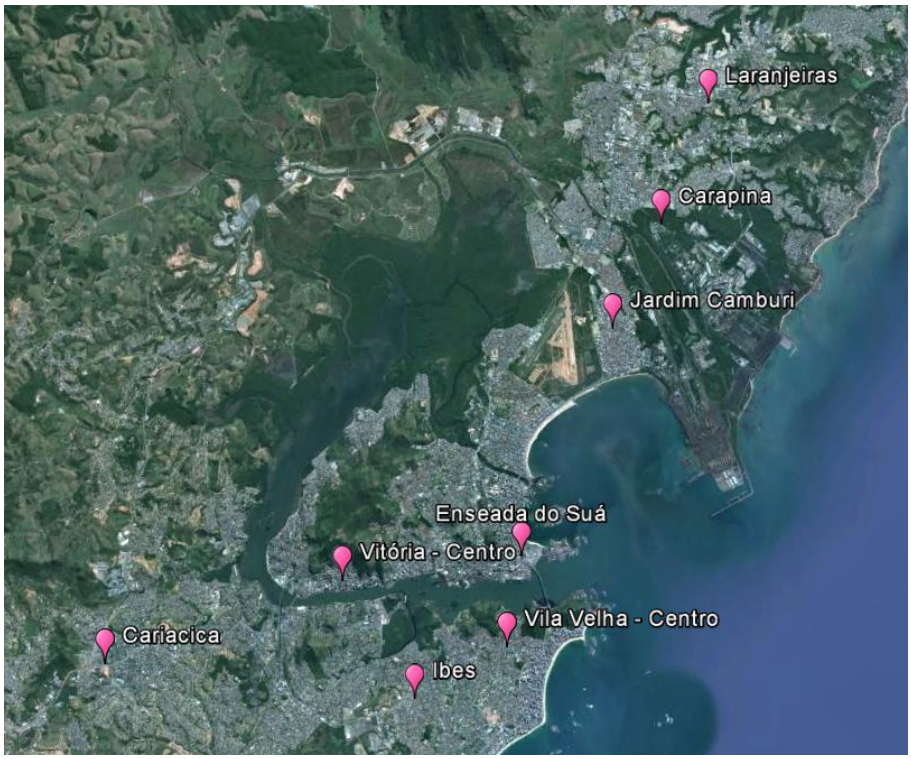


Figura 1: Localização espacial das estações da RAMQAr.

partículas totais em suspensão, 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 as estações. Os poluentes e parâmetros monitorados por cada estação estão mostrados 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

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

Este trabalho foi realizado utilizando os dados de SO_2 e de PM_{10} medidos na RAMQAR. O período de medição foi de janeiro de 2005 a dezembro de 2009. Os dados são fornecidos em médias horárias de concentrações medidas em $\mu g/m^3$.

3.3 MODELAGEM UTILIZADA

Inicialmente o comportamento dos dados da RAMQAR foi avaliado estatísticamente. Foram escolhidos os dados de SO_2 e de PM_{10} que apresentaram comportamentos específicos, necessitando de desenvolvimento teórico para serem caracterizados e para estimação de previsões mais precisas.

Em seguida foi adotado um modelo SARFIMA na presença de valores extremos (outliers), para modelar as médias diárias de concentrações de SO_2. Para estimar os parâmetros fracionais robustos d_R and D_R foram combinados os métodos sugeridos por Reisen et al. (2014b), Lévy-Leduc et al.(2011c) and Molinares et al. (2009). Baseado nos estimadores robustos $\hat{d}_R$ e $\hat{D}_R$ foi ajustado o modelo SARFIMA. O modelo SARFIMA foi utilizado para fazer a previsão das concentrações de SO_2, que foram comparadas com o modelo SARMA, através do erro quadrático médio percentual.

Posteriormente foi proposto um modelo fatorial robusto, aplicado em séries temporais de grandes dimensões com propriedades de curta e longa dependência, na presença de observações atípicas (outliers) e de sazonalidade. Alguns resultados teóricos foram discutidos e investigados, empiricamente, através dos experimentos de Monte Carlo. E por fim, o modelo fatorial foi aplicado para identificar o comportamento das concentrações de PM_{10} na Região da Grande Vitória e utilizado para prever os níveis de concentrações deste poluente. Estas previsões foram comparadas com o modelo vetorial autoregressivo (VAR).

4 RESULTADOS

Nesta seção estão os resultados desta tese, apresentados em dois artigos: o artigo 1, Fractional seasonal process with outliers to model and forecast daily average SO_2 concentrations e o artigo 2, Robust Factor Modeling for High-Dimensional Time Series with short and long memory: an Application to Air Pollution Data.

O artigo 1 aborda o efeito da sazonalidade e de memória longa, em séries temporais que apresentam outliers. O estudo é feito propondo um estimador fracional robusto para caracterizar a longa dependência na presença de sazonalidade. O efeito da ocorrência de observações atípicas também é incluído no processo de estimação da matriz de covariância clássica e de covariância robusta. A contribuição prática deste artigo, na área de poluição atmosférica, é feita através da utilização do modelo SARFIMA em previsões de séries de concentrações de SO_2 da RAMQAR. Além disso, é feita uma comparação do modelo SARFIMA com o modelo SARMA no processo de previsão da série de SO_2. O artigo foi submetido à publicação no European Journal of Operational Research (classificação A1 no sistema de avaliação WEB-QUALIS na área de Engenharias I).

O artigo 2 agrega as propriedades de sazonalidade, outliers e longa dependência ao modelo fatorial, e contribui teoricamente com a análise fatorial para séries temporais de grandes dimensões. Isto permite uma melhoria significativa na estimação dos modelos fatoriais e, consequentemente, na aplicação do modelo fatorial na área de poluição atmosférica. O modelo fatorial foi aplicado na previsão de concentrações de PM_{10} na Região da Grande Vitória, e os resultados mostraram que ele apresentou melhor acurácia na previsão, em relação ao modelo vetorial autoregressivo (VAR).

O artigo1 e o artigo 2 estão apresentados a seguir.

Fractional seasonal process with outliers to model and forecast daily average SO_2 concentrations

Valdério Anselmo Reisen[a,b,*], Adriano Marcio Sgrancio[b], Edson Zambon Monte[b], Fabio Alexander Fajardo Molinares[a], Glaura da Conceição Franco[c], Flávio Augusto Ziegelmann[d]

[a] *Department of Statistics, Federal University of Espírito Santo, Espírito Santo, Brazil*
[b] *Graduate Program in Environmental Engineering, Federal University of Espírito Santo, Espŕito Santo, Brazil*
[c] *Department of Statistics, Federal University of Minas Gerais, Minas Gerais, Brazil*
[d] *Department of Statistics, Federal University of Rio Grande do Sul, Rio Grande do Sul, Brazil*

Abstract

This paper combines the SARFIMA model with a robust autocovariance estimator to suggest robust fractional estimators at long-run and seasonal frequencies. The model and the estimation method are discussed and the usefulness of the proposed methodology is applied to SO_2 concentrations. The additive outlier effects are also discussed when dealing with a non-robust model estimation. In addition, forecasts using the proposed model were found to be superior to a robust short-memory SARMA variant, especially for long lead time forecasting. Therefore, the proposed robust estimation method and the SARFIMA model well captured the features of long-memory, seasonality and additive outliers effect presented in the SO_2 concentrations.

Keywords: Stochastic processes; Robust periodogram; Outliers; SO_2 contaminant; Long-memory.

*Corresponding author. Department of Statistics, Federal University of Espírito Santo, 29075-910, 514, Vitoria, ES, Brazil. Tel.: +5502740092903.
E-mail address: valderioanselmoreisen@gmail.com (V. A. Reisen).

1. Introduction

The sulfur dioxide (SO_2) is part of a group of highly reactive gases known as "oxides of sulfur". The main sources of SO_2 emissions into the atmosphere come from fossil fuels (coal and petroleum products), power plants and other industrial plants. Other emission sources of SO_2 include industrial processes such as steel and mining; burning fuels containing high sulfur by transport vehicles such as locomotives and large ships; pulp industry and natural sources such as volcanic emissions (Mallik et al., 2013; Andersson et al., 2013). High levels of pollution emissions can result in peaks of concentration of SO_2 in the atmosphere.

The high concentrations of SO_2 are associated with several effects on population health, as the increase of chronic bronchitis (Kanaroglou et al., 2013). The exposure to SO_2, even for short periods of time (ranging from 5 minutes to 24 hours), may cause respiratory problems, including bronchoconstrictor and increased asthma symptoms. Consequently, these can cause an increase in hospital admissions of the population at hight risk: elderly, children and asthmatics (EPA, 2014). Epidemiological studies have also shown that especially at high temperatures there is considerable rise in the association between mortality and air pollution, mainly from pollution caused by SO_2 (Basu & Samet, 2002).

Furthermore, SO_2 is oxidized in the troposphere forming sulfuric acid (H_2SO_4) which is deposited on the earth surface as acid rain (Georgoulias et al., 2009). This phenomenon causes many ecological problems, such as the reduction of water quality, the change in the growth of trees, the increase of the acidity of the soil and aquatic ecosystems and the increase of the mortality of fish and wildlife. In addition, sulfur dioxide contributes to the formation of ground-level ozone and fine particle pollution. Thus, it is important to developed statistical methods to model and forecast the SO_2 concentration, especially to the formulation of preventive and control measures of air quality by the rulers.

Several methods have been adopted to study the impacts of SO_2 pollution at the environment, such as: principal component analysis and cluster analysis (Cheng & Lam, 2000); artificial neural networks for predicting sulfur dioxide

concentrations (Saral & Ertürk, 2003) and other pollutants (Cook et al., 2006; Brunelli et al., 2007); multiple regression models to investigate the influence of emission sources and meteorological conditions on SO_2 pollution (Luvsan et al., 2012); among others. Especially for modeling and forecasting, statistical models based on multiple regression and time series tools, such as the Autoregressive Integrated Moving Average (ARIMA) model (Box & Jenkins, 1970), have been widely used for this class of problems (Prybutok et al., 2000; Hassanzadeh et al., 2009).

Many time series have a pattern called seasonality that repeats itself after a regular period of time. For pollutants, seasonal variation is often associated, especially with the changes in meteorological parameters. Furthermore, the reduction of emission levels on the weekends can also provoke the seasonality in the series. Thus, it is very important to consider statistical tools which take into account the seasonality effect.

Recently, several authors have studied time series with long-term dependency. In time domain analysis, the long-term dependence is generally characterized by a slow decay of the autocorrelation, even for observations separated by a large period of time. Granger & Joyeux (1980) and Hosking (1981) proposed the ARFIMA model do describe the long-memory feature of a time series. The ARFIMA process is an extension of the ARIMA model, where the parameter of integration (d) assumes fractional values. The methods proposed for estimating the parameters of the ARFIMA model are classified into parametric and semiparametric. Parametric methods consist of simultaneous estimation of model parameters, usually by maximum likelihood. In the semiparametric method, the estimation is performed in two steps: firstly, the parameter d is estimated; secondly, the autoregressive and moving average parameters are estimated. The most popular semiparametric estimator was proposed by Geweke & Porter-Hudak (1983). For a recent review of this subject, see Palma (2007). Modifications of this estimator have been developed by Reisen (1994), Lobato & Robinson (1996), Delgado & Robinson (1996), Velasco (2000), among others. More specifically, Sena Jr. et al.

(2006) and Molinares et al. (2009) investigated the estimator under various model specifications, such as the presence of non-Gaussian errors and outliers.

Due to the important model features of the ARFIMA process, it is already being used in environmental problems. As an example, Iglesias et al. (2006) adopted an ARFIMA model to analyse time series of pollutant concentrations that present long memory and missing values. These phenomena are widely found in time series of different areas of interest. For instance, Windsor & Toumi (2001) studied the variability of ozone and particulate matter concentration adopting the long-memory methodology.

An extension of ARFIMA to handle time series with seasonality is the SARF IMA model. Porter-Hudak (1990) applied the seasonal fractionally differenced model to study the monetary aggregates of the United States. Koopman et al. (2007) combined the SARFIMA and GARCH models to analyze the electricity price. Reisen et al. (2014a) have estimated the fractional and seasonal parameters of SARFIMA models by means of a semiparametric procedure, considering a non-constant conditional error variance. Reisen et al. (2014b) studied the properties of the SARFIMA model when the data presents one and two seasonal periods and short-memory components. Ye et al. (2015) proposed a new method to estimate the fractionally differencing parameter in the SARFIMA model using tapered periodogram.

Furthermore, environmental series usually contain observations influenced by external events that can cause changes in their dynamics, sometimes transient and sometimes permanently. These observations are known in the literature as outliers and, depending on its nature, its effects on inferential processes can be substantial. For example, in the presence of outliers there is an increase in the variance of the stochastic process, which implies a decrease in the values of the autocorrelations and loss of information about the structure of the autocorrelation of the process. The property of memory loss presented by the sample autocovariance function is discussed by Chan (1992, 1995). Although the study of long memory models in the presence of outliers has recently been a subject of much interest

to researchers, especially in the areas of economics and finance (Beran, 1994; Franses et al., 1999; Tolvi, 2003), very few works have been devoted to this topic in the environmental field. The proposal study in this paper, which concatenates long-memory model, with more than one fractional parameter, robust estimation and application to pollution data, has not been shown in other work.

Thus, the main purpose of this paper is to propose a robust semiparametric estimator for the non-seasonal and seasonal memory parameters in SARFIMA models based on a robust autocovariance estimators. The asymptotical results and the effect of additive outliers in the fractional estimates are discussed. In addition, the usefulness of the methods is applied to model and forecast the SO_2 concentration, measured at the Air Quality Automatic Monitoring Network (AQAMN) of the Greater Vitoria Region (GVR), ES, Brazil.

This paper is organized as follows. Section 2 introduces the model, discusses its properties and summarizes the parameter estimation methods. Section 3 deals with the analysis and modeling of the SO_2 contaminant and forecasting issues. Some conclusions are drawn in Section 4.

2. The model and parameter estimation

2.1. SARFIMA model and robust spectral estimation

A process $X_t \equiv \{X_t\}_{t\in\mathbb{Z}}$ is defined as a zero-mean SARFIMA$(p, d, q)\times(P, D, Q)_s$ model with non-seasonal orders p and q, seasonal orders P and Q, difference parameters d and D, and seasonal length $s \in \mathbb{N}^* = \mathbb{N} - \{0\}$ if

$$\eta_t = \nabla^{\mathbf{d}} X_t \tag{1}$$

is a SARMA $(p, q)\times(P, Q)_s$ process (see, Brockwell & Davis (2006)). That is, the process $\{\eta_t\}_{t\in\mathbb{Z}}$ satisfies

$$\eta_t = \frac{\Theta(B^s)\theta(B)\epsilon_t}{\Phi(B^s)\phi(B)}, \tag{2}$$

where $\{\epsilon_t\}_{t\in\mathbb{Z}}$ is a white noise with $\mathbb{E}(\epsilon_t) = 0$; $\mathrm{Var}(\epsilon_t) = \sigma^2_\epsilon$; and B is the backward operator satisfying $B^k Y_t = Y_{t-k}$ for any process $\{Y_t\}_{t\in\mathbb{Z}}$.

In (1), the operator $\nabla^{\mathbf{d}}$ is defined by:

$$\nabla^{\mathbf{d}} = (1 - B)^d (1 - B^s)^D, \tag{3}$$

where $\mathbf{d} = (d, D) \in \mathbb{R}^2$ is the memory parameter vector; d and D are the fractional parameters at the zero (or long-run) and seasonal frequencies, respectively. $D = 0$ implies that the process does not have seasonal poles. In addition, the fractional filters are

$$(1 - B^k)^x = \sum_{j=0}^{\infty} \binom{x}{j} \left(-B^k\right)^j, k = 1, s, \text{ and } x = d, D,$$

where

$$\binom{x}{j} = \frac{\Gamma(x+1)}{\Gamma(j+1)\Gamma(x-j+1)},$$

and $\Gamma(\cdot)$ is the well-known Gamma function.

In (2), the polynomials $\Phi(\cdot)$, $\Theta(\cdot)$, $\phi(\cdot)$ and $\theta(\cdot)$ are given, respectively, by

$$\begin{aligned}
\Phi(z^s) &= 1 - \Phi_1 z^s - \Phi_2 z^{2s} - \cdots - \Phi_P z^{Ps}, \\
\Theta(z^s) &= 1 - \Theta_1 z^s - \Theta_2 z^{2s} - \cdots - \Theta_Q z^{Qs}, \\
\phi(z) &= 1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p, \\
\theta(z) &= 1 - \theta_1 z - \theta_2 z^2 - \cdots - \theta_q z^q.
\end{aligned}$$

It is assumed that these polynomials have no common roots and satisfy the conditions $\Phi(z^s)\phi(z) \neq 0$ and $\Theta(z^s)\theta(z) \neq 0$, for $|z| = 1$. Furthermore, in the above equations, $(\Phi_i)_{1 \leq i \leq P}$, $(\Theta_j)_{1 \leq j \leq Q}$, $(\phi_k)_{1 \leq k \leq p}$ and $(\theta_\ell)_{1 \leq \ell \leq q}$ are unknown parameters. For more details, see, for example, Palma & Chan (2005), Giraitis & Leipus (1995), among others. If $|d + D| < 1/2$ and $|D| < 1/2$, X_t is a stationary and invertible process and, at seasonal frequencies $\omega_s \in [-\pi, \pi]$, the spectral density becomes unbounded and behaves as

$$f(\omega + \omega_s) \sim C|s\omega|^{-2D} \left|2 \sin \frac{\omega_s}{2}\right|^{-2d} \quad \omega \to 0, \tag{4}$$

where C is a non-negative constant (more details see, for example, Proposition 1 in Reisen et al. (2014b)).

If, in addition to long-memory and periodicity features, X_t also presents outliers, it is necessary to build a model that encompasses this type of characteristics. To achieve this, the estimation procedure suggested here concatenates the methods given in Reisen et al. (2014b), Lévy-Leduc et al. (2011c) and Molinares et al. (2009) to estimate the parameters of the model presented in Equation 1.

The robust estimation is summarized below. More details can be found in Lévy-Leduc et al. (2011c,b,a).

Let $\{X_1, \ldots, X_n\}$ be a sample from the process X_t (Eq. (1)). The periodogram function of the process X_t is given by

$$I_n(\omega_j) = \frac{1}{2\pi n} \left| \sum_{t=1}^{n} X_t e^{i\omega_j t} \right|^2, \tag{5}$$

where $\omega_j = \frac{2\pi_j}{n}$, $j = 1, \ldots, (\frac{n}{2} - 1)$, is the Fourier frequency.

Under the assumption of stationarity of the process X_t, $I_n(\omega_j)$ can be also written as follows:

$$I_n(\omega_j) = \frac{1}{2\pi} \sum_{k=-(n-1)}^{n-1} \hat{\gamma}_k e^{-i\omega_j k}, \tag{6}$$

where $\hat{\gamma}_k$ is the sample autocovariance function of $\{X_1, \ldots, X_n\}$.

An alternative robust spectral estimator proposed by Molinares et al. (2009) uses the robust autocovariance function for short-memory processes given in Ma & Genton (2000) to replace the classical sample autocovariance $\hat{\gamma}_k$ in Equation (6), obtaining a robust fractional estimator. The main asymptotic results of the robust autocorrelation function (ACF) for long-memory processes are discussed in Lévy-Leduc et al. (2011c). Here, as mentioned previously, the robust ACF estimator will be used to obtain fractional seasonal estimates by combining the method given by Reisen et al. (2014b) with the robust ACF discussed in Lévy-Leduc et al. (2011c). Therefore, the estimation approach proposed here is an extension to the methods discussed above.

Ma & Genton (2000) suggested the following robust sample autocovariance function

$$\widehat{\gamma}_Q(h) = \frac{1}{4}\left[Q^2_{n-h}(\mathbf{u}+\mathbf{v}) - Q^2_{n-h}(\mathbf{u}-\mathbf{v})\right], \tag{7}$$

where $\mathbf{u}$ and $\mathbf{v}$ are vectors containing the initial $n-h$ and the final $n-h$ observations, respectively, and the robust estimator for the autocorrelation function (ACF) is

$$\widehat{\rho}_Q(h) = \frac{Q^2_{n-h}(\mathbf{u}+\mathbf{v}) - Q^2_{n-h}(\mathbf{u}-\mathbf{v})}{Q^2_{n-h}(\mathbf{u}+\mathbf{v}) + Q^2_{n-h}(\mathbf{u}-\mathbf{v})}. \tag{8}$$

It can be shown that $|\widehat{\rho}_Q(h)| \leq 1$ for all h. The function $Q_n(\cdot)$ is a robust scale estimator which is based on the τth order statistic of $\binom{n}{2}$ distances $\{|\eta_j - \eta_k|, j < k\}$, and can be written as

$$Q_n(\eta) = c \times \{|\eta_j - \eta_k|; j < k\}_{(\tau)}, \tag{9}$$

where $\eta = (\eta_1, \eta_2, \ldots, \eta_n)'$, c is a constant used to guarantee consistency ($c = 2.2191$ for the normal distribution) and $\tau = \left\lfloor \frac{\binom{n}{2}+2}{4} \right\rfloor + 1$. The asymptotic properties of $\widehat{\gamma}_Q(h)$ for short and long-memory process are discussed in Lévy-Leduc et al. (2011c,b,a). Some of their results are addressed in the following remarks. Note that, since the SARFIMA model satisfies the model conditions given in these references, the asymptotical results for the robust autocovariance discussed below can be also extended for this process.

Remark 1. *Under Gaussian distribution of the process, for short and long memory properties, Lévy-Leduc et al. (2011c) showed asymptotical results for $Q_n(\eta)$ and $\widehat{\gamma}_Q(h)$, with a special attention to long-memory time series, that is, processes that have autocovariance funcion $\gamma(h)$ satisfying the assumption*

$$\gamma(h) = h^{2d-1}L(h),$$

where L is a slowly varying function at infinity and is positive for large h. The authors showed that for $0 < d < 1/4$, the robust autocovariance estimator $\widehat{\gamma}_Q(h)$

has the same asymptotic behavior as the classical autocovariance estimator $\hat{\gamma}_k$. In this case, there is no loss of efficiency. The standard Gaussian ARFIMA(p, d, q) is one particular case of the model discussed by the authors.

Remark 2. *For $1/4 < d < 1$, $\widehat{\gamma}_Q(h)$ has the rate of convergence equal to $n^{1-2d}/L(n)$, where L is a slowly varying function defined previously; the limiting distribution is non-Gaussian and belongs to the second Wiener Chaos.*

Based on the previous discussion and on Molinares et al. (2009), a robust spectral estimator may be computed as follows

$$I_{Q_n}(\omega) = \frac{1}{2\pi} \sum_{|h|<n} \kappa(h)\widehat{\gamma}_Q(h) \cos(h\omega), \tag{10}$$

where $\kappa(h)$ is defined as

$$\kappa(h) = \begin{cases} 1, & |h| \leq M, \\ 0, & |h| > M. \end{cases}$$

$\kappa(h)$ is a particular case of the *lag window* functions used in classical spectral theory to obtain a consistent spectral estimator, and M is the truncation point which is a function of n, say $M = G(n)$, where $G(n)$ must satisfy $G(n) \to \infty$, $n \to \infty$, with $\frac{G(n)}{n} \to 0$. $G(n)$ is usually chosen to be $G(n) = n^{\beta}$, where $0 < \beta < 1$ (see, e.g. (Priestley, 1981, pp. 433–437)). Note that, equivalently to the classical spectral estimation theories, other *lag window* functions can be used to obtain a robust spectral estimator, as discussed in Molinares et al. (2009).

The robust spectral estimator given in Equation (10) does not have the same finite-sample properties as the periodogram. For large h, the number of observations in the calculation of $\widehat{\gamma}_Q(h)$ is very small and, consequently, this function becomes very unstable. Then, to avoid these undesirable covariance estimates in the calculation of the estimator given in Equation (10), Molinares et al. (2009) justify the use of a truncation point M in the calculation of this sample function. Based on empirical studies, Molinares et al. (2009) suggest $M = n^{\alpha}$, where $0 < \alpha < 1$.

2.2. Robust semiparametric estimation method for $\mathbf{d}$

Reisen et al. (2014b) suggest the estimator $\hat{\mathbf{d}} = (\hat{d}, \hat{D})$ computed from the approximated multiple linear regression equation

$$\log I_n(\omega_{kj}) \cong a_0 - D \log \left[2 \sin\left(\frac{s\omega_{kj}}{2}\right)\right]^2 - d \log \left[2 \sin\left(\frac{\omega_{kj}}{2}\right)\right]^2 + u_{kj}, \qquad (11)$$

where a_0 is a constant and

$$u_{kj} = \log \frac{I_n(\omega_{kj})}{f_X(\omega_{kj})} - \mathbb{E}\left[\log \frac{I_n(\omega_{kj})}{f_X(\omega_{kj})}\right].$$

The frequencies ω_{kj}, $k = 0, 1, ..., [\frac{s}{2}]$, $1 \leq j \leq M$, are defined as

$$\omega_{kj} = \begin{cases} \frac{2\pi k}{s} + \frac{2\pi j}{n}, & k = 0; \\ \frac{2\pi k}{s} \pm \frac{2\pi j}{n}, & k = 1, ..., [\frac{s}{2}] - 1. \end{cases} \qquad (12)$$

In the above equations, M is the bandwidth that has to satisfy

$$\left(\frac{M}{n}\right)^{\alpha} \log M + \frac{1}{M} \to 0 \ , \ \text{ as } \ n \to \infty \ ,$$

for some $\alpha > 0$ (see, e.g., Reisen et al. (2014b)).

Under some model conditions, which also includes the Gaussian distribution for the innovations $\{\epsilon_t\}_{t\in\mathbb{Z}}$, the authors establish that

$$\sqrt{M}(\hat{\mathbf{d}} - \mathbf{d}) \to \mathcal{N}\left(W^{-1}b, \frac{\pi^2}{6} W^{-1}\right) \ , \qquad (13)$$

where b and W are a vector and a 2×2 matrix of constants, respectively.

The robust estimator of $\hat{\mathbf{d}}$ proposed here is obtained directly from Equation (11) by replacing $I_n(\omega_{kj})$ by $I_{Q_n}(\omega_{kj})$. This estimator is denoted here as $\hat{\mathbf{d}}_R = (\hat{d}_R, \hat{D}_R)$. The choice of the bandwidths will be based on the empirical evidence given in Molinares et al. (2009) under the restrictions of Equation (12). The estimator is implemented in R-project (R Development Core Team (2014)) and the

code can be obtained under request.

3. Analysis and results of modeling SO_2 concentration based on the robust estimator

As previously mentioned, the daily average SO_2 concentration is the data set analyzed to show the effect of high levels of the pollutant, considered here as outliers, on the parameter estimates and also on the empirical performance of the robust estimator. The series is expressed in μg/m^3 and was measured at the Air Quality Automatic Monitoring Network (AQAMN) of Cariacica, which belongs to the Metropolitan area of the Great Vitória Region (GVR) - ES - Brazil. GVR is comprised of seven cities with a population of about 1.7 million inhabitants in an area of 2,331 km^2. The region is situated along the South Atlantic coast of Brazil (latitude 20°19S, longitude 40°20W) and has a warm tropical climate, with average temperatures ranging from 24°C (Celsius) to 30°C.

The raw series has a sample size of 1826 observations, measured from January 1st 2005 to December 31st 2009. The series has mean $\bar{X} = 4.87\mu g/m^3$ and standard deviation equals to 2.26 $\mu g/m^3$. Maximum concentration is generally observed in the winter months from July to September. The series has many observations with high SO_2 values compared with the whole data. These high levels can be viewed here as outliers, since their values can provoke serious damage to the statistical functions such as the mean and the standard deviation and, therefore, may destroy the correlation structure of the series, leading to spurious data analyses.

Since the series presents large variability, the log transformation was used ($Y_t = log(X_t)$) and the new series was centered on the mean ($Z_t = Y_t - \bar{Y}$). Moreover, the series was divided into two parts: learning and prediction sets. The 1626 observations from January 1st 2005 to June 14th 2009 are considered as learning set and the remaining 200 observations are considered for the forecasting study.

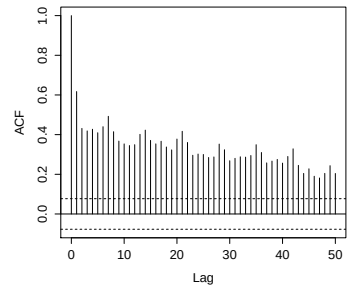
The discussion of the possible effect of high levels of the pollutant in the classical sample statistical time series functions is now addressed. The sample autocor-

relation (ACF), the partial autocorrelation (PACF) and the periodogram functions of Z_t are shown in Figure 1. These plots indicate possible seasonal behavior with a period equal to seven, which is an expected result since the data are daily mean levels. However, due to the hight levels of SO_2, any conclusion based from these plots may lead to an erroneous analysis.

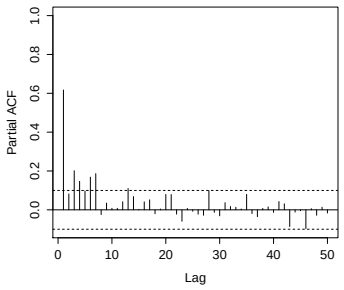
In this paper, the way to verify whether or not there is any damage in the statistical procedures caused by atypical observations is to compare, for example, the plots of the classical ACF, PACF and periodogram functions with the robust ones displayed in Figures 2(a), 2(b) and 2(c), respectively. The high levels of the concentration reduce the size of the classical ACF and PACF functions while increase the peaks of the periodogram, that is, the classical periodogram across the frequencies close to zero are much higher than the robust ones. These are expected results which are theoretically justified in Corollary 1 of Molinares et al. (2009). Note that the long memory property of the series is well observed by looking at the periodogram plots in Figures 1(c) and 2(c). Both plots indicate high values for the frequencies close to zero.

Additionally, in Figures 1(d) and 2(d) the log-periodogram is plotted against the log of the frequencies, for classical and robust cases, respectively. The figures also present the ordinary least square estimator of β_i in the model $log[I(\omega_j)] = \beta_0 + \beta_i log(\omega_j)$, where $i = 1$, if classical, $i = 2$, if robust, and $j = 1, \ldots, M$, with $M = n^{0.80}$. Comparing the intensity of the long memory dependency (the slopes of the regressions in Figures 1(d) and 2(d)), it can be observed that the dependency is larger for the robust estimator ($|\hat{\beta}_1| < |\hat{\beta}_2|$). As the SO_2 concentrations present seasonality, the conclusions about the estimator β_i should be take with care.

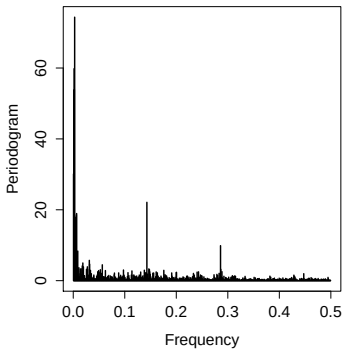
The evidences of seasonality, long memory behavior and outliers, corroborated by the above discussion, motivates the use of the SARFIMA model with robust estimation. The robust SARFIMA modeling strategy follows similar steps suggested in Hosking (1981) and investigated empirically by Reisen (1994) and Reisen & Lopes (1999). In semiparametric procedures, the estimation of the model parameters is performed in two steps: firstly, the parameter vector **d** is



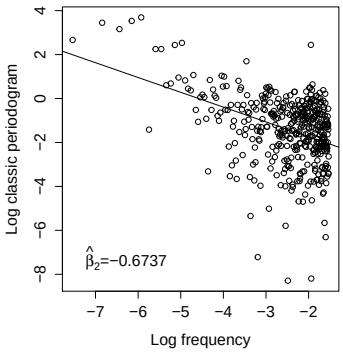
(a) ACF.



(b) PACF.



(c) Periodogram.



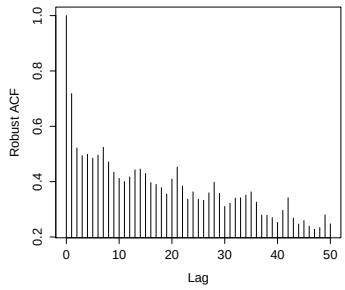
(d) Log(per.) versus log(freq.).

Figure 1: ACF, PACF, periodogram of Z_t and and log(classic periodogram) versus log(frequency).

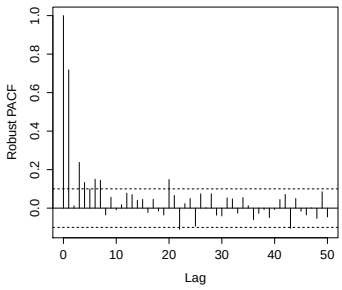
estimated based on the procedures presented in Section 2.2. Secondly, the truncated filter $(1 - B)^{\hat{d}}(1 - B^s)^{\hat{D}}$ is used to filter the observations. This new series is used to estimate the autoregressive and moving average parameters. The fitted models and their accuracy are discussed in the next subsections.

3.1. Adjusted models

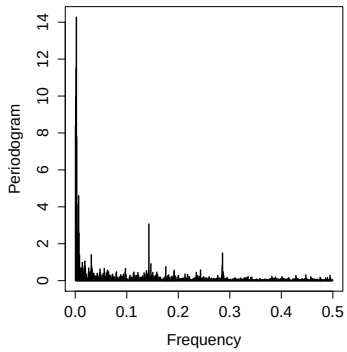
The robust and classic estimates of the parameter vector **d** are displayed in Table 1 for different bandwidths (M) (as discussed in Sections 2.1 and 2.2). The values in brackets correspond to the standard deviations. Note that the estimates do not satisfy all the stationary conditions, that is, $0.5 < |\hat{d} + \hat{D}| < 1.0$, but the



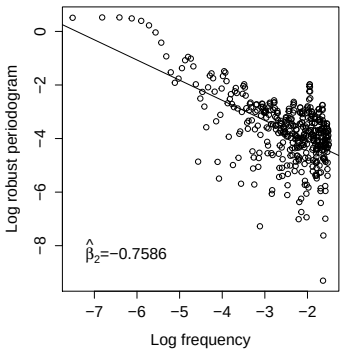
(a) Robust ACF.



(b) Robust PACF.



(c) Robust periodogram.



(d) Log(per.) versus log(freq.).

Figure 2: Robust ACF, PACF, periodogram of Z_t and log(robust periodogram) versus log(frequency).

model still has the mean reverting property in the sense that its cumulative impulse response weights sum to a finite number.

In the semiparametric approach, the choice of estimates depends on the size of the bandwidth. For example, large M gives more bias to the fractional estimators, when there are short-run components in the model. It can be noted that increasing M leads to fractional estimators with less power. As expected, all estimates using the classic periodogram were lower than those adopting the robust periodogram. This is due to presence of aberrant observations in the SO_2 concentrations, as discussed in Section 2.1. In this work, α was chosen equal to 0.80 and thus, using

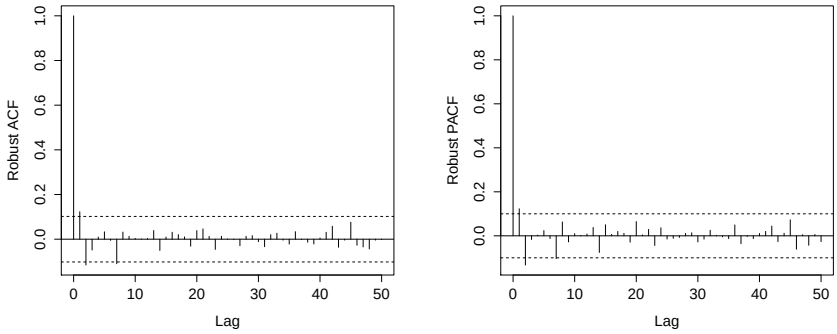
the robust periodogram, $\hat{d}_R = 0.4510$ and $\hat{D}_R = 0.2610$.

Table 1: Estimates of d_R and D_R for different bandwidths (M) for Z_t, using the robust and classical periodogram

Robust periodogram					
α	M	$\hat{d}_R$	$sd(\hat{d}_R)$	$\hat{D}_R$	$sd(\hat{D}_R)$
0.76	25	0.4706	(0.0465)	0.2934	(0.0304)
0.78	28	0.4675	(0.0423)	0.2723	(0.0285)
0.80	33	0.4510	(0.0398)	0.2610	(0.0280)
0.82	38	0.4565	(0.0359)	0.2391	(0.0262)
0.84	43	0.4534	(0.0337)	0.2202	(0.0254)
0.86	49	0.4511	(0.0307)	0.2100	(0.0240)
0.88	57	0.4539	(0.0281)	0.1724	(0.0229)
0.90	65	0.4596	(0.0262)	0.1594	(0.0221)
Classical periodogram					
α	M	$\hat{d}$	$sd(\hat{d})$	$\hat{D}$	$sd(\hat{D})$
0.76	25	0.4295	(0.0578)	0.2214	(0.0378)
0.78	28	0.4303	(0.0552)	0.2080	(0.0372)
0.80	33	0.4301	(0.0500)	0.1925	(0.0351)
0.82	38	0.4221	(0.0481)	0.1923	(0.0351)
0.84	43	0.4099	(0.0451)	0.1952	(0.0341)
0.86	49	0.3983	(0.0420)	0.1661	(0.0329)
0.88	57	0.3876	(0.0386)	0.1440	(0.0315)
0.90	65	0.3815	(0.0362)	0.1456	(0.0305)

Now, using the fractional differences, showed in Table 1 with $\alpha = 0.80$, the series $\hat{\eta}_t = (1-B)^{\hat{d}_R}(1-B^s)^{\hat{D}_R}Z_t, \quad t = 1, ..., 1625$, was obtained. Figure 3 shows the sample robust ACF and robust PACF functions of $\hat{\eta}_t$.

Based on the sample robust ACF and PACF functions displayed in Figure 3 and also on the Akaike information criterion (AIC), that was equal to 1,557.30, the MA(1) model was chosen to fit $\hat{\eta}_t$. Table 2 shows the estimated model. The MA(1) estimate was computed by the Hannan-Rissanen algorithm, in which the classical autocovariance was replaced by the robust one. Although the MA(1) component is adding a small contribution, the SARFIMA$(0, d, 1) \times (0, D, 0)_7$ was chosen for the SO_2 average data.



(a) Robust ACF. (b) Robust PACF.

Figure 3: Robust ACF and PACF of $\hat{\eta}_t$.

Table 2: Adjusted SARFIMA model for SO_2 concentration

Parameter	*Estimate*	*sd*
d	0.4511	0.0398
D	0.2610	0.0228
θ_1	0.0225	0.0145

Model adequacy is now addressed (Table 3). The Box-Pierce and Ljung-Box statistics (robust tests) demonstrated that the sample residuals are not time-correlated. In addition, the results also showed that the residuals are not normally distributed, which was an expected result since the original data is skewed to the right.

Table 3: Tests for normality* and non correlation (robust tests)**

Shapiro-Wilk*	Jarque-Bera*	Box-Pierce**	Ljung-Box**
<0.0001	<0.0001	0.8403	0.8404

Note: the p-values correspond to the robust Box-Pierce and robust Ljung-Box test statistics with lag 1. Other lags were tested and they presented similar conclusions.

As an alternative way to show the usefulness, the quality and the forecasting performance of the proposed model is to compare it to standard SARMA

models. Based on the standard Box-Jenkins strategies (Box et al., 2008) and also on the robust autocovariance, the SARMA$(1,0)\times(1,0)_7$ model (with AIC equal to 1,711.16) was chosen to fit the raw data among other candidates in the SARMA class. The residual analysis of SARMA model shows similar conclusion to the SARFIMA model, that is, the residuals were right-skewed and uncorrelated. The forecast performance of the SARFIMA$(0,d,1)\times(0,D,0)_7$ and SARMA$(1,0)\times(1,0)_7$ models is discussed in the next subsection.

3.2. Forecasting issues

In this subsection, the performance of the SARMA and SARFIMA models are compared considering forecasts one to ten-step ahead. As stated before, the observations from June 15th 2009 to December 31st 2009 were discarded from the modeling stage (200 observations) to be used in the out-of-sample forecast study.

To measure the accuracy of the forecasts, the criteria used was the Prediction Mean Square Error (PMSE) and the values are displayed in Table 4. From this table, It can be seen that the SARFIMA model presented more accurate forecasts than the SARMA model, especially for long-term forecasts. The forecast percentage rates (in the last column in the table) clearly show the superiority of the SARFIMA model. In addition, the Diebold-Mariano test (Diebold & Mariano, 1995) showed that the forecast of the two models are significantly different and confirmed the superiority forecasting performance of the SARFIMA model. For example, for $h = 1$ and 10, the statistic test gave p-values of 0.5×10^{-3} and 0.025, respectively, showing with high probability the rejection of the null hypothesis in favor of the SARFIMA forecasting performance.

Figure 4 presents the visual analysis of the one-step-ahead forecast values of the SARFIMA model, that is, from June 15th 2009 to December 31st 2009. It can be observed that it indicates a reasonably good performance of the model proposed here. The SARFIMA model was able to capture the dynamic of the SO_2 series for one-step-ahead forecasts. These results corroborate the power of the

fitted robust SARFIMA model to forecast the SO_2 concentration over the robust SARMA model, especially for long-term forecasting.

Table 4: PMSE of the two fitted models values to the SO_2 concentration

Horizon	SARMA (A)	SARFIMA (B)	[(A/B)-1]*100
1	0.0955	0.0857	11.48%
2	0.1220	0.1034	18.02%
3	0.1322	0.1079	22.56%
4	0.1393	0.1126	23.72%
5	0.1435	0.1140	25.85%
6	0.1463	0.1156	26.59%
7	0.1479	0.1161	27.33%
8	0.1635	0.1264	29.38%
9	0.1706	0.1240	37.54%
10	0.1736	0.1236	40.43%

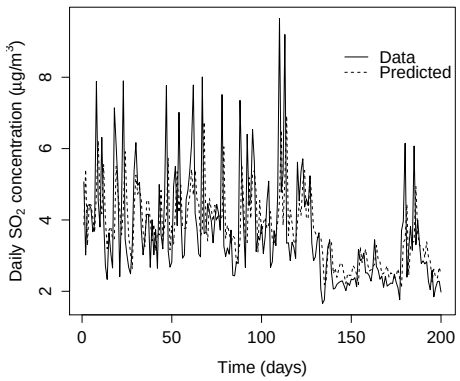


Figure 4: Forecasted values by the SARFIMA model and SO_2 concentration ($\mu g/m^3$) from June 15th of 2009 to December 31st of 2009, one-step-ahead.

4. Conclusions

In this article was adopted a SARFIMA model under effects of outliers to model daily average SO_2 concentrations. The statistical analyses showed that the

SO_2 series presents seasonality and long memory behavior. To estimate the robust fractional parameter d_R and D_R, this paper combined the methods suggests in Reisen et al. (2014b), Lévy-Leduc et al. (2011c) and Molinares et al. (2009). Based on the robust estimators of $\hat{d}_R$ and $\hat{D}_R$ it was adjusted a SARFIMA$(0, d, 1)\times(0, D, 0)_7$ model. The results suggested that residuals are uncorrelated and not normally distributed. Standard SARMA$(1, 0) \times (1, 0)_7$ model was also considered to measure the forecasting quality of the SARFIMA model. The PMSE indicated that the robust SARFIMA model presented a high level of accuracy, especially to forecast for long lead times. The robust fractional parameters is an attractive procedure for estimating the parameters of the SARFIMA model with long memory, seasonality and outliers and can be easily used in a real application in several areas of study. The results in this paper will hopefully stimulate further research on using robust estimation method and long-memory models to represent and forecast environmental time series.

5. Acknowledgements

The authors gratefully acknowledge partial financial support from FAPES/ES, FAPEMIG-MG and CNPq/Brazil.

References

Andersson, S. M., Martinsson, B. G., Friberg, J., Brenninkmeijer, C. A. M., Rauthe-Schöch, A., Hermann, M., van Velthoven, P. F. J., & Zahn, A. (2013). Composition and evolution of volcanic aerosol from eruptions of Kasatochi, Sarychev and Eyjafjallajükull in 2008-2010 based on CARIBIC observations. *Atmospheric Chemistry and Physics*, *13*, 1781–1796. doi:10.5194/acp-13-1781-2013.

Basu, R., & Samet, J. M. (2002). Relation between elevated ambient temperature and mortality: a review of the epidemiologic evidence. *Epidemiologic Reviews*, *24*, 190–202. doi:10.1093/epirev/mxf007.

Beran, J. (1994). On a class of M-estimators for gaussian long-memory models. *Biometrika*, *81*, 755–766.

Box, G. E. P., & Jenkins, G. (1970). *Time Series Analysis, Forecasting and Control*. San Francisco: Holden-Day, Incorporated.

Box, G. E. P., Jenkins, G., & Reinsel, G. (2008). *Time series analysis: forecasting and control*. (4th ed.). New Jersey: Prentice Hall.

Brockwell, P. J., & Davis, R. A. (2006). *Time series: theory and methods*. (2nd ed.). New York: Springer Series in Statistics.

Brunelli, U., Piazza, V., Pignato, L., Sorbello, F., & Vitabile, S. (2007). Two-days ahead prediction of daily maximum concentrations of SO_2, O_3, PM_{10}, NO_2, CO in the urban area of Palermo, Italy. *Atmospheric Environment*, *41*, 2967–2995. doi:http://dx.doi.org/10.1016/j.atmosenv.2006.12.013.

Chan, W. (1992). A note on time series model specification in the presence outliers. *Journal of Applied Statistics*, *19*, 117–124. doi:10.1080/02664769200000010.

Chan, W. (1995). Outliers and financial time series modelling: a cautionary note. *Mathematics and Computers in Simulation*, *39*, 425–430. doi:http://dx.doi.org/10.1016/0378-4754(94)00094-7.

Cheng, S., & Lam, K. (2000). Synoptic typing and its application to the assessment of climatic impact on concentrations of sulfur dioxide and nitrogen oxides in Hong Kong. *Atmospheric Environment*, *34*, 585–594. doi:ttp://dx.doi.org/10.1016/S1352-2310(99)00194-6.

Cook, D. F., Zobel, C. W., & Wolfe, M. L. (2006). Environmental statistical process control using an augmented neural network classification approach. *European Journal of Operational Research*, *174*, 1631–1642. doi:http://dx.doi.org/10.1016/j.ejor.2005.04.035.

Delgado, M. A., & Robinson, P. M. (1996). Optimal spectral bandwidth for long memory. *Statistica Sinica*, *6*, 97–112.

Diebold, F. X., & Mariano, R. S. (1995). Comparing Predictive Accuracy. *Journal of Business & Economic Statistics*, *13*, 253–63.

EPA (2014). *Sulfur dioxide: health*. EPA - United State Environmental Protection Agency. URL: `http://www.epa.gov/airquality/sulfurdioxide/health.html` accessed: 2014.05.15.

Franses, P. H., Ooms, M., & Bos, C. S. (1999). Long memory and level shifts: re-analyzing inflation rates. *Empirical Economics*, *24*, 427–449.

Georgoulias, A., Balis, D., Koukouli, M., Meleti, C., Bais, A., & Zerefos, C. (2009). A study of the total atmospheric sulfur dioxide load using ground-based measurements and the satellite derived sulfur dioxide index. *Atmospheric Environment*, *43*, 1693–1701. doi:`http://dx.doi.org/10.1016/j.atmosenv.2008.12.012"`.

Geweke, J. S., & Porter-Hudak (1983). The estimation and application of long memory times series model. *Journal of Time Series Analysis*, *4*, 221–238. doi:`10.1111/j.1467-9892.1983.tb00371.x`.

Giraitis, L., & Leipus, R. (1995). A generalized fractionally diferencing approach in long-memory modeling. *Lithuanian Mathematical Journal*, *35*, 53–65. doi:`10.1007/BF02337754`.

Granger, C. W. J., & Joyeux, R. (1980). An introduction to long-memory times series models and fractional differencing. *Journal of Time Series Analysis*, *1*, 15–29.

Hassanzadeh, S., Hosseinibalam, F., & Alizadeh, R. (2009). Statistical models and time series forecasting of sulfur dioxide: a case study Tehran. *Environmental Monitoring and Assessment*, *155*, 149–155. doi:`10.1007/s10661-008-0424-1`.

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

Iglesias, P., Jorqueira, H., & Palma, W. (2006). Data analysis using regression model with missing observations and long memory. *Computational Statistics and Data Analysis*, *50*, 2028–2043. doi:http://dx.doi.org/10.1016/j.csda.2005.03.007.

Sena Jr., M. R., Reisen, V. A., & Lopes, S. R. (2006). Correlated error in the parameters estimation of the arfima model: a simulated study. *Communications in Statistics - Simulation and Computation*, *35*, 789–802. doi:10.1080/03610910600716928.

Kanaroglou, P. S., Adams, M. D., Luca, P. F. D., Corr, D., & Sohel, N. (2013). Estimation of sulfur dioxide air pollution concentrations with a spatial autoregressive model. *Atmospheric Environment*, *79*, 421–427. doi:http://dx.doi.org/10.1016/j.atmosenv.2013.07.014.

Koopman, S. J., Ooms, M., & Carnero, M. A. (2007). Periodic seasonal reg-arfima-garch models for daily electricity spot prices. *Journal of the American Statistical Association*, *102*, 16–27. doi:10.1198/016214506000001022.

Lévy-Leduc, C., Boistard, H., Moulines, E., Taqqu, M. S., & Reisen, V. A. (2011a). Asymptotic properties of U-processes under long-range dependence. *The Annals of Statistics*, *39*, 1399–1426. doi:10.1214/10-AOS867.

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

Lévy-Leduc, C., Boistard, H., Moulines, E., Taqqu, M. S., & Reisen, V. A. (2011c). Robust estimation of the scale and the autocovariance function of Gaussian short and long-range dependent processes. *Journal of Time Series Analysis*, *32*, 135–156. doi:10.1080/02331888.2011.539442.

Lobato, I., & Robinson, P. M. (1996). Averaged periodogram estimation of long memory. *Journal of Econometrics*, *73*, 303–324.

Luvsan, M., Shie, R., Purevdorj, T., Badarch, L., Baldorj, B., & Chan, C. (2012). The influence of emission sources and meteorological conditions on SO_2 pollution in Mongolia. *Atmospheric Environment*, *61*, 542–549. doi:http://dx.doi.org/10.1016/j.atmosenv.2012.07.044.

Ma, Y., & Genton, M. G. (2000). Highly robust estimation of the autocovariance function. *Journal of Time Series Analysis*, *21*, 663–684. doi:10.1111/1467-9892.00203.

Mallik, C., Lal, S., Naja, M., Chand, D., Venkataramani, S., Joshi, H., & Pant, P. (2013). Enhanced SO_2 concentrations observed over northern India: role of long-range transport. *International Journal of Remote Sensing*, *34*, 2749–2762. doi:10.1080/01431161.2012.750773.

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*, 2511–2525. doi:http://dx.doi.org/10.1016/j.jspi.2008.12.014.

Palma, W. (2007). *Long-Memory time series: theory and methods*. Wiley.

Palma, W., & Chan, N. H. (2005). Efficient estimation of seasonal long-range-dependent processes. *Journal of Time Series Analysis*, *26*, 863–892. doi:10.1111/j.1467-9892.2005.00447.x.

Porter-Hudak, S. (1990). An application of the seasonal fractionally differenced model to the monetary aggregates. *Journal of the American Statistical Association*, *85*, 338–344.

Priestley, M. B. (1981). *Spectral Analysis and Time Series*. London: Academic Press.

Prybutok, V. R., Yi, J., & Mitchell, D. (2000). Comparison of neural network models with ARIMA and regression models for prediction of Houston's daily maximum ozone concentrations. *European Journal of Operational Research*, *122*, 31–40. doi:http://dx.doi.org/10.1016/S0377-2217(99)00069-7.

R Development Core Team (2014). *R: A Language and Enviroment for Statistical Computing*. Vienna, Austria. URL: http://cran.r-project.org/.

Reisen, V. A. (1994). Estimation of the fractional difference parameter in the ARIMA(p, d, q) model using the smoothed periodogram. *Journal of Time Series Analysis*, *15*, 335–350. doi:10.1111/j.1467-9892.1994.tb00198.x.

Reisen, V. A., & Lopes, S. (1999). Some simulations and applications of forecasting long-memory time-series models. *Journal of Statistical Planning and Inference*, *80*, 269–287. doi:http://dx.doi.org/10.1016/S0378-3758(98)00254-7.

Reisen, V. A., Sarnaglia, A. J. Q., Reis Jr, N. C., Lévy-Leduc, C., & Santos, J. M. (2014a). Modeling and forecasting daily average PM_{10} concentrations by a seasonal long-memory model with volatility. *Environmental Modelling & Software*, *51*, 286–295. doi:http://dx.doi.org/10.1016/j.envsoft.2013.09.027.

Reisen, V. A., Zamprogno, B., Palma, W., & Arteche, J. (2014b). A semiparametric approach to estimate two seasonal fractional parameters in the SARFIMA model. *Mathematics and Computers in Simulation*, *98*, 1–17. doi:http://dx.doi.org/10.1016/j.matcom.2013.11.001.

Saral, A., & Ertürk, F. (2003). Prediction of ground level SO_2 concentration using artificial neural networks. *Water, Air and Soil Pollution: Focus*, *3*, 307–316. doi:10.1023/A:1026081901947.

Tolvi, J. (2003). Long memory and outliers in stock market returns. *Applied Financial Economics*, *13*, 495–502. doi:10.1080/09603100210161983.

Velasco, C. (2000). Non-Gaussian log-periodogram regression. *Econometric Theory*, *16*, 44–79.

Windsor, H., & Toumi, R. (2001). Scaling and persistence of UK pollution. *Atmospheric Environment*, *35*, 4545–4556. doi:http://dx.doi.org/10.1016/S1352-2310(01)00208-4.

Ye, X., Gao, P., & Li, H. (2015). Improving estimation of the fractionally differencing parameter in the SARFIMA model using tapered periodogram. *Economic Modelling*, *46*, 167 – 179.

Robust Factor Modeling for High-Dimensional Time Series with short and long memory: an Application to Air Pollution Data

Adriano M. Sgrancio[1], Valderio A. Reisen[1,2], Flávio A. Ziegelmann[3]
[2]DEST-CCE, [1]PPGEA-CT -Universidade Federal do Espírito Santo,
[3] DEST- Universidade Federal do Rio Grande do Sul, BRAZIL

Abstract

This paper considers the factor modeling for high-dimensional time series with short and long-memory in the presence of additive outlier. The factor model studied by Lam & Yao (2012) is extended to consider the long-memory (VARFIMA model) and the presence of outliers. The estimators of the number of factors are obtained by an eigenanalysis of a non-negative definite matrix, i.e, the covariance matrix or the robust covariance matrix. The proposed methodology is analyzed in terms of the convergence rate of the number factors by means of asymptotic properties and Monte Carlo simulations. As an example of application, the robust factor analysis is utilized to identify pollution behavior for the pollutant PM_{10} in the Greater Region of Vitória aiming to reduce the dimensionality of the data.

Keywords: air pollution, factor analysis, autocorrelation, robustness, long-memory, eigenvalues.

1 Introduction

In the last 50 years, issues related to air pollution problems have increased considerably and these have been growing, especially in developing countries, where the air quality has been degraded as a result of industrialization, population growth, high rates of urbanization, and inadequate or non-existent policies to control air pollution, among other reasons. The problems caused by air pollution produce local, regional and global impacts. Among different environmental problems, air pollution is reported to cause the greatest damage to health and loss of quality of life. The most common human health problems caused by the air pollution are asthma, rhinitis, burning eyes , fatigue, dry cough, heart and lung diseases and heart failure. For example, among the other the large amount research in the area, the works by Brunekreef & Holgate (2002), Maynard (2004), WHO (2005), Curtis et al. (2006) showed the relationship between the legislated pollutants (inhalable particles with smaller diameter than 10 micrometers (PM_{10}), carbon monoxide (CO), sulfur dioxide (SO_2), nitrogen oxides (NO_x) and ozone (O_3)) and health problems. Issues relating to air quality have become increasingly important, since several health problems stem from air pollution, such as: asthma, rhinitis, eyes burning, fatigue, dry cough, heart and lung diseases, heart failure , and, etc. Authors as Brunekreef & Holgate (2002), Maynard (2004), WHO (2005), Curtis et al. (2006), among others, showed the relationship between the legislated pollutants (inhalable particles with smaller diameter than 10 micrometers (PM_{10}), carbon monoxide (CO), sulfur dioxide (SO_2), nitrogen oxides (NO_x) and ozone (O_3)) and health problems. In 2012, for example, the death of 4.3 million people have been attributed to air pollution (WHO 2014). In addition, air pollution contributes to the degradation of the environment, contributing to the greenhouse effect.

In the recent studies related to air pollution much attention has been paid to the mathematical models named receptor models, which attempt to measure and analyses concentrations at their sources from a given site without reconstructing the dispersion patterns of the pollutants, such as Particular Matter (PM), Sulfur Dioxide (SO_2) among others. These models have mathematical and statistical tools which are mainly used to provide the identification of the pollutant emission sources from chemical characteristics of the particles on the receiver and the pollutant emission sources (Seinfeld & Pandis 2006). In the literature, the majority of receptor models studied are: chemical balance of mass (CBM), multivariate analysis, principal component analysis techniques (PCA), factor analysis model (AF), multiple linear regression, cluster analysis, factoring positive matrix (FPM), among others (Watson et al. 2002). Regarding the classical factor analysis, this technique has been widely used in the area of air pollution, especially for the identification of emission sources, the management of monitoring networks, regression analysis, cluster analysis and prediction, among others.

Many time series arising in practice are best considered as a component of multivariate time series models, which accommodate the serial dependence of each component and, also, the interdependence between different components. However, it should be noted that, among the studies that adopted the classical principal component and techniques factorial analysis, especially in the area of air pollution , the time-dependence of the data is a common feature neglected, since the standard assumption of these multivariate statistical tools is the independency of the data (see, for example, Anderson (2003) and Johnson & Wichern (2007)). For example, the recent PhD thesis of Zamprogno (2013) shows ,empirically, spurious statistical analysis when applying PCA specifically in multivariate time series with more than a weak dependence property. To deal with this problem, Peña & Box (1987), Stock & Watson (2002), Lam et al. (2011) and Lam & Yao (2012) studied the factor modeling for multivariate time series from a dimension-reduction point of view. Differently from the PCA and factor analysis for independent observations, these papers look for factors which drive the serial dependence of the original time series. Further discussions and additional references in this direction can be found in the Introduction section of Lam & Yao (2012).

Apart from the purpose of dimension reduction, factor analysis has been widely used with the aim of forecasting in the sense this technique can drastically reduce the order of the estimated model which is one of the main justifications of the use of this method in real applications.

According to Stock & Watson (2002), in forecasting investigation, where the number of candidate predictor series (say, k) can be very large this can make the impractical viable such as, for example, in the use of vector autoregressive moving average (VARMA) models with a large number of parameters. This high-dimensional problem is simplified by modeling the covariability of the series in terms of a relatively small number of unobserved latent factors. Forecasting can then be carried out in a two-step process: first, a time series of the factors is estimated from the predictors; second, the relationship between the variable to be forecast and the factors is estimated by a linear regression, for example.

The environmental time series are often of a high dimension due to a large number of indices monitored across many different locations, these data may also present interesting phenomena to be considered from applied and theoretical statistical points of view. As investigated in Reisen et al. (2015), Reisen, Sarnaglia, Reis, Lévy-Leduc & Santos (2014), and Reisen, Zamprogno, Palma & Arteche (2014) the pollution data, in general, presents the long-memory behaviour at zero and seasonal frequencies, non-stationarity (local trend, unit root etc), periodicity, asymmetry, volatility, among other characteristics. In time domain analysis, the long-term dependence is generally characterized by a slow decay of the autocorrelation, even for observations separated by a large period of time. If this is not

considered in the model estimation, the series can cause a loss of information of the data correlation structure (Hosking (1981), Granger (1980, 1981), Granger & Joyeux (1980), Sowell (1992*a*,*b*),Reisen (1994) and Chung (2002)).

In addition to the above features, pollution data may have high level observations which can be defined as outliers from the statistical point of view. The outlier effect on the model estimation in univariate long memory processes has been the motivation of the papers Molinares et al. (2009), Lévy-Leduc et al. (2011*b*,*c*), Cotta & Reisen (2015) and references therein.

As is well known (see, for example, Chang et al. (1988), Tsay (1988), Chen & Liu (1993) and the references therein), the outliers can destroy the statistical properties of sample functions such as the standard mean and covariance. Since the estimation of time series models is connected with these sample functions, the final estimated model can be strongly affected by large peaks of the concentration. One way to deal with model estimation with outliers is to use the robust the robust ACF function based on the robust scale function $Q_n(.)$ proposed by Rousseeuw & Croux (1993*a*). The extension of this statistics for multivariate ARMA model is the recent paper by Cotta & Reisen (2015)).

This paper considers most of the above phenomena while using factor analysis for dimension reduction and forecasting issues with special attention to air pollution variables. In this context, the paper extends the dimension reduction test proposed in Lam & Yao (2012) for a fixed model order k and when the sample n going to infinity, but considering the presence of possible atypical observations (high levels of the pollutants) and the short and long-memory phenomena characterized by VARFIMA models. In this direction, this papers proposes, as a novelty, the use of the robust long-memory estimation method, given in Reisen et al. (2015), which make uses of $Q_n(.)$ proposed by Rousseeuw & Croux (1993*a*) and considered in Ma & Genton (2000) for ARMA models, to identify the number of the factors under VARFIMA processes with additive outliers. Some theoretical results are discussed and their performance, for finite sample sizes, is investigated trough Monte Carlo simulations. The usefulness of the proposed methodology is applied with aim of reducing the dimension of data observed at the Air Quality Automatic Monitoring Network (AQAMN) of the Greater Vitoria Region (GVR) Brazil, and to obtain forecasting observations.

Besides the introduction, this paper is divided as follows. In the Section 2, the model and the estimation methods are presented. The asymptotic properties of the estimation methods are investigated in Section 3. Section 4 presents some Monte Carlo experiments so as to support our theoretical claims. The data obtained from AQAMN stations are studied as an example of application in Section 5. Some concluding remarks are provided in Section 6.

2 Factor model in time series

2.1 The factor model with independent factors and long memory

Let $\boldsymbol{Z}_t$, $t \in \mathbb{Z}$, be a k-dimensional vector of an observed time series, and let $\boldsymbol{z}_t = \boldsymbol{Z}_t - \boldsymbol{\mu}_z$ be the vector of deviations from some origin $\boldsymbol{\mu}_z$ that will be mean if the series are stationary. Here, it is assumed without loss of generality that $\boldsymbol{\mu}_z = 0$. Also, let $\boldsymbol{x}_t$ be an unobserved r-dimensional vector of common factors. It is assumed that these series are generated by r ($r \leq k$) factors, $\boldsymbol{x}_t$, plus a measurement error $\boldsymbol{\epsilon}_t$ as

$$\boldsymbol{z}_t = \boldsymbol{P}\boldsymbol{x}_t + \boldsymbol{\epsilon}_t, \tag{1}$$

where $\boldsymbol{P}$ is a $k \times r$ matrix of parameters of rank r, and $\boldsymbol{\epsilon}_t$ is a k-dimensional white-noise sequence with full-rank covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ whose spectral density $f_a(\lambda)$ is bounded away from zero at the zero frequency $\lambda = 0$. Thus, all the common dynamic structure comes through the common factors, $\boldsymbol{x}_t$.

Suppose that the vector of common factors $\boldsymbol{x}_t = (x_{1,t}, ..., x_{r,t})^{'}$, $t \in \mathbb{Z}$, follows a zero-mean r-dimensional stationary vector Fractional Autoregressive Moving Average process (VARFIMA(p_x, d_x, q_x)) given by

$$\boldsymbol{\phi}_x(B)\boldsymbol{D}_x[(1-B)^d]\boldsymbol{x}_t = \boldsymbol{\theta}_x(B)\boldsymbol{a}_t, \tag{2}$$

where

$$\boldsymbol{\phi}_x(B) = \boldsymbol{I} - \boldsymbol{\phi}_1 B - \cdots - \boldsymbol{\phi}_p B^p$$

and

$$\boldsymbol{\theta}_x(B) = \boldsymbol{I} + \boldsymbol{\theta}_1 B + \cdots + \boldsymbol{\theta}_q B^q$$

are matrix polynomials in the backshift operator B, the $\boldsymbol{\phi}'s$ and the $\boldsymbol{\theta}'s$ are $r \times r$ matrices, the roots of the determinant polynomial $|\boldsymbol{\phi}_x(B)|$ are all outside the unit circle, and those of $|\boldsymbol{\theta}_x(B)|$ are all outside the unit circle. In addition, $\boldsymbol{a}_t$ is a sequence of $r \times r$ vector Gaussian White noise with zero mean and covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{a}}$. It is assumed that the r factors are independent and that all of the $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ matrices are diagonals. Later, this assumption will be replaced by a more general one; that is, the contemporaneous dependency in the noise matrix $\boldsymbol{\Sigma}_{\boldsymbol{a}}$ will be also considered, however, the only assumption is that $\boldsymbol{\Sigma}_{\boldsymbol{a}}$ is positive definite. The operator $\boldsymbol{D}_x[(1-B)^d]$ is a $r \times r$ diagonal matrix with $(1-B)^{d_1}, \ldots, (1-B)^{d_r}$ on the diagonal (see, for example, Chung (2002)). As in the univariate case; that is, in the ARFIMA (p, d_i, q) model, the parameter d_i governs the dependency of the process and if $-0.5 < d_i < 0.5$, for all $i = 1, \ldots, r$, then the process $\boldsymbol{x}_t$ is both stationary and invertible, which are the assumptions here considered for the model (2) (Hosking (1981), Granger (1980, 1981),Granger & Joyeux (1980), Sowell (1992a,b) and Reisen (1994)). As in the case of univariate ARFIMA process, the type of dependence of the VARFIMA process, follows the general definition of the memory of a univariate stationary time series process $y_t, t \in \mathbb{Z}$, with finite variance (see, for example, Taqqu (2003)). The following are common definitions of short, long and intermediate dependency.

- The process y_t has absolutely summable autocovariance; that is, the ACF decays at a geometrical rate in the sense that there is an upper bound $\mid \rho_y(h) \mid \leq ba^h$, such that,$0 < b < \infty$ and $0 < a < 1$ are constants. This implies that the process belongs to the short-range dependence class and the spectral density satisfies $0 < f_y(0) < \infty$. In the case of the ARFIMA process, when $d_i = 0$ the resulting process is short-memory. If this is valid for all i, then the VARFIMA model becomes an VARMA model (short-memory);
- The process does not have absolutely summable autocovariance; that is, the ACF decays at a hyperbolic rate such as $\rho_y(h) \simeq h^{-\alpha}$ (for some $0 < \alpha < 1$). In this situation, the process is defined to have long-memory property and $f_y(0) = \infty$. In the ARFIMA case, the long-memory is defined when $0 < d < 0.5$, whereas for the VARFIMA model there is at least one i such that $0 < d_i < 0.5$;

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

Remark 1. *The VARFIMA process x_t, $t \in \mathbb{Z}$, can be written as an infinite stationary second-order moving average representation as follows:*

$$\boldsymbol{x}_t = \sum_{j=0}^{\infty} \boldsymbol{\Psi}_j \boldsymbol{a}_{t-j}, \tag{3}$$

where the innovations $\boldsymbol{a}_t = [a_{1,t}, \ldots, a_{r,t}]'$ are r-dimensional martingale differences with respect to an increasing sequence of σ-fields $\{F_t\}$ such that for some $\lambda > 0$, $sup_t E(|a_{i,t}|^{2+\lambda} \| F_{t-1}) < \infty$, a.s., for all $i = 1, \ldots, r$. Let $E(\boldsymbol{a}_t \boldsymbol{a}_t' \| F_{t-1}) = \boldsymbol{\Sigma}_a$, a.s. The $r \times r$ matrix coefficients $\boldsymbol{\Psi}_j$ are often referred to as impulse responses. The main characterization of the process $\boldsymbol{x}_t$ considered in this paper is that impulses responses $\boldsymbol{\Psi}_j$ converge at slow hyperbolic rates as $j \longrightarrow \infty$. More precisely, there are r memory parameters $d_1, d_2, \ldots, d_r$, whose values lie in $(0, 0.5)$ such that the impulse responses $\boldsymbol{\Psi}_j$ can be approximated by

$$\boldsymbol{\Psi}_j \backsim \boldsymbol{D}\left[\frac{1}{\Gamma(d)} j^{d-1}\right] \boldsymbol{\Pi}, \quad as \;\; j \longrightarrow \infty \,, \tag{4}$$

where $\Gamma(\cdot)$ is a gamma function and $\boldsymbol{\Pi}$ is a nonsingular $r \times r$ matrix of constants that are independent of j and may be functions of a smaller set of unknown parameters. The notation $\boldsymbol{D}[j^{d-1}/\Gamma(j)]$ represents a $r \times r$ diagonal matrix with $j^{d_1-1}/\Gamma(d_1), \ldots, j^{d_r-1}/\Gamma(d_r)$ on the diagonal. In fact, for any univariate function f of a single variable, the notation $\boldsymbol{D}[f(d)]$ represents $r \times r$ diagonal matrix with $f(d_1), \ldots, f(d_r)$ on the diagonal. Also, the notation $\sim$ is defined as follows: given two sequences of matrices $\boldsymbol{U}_j$ and $\boldsymbol{V}_j$, as $j \longrightarrow \infty$, for i and r, where $u_{i,r,j}$ and $v_{i,r,j}$ are the (i,r)th elements of $\boldsymbol{U}_j$ and of $\boldsymbol{V}_j$, respectively. Let $\boldsymbol{\psi}_{i,j}'$ and $\boldsymbol{\pi}_i$ be the ith rows of $\boldsymbol{\Psi}_j$ and $\boldsymbol{\Pi}$, respectively; then Equation (4) implies that $\boldsymbol{\psi}_{i,j}' \sim j^{d_1-1}\Gamma(d_i)^{-1}\boldsymbol{\pi}_i'$, as $j \longrightarrow \infty$, for all $i = 1, \ldots, r$. Note that the conditions on the memory parameters $d_i \in (0, 0.5)$, for $i = 1, \ldots, r$, ensure that the impulse responses are square-summable and the infinite sum in Equation (3) exists.

Remark 2. *Given the hyperbolic convergence rates of the impulse responses (Equation (4)), the cumulative impulse responses also progress at hyperbolic rates as follows (see Lemma (1) in Chung (2002)).*

$$\sum_{j=0}^{j} \boldsymbol{\Psi}_j = \boldsymbol{D}\left[\frac{1}{\Gamma(d+1)} j^{d}\right] \boldsymbol{\Pi}, \quad as \;\; j \longrightarrow \infty \,, \tag{5}$$

Remark 3. *The autocovariances $\boldsymbol{\Gamma}(j) \equiv Cov(\boldsymbol{x}_t, \boldsymbol{x}_{t+j})$ of the process x_t must also converge at hyperbolic rates as follows (Chung (2002)).*

$$\Gamma(j) \sim \boldsymbol{D}(j^{d-0.5}).\boldsymbol{A}.\boldsymbol{D}(j^{d-0.5}), \quad as \;\; j \longrightarrow \infty \,, \tag{6}$$

where the (i,r)th element of the $r \times r$ matrix $\boldsymbol{A}$ is $Gamma(1 - d_i - d_r)/[\Gamma(d_r) \times \Gamma(1 - d_r)].\boldsymbol{\pi_i'\Sigma_a\pi_r}$.

Hence, as $j \longrightarrow \infty$, not only do the autocovariances $\gamma_{i,i}(j)$ of each $x_{i,t}$ die out slowly at a hyperbolic rate, i.e, $\gamma_{i,i}(j) \sim j^{2d_i-1}.\Gamma(1-2d_i)/[\Gamma(d_i)\Gamma(1-d_i)].\boldsymbol{\pi_i'\Sigma_a\pi_i}$, the covariances $\boldsymbol{\gamma}_{i,r}(j)$ between the current $x_{i,t}$ and the future $x_{r,t+j}$, for $i \neq r$, also vanish at hyperbolic rates, i.e., $\gamma_{i,r}(j) \sim j^{d_i+d_r-1}\Gamma(1-d_i-d_r)/[\Gamma(d_r) \times \Gamma(1-d_r)].\boldsymbol{\pi_i'\Sigma_a\pi_r}$. Hosking (1996) presents the result for the univariate case.

Returning to Model 1, the matrix $\boldsymbol{P}$ will be called the factor-loadings or factor-weights matrix, and its elements p_{ij} represent the weight of the factor j in the observed ith component. If $\boldsymbol{C}$ is any $r \times r$ nonsingular matrix, the generating equation could also be written as

$$\boldsymbol{z}_t = \boldsymbol{P}^* \boldsymbol{x}_t^* + \boldsymbol{\epsilon}_t,$$

where $\boldsymbol{P}^* = \boldsymbol{P}\boldsymbol{C}^{-1}$ is the new rectangular matrix of coefficients and $\boldsymbol{x}_t^* = \boldsymbol{C}\boldsymbol{x}_t$ is a linear transformation of the factors. Multiplying the Equation (2) by $\boldsymbol{C}$ gives

$$\boldsymbol{C}\boldsymbol{\phi}(B)\boldsymbol{D}_x[(1-B)^d]\boldsymbol{C}^{-1}\boldsymbol{C}\boldsymbol{x}_t = \boldsymbol{C}\boldsymbol{\theta}(B)\boldsymbol{C}^{-1}\boldsymbol{C}\boldsymbol{a}_t,$$

$$\boldsymbol{\phi}^*(B)\boldsymbol{x}_t^* = \boldsymbol{\theta}^*(B)\boldsymbol{a}_t^*$$

so that the model for the new set of factors is again an r-dimensional VARFIMA(p_x, d_x, q_x) model whose parameters are given by

$$\boldsymbol{\phi}^*(B) = \boldsymbol{C}\boldsymbol{\phi}(B)\boldsymbol{D}[(1-B)^d]\boldsymbol{C}^{-1},$$

$$\boldsymbol{\theta}^*(B) = \boldsymbol{C}\boldsymbol{\theta}(B)\boldsymbol{C}^{-1},$$

$$\boldsymbol{\Sigma}_a^* = \boldsymbol{C}\boldsymbol{\Sigma}_a\boldsymbol{C}'.$$

To remove this source of indeterminacy, it is assumed $\boldsymbol{P}'\boldsymbol{P} = \boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix.

Lemma 1. *Let $\boldsymbol{z}_t = \boldsymbol{P}\boldsymbol{x}_t + \boldsymbol{\epsilon}_t$ as defined in Equations (1) and (2) and assume that $\boldsymbol{\Gamma}_z(h) = E[\boldsymbol{z}_{t-h}\boldsymbol{z}_t']$ are the covariance matrices of the process $\boldsymbol{z}_t$ and $\boldsymbol{\Gamma}_x(h) = E[\boldsymbol{x}_{t-h}\boldsymbol{x}_t']$ are the covariance matrices for the generating vector $\boldsymbol{x}_t$. Then*

$$\boldsymbol{\Gamma}_z(0) = \boldsymbol{P}\boldsymbol{\Gamma}_x(0)\boldsymbol{\Gamma}' + \boldsymbol{\Sigma}_\epsilon, \tag{7}$$

$$\boldsymbol{\Gamma}_z(h) = \boldsymbol{P}\boldsymbol{\Gamma}_x(h)\boldsymbol{P}', \quad h \geq 1, \tag{8}$$

where $rank(\boldsymbol{\Gamma}_z(h)) = rank(\boldsymbol{\Gamma}_x(h))$, as $h \geq 1$.

Lemma 2. *If the factors are independent for all lags and the matrix $\boldsymbol{\Sigma}_a$ is diagonal, then*

a) All of the covariance matrices $\boldsymbol{\Gamma}_x(h)$ are diagonal;

b) The matrices $\boldsymbol{\Gamma}_z(h)$ are symmetric for $h \geq 1$;

c) By Spectral Decomposition (see Result 2A.14, Johnson & Wichern (2007, p. 100)), the columns of $\boldsymbol{P}$ will be eigenvectors of $\boldsymbol{\Gamma}_z(h)$ with eigenvalues $\gamma_i(h)$, where $\gamma_i(h)$ are the diagonal elements of $\boldsymbol{\Gamma}_x(h)$.

Theorem 1. *Suppose $\boldsymbol{z}_t = \boldsymbol{P}\boldsymbol{x}_t + \boldsymbol{\epsilon}_t$, where $\boldsymbol{x}_t$ is a r-dimensional VARFIMA(p_x, d_x, q_x) process, $\boldsymbol{P}$ is a $k \times r$ matrix ($k \geq r$) of rank r, and $\boldsymbol{\epsilon}_t$ is a k-dimensional white noise sequence with covariance $\boldsymbol{\Sigma}_\epsilon$. Then $\boldsymbol{z}_t$ follows a k-dimensional VARFIMA(p_z, d_z, q_z) with $p_z = p_x$, $d_z = d_x$ and $q_z = max(p_x, q_x)$.*

Remark 4. *In Equation (2), consider* $\boldsymbol{y}_t = \boldsymbol{D}_x[(1-B)^d]\boldsymbol{x}_t$. *Then, if the parameters of long memory* (d_i) *are all equal to zero, the model presented in Equation 2 becomes the short memory model described by Peña & Box (1987), i.e., a VARMA*(p,q). *Thus,* $\boldsymbol{\phi}_x(B)\boldsymbol{D}_x[(1-B)^d]\boldsymbol{x}_t = \boldsymbol{\theta}_x(B)\boldsymbol{a}_t$ *becomes* $\boldsymbol{\phi}_y(B)\boldsymbol{y}_t = \boldsymbol{\theta}_y(B)\boldsymbol{a}_t$.

An important conclusion from Theorem (1) is that the autoregressive matrices of the multivariate observed process $\boldsymbol{z}_t$ must satisfy

$$\boldsymbol{\phi}_z^\star(k)\boldsymbol{P} = \boldsymbol{P}\boldsymbol{\phi}_x^\star(k), \tag{9}$$

where $\boldsymbol{\phi}_z^\star(k) = \boldsymbol{\phi}_z(k)\boldsymbol{D}_z[(1-B)^d]$ and $\boldsymbol{\phi}_x^\star(k) = \boldsymbol{\phi}_x(k)\boldsymbol{D}_x[(1-B)^d]$, and which has the general solution

$$\boldsymbol{\phi}_z^\star(k) = \boldsymbol{P}\boldsymbol{\phi}_x^\star(k)\boldsymbol{P}^- + \boldsymbol{A}(\boldsymbol{I}-\boldsymbol{P}\boldsymbol{P}^-), \tag{10}$$

where $\boldsymbol{P}^-$ is a generalized inverse of $\boldsymbol{P}$ that satisfies $\boldsymbol{P}\boldsymbol{P}^-\boldsymbol{P} = \boldsymbol{P}$ and $\boldsymbol{A}$ is any arbitrary matrix (nonull matrix of rank $= k-r$), with the only restriction being that the roots of $|\boldsymbol{\phi}_z|$ are outside the unit circle. As the $\boldsymbol{\phi}_x^\star(k)$ is diagonal, Equation (9) shows that the columns of $\boldsymbol{P}$ are eigenvectors of $\boldsymbol{\phi}_z^\star(k)$, with eigenvalues as the diagonal elements of $\boldsymbol{\phi}_x^\star(k)$. The matrix $\boldsymbol{\phi}_z^\star(k)$ can have any rank, because of the presence of the arbitrary matrix $\boldsymbol{A}$.

2.2 The factor model with dependent factors

The main difference between the basic properties of the independent factors and dependent factors is related to the eigenvectors of the covariance and autoregressive parameter matrices that are not columns of the factor-loading matrix in the second case. Equations (8) and (9) are held in this case, but same conclusions are different. For instance, assume starting with Equation (8),

$$\boldsymbol{\Gamma}_z(h) = \boldsymbol{P}\boldsymbol{\Gamma}_x(h)\boldsymbol{P}', \quad h \geq 1,$$

Let $\boldsymbol{\Gamma}_x(h) = \boldsymbol{U}_h\boldsymbol{D}_h\boldsymbol{U}_h^{-1}$, where $\boldsymbol{D}_h$ is diagonal, it is clear that $\boldsymbol{\Gamma}_z(h)$ has the same eigenvalues as $\boldsymbol{\Gamma}_x(h)$ but with eigenvectors $\boldsymbol{P}\boldsymbol{U}_h$ and $\boldsymbol{V}$, where the columns of $\boldsymbol{V}$ belong to the null space of $\boldsymbol{P}\boldsymbol{P}'$. To see this, write

$$\boldsymbol{\Gamma}_z(h) = [\boldsymbol{P}\boldsymbol{U}_h : \boldsymbol{V}] \left| \begin{array}{cc} \boldsymbol{D}_h & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{array} \right| \left| \begin{array}{c} \boldsymbol{U}_h^{-1}\boldsymbol{P}' \\ \boldsymbol{V}' \end{array} \right|, \tag{11}$$

which shows the eigenstructure of $\boldsymbol{\Gamma}_z(h)$ clearly. Of course, if $\boldsymbol{\Gamma}_x(h)$ is diagonal, then $\boldsymbol{U}_h = \boldsymbol{I}$ and the independent factor model it is obtained.

As for the autoregressive parameter matrices, the condition is

$$\boldsymbol{\phi}_z^\star(h)\boldsymbol{P} = \boldsymbol{P}\boldsymbol{\phi}_x^\star(h),$$

assuming that $\boldsymbol{\phi}_x^\star(h)$ has linearly independent eigenvectors and writing $\boldsymbol{\phi}_x^\star(h) = \boldsymbol{W}_h\boldsymbol{F}_h\boldsymbol{W}_h^{-1}$, where $\boldsymbol{F}_h$ is diagonal and contains the eigenvalues of $\boldsymbol{\phi}_x^\star(h)$. Then

$$\boldsymbol{\phi}_z^\star(h)\boldsymbol{P}\boldsymbol{W}_h = \boldsymbol{P}\boldsymbol{W}_h\boldsymbol{F}_h, \tag{12}$$

shows that the eigenvalues of $\boldsymbol{\phi}_x^\star(h)$ are eigenvalues of $\boldsymbol{\phi}_z^\star(h)$ as well, with eigenvectors $\boldsymbol{PW}_h$.

Based on the discussion above, the following remarks summarize the differences between the basic properties of the independent-factors model and the dependent-factors model.

Remark 5. *The main differences between the properties of the independent and dependent factors are presented (**yes** means that the property is present; and, **no** means that the property is not present):*

a) Rank$(\boldsymbol{\Gamma}_z(h)) = r(h \geq 1)$*: independent factors (**yes**); dependent factors (**yes**);*

b) $\boldsymbol{\Gamma}_z(h)$ *is symmetric: independent factors (**yes**); dependent factors (**no**);*

c) Eigenvectors of $\boldsymbol{\Gamma}_z(h)$ *are columns of* $\boldsymbol{P}$*: independent factors (**yes**); dependent factors (**no**);*

d) Eigenvalues of $\boldsymbol{\Gamma}_z(h)$ *are eigenvalues of* $\boldsymbol{\Gamma}_x(h)$*: independent factors (**yes**); dependent factors (**yes**).*

e) Rank$(\boldsymbol{P}(l) = r)$*: independent factors (**yes**); dependent factors (**yes**);*

f) $\boldsymbol{z}_t \sim ARFIMA(p_z = p_x, d_z = d_x, q_z = max(p_x, q_x))$*: independent factors (**yes**); dependent factors (**yes**);*

g) Rank$(\boldsymbol{\psi})_i = r, (i \geq 1)$*, where* $\boldsymbol{\psi}(B) = \boldsymbol{\phi}^{-1}(B)\boldsymbol{D}_x[(1-B)^d]^{-1}\boldsymbol{\theta}(B)$*: independent factors (**yes**); dependent factors (**yes**);*

h) Eigenvalues of $\boldsymbol{\phi}_z^\star(k)$ *are eigenvalues of* $\boldsymbol{\phi}_x^\star(k)$*: independent factors (**yes**); dependent factors (**yes**);*

i) Eigenvectors of $\boldsymbol{\phi}_z^\star(k)$ *are eigenvectors of* $\boldsymbol{\phi}_x^\star(k)$*: independent factors (**yes**); dependent factors (**no**).*

The proofs of the above statements are straightforwardly obtained from Algebra Matrix Analysis and, these are, therefore, here omitted, but available upon request.

2.3 Models and estimation

2.3.1 Short and long-memory cases

Let Equation (1) be the short or long-memory model described previously for independent factors. Since none of the elements on the RHS of Equation (1) are observable, these have to be characterized further to make them identifiable. Following the same lines of Lam & Yao (2012), it is assumed that no linear combinations of $\boldsymbol{x}_t$ are white noise, as any such components can be absorbed into $\boldsymbol{\epsilon}_t$ [see condition (C1) below]. It is also assumed that the rank of $\boldsymbol{P}$ is r. Otherwise Equation (1) may be expressed equivalently in terms of a lower-dimensional factor. Furthermore, since Equation (1) is unchanged if we replace ($\boldsymbol{P}$, $\boldsymbol{x}_t$) by ($\boldsymbol{PH}$, $\boldsymbol{H}^{-1}\boldsymbol{x}_t$) for any invertible $r \times r$ matrix $\boldsymbol{H}$, it can assume that the columns of $\boldsymbol{P} = (\boldsymbol{p}_1, \ldots, \boldsymbol{p}_r)$ are orthonormal; that is, $\boldsymbol{P}'\boldsymbol{P} = \boldsymbol{I}_r$, where $\boldsymbol{I}_r$ denotes the $r \times r$ identity matrix. Note that even with this constraint, $\boldsymbol{P}$ and $\boldsymbol{x}_t$ are not uniquely determined in Equation (1), as the aforementioned replacement is still applicable for any orthogonal $\boldsymbol{H}$. However, the factor loading space; that is, the r-dimensional linear space spanned by the columns of $\boldsymbol{P}$, denoted by $\Omega(\boldsymbol{P})$, is uniquely defined.

The condition (C1) summarize all the assumptions introduced so far:

(C1) In model presented in Equation (1), $\boldsymbol{\epsilon}_t \sim WN(\mathbf{0}, \boldsymbol{\Sigma}_\epsilon)$. If $\boldsymbol{c}'\boldsymbol{x}_t$ is white noise for a constant $\boldsymbol{c} \in \mathbb{R}^k$, then $\boldsymbol{c}'Cov(\boldsymbol{x}_{t+h}, \boldsymbol{\epsilon}_t) = \mathbf{0}$ for any nonzero integers h. Furthermore $\boldsymbol{P}'\boldsymbol{P} = \boldsymbol{I}_r$.

This assumption relaxes the independence assumption between $\boldsymbol{x}_t$ and $\boldsymbol{\epsilon}_t$ imposed in most factor model literature.

The key to the inference for the model in Equation (1) is to determine the number of factors r and to estimate the $k \times r$ factor loading matrix $\boldsymbol{P}$, or more precisely the factor loading space $\Omega(\boldsymbol{P})$. Once an estimator is obtained, say, $\hat{\boldsymbol{P}}$, a natural estimator for the factor process is

$$\hat{\boldsymbol{x}}_t = \hat{\boldsymbol{P}}'\boldsymbol{z}_t, \tag{13}$$

and the resulting residuals are

$$\hat{\boldsymbol{\epsilon}}_t = (\boldsymbol{I}_d - \hat{\boldsymbol{P}}\hat{\boldsymbol{P}}')\boldsymbol{z}_t. \tag{14}$$

The dynamic modeling for $\boldsymbol{z}_t$ is achieved via such a modeling for $\hat{\boldsymbol{x}}_t$ and the relationship $\hat{\boldsymbol{z}}_t = \hat{\boldsymbol{P}}\hat{\boldsymbol{x}}_t$. A parsimonious fitting for $\hat{\boldsymbol{x}}_t$ may be obtained by rotating $\hat{\boldsymbol{x}}_t$ appropriately (Tiao & Tsay 1989). Such a rotation is equivalent to replacing $\hat{\boldsymbol{P}}$ by $\hat{\boldsymbol{P}}\boldsymbol{H}$ for an appropriate $r \times r$ orthogonal matrix $\boldsymbol{H}$. Note that $\Omega(\hat{\boldsymbol{P}}) = \Omega(\hat{\boldsymbol{P}}\boldsymbol{H})$, and the residuals in Equation (14) are unchanged with such a replacement.

2.3.2 Estimation for P and r

The estimation method proposed by Lam & Yao (2012) for estimating $\boldsymbol{P}$ for stationary cases is similar to the method described in Pan & Yao (2008). The goal is to estimate $\Omega(\boldsymbol{P})$, or, equivalently, its orthogonal complement $\Omega(\boldsymbol{B})$, where $\boldsymbol{B} = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_{k-r})$ is a $k \times (k-r)$ matrix for which $(\boldsymbol{P}, \boldsymbol{B})$ forms $k \times k$ orthogonal matrix; that is, $\boldsymbol{B}'\boldsymbol{P} = 0$ and $\boldsymbol{B}'\boldsymbol{B} = \boldsymbol{I}_{k-r}$ [see also (C1)]. It follows from Equation (1) that

$$\boldsymbol{B}'\boldsymbol{z}_t = \boldsymbol{B}'\boldsymbol{\epsilon}_t, \tag{15}$$

implying that for any $1 \le j \le k-r$, $\{\boldsymbol{b}_j'\boldsymbol{x}_t, t = 0, \pm 1, \ldots\}$ is a white-noise process. Hence, it is possible to search for mutually orthogonal directions $\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots$ one by one such that the projection of $\boldsymbol{z}_t$ on each of those directions is a white noise. The search is stopped when such a direction is no longer available, and take $k-g$ as the estimated value of r, where g is the number of directions obtained in the search. This is essentially how Pan & Yao (2008) accomplish the estimation. It is irrelevant in the above derivation if $\boldsymbol{y}_t$ is stationary or not.

However, a much simpler method is available when $\boldsymbol{x}_t$, and, therefore also $\boldsymbol{z}_t$, is stationary:

(C2) $\boldsymbol{x}_t$ is stationary model defined in Equation (2), and $Cov(\boldsymbol{x}_t, \boldsymbol{\epsilon}_{t+h}) = 0$ for any $h \ge 0$.

In most factor modeling literature, $\boldsymbol{x}_t$ and $\boldsymbol{\epsilon}_s$ are assumed to be uncorrelated for any t and s. Condition (C2) requires only that the future white-noise components are uncorrelated with the factors up to the present. This enlarges the model capacity substantially. Put

$$\boldsymbol{\Gamma}_z(h) = Cov(\boldsymbol{z}_{t+h}, \boldsymbol{z}_t), \qquad \boldsymbol{\Gamma}_x(h) = Cov(\boldsymbol{x}_{t+h}, \boldsymbol{x}_t),$$

$$\boldsymbol{\Gamma}_{x,\epsilon}(h) = Cov(\boldsymbol{x}_{t+h}, \boldsymbol{\epsilon}_t).$$

It follows from Equations (1) and (C2) that

$$\boldsymbol{\Gamma}_z(h) = \boldsymbol{P}\boldsymbol{\Gamma}_x(h)\boldsymbol{P}' + \boldsymbol{P}\boldsymbol{\Gamma}_{x,\epsilon}(h),\ h \geq 1. \tag{16}$$

For a prescribed integer $h_0 \geq 1$, define

$$\boldsymbol{M} = \sum_{h=1}^{h_0} \boldsymbol{\Gamma}_z(h)\boldsymbol{\Gamma}_z(h)', \tag{17}$$

Then $\boldsymbol{M}$ is a $k \times k$ non-negative matrix. It follows from Equation (16) that $\boldsymbol{M}\boldsymbol{B} = 0$; that is, the columns of $\boldsymbol{B}$ are the eigenvectors of $\boldsymbol{M}$ corresponding to zero-eigenvalues. Hence conditions (C1) and (C2) imply:

The factor loading space $\Omega(\boldsymbol{P})$ is spanned by the eigenvectors of $\boldsymbol{M}$ corresponding to its nonzero eigenvalues, and the number of the nonzero eigenvalues is r.

$\boldsymbol{M}$ accumulates the information from different time lags. This is useful especially when the sample size n is small. $\boldsymbol{M}$ is a non-negative definite matrix. $\boldsymbol{\Gamma}_z(h)\boldsymbol{\Gamma}_z(h)'$ is used [instead of $\boldsymbol{\Gamma}_z(h)$] to avoid cancellation of the information from different lags. This is guaranteed by the fact that for any matrix $\boldsymbol{C}$, $\boldsymbol{M}\boldsymbol{C} = 0$ if and only if $\boldsymbol{\Gamma}_z(h)'\boldsymbol{C} = 0$ for all $1 \leq h \leq h_0$. Since the strongest autocorrelations are often at small time lags, the tendency is to use small h_0. On the other hand, adding more terms will not alter the value of r, although the estimation for $\boldsymbol{\Gamma}_z(h)$ with large h is less accurate. For the short-memory model, the simulation results reported in Lam et al. (2011) also confirm that the estimation for $\boldsymbol{P}$ and r, defined below, is not sensitive to the choice of h_0.

Following the same lines in Lam & Yao (2012), the estimate $\Omega(P)$ is obtained by performing an eigenanalysis on

$$\hat{\boldsymbol{M}} = \sum_{h=1}^{h_0} \hat{\boldsymbol{\Gamma}}_z(h)\hat{\boldsymbol{\Gamma}}_z(h)'. \tag{18}$$

where $\hat{\boldsymbol{\Gamma}}_z(h)$ denotes the sample covariance matrix of $\boldsymbol{z}_t$ at lag h. Then the estimator $\hat{r}$ for the number of factors is defined in Equation (19) below. The columns of the estimated factor loading matrix $\hat{\boldsymbol{P}}$ are the $\hat{r}$ orthonormal eigenvectors of $\hat{\boldsymbol{M}}$ corresponding to its $\hat{r}$ largest eigenvalues. Note that the estimator $\hat{\boldsymbol{P}}$ is essentially the same as that defined in Section 2.4 of Lam et al. (2011), although a canonical form of the model is used in theirs study in order to define the factor loading matrix uniquely.

Due to the random fluctuation in a finite sample, the estimates for the zero-eigenvalues of $\boldsymbol{M}$ are unlikely to be exactly zero. A common practice is to plot all the estimated eigenvalues in descending order, and look for a cut-off value $\hat{r}$ such that the $(\hat{r}+1)$th largest eigenvalue is substantially smaller than the $\hat{r}$ largest eigenvalues. This is effectively an eyeball test. The ratio-based estimator defined below may be viewed as an enhanced eyeball test, based on the same idea as in Wang (2010). In fact this ratio-based estimator benefits from the faster convergence rates of the estimators for the zero-eigenvalues (for more details see Lam & Yao (2012)).

A ratio-based estimator for r. The estimator for the number of factors r is given by:

$$\hat{r} = \underset{1 \leq i \leq R}{\operatorname{argmin}} \hat{\lambda}_{i+1}/\hat{\lambda}_i, \tag{19}$$

where $\hat{\lambda}_1 \geq \ldots \geq \hat{\lambda}_k$ are the eigenvalues of $\hat{\boldsymbol{M}}$, and $r < R < k$ is a constant.

As suggested in Lam & Yao (2012)), in practice, $R = p/2$ can be the starting point. The search can not be extended up to p, as the minimum eigenvalue of $\hat{\boldsymbol{M}}$ is likely to be practically 0, especially when n is small and k is large. It is worth noting that when k and n are on the same order, the estimators for eigenvalues are no longer consistent. However, the ratio-based estimator in Equation (19) still works well.

Note that, although the above method is here considered for stationary processes, this can also be applied for non-stationary VARFIMA model with mean-reverting property; that is, $0.5 < d_i < 1.0$, for at least one $i = 1, ..., r$.

Remark 6. *Based on Remarks 5a and 5d, the ratio-based estimator presented in Equation (19) holds for independent and dependent factors.*

2.3.3 Robust M estimator ($\hat{M}_{Q_n}$)

Let $x_1, ..., x_n$ be a sample of X. The $Q_n(\cdot)$ 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 $Q_n(\cdot)$ function. Rousseeuw & Croux (1993*b*) showed that the asymptotic breakdown point of $Q_n(\cdot)$ 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

$$Cov(X, Y) = \frac{\alpha\beta}{4}\left[Var\left(\frac{X}{\alpha} + \frac{Y}{\beta}\right) - Var\left(\frac{X}{\alpha} - \frac{Y}{\beta}\right)\right], \tag{21}$$

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

Now, let X_t, $t \in \mathbb{Z}$, a time series with $Var(X_t) = \gamma_{X_t}(0)$. Let $\alpha = \beta = 1/\sqrt{\gamma_{X_t}(0)}$, based on Equation 21 and replacing the $Var(\cdot)$ by $Q_n^2(\cdot)$, 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 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 on the univariate case given by Equation 22 and the $\hat{\gamma}_{Q_n}(\cdot)$ 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

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

where $\hat{\gamma}_{Q_{n-h}}(X_{i,t}, X_{j,t})$ is estimated from

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

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

The robust correlation matrix function is suggested below

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

where $\hat{\rho}_{Q_{n-h}}(X_{i,t}, X_{j,t})$ is estimated from

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

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

Based on Equation 18 and on the robust ACF estimator, the robust M estimator is here suggested as

$$\hat{\boldsymbol{M}}_{Q_n} = \sum_{h=1}^{h_0} \hat{\boldsymbol{\Gamma}}_{z,Q_n}(h)\hat{\boldsymbol{\Gamma}}_{z,Q_n}(h)'. \quad (28)$$

where $\hat{\boldsymbol{\Gamma}}_{z,Q_n}(h)$ denotes the sample robust covariance matrix of $\boldsymbol{z}_t$ at lag h.

Therefore, the estimator $\hat{r}_{Q_n}$ for the number of factors is similarly obtained from Equation (19).

Note that, the asymptotic properties of the robust ACF function in univariate short and long-memory time series models were well established in Lévy-Leduc et al. (2011*a*,*b*,*c*) and some are summarized below. Cotta & Reisen (2015) extended the univariate robust ACF function to multivariate VARMA model. Their study was concentred on empirical investigations. Therefore, the asymptotic properties of the multivariate Robust ACF , as well as, the robust factor test based on the eigenvalues of $\hat{\boldsymbol{M}}_{Q_n}$ remain open problems and these are one of current research themes of V. A. Reisen.

Remark 7. *Under Gaussian distribution of the process, for short and long memory properties, Lévy-Leduc et al. (2011*c*) showed asymptotical results for* $\widehat{\gamma}_{Q_n}(h,\cdot)$*, with a special attention to long-memory time series; that is, processes that have autocovariance funcion* $\gamma(h)$ *satisfying the assumption*

$$\gamma(h) = h^{2d-1}L(h),$$

where L *is a slowly varying function at infinity and is positive for large* h*. The authors showed that for* $0 < d < 1/4$*, the robust autocovariance estimator* $\widehat{\gamma}_{Q_n}(_h,\cdot)$ *has the same asymptotic behavior as the classical autocovariance estimator* $\hat{\gamma}_k$*. In this case, there is no loss of efficiency. The standard Gaussian ARFIMA*(p,d,q) *is one particular case of the model discussed by the authors.*

Remark 8. *For* $1/4 < d < 1$*,* $\widehat{\gamma}_{Q_n}(h,\cdot)$ *has the rate of convergence equal to* $n^{1-2d}/L(n)$*, where* L *is a slowly varying function defined previously; the limiting distribution is non-Gaussian and belongs to the second Wiener Chaos.*

3 Asymptotic property of the eigenvalues of $\hat{\boldsymbol{M}}$

In the following section, the asymptotic properties of the eigenvalues of the matrix $\hat{\boldsymbol{M}}$ in Equation (18) are summarized according to the results given Lam & Yao (2012) for short-memory process associated with the classical autocovariance matrix

$$\hat{\boldsymbol{S}}_z(h) = n^{-\alpha}\sum_{t=1}^{n-1}(\boldsymbol{z}_{t+h} - \bar{\boldsymbol{z}})(\boldsymbol{z}_t - \bar{\boldsymbol{z}})', \tag{29}$$

which exists for $1 \leq h \leq h_0$, where $\alpha > 1$ is a constant. The asymptotical theory of the eigenvalues of the matrix $\hat{\boldsymbol{M}}$ when it is assumed a VARFIMA model and the use of the robust autocovariance estimator here proposed will be left as an open problem for a future research topic.

Here, the asymptotic properties are considered under assumption that $n \longrightarrow \infty$ and k is fixed. Let, $\lambda_1, \ldots, \lambda_k$ be the eigenvalues of the matrix $\hat{\boldsymbol{M}}$:

(C3) $\boldsymbol{z}_t$ is strictly stationary and ψ-mixing with the mixing coefficients $\psi(\cdot)$ satisfying the condition that $\sum_{t\geq 1}\psi(t)^{1/2} < \infty$. Furthermore, $E\{|\boldsymbol{y}_t|^4\} < \infty$ element-wise.

(C4) $\lambda_1, \ldots, \lambda_r > 0 = \lambda_{r+1} = \ldots = \lambda_k$.

Denoted by $(\hat{\lambda}_1, \hat{\boldsymbol{\gamma}}_1), \ldots, (\hat{\lambda}_k, \hat{\boldsymbol{\gamma}}_k)$, the k pairs of eigenvalue and eigenvectors of matrix $\hat{\boldsymbol{M}}$: the eigenvalues $\hat{\lambda}_i$ are arranged in descending order, and the eigenvactors $\hat{\boldsymbol{\gamma}}_i$ are orthonormal. Moreover, it may go without explicit statement that $\hat{\boldsymbol{\gamma}}_i$ may be replaced by $-\hat{\boldsymbol{\gamma}}_i$ in order to match the direction of $\boldsymbol{\gamma}_i$ for $1 \leq i \leq r$.

According to Lam & Yao (2012) the following proposition can be described:

Proposition 1. *Let conditions (C1)-(C4) hold. Then as* $n \longrightarrow \infty$ *(but* r *is fixed), it is holds that:*

(i) $|\hat{\lambda}_i - \lambda_i| = O_P(n^{-1/2})$ *and* $\|\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i\| = O_P(n^{-1/2})$*, for* $i = 1, \ldots, r$*, and*

(ii) $\hat{\lambda}_i = O_P(n^{-1})$*, for* $i = r+1, \ldots, k$*.*

Remark 9. *Based on Remarks* 5*a,* 5*d and* 6*, the Proposition 1 holds for independent and dependent factors.*

As previously stated, the asymptotical results of the eigenvalues of $\hat{\boldsymbol{M}}_{Q_n}$ are being left as open problems for future work.

4 Monte Carlo studies

This section reports the results of several Monte Carlo experiments to analyze the effect of high-dimensional time series with additive outliers on the factor modeling and time series with long and short-memory dependency. In this context, the empirical study is divided into two cases of $\boldsymbol{x}_t$ (Equation 2), which follows a VARFIMA model with $r = 3$: (1) short-memory process where $\mathbf{d} = (0, 0, 0)'$; that is, $\boldsymbol{x}_t$ is a VARMA process; (2) long-memory process where $\mathbf{d} = (d_1, d_2, d_3)'$, for at least one $d_1, d_2, d_3 \in (0, 0.5)$. The VARFIMA model was generated with independent a_t from $\mathrm{N}(\mathbf{0}, \boldsymbol{I})$ and Φ coefficients, which are displayed in Table 1. The sample size is $n = 50, 100, 200, 400, 800$ and 1600, and $k = 0.2n, 0.5n, 0.8n$ and $1.2n$. The factor model (Equation 1) was generated as follows: first, all $k \times r$ elements of matrix $\mathbf{P}$ were generated as independent observations from the uniform distribution on the interval [-1,1], and, all simulated $\mathbf{P}$ matrix were divided by $k^{\delta/2}$ to make all three factors of the strength δ; that is, δ measures the strength of the factor $x_{j,t}, j = 1, ..., r$ (see Equation 3.1 in Lam & Yao (2012)). The process ϵ_t in 1 consists of independent $\mathrm{N}(0, 1)$ components and they are also independent across t.

The matrices Γ_z and $\Gamma_{\boldsymbol{z},Q_n}$ denote the standard ACOVF and the robust ACOVF estimators of $\boldsymbol{z}_t$, respectively.

Table 1: Φ matrices for VARFI(1,$\boldsymbol{d}$) process.

Φ_1 (**Model 1**)			Φ_1 (**Model 2**)			Φ_1 (**Model 3**)		
0.6	0.0	0.0	0.6	0.35	0.1	0.2	0.0	0.6
0.0	-0.5	0.0	0.05	-0.5	0.65	0.0	0.3	0.0
0.0	0.0	0.3	0.8	0.0	0.3	0.2	0.0	0.5

Note that, in Table 1 each model has its particularities. Model 1 corresponds to a process with no temporal correlation outside the diagonal; that is, each $x_{i,t}$, $i = 1, 2, 3$, has serial dependence only, while in Models 2 and 3, $x_{i,t}$ has not only serial dependence, but also the interdependence between different component series $x_{i,t}$ and $x_{j,t}$.

The main interest in this empirical study is to verify the performance of the statistic $\hat{r}$, given by Equation 19, in the context of VARFIMA models with and without outliers. For this, the relative frequencies of $\hat{r} = r$, denoted here as $f_{rel.}(\hat{r} = r)$, were computed, where $\hat{r}$ is the estimator of r. The statistical quantities were computed based on 200 replications.

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

$$\mathbf{z}_t = \boldsymbol{P}\boldsymbol{x}_t + \boldsymbol{\omega} \circ \boldsymbol{\delta}_t, \tag{30}$$

where "$\circ$" is the Hadamard product (Johnson 1989). $\boldsymbol{\omega} = [\omega_1, ..., \omega_k]'$ is a magnitude vector of additive outliers. $\boldsymbol{\delta}_t = [\delta_{1t}, ..., \delta_{kt}]'$ is a random vector indicating the occurrence of an outlier at time t, in variable k, such as $\mathbb{P}(\delta_{k,t} = -1) = \mathbb{P}(\delta_{k,t} = 1) = p/2$ and $\mathbb{P}(\delta_{k,t} = 0) = 1 - p$, where $\mathbb{E}[\delta_{k,t}] = 0$ and $\mathbb{E}[\delta_{k,t}^2] = \mathbb{V}ar(\delta_{k,t}) = p$. The model described above assumes that $\{\mathbf{Z}_t\}$ and $\{\boldsymbol{\delta}_t\}$ are independent processes. Also, it is assumed that the elements of $\boldsymbol{\delta}_t$ are not correlated and temporally uncorrelated, i.e, $\mathbb{E}(\boldsymbol{\delta}_t\boldsymbol{\delta}_t^{'}) = \Sigma_{\boldsymbol{\delta}} = diag(p, ..., p)$ and $\mathbb{E}(\boldsymbol{\delta}_t\boldsymbol{\delta}_{t+h}^{'}) = 0$ for $h \neq 0$.

Remark 10. *$\boldsymbol{\delta}_{kt}$ is the product of Bernoulli(p) random variable with Rademacher random variable, the latter equals 1 or -1, both with probability 1/2.*

Here, in the empirical investigation, the probability of an outlier occurring at the time t is $p = 0.05$ and, without loss of generality, it is also assumed that $\boldsymbol{\omega} = [\omega, ..., 0]'$; that is, $z_{1,t}$, $t = 1, ..., n$, is the only one process in $\mathbf{z}_t = (z_{1,t}, z_{2,t}, z_{3,t})^{'}$ contaminated with outliers and $\omega = 15$.

4.1 Case 1: short memory (d = 0) with and without outliers

Table 2 reports the relative frequency estimates $f_{rel.}(\hat{r} = 3)$ with $\delta = 0, 0.5$ for Model 1. As expected, the estimation performs better when the factors are stronger(i.e., $\delta = 0$). Observe that the test improves when n is very large. Similar performance is observed when the dimension k increases. When the factors are weak (i.e., $\delta = 0.5$), the test is also improved for large sample size (fixed k) but it presents poorer performance than when $\delta = 0$; that is, the convergence of $\hat{r}$ is slower.

Table 3 displays the empirical investigation when the VAR(1) is Model 2; that is, now the process is generated with a no-diagonal Φ matrix. As expected the convergence of the estimated relative frequencies to 1 is much slower compared to the results related to Model 1. In this case, the estimation performs better again when the factors are stronger (i.e., $\delta = 0$). The estimation for r is very accurate for $n \geq 800$. An interesting phenomenon is observed when the strength of the factor is weak ($\delta = 0.5$). In this context, the test presents a radical reduction of the relative frequencies $f_{rel.}(\hat{r} = 3)$. Therefore, the test has difficulty addressing the number of factors correctly when there is inter-correlation among variables which provokes another phenomenon, coined as "non-blessing of dimensionality". This can be explained by considering the Remark 1 in Lam & Yao (2012), where $\| \Sigma_x(h)|| \asymp k^{1-\delta}$ and $\| A \|^2 \asymp k^{1-\delta}$. From this, if $\delta > 0$ the test leads to under reduction phenomenon.

This phenomenon is more evident in the case of Model 3 (the simulation is not presented here but is available upon request). These results indicate that the presence of a more complex structure of correlation leads to an incorrect estimation of the dimensional reduction and is also confirms the discussion in Section 3.

Now, the investigation is directed to the case where the process $\boldsymbol{z}_t$ contains additive outliers. Table 4 shows the relative frequency estimates for the dimensional reduction ($\widehat{r}$ and $\widehat{r}_{Q_n} = 1, 2$ or 3) when $r = 3$ for Model 1 considering the presence of outliers. The standard case ($\widehat{\Gamma}_{\boldsymbol{z}}$ and $p = 0$) is in accord with the results given by Table 2. The third column gives the simulation results using $\widehat{\Gamma}_{\boldsymbol{z},Q_n}$ when $p = 0$. As one can see, the $\widehat{r}$ estimates using $\widehat{\Gamma}_{\boldsymbol{z},Q_n}$ present similar results of $\widehat{\Gamma}_{\boldsymbol{z}}$ when $p = 0$, which is in accordance with the results presented in

Table 2: Relative frequency estimates for $f_{rel.}(\hat{r}=3)$ in the simulation with 200 replications - Model 1.

	n	**50**	**100**	**200**	**400**	**800**	**1600**
$\delta=0$	$k=0.2n$	0.170	0.585	0.870	0.995	1	1
	$k=0.5n$	0.395	0.710	0.975	1	1	1
	$k=0.8n$	0.435	0.740	0.960	1	1	1
	$k=1.2n$	0.470	0.785	0.960	1	1	1
$\delta=0.5$	$k=0.2n$	0.090	0.130	0.265	0.470	0.840	1
	$k=0.5n$	0.070	0.155	0.315	0.585	0.905	1
	$k=0.8n$	0.070	0.165	0.385	0.635	0.945	1
	$k=1.2n$	0.070	0.215	0.365	0.700	0.915	1

Table 3: Relative frequency estimates for $f_{rel.}(\hat{r}=3)$ in the simulation with 200 replications - Model 2.

	n	**50**	**100**	**200**	**400**	**800**	**1600**
$\delta=0$	$k=0.2n$	0.080	0.095	0.145	0.360	0.815	1
	$k=0.5n$	0.180	0.155	0.205	0.450	0.875	1
	$k=0.8n$	0.155	0.165	0.250	0.455	0.830	1
	$k=1.2n$	0.180	0.160	0.285	0.465	0.915	1
$\delta=0.5$	$k=0.2n$	0.050	0.000	0.000	0.050	0.000	0.000
	$k=0.5n$	0.025	0.000	0.000	0.000	0.000	0.000
	$k=0.8n$	0.000	0.010	0.000	0.000	0.000	0.000
	$k=1.2n$	0.030	0.005	0.000	0.000	0.000	0.000

Cotta & Reisen (2015). This fact indicates that the robust methodology may be used when one is uncertain of the presence of outliers in the series. The impact of additive outliers in the number of estimated factors can be verified from the second column where the presence of atypical observations in the data leads to a reduction of the estimated frequencies when $\widehat{r}=3$ for all values of k. This does not occur when the robust estimator is utilized, the results are quite close to the ones from the first column. The percentage of outliers in only one vector seems to be, in general, not strong enough to destroy the robustness of the proposed method.

Table 4: Relative frequency estimates for dimensional reduction - Model 1's $\mathbf{\Phi}$ coefficients.

		$\widehat{\Gamma}_{\boldsymbol{z}}$			$\widehat{\Gamma}_{\boldsymbol{z}}$			$\widehat{\Gamma}_{\boldsymbol{z},Q_n}$			$\widehat{\Gamma}_{\boldsymbol{z},Q_n}$		
		$p=0$			$p=0.05$ and $\omega=15$			$p=0$			$p=0.05$ and $\omega=15$		
		$n=100$			$n=100$			$n=100$			$n=100$		
		$\widehat{r}=1$	$\widehat{r}=2$	$\widehat{r}=3$	$\widehat{r}=1$	$\widehat{r}=2$	$\widehat{r}=3$	$\widehat{r}_{Q_n}=1$	$\widehat{r}_{Q_n}=2$	$\widehat{r}_{Q_n}=3$	$\widehat{r}_{Q_n}=1$	$\widehat{r}_{Q_n}=2$	$\widehat{r}_{Q_n}=3$
$\delta=0$	$k=0.2n$	0.110	0.330	0.585	0.250	0.230	0.290	0.140	0.410	0.450	0.180	0.380	0.440
	$k=0.5n$	0.100	0.280	0.710	0.240	0.240	0.260	0.130	0.320	0.550	0.160	0.310	0.530
	$k=0.8n$	0.040	0.200	0.785	0.130	0.120	0.210	0.040	0.270	0.690	0.060	0.290	0.650

In Table 5, the relative frequency estimates of $\widehat{r}$ for Model 3 are presented. Note that for Model 3 all correlations are positive. The first column shows that the complex structure of correlation leads to a decrease in the number of times that the dimension reduction was correctly estimated; that is, the values when $\widehat{r}=1$ is much bigger then $\widehat{r}=3$. This result is consistent with the ones from Table 3 and, in Zamprogno (2013) and Cotta & Reisen (2015) and, discussed in Section 3. The occurrence of outliers seems also to increase the frequency of $\widehat{r}=1$. When the robust estimator $\widehat{\Gamma}_{\boldsymbol{z},Q_n}$ was applied, the results of the 4th column are closer to the values of the first column showing the possibility of applying the

proposed robust methodology in this context.

Table 5: Relative frequency estimates for dimensional reduction - Model 3's $\mathbf{\Phi}$ coefficients.

		$\widehat{\Gamma}_{\boldsymbol{z}}$			$\widehat{\Gamma}_{\boldsymbol{z}}$			$\widehat{\Gamma}_{\boldsymbol{z},Q_n}$			$\widehat{\Gamma}_{\boldsymbol{z},Q_n}$		
		$p=0$			$p=0.05$ and $\omega=15$			$p=0$			$p=0.05$ and $\omega=15$		
		$n=100$			$n=100$			$n=100$			$n=100$		
		$\widehat{r}=1$	$\widehat{r}=2$	$\widehat{r}=3$	$\widehat{r}=1$	$\widehat{r}=2$	$\widehat{r}=3$	$\widehat{r}_{Q_n}=1$	$\widehat{r}_{Q_n}=2$	$\widehat{r}_{Q_n}=3$	$\widehat{r}_{Q_n}=1$	$\widehat{r}_{Q_n}=2$	$\widehat{r}_{Q_n}=3$
$\delta=0$	$k=0.2n$	0.950	0.050	0.000	0.950	0.050	0.000	0.600	0.225	0.155	0.850	0.075	0.075
	$k=0.5n$	0.825	0.125	0.050	0.900	0.100	0.000	0.525	0.125	0.225	0.875	0.100	0.025
	$k=0.8n$	0.850	0.075	0.075	0.900	0.050	0.050	0.525	0.300	0.075	0.775	0.100	0.100

4.2 Case 2: long memory com $\mathbf{d} \in (0, 0.5)$ with and without outliers

In the following tables, the relative frequency estimates for estimated number of dimensions is presented in order to verify the asymptotic properties discussed in Section 3 for time series with long-memory features. The effect of additive outliers and the usefulness of the proposed robust methodology are also considered.

For the long-memory experiments the additional $\boldsymbol{d} = [0.1, 0.2, 0.4]'$ parameter is set for the VARFIMA model in Equation 2. Table 6 shows the relative frequency estimates for the dimensional reduction ($\widehat{r}$ and $\hat{r}_{Q_n} = 1, 2$ or 3) when $r = 3$ using Model 1's $\mathbf{\Phi}$ coefficients from Table 2. Comparing the first columns of Tables 4 and 6, it is possible to see that the long memory property produces an improvement in the estimation of the number of factors, compare the values when $\widehat{r} = 3$. The robust methodology performs similarly to the short-memory case in long-memory context and the same conclusions are held.

Table 6: Relative frequency estimates for dimensional reduction - $\boldsymbol{d} = [0.1, 0.2, 0.4]'$ and Model 1's $\mathbf{\Phi}$ coefficients.

		$\widehat{\Gamma}_{\boldsymbol{z}}$			$\widehat{\Gamma}_{\boldsymbol{z}}$			$\widehat{\Gamma}_{\boldsymbol{z},Q_n}$			$\widehat{\Gamma}_{\boldsymbol{z},Q_n}$		
		$p=0$			$p=0.05$ and $\omega=15$			$p=0$			$p=0.05$ and $\omega=15$		
		$n=100$			$n=100$			$n=100$			$n=100$		
		$\widehat{r}=1$	$\widehat{r}=2$	$\widehat{r}=3$	$\widehat{r}=1$	$\widehat{r}=2$	$\widehat{r}=3$	$\widehat{r}_{Q_n}=1$	$\widehat{r}_{Q_n}=2$	$\widehat{r}_{Q_n}=3$	$\widehat{r}_{Q_n}=1$	$\widehat{r}_{Q_n}=2$	$\widehat{r}_{Q_n}=3$
$\delta=0$	$k=0.2n$	0.110	0.260	0.630	0.310	0.270	0.260	0.150	0.310	0.540	0.160	0.320	0.520
	$k=0.5n$	0.080	0.110	0.810	0.100	0.200	0.320	0.140	0.110	0.750	0.160	0.130	0.710
	$k=0.8n$	0.010	0.150	0.840	0.140	0.160	0.280	0.020	0.160	0.820	0.020	0.200	0.780

Table 7 shows the relative frequency estimates for the dimensional reduction using Model 3's Φ coefficients with long-memory. Although, like the results in Table 3 and 5, the results are affected by a complex correlation structure, the presence of the long-memory property seems to slightly improve the estimates when k is small. The robust methodology performs similarly as in the results of others tables.

Table 7: Relative frequency estimates for dimensional reduction - $\boldsymbol{d} = [0.1, 0.2, 0.4]'$ and Model 3's $\mathbf{\Phi}$ coefficients.

		$\widehat{\Gamma}_{\boldsymbol{z}}$			$\widehat{\Gamma}_{\boldsymbol{z}}$			$\widehat{\Gamma}_{\boldsymbol{z},Q_n}$			$\widehat{\Gamma}_{\boldsymbol{z},Q_n}$		
		$p=0$			$p=0.05$ and $\omega=15$			$p=0$			$p=0.05$ and $\omega=15$		
		$n=100$			$n=100$			$n=100$			$n=100$		
		$\widehat{r}=1$	$\widehat{r}=2$	$\widehat{r}=3$	$\widehat{r}=1$	$\widehat{r}=2$	$\widehat{r}=3$	$\widehat{r}_{Q_n}=1$	$\widehat{r}_{Q_n}=2$	$\widehat{r}_{Q_n}=3$	$\widehat{r}_{Q_n}=1$	$\widehat{r}_{Q_n}=2$	$\widehat{r}_{Q_n}=3$
$\delta=0$	$k=0.2n$	0.850	0.125	0.025	0.650	0.125	0.200	0.950	0.025	0.025	0.900	0.050	0.050
	$k=0.5n$	0.875	0.100	0.025	0.475	0.175	0.150	0.900	0.075	0.025	0.900	0.100	0.000
	$k=0.8n$	0.900	0.050	0.500	0.400	0.125	0.200	0.950	0.050	0.000	0.925	0.050	0.025

5 Application to PM_{10}

This section presents an application of the methodology discussed in Sections 2 and 3 for PM_{10} concentrations measured at the Air Quality Automatic Monitoring Network (AQAMN) of the Greater Vitória Region (GVR). The application was divided into two parts: 1) reduction of the dimensions, and 2) forecasting. The AQAMN consists of eight monitoring stations distributed in the cities of GVR as follows: two stations in Serra, Laranjeiras and Carapina. The city of Vitória has three stations, Jardim de Camburi, Suá and Centro (Vix-Centro). Vila Velha has two stations, Centro (VVCentro) and Ibes. The city of Cariacica has one station at Ceasa. The PM_{10} in $\mu g/m^3$ is monitored in all stations, therefore, $k = 8$. The PM_{10} series corresponds to the daily average observed at all stations from January 2005 to December 2009 in a total of $n = 1826$ observations

In Table 8, the descriptive statistics of the PM_{10} concentrations in the stations are summarized. From this table, one can observe that the mean and median of the pollutants are between $19\mu g/m^3$ and $44\mu g/m^3$ which are in accordance with parameters set the national regulatory agency. In addition, the maximum values are much larger than the third quartile quantities. This may be an indication that the data possesses aberrant observations.

Furthermore, Figure 1 displays the plots of the PM_{10} concentrations from January 2005 to December 2009 and Figure 2 shows the box-plots of the series. Based on these plots, the series indicated that they have high levels of concentrations which can be identified, from statistical point of view, as being outliers (additive). This empirical evidence justifies the use of both robust and non-robust methods to verify whether or not these high levels make any impact in the present investigation.

Table 8: The descriptive statistics for all AQAMN'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**

As discussed by the literature, the high concentration levels may lead to a reduction of the values of the classical ACF, thus, in Figure 3, the robust ACFs of the series are displayed. The plots of the classical ACF are not presented here to save space. However, these values were very similar to the robust ACF ones. As one way to visualize this, Figure 4 shows plots of $\widehat{\rho}_{\boldsymbol{z}}$ against $\widehat{\rho}_{\boldsymbol{z},Q_n}$ for the PM_10 concentrations recorded at the Vila Velha-Ibes station. As can be seen, there is an approximate linear relationship between both ACFs, which corroborate to the fact that there is no indication of outliers in the data.

Besides, from the robust ACFs plots, it is possible to note that the long memory property is well observed. To corroborate the similarity of standard ACF and robust ACF, Table 9 displays the estimates of the long-memory parameter d, at zero frequency, for the standard fractional ($\hat{d}$) and the robust fractional ($\hat{d}_r$) estimators. These estimates were obtained following the steps suggested in Molinares et al. (2009) and were computed near zero frequency to avoid any contamination from the seasonal counterparts. This empirical study confirms the evidence previously observed, that is, if there are any aberrant observations, they were

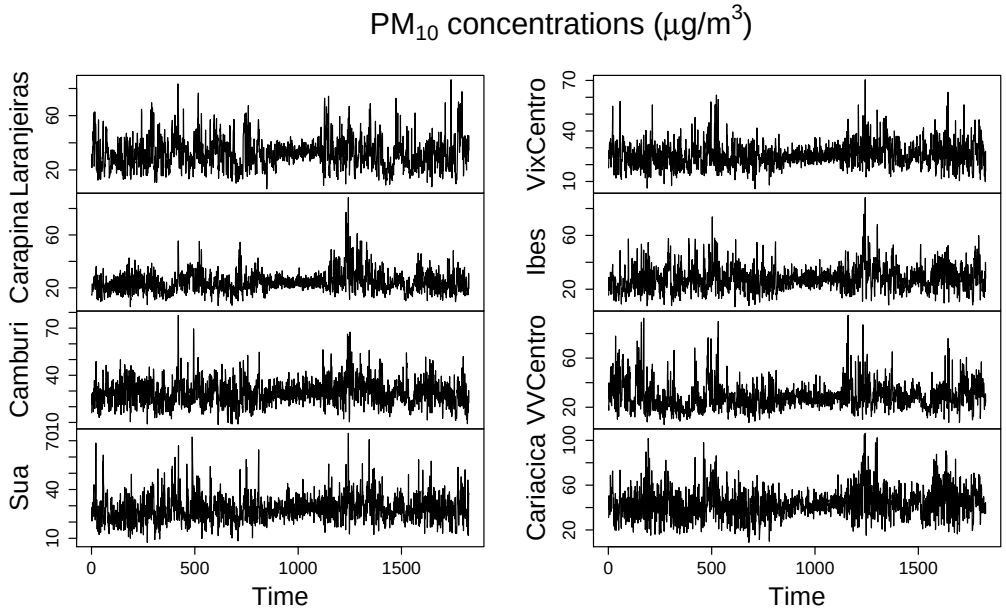


Figure 1: PM_{10}'s concentrations of RAMQAr's stations.

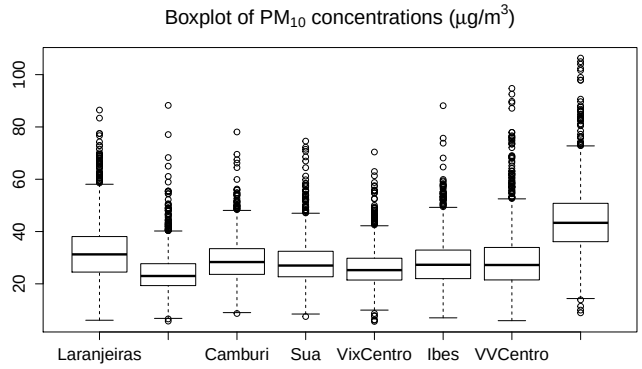


Figure 2: Boxplot of PM_{10}'s of RAMQAr's stations.

not strong enough to provoke interference with estimates.

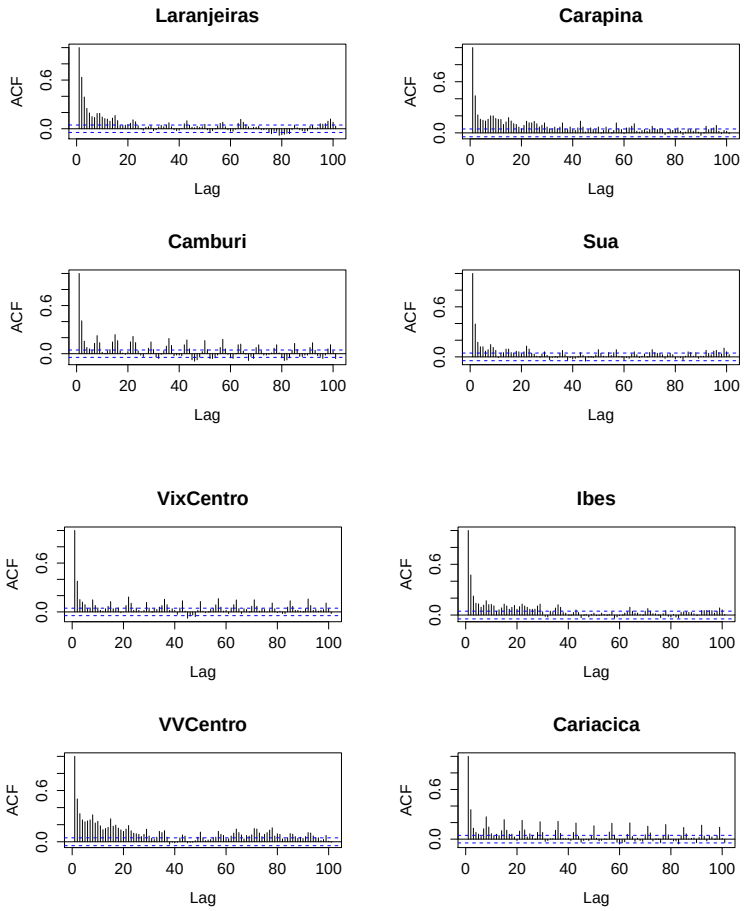


Figure 3: PM_{10} $\widehat{\rho}_{\boldsymbol{Z},Q_n}(h)$ function for AQAMN's Stations.

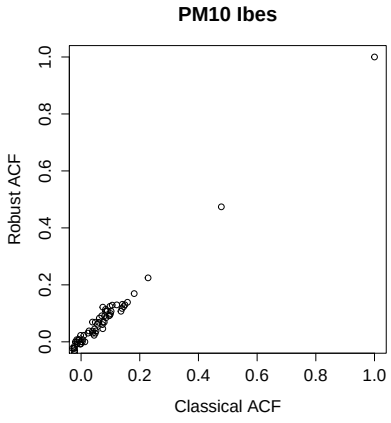


Figure 4: Scatter plot of $\widehat{\rho}_{\boldsymbol{z}}$ against $\widehat{\rho}_{\boldsymbol{z},Q_n}$ for Vila Velha-Ibes station.

Table 9: Estimates of standard d and robust d_r for different alphas (α)

α	$\hat{d}$	$sd(\hat{d})$	$\hat{d}_r$	$sd(\hat{d}_r)$
0.70	0.2564	0.0507	0.2664	0.0258
0.71	0.2554	0.0486	0.2525	0.0252
0.72	0.2275	0.0466	0.2275	0.0242
0.73	0.2257	0.0442	0.2353	0.0239
0.74	0.2167	0.0418	0.2355	0.0230
0.75	0.2254	0.0411	0.2232	0.0215
0.76	0.2347	0.0389	0.2382	0.0204
0.77	0.2405	0.0376	0.2392	0.0198
0.78	0.2310	0.0356	0.2456	0.0192
0.79	0.2592	0.0345	0.2630	0.0185
0.80	0.2639	0.0342	0.2510	0.0175

From the above discussion, it is expected that the next empirical investigation, which is related to estimate AF model and forecasting issues, will show similar performance for both methodologies, that is, the standard and robust ones. In this context, the attention is directed toward the application of the methodology to reduce the dimensionality of PM_{10} data from AQMN of RGV. The analysis was carried out with $h_0 = 14$ to consider long-memory and seasonality for the eigenanalysis on $\widehat{\boldsymbol{M}}$ and on $\widehat{\boldsymbol{M}}_{Q_n}$ of Equations 18 and 28, respectively. The eigenvalues obtained (in decreasing order) and their ratios obtained using $\widehat{\Gamma}_{\boldsymbol{z}}$ are shown in Figure 5 (first two panels, respectively). The corresponding robust version; i.e using $\widehat{\Gamma}_{\boldsymbol{z},Q_n}$ is shown in Figure 6. The plots show similar results which is, as previously stated, an expected result. Both plots indicate that the number of factors is $r = 1$, and this can also be seen from the ratio-based estimator that leads to a reduction to one factor. The reduction was not affected when varying the value of h_0.

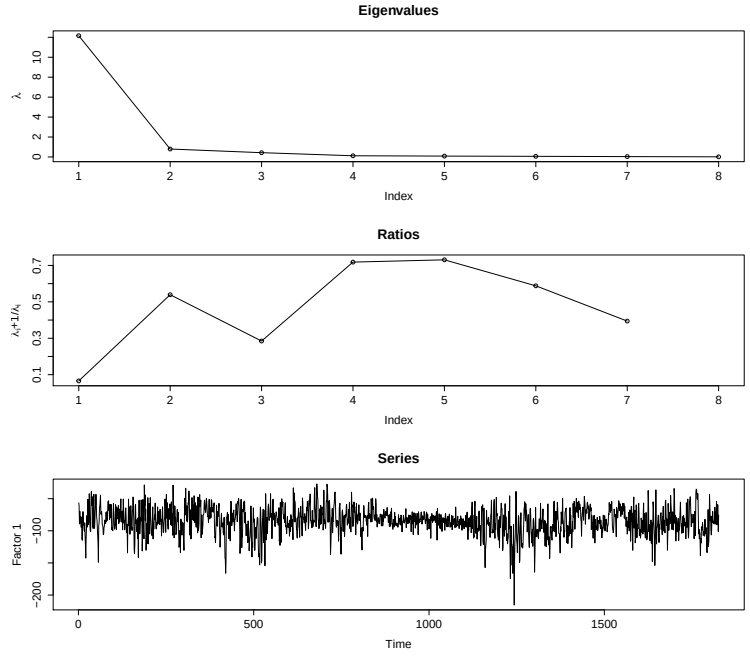


Figure 5: Plots of estimated eigenvalues, ratios of estimated eigenvalues of $\hat{\boldsymbol{M}}$ and estimated first factor.

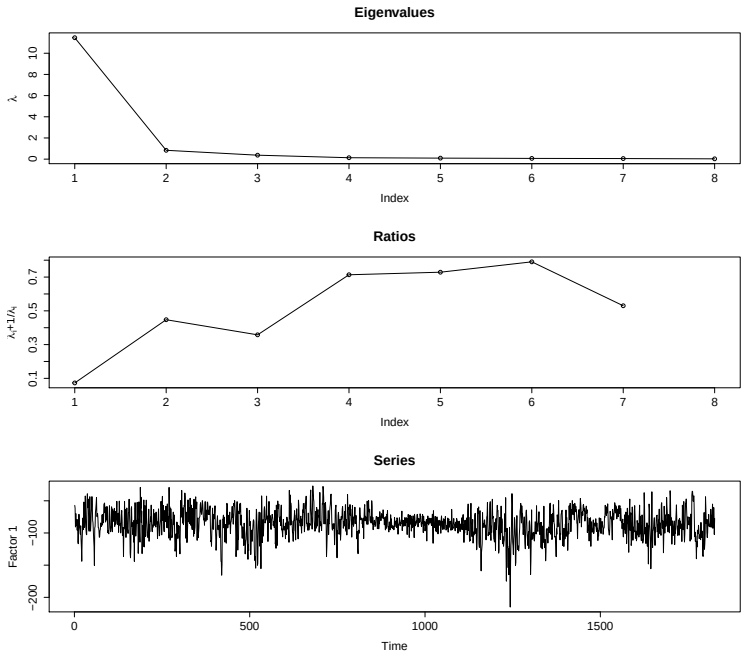


Figure 6: Plots of estimated eigenvalues, ratios of estimated eigenvalues of $\widehat{\boldsymbol{M}}_{Q_n}$ and estimated first factor.

The last panels in Figures 5 and 6 display the time series plots of the estimated $\widehat{\boldsymbol{z}}_t$ defined in Equation 13, for $\widehat{\Gamma}_{\boldsymbol{z}}$ and $\widehat{\Gamma}_{\boldsymbol{z},Q_n}$, respectively. Their corresponding ACFs are shown in Figure 7. As one can see, the ACFs from both methods present significant autocorrelations (decaying at a hyperbolic rate) and stochastic seasonal behavior remained from the original data set. Thus, these features must be also considered in a subsequent application of the proposed methodology.

As previously mentioned, the use of standard VARMA model to estimate the multivariate series may pose difficulties to model estimation and forecasting due to the dimension of the data. However, to circumvent this problem, the use of the factorial analysis proposed in this paper (Equation (1)), with $\hat{\boldsymbol{P}}$ and the estimated factor series $\hat{x}_t$, is encouraged. Based on this, the h-step ahead forecast for the $\boldsymbol{z}_t$ series can be simplified using the formula $\hat{\boldsymbol{z}}_{T+h}^{(h)} = \hat{\boldsymbol{P}}\hat{\boldsymbol{x}}_{T+h}^{(h)}$, where $\hat{\boldsymbol{x}}_{T+h}^{(h)}$ is an h-step ahead forecast for f_t, based on the estimated past values $\hat{\boldsymbol{x}_1}, \ldots, \hat{\boldsymbol{x}_T}$. This can be obtained, for example, by fitting a vetor-autoregressive model to $\hat{\boldsymbol{x}_1}, \ldots, \hat{\boldsymbol{x}_T}$ (Lam et al. 2011).

The factor model indicated the use of one factor ($\hat{r} = 1$) and the standard ACF and robust ACF (Figure (7)) demonstrated that the factor may presents long memory property at seasonal and non-seasonal periods of approximately 7 days. These oscillations are due to the fact that the series corresponds of the daily mean observations. Although, the proposed methodology discussed in the previous sections does not embrace the seasonal behaviour, this characteristic is considered here in subsequent analysis in order to produce accurate predictions. Thus, the $\hat{\boldsymbol{x}}_{T+h}^{(h)}$ was obtained by means of robust SARFIMA model (denoted by Model 1) proposed by Reisen et al. (2015). In addition, $\hat{\boldsymbol{x}}_{T+h}^{(h)}$ was also obtained using the standard SARMA model (Model 2). The performance of both methods will be compared to the performance of the standard VARMA model (Model 3).

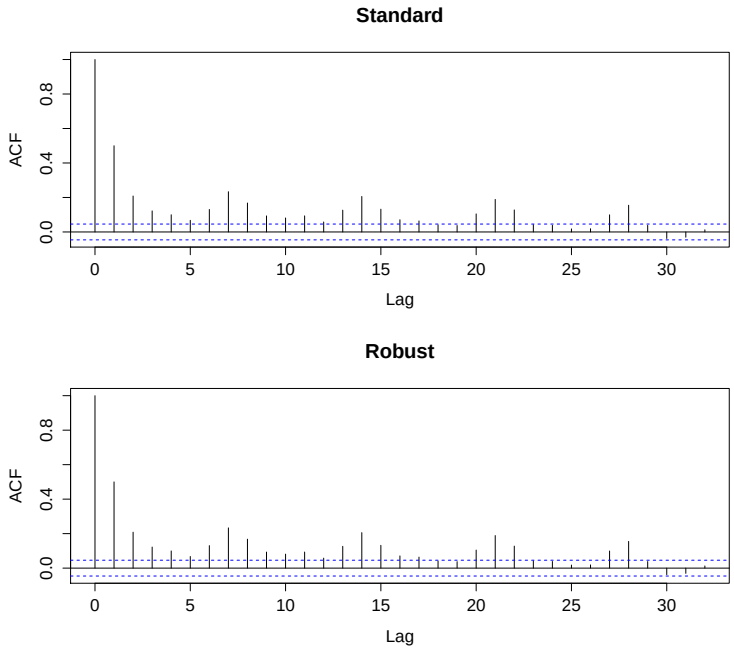


Figure 7: The autocorrelations of the estimated factor for $\widehat{\Gamma}_{\boldsymbol{z}}$ and $\widehat{\Gamma}_{\boldsymbol{z},Q_n}$, respectively.

For the subsequent analysis, the resulting factor series was divided into two parts: learning and prediction sets. The 1626 observations from January 1st, 2005 to June 14th, 2009 are considered to be the learning set and the remaining 200 observations are considered for the forecasting study. Based on statistical analysis, the SARFIMA$(1, 0.3315, 0) \times (0, 0.1623, 0)_7$, where the estimates were computed for the bandwidth $m' = 33$ in accord with Reisen et al. (2015), and SARMA$(1, 0) \times (1, 0)_7$ were chosen for the factor series. The ACFs of the residuals of both models are presented in Figure 8, where it can be observed that the filters captured the long-memory correlation and seasonality of the factor series quite well. The Box-Pierce and Ljung-Box statistics (robust tests) demonstrated that the sample residuals are not time-correlated (the results are available upon request).

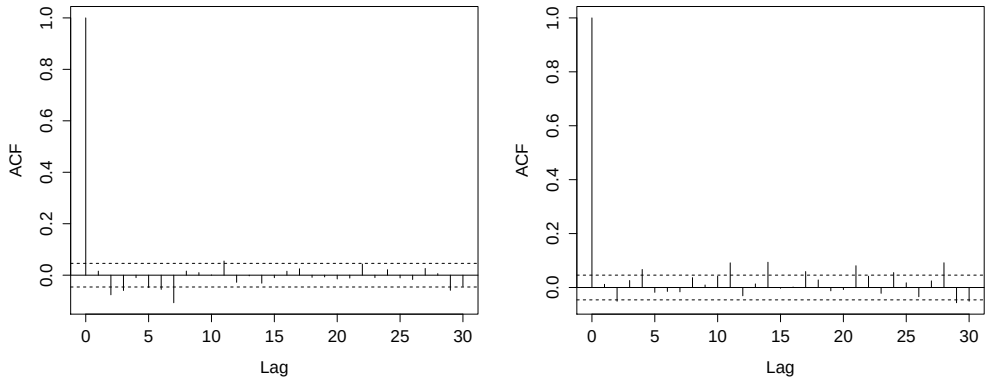


Figure 8: ACFs values of the residuals of SARFIMA and SARMA models, respectively, for the resulting factor series

As stated above, the 200 observations from June 15th, 2009 to December 31st, 2009 were discarded from the modeling stage to be used in the out-of-sample forecast study. Figure 9 presents the visual analysis of the one-step-ahead forecast values of the Model 1 (first panel) and Model 2 models (second panel) for the estimated factor, i.e, from June 15th, 2009 to December 31st, 2009. It can be observed that both plots indicate a reasonably good performance. That is, the models were able to capture the dynamic behavior of the factor series for one-step-ahead forecasts.

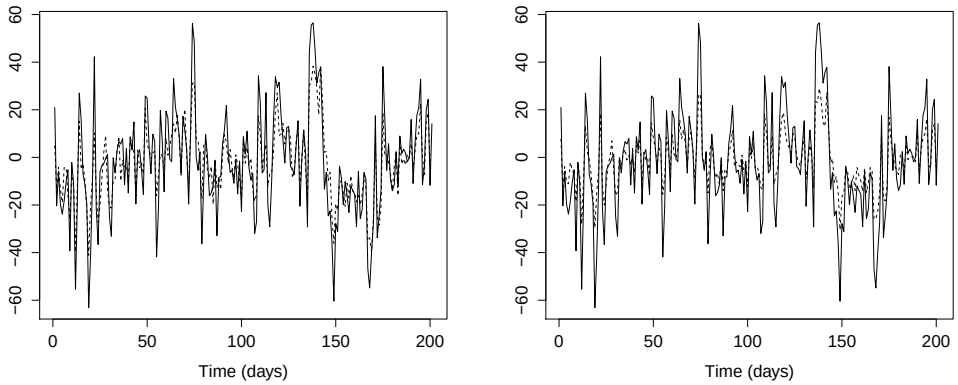


Figure 9: Forecasted values by the SARFIMA and SARMA models, respectively, for the resulting factor series

Now, motivated by the above results, the next step is to use the estimated factor series to forecast values of the observed series. Let $\hat{\boldsymbol{z}}_{T+h}^{(h)} = \hat{\boldsymbol{P}}\hat{\boldsymbol{x}}_{T+h}^{(h)}$, thus, the forecasts for the original series were performed based on the Model 1 and Model 2. For all series (stations), the one-step-ahead forecast values performed well. In order to exemplify, Figure (10) presents the visual analysis of the one-step-ahead forecast values for PM_{10} concentrations of the Vitória-Suá station obtained by Model 1 and 2 from the resulting factor series and Model 3 (using the standard VAR(1) model; estimated model parameters are available upon request) for

the original vector series, which is also considered in the study for the purpose of comparison between these models.

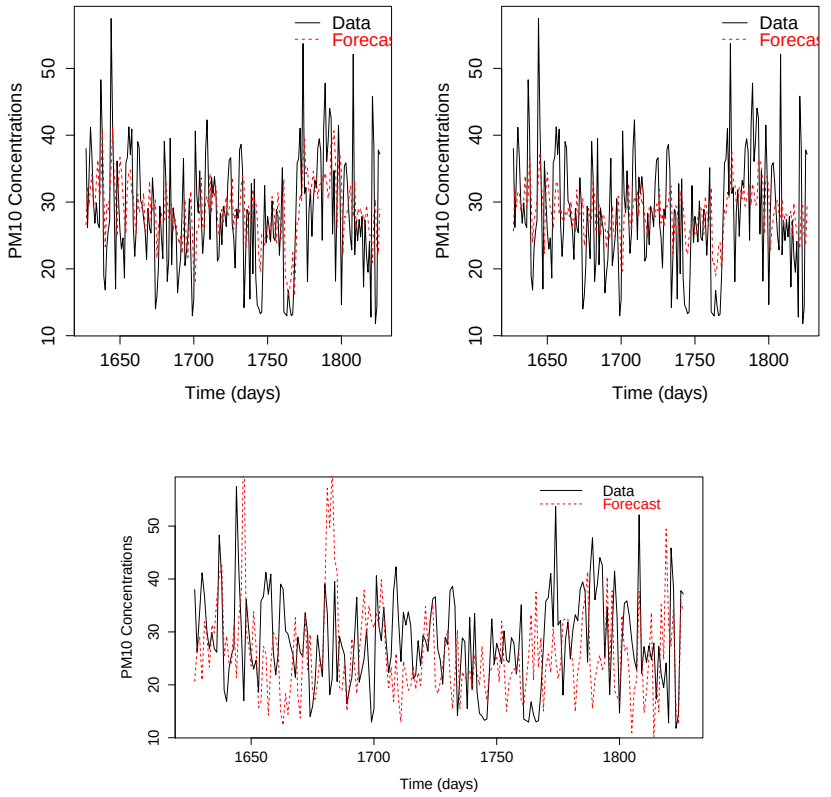


Figure 10: Forecasted values of Vitória-Suá obtained by the Model 1 and Model 2 (from the resulting factor series) and Model 3 (from the original series).

To measure the accuracy of the forecasts, the criteria used were the Mean Square prediction Error (MSPE), Mean Percent Prediction Error (MPPE), Mean Absolute Percent Prediction Error (MAPPE), the values are displayed in Table 10 for the Vitória-Suá station. From the table, it can be seen that the AF-SARFIMA model presented more accurate forecasts than the AF-SARMA and VAR models.

Table 10: MSPE, MPPE and MAPPE of the fitted models.

	MSPE	MPPE	MAPPE
AF-SARFIMA	8.22	6.49	22.83
AF-SARMA	8.22	6.52	23.16
VARMA	12.34	9.27	40.28

To summarize, the application discussed in this section corroborates with the usefulness of applying the proposed methodology in the context of additive outliers and long-memory in multivariate time series.

6 Conclusions

In this article a robust factor model for high-dimensional time series with short and long-memory and additive outliers is proposed. Some theoretical results are discussed and these were empirically investigated by Monte Carlo experiments under different scenarios, which showed evidence that, for finite sample, the effect of the additive outliers on the reduction order factor dimension test. The proposed robust estimator performed quite well and this indicates that the robust test can be very useful in practical applications where there is any evidence of aberrant observations, such as, high levels of concentrations in the pollution area. In addition, the proposed methodology was used to identify pollution behavior of the pollutant PM_{10} in the Greater Region of Vitória and to forecast the observations, which can be very useful for the management of the air quality network. The results in this paper will hopefully stimulate further research on this theme.

7 Acknowledgment

The authors would like to thank CNPq, CAPES and FAPES for their financial support.

References

Anderson, T. W. (2003), *An introduction to multivariate statistical analysis*, 3rd edn, John Wiley & Sons, New Jersey.

Brunekreef, B. & Holgate, S. T. (2002), 'Air pollution and health', *The Lancet* **360**(9341), 1233–1242.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/12401268*

Chang, I., Tiao, G. C. & Chen, C. (1988), 'Estimation of time series parameters in the presence of outliers', *Technometrics* **30**(2), 193–204.

Chen, C. & Liu, L.-M. (1993), 'Joint estimation of model parameters and outlier effects in time series', *Journal of the American Statistical Association* **88**(421), 284–297.

Chung, C. (2002), 'Sample means, sample autocovariances, and linear regression of stationary multivariate long memory processes', *Econometric Theory* **18**, 51–78.

Cotta, H. H. A. & Reisen, V. A. (2015), Robust principal component analysis with air pollution data: an application to the clustering of RAMQAr. Unpublished manuscript.

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

Curtis, L., Rea, W., Smith-Willis, P., Fenyves, E. & Pan, Y. (2006), 'Adverse health effects of outdoor air pollutants', *Environment International* **32**(6), 815–830.
URL: *http://www.sciencedirect.com/science/article/pii/S0160412006000444*

Granger, C. (1980), 'Long memory relationships and the aggregation of dynamic models', *Journal of Econometrics* **14**(2), 227–238.

Granger, C. (1981), 'Some properties of time series data and their use in econometric model specification', *Journal of Econometrics* **16**(1), 121–130.

Granger, C. W. J. & Joyeux, R. (1980), 'An introduction to long-memory times series models and fractional differencing', *Journal of Time Series Analysis* **1**, 15–29.

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

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

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. (2007), *Applied multivariate statistical analysis*, 6rd edn, Prentice Hall, New Jersey.

Lam, C. & Yao, Q. (2012), 'Factor modeling for high-dimensional time series: Inference for the number of factors', *Ann. Statist.* **40**(2), 694–726.
URL: *http://dx.doi.org/10.1214/12-AOS970*

Lam, C., Yao, Q. & Bathia, N. (2011), Estimation of latent factors for high-dimensional time series, Lse research online documents on economics, London School of Economics and Political Science, LSE Library.
URL: *http://EconPapers.repec.org/RePEc:ehl:lserod:31549*

Lévy-Leduc, C., Boistard, H., Moulines, E., Taqqu, M. S. & Reisen, V. A. (2011*a*), 'Asymptotic properties of u-processes under long-range dependence', *The annals of statistics* pp. 1399–1426.

Lévy-Leduc, C., Boistard, H., Moulines, E., Taqqu, M. S. & Reisen, V. A. (2011*b*), '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*c*), '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.

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.

Maynard, R. (2004), 'Key airborne pollutants: the impact on health', *Science of The Total Environment* **334-335**(0), 9–13.
URL: *http://www.sciencedirect.com/science/article/pii/S0048969704003493*

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.

Pan, J. & Yao, Q. (2008), 'Modelling multiple time series via common factors', *Biometrika* **95**(2), pp. 365–379.
URL: *http://www.jstor.org/stable/20441470*

Peña, D. & Box, G. E. P. (1987), 'Identifying a simplifying structure in time series', *Journal of the American Statistical Association* **82**(399), pp. 836–843.
URL: *http://www.jstor.org/stable/2288794*

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

Reisen, V. A., Sarnaglia, A. J. Q., Reis, N. C., Lévy-Leduc, C. & Santos, J. M. (2014), 'Modeling and forecasting daily average pm 10 concentrations by a seasonal long-memory model with volatility', *Environmental Modelling & Software* **51**, 286–295.

Reisen, V. A., Sgrancio, A. M., Monte, E. Z., Molinares, F. A. F., da Conceição Franco, G. & Ziegelmann, F. A. (2015), Fractional seasonal process with outliers to model and forecast daily average SO_2 concentrations. Preprint submitted to European Journal of Operational Research.

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

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

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

Seinfeld, J. H. & Pandis, S. N. (2006), *Atmospheric chemistry and physics: from air pollution to climate change*, J. Wiley, New York.

Sowell, F. (1992*a*), 'Maximum likelihood estimation of stationary univariate fractionally integrated time series models', *Journal of Econometrics* **53**(1–3), 165–188.

Sowell, F. (1992*b*), 'Modeling long-run behavior with the fractional ARIMA model', *Journal of Monetary Economics* **29**(2), 277–302.

Stock, J. H. & Watson, M. W. (2002), 'Forecasting using principal components from a large number of predictors', *Journal of the American Statistical Association* **97**(460), 1167–1179.
URL: *http://www.jstor.org/stable/3085839*

Taqqu, M. S. (2003), Fractional brownian motion and long-range dependence, *in* P. Doukhan, G. Oppenheim & M. Taqqu, eds, 'Theory and applications of long-range dependence', Birkhäuser, Boston.

Tiao, G. C. & Tsay, R. S. (1989), 'Model specification in multivariate time aeries', *Journal of the Royal Statistical Society. Series B (Methodological)* **51**(2), pp. 157–213.
URL: *http://www.jstor.org/stable/2345602*

Tsay, R. S. (1988), 'Outliers, level shifts, and variance changes in time series.', *Journal of forecasting* **7**(1), 1–20.

Wang, H. (2010), 'Factor profiling for ultra high dimensional variable selection'.
URL: *http://ssrn.com/abstract=1613452*

Watson, J. G., Zhu, T., Chow, J. C., Engelbrecht, J., Fujita, E. M. & Wilson, W. E. (2002), 'Receptor modeling application framework for particle source apportionment', *Chemosphere* **49**(9), 1093–1136.
URL: *http://www.sciencedirect.com/science/article/pii/S0045653502002436*

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

WHO (2014), *Air pollution estimates.*

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

5 CONCLUSÕES

As conclusões desta tese foram apresentadas no artigo 1 e no artigo 2. E estão relatadas a seguir.

No artigo 1 foi adotado um modelo SARFIMA na presença de valores extremos (outliers) para modelar as médias diárias de concentrações de SO_2. As análises estatísticas mostraram que a série SO_2 apresenta o comportamento de sazonalidade, estacionariedade e de memória longa. Para estimar os parâmetros fracionais robustos d_R and D_R foram combinados os métodos sugeridos por Reisen et al. (2014b), Lévy-Leduc et al.(2011c) and Molinares et al. (2009).

Baseado nos estimadores robustos $\hat{d}_R$ e $\hat{D}_R$ foi ajustado o modelo SARFIMA$(0, d, 1) \times (0, D, 0)_7$. Os resultados mostraram que os resíduos do modelo ajustado são não-correlacionados e normalmente distribuídos. Um modelo padrão SARMA$(1, 0) \times (1, 0)_7$ também foi ajustado às concentrações de SO_2, para mensurar a qualidade da previsão do modelo SARFIMA em relação ao modelo SARMA. A previsão, utilizando o erro quadrático médio percentual, indicou que o modelo SARFIMA robusto apresentou um alto nível de precisão, especialmente para previsões em períodos de tempo maiores.

Os parâmetros fracionários robustos são uma forma atraente de estimar os parâmetros do modelo SARFIMA com memória longa, sazonalidade e valores atípicos, e podem ser facilmente utilizados em aplicações práticas reais, em diversas áreas do conhecimento. Através dos resultados deste trabalho, esperamos estimular a investigação sobre o uso de métodos de estimação robusta e de memória longa, para representar séries ambientais e estimar as previsões de concentrações dos poluentes de forma mais precisa.

No artigo 2 foi proposto um modelo fatorial robusto aplicado em séries temporais de grandes dimensões, com propriedades de curta e longa dependência, na presença de observações atípicas (outliers) e de sazonalidade.

Alguns resultados teóricos foram discutidos e investigados, empiricamente, através dos experimentos de Monte Carlo em diferentes cenários, que evidenciaram, para a amostra finita, o efeito dos outliers aditivos na redução da ordem da dimensão do modelo fatorial estimado. O estimador robusto proposto apresentou bom desenpenho, indicando que o teste robusto pode ser muito útil em aplicações práticas, onde não há qualquer evidência de observações atípicas, tais como altos níveis de concentrações na área de poluição.

Além disso, esta metodologia proposta foi utilizada em uma situação real de poluentes at-

mosféricos. O modelo fatorial foi aplicado para identificar o comportamento das concentrações de PM_{10} na Região da Grande Vitória, e para fazer previsões dos níveis de concentrações deste poluente. Os resultados mostraram que o modelo fatorial apresentou melhor acurácia na previsão, em relação ao modelo vetorial autoregressivo (VAR).

A conclusões feitas aqui mostraram que a metodologia apresentada nesta tese pode ser muito útil para a gestão da rede de monitoramento da qualidade do ar. Esperamos que os pesquisadores tenham sido estimulados com o tema.

REFERÊNCIAS

AMENGUAL, D.; WATSON, M. W. Consistent estimation of the number of dynamic factors in a large n and t panel. **Journal of Business & Economic Statistics**, Taylor & Francis, v. 25, n. 1, p. 91–96, 2007.

ANDERSON, M. J. et al. Source apportionment of exposures to volatile organic compounds: Ii. application of receptor models to team study data. **Atmospheric Environment**, Elsevier, v. 36, n. 22, p. 3643–3658, 2002.

ANDERSON, T. W. et al. **An introduction to multivariate statistical analysis**. [S.l.]: Wiley New York, 1958. v. 2.

ANDERSSON, S. et al. Composition and evolution of volcanic aerosol from eruptions of kasatochi, sarychev and eyjafjallajökull in 2008–2010 based on caribic observations. **Atmospheric Chemistry and Physics**, Copernicus GmbH, v. 13, n. 4, p. 1781–1796, 2013.

BAI, J. Inferential theory for factor models of large dimensions. **Econometrica**, JSTOR, p. 135–171, 2003.

BAI, J.; NG, S. Determining the number of factors in approximate factor models. **Econometrica**, Wiley Online Library, v. 70, n. 1, p. 191–221, 2002.

BAIRD, C. **Química ambiental**. [S.l.]: Reverté, 2001.

BELIS, C. et al. Critical review and meta-analysis of ambient particulate matter source apportionment using receptor models in europe. **Atmospheric Environment**, Elsevier, v. 69, p. 94–108, 2013.

BOX, G. E.; TIAO, G. C. A canonical analysis of multiple time series. **Biometrika**, Biometrika Trust, v. 64, n. 2, p. 355–365, 1977.

BRAGA, B. et al. **Introdução à Engenharia Ambiental: o Desafio do Desenvolvimento Sustentável, 2ª edição**. [S.l.]: São Paulo: Editora Pearson, 2005.

COTTA, H. H. A. **Analise de Componentes Principais Robusta em Dados de Poluição do Ar: Aplicação à Otimização de uma Rede do Monitoramento**. Dissertação (Mestrado) — Departamento de Engenharia Ambiental, Universidade Federal do Espírito Santo, Vitória, 2014.

COULTER, T. C. **EPA-CMB8. 2 users manual**. [S.l.]: US Environmental Protection Agency, Office of Air Quality Planning & Standards, Emissions, Monitoring & Analysis Division, Air Quality Modeling Group, 2004.

FORNI, M. et al. The generalized dynamic factor model consistency and rates. **Journal of Econometrics**, Elsevier, v. 119, n. 2, p. 231–255, 2004.

GEORGOULIAS, A. et al. A study of the total atmospheric sulfur dioxide load using ground-based measurements and the satellite derived sulfur dioxide index. **Atmospheric Environment**, Elsevier, v. 43, n. 9, p. 1693–1701, 2009.

GODISH, T. **Air quality**. [S.l.]: Boca Raton: CRC Press, LLC, 1997.

HAMILTON, J. D. **Time series analysis**. [S.l.]: Princeton university press Princeton, 1994. v. 2.

HARMAN, H. H. Modern factor analysis. Univ. of Chicago Press, 1960.

HASSANZADEH, S.; HOSSEINIBALAM, F.; ALIZADEH, R. Statistical models and time series forecasting of sulfur dioxide: a case study tehran. **Environmental monitoring and assessment**, Springer, v. 155, n. 1-4, p. 149–155, 2009.

HOLGATE, S. T. et al. **Air pollution and health**. [S.l.]: Academic Press, 1999.

HOPKE, P. K. **Receptor modeling for air quality management**. [S.l.]: Elsevier, 1991. v. 7.

HOSKING, J. R. Fractional differencing. **Biometrika**, Biometrika Trust, v. 68, n. 1, p. 165–176, 1981.

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, 2013.

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

JACOBSON, M. **Atmospheric Pollution: History, Science, and Regulation**. Cambridge University Press, 2002. ISBN 9780521010443. Disponível em: ⟨https://books.google.com.br/books?id=NN5S0_3dEvkC⟩.

JOHNSON, R. A.; WICHERN, D. W. et al. **Applied multivariate statistical analysis**. [S.l.]: Prentice hall Englewood Cliffs, NJ, 1992. v. 4.

KAISER, H. F. The varimax criterion for analytic rotation in factor analysis. **Psychometrika**, Springer, v. 23, n. 3, p. 187–200, 1958.

KANAROGLOU, P. S. et al. Estimation of sulfur dioxide air pollution concentrations with a spatial autoregressive model. **Atmospheric Environment**, Elsevier, v. 79, p. 421–427, 2013.

LAM, C.; YAO, Q. et al. Factor modeling for high-dimensional time series: inference for the number of factors. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 40, n. 2, p. 694–726, 2012.

LEE, S. et al. Source apportionment of pm 2.5: Comparing pmf and cmb results for four ambient monitoring sites in the southeastern united states. **Atmospheric Environment**, Elsevier, v. 42, n. 18, p. 4126–4137, 2008.

LÉVY-LEDUC, C. et al. Robust estimation of the scale and of the autocovariance function of gaussian short-and long-range dependent processes. **Journal of Time Series Analysis**, Wiley Online Library, v. 32, n. 2, p. 135–156, 2011.

LUVSAN, M.-E. et al. The influence of emission sources and meteorological conditions on so 2 pollution in mongolia. **Atmospheric Environment**, Elsevier, v. 61, p. 542–549, 2012.

MALLIK, C.; VENKATARAMANI, S.; LAL, S. Study of a high so2 event observed over an urban site in western india. **Asia-Pacific Journal of Atmospheric Sciences**, Springer, v. 48, n. 2, p. 171–180, 2012.

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

MONROY, N. A. J. **Modelo ARFIMA Espaço-Temporal em Estudos de Poluição do Ar**. Tese (Doutorado) — Programa de Pós-graduação em Engenharia Ambiental - Universidade Federal do Espírito Santo., 2013.

PAATERO, P. et al. Understanding and controlling rotations in factor analytic models. **Chemometrics and intelligent laboratory systems**, Elsevier, v. 60, n. 1, p. 253–264, 2002.

PENA, D.; BOX, G. E. Identifying a simplifying structure in time series. **Journal of the American Statistical Association**, Taylor & Francis Group, v. 82, n. 399, p. 836–843, 1987.

PEÑA, D.; PONCELA, P. Nonstationary dynamic factor analysis. **Journal of Statistical Planning and Inference**, Elsevier, v. 136, n. 4, p. 1237–1257, 2006.

PINTO, W. P. **O uso da Metodologia de Dados Faltantes em Séries Temporais com Aplicação a dados de Concentração de (PM10) observados na região da Grande Vitória**. Dissertação (Mestrado) — Departamento de Engenharia Ambiental, Universidade Federal do Espírito Santo, Vitória, 2013.

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.

PRIESTLEY, M.; RAO, T. S.; TONG, H. Applications of principal component analysis and factor analysis in the identification of multivariable systems. **Automatic Control, IEEE Transactions on**, IEEE, v. 19, n. 6, p. 730–734, 1974.

PRYBUTOK, V. R.; YI, J.; MITCHELL, D. Comparison of neural network models with arima and regression models for prediction of houston's daily maximum ozone concentrations. **European Journal of Operational Research**, Elsevier, v. 122, n. 1, p. 31–40, 2000.

REISEN, V. A. Estimation of the fractional difference parameter in the arima (p, d, q) model using the smoothed periodogram. **Journal of Time Series Analysis**, Wiley Online Library, v. 15, n. 3, p. 335–350, 1994.

REISEN, V. A. et al. Modeling and forecasting daily average pm 10 concentrations by a seasonal long-memory model with volatility. **Environmental Modelling & Software**, Elsevier, v. 51, p. 286–295, 2014.

SEINFELD, J. H.; PANDIS, S. N. **Atmospheric chemistry and physics: from air pollution to climate change**. [S.l.]: John Wiley & Sons, 2006.

SOARES, I. P. **Avaliação do uso de diferentes modelos receptores para determinação da contribuição das fontes de partículas totais em suspensão**. Dissertação (Mestrado) — Departamento de Engenharia Ambiental, Universidade Federal do Espírito Santo, Vitória, 2011.

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 RGV, ES.** Dissertação (Mestrado) — Departamento de Engenharia Ambiental, Universidade Federal do Espírito Santo, Vitória, 2013.

SPEARMAN, C. "general intelligence,"objectively determined and measured. **The American Journal of Psychology**, JSTOR, v. 15, n. 2, p. 201–292, 1904.

STOCK, J. H.; WATSON, M. W. Forecasting with many predictors. **Handbook of economic forecasting**, Elsevier, v. 1, p. 515–554, 2006.

TRINDADE, C. C. **Avaliação do Uso de Diferentes Modelos Receptores com Dados de PM2.5: Balanço Químico de Massa (BQM) e Fatoração de Matriz Positiva (FMP).** Dissertação (Mestrado) — Departamento de Engenharia Ambiental, Universidade Federal do Espírito Santo, Vitíria, 2009.

USERO, J.; GRACIA, I. Chemical element balances and identification of dustfall sources from the seville atmospheric environment? **Toxicological & Environmental Chemistry**, Taylor & Francis, v. 11, n. 1, p. 51–60, 1986.

VIANA, M. et al. Identification of pm sources by principal component analysis (pca) coupled with wind direction data. **Chemosphere**, Elsevier, v. 65, n. 11, p. 2411–2418, 2006.

WATSON, J. G. et al. Receptor modeling application framework for particle source apportionment. **Chemosphere**, Elsevier, v. 49, n. 9, p. 1093–1136, 2002.

WEI, W. W.-S. **Time series analysis**. [S.l.]: Addison-Wesley Redwood City, California, 1994.

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

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