UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO

Centro Tecnológico

Programa de Pós-graduação em Engenharia Ambiental

Tese de Doutorado

Modelo ARFIMA Espaço-Temporal em Estudos de Poluição do Ar

Orientador:
Prof. Valdério A. Reisen, PhD.

Aluno:
Nátaly A. Jiménez Monroy

Co-orientador:
Prof. Tata Subba Rao, PhD.

Vitória
2013

Nátaly Adriana Jiménez Monroy

MODELO ARFIMA ESPAÇO-TEMPORAL EM ESTUDOS DE POLUIÇÃO DO AR.

Tese apresentada ao Programa de Pós-graduação em Engenharia Ambiental do Centro Tecnológico da Universidade Federal do Espírito Santo, como requisito parcial para obtenção do título de Doutora em Engenharia Ambiental, na área de concentração Poluição do Ar.
Orientador: Prof. Valdério Reisen, PhD.
Co-orientador: Prof. Tata Subba Rao, PhD.

Vitória
2013

Aos meus amores, Sara e Fabio.

Agradecimentos

A Deus por me dar a vida, a família e as ótimas oportunidades que tenho aproveitado.

À minha adorada filha Sara, só o teu sorriso me faz esquecer dos momentos difíceis.

Ao meu amado esposo Fabio, pelo constante apoio, incentivo e paciência os quais foram fundamentais para finalizar mais esta travessia.

Aos meus pais Salvador e Teresa, às minhas irmãs Teisy e Gigi, meu cunhado Wilson e minha linda sobrinha Keyla, por sua constante voz de ânimo. Mesmo estando longe, seu amor e força me acompanham aonde quer que eu vá.

Ao professor Valdério A. Reisen pela orientação, sugestões e valiosas recomendações que tornaram possível a finalização desta Tese.

Ao professor Tata Subba Rao, pelas valiosíssimas intervenções que contribuiram grandemente para o melhoramento da qualidade desta pesquisa. Thanks a lot!

Aos amigos Alyne, Bart, Márcia, Alessandro, Marcelo, Melina, Rita e Mayana, pela amizade e os momentos de diversão que tornaram mais amenos estes anos.

A todos aqueles que participaram direta ou indiretamente na concretização deste sonho. Meus tios e primos na Colômbia e meus amigos da UNAL, especialmente Luz Clarita e Edwin.

Aos colegas do PPGEA e do NuMEs, pela solidariedade e as experiências compartidas.

À Rose, pela presteza e carinho com que sempre me ofereceu sua ajuda.

À CAPES, pelo apoio financeiro.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Símbolos e Abreviaturas

ACF	Função de Autocorrelação
ARMA(p, q)	Autorregressivo de Média Móvel com parâmetros p e q
ARFIMA(p, d, q)	Autorregressivo Integrado Fracionário de Média Móvel com parâmetros p, d e q
CO	Monóxido de Carbono
d	Parâmetro de diferenciação fracionária
EQM ou MSE	Erro Quadrático Médio
IBGE	Instituto Brasileiro de Geografia e Estatística
IEMA	Instituto Estadual de Meio Ambiente e Recursos Hídricos
IJSN	Instituto Jones dos Santos Neves
MAE	Erro Médio Absoluto
NO_2	Dióxido de Nitrogênio
p	Parâmetro autorregressivo
PACF	Função de Autocorrelação Parcial
PM_{10}	Material Particulado inalável. Diâmetro inferior a 10 mícrons
$PM_{2,5}$	Material Particulado com diâmetro inferior a $2,5$ mícrons
PTS	Partículas Totais em Suspensão
q	Parâmetro de média móvel
RAMQAr	Rede automática de monitoramento da qualidade do ar
RMSE	Raiz do Erro Quadrático Médio
SO_2	Dióxido de enxofre
STACF	Função de Autocorrelação Espaço-Temporal
STPACF	Função de Autocorrelação Parcial Espaço-Temporal
STARFIMA$(p_{\lambda_1,\lambda_2,\ldots,\lambda_p}, \boldsymbol{d}, q_{m_1,m_2,\ldots,m_q})$	Espaço-Temporal Autorregressivo Integrado Fracionário de Média Móvel com parâmetros p, $\lambda_1, \lambda_2, \ldots, \lambda_p$, $\boldsymbol{d} = (d_1, \ldots, d_N)$, q e $m_1, m_2, \ldots, m_q$
WHO	Organização Mundial da Saúde
d_{ij}	Distância Euclidiana entre os lugares i e j
$\mathcal{D}(B)$	Matriz diagonal de operadores de diferença fracionária
E$[X]$	Valor esperado da variável aleatória X
$f(\omega)$	Função de densidade espectral na frequência ω
$\boldsymbol{\varepsilon}(t) = [\epsilon_1(t), \ldots, \epsilon_N(t)]'$	Termo de erro aleatório no tempo $t = 1, \ldots, T$
$\boldsymbol{G}$	Matriz de variâncias e covariâncias do erro aleatório
$\gamma_{lk}(s)$	Função de covariância espaço-temporal
$\boldsymbol{I_N}$	Matriz identidade de tamanho N
λ_k	Ordem espacial do $k-$ésimo termo AR
m_k	Ordem espacial do $k-$ésimo termo MA
μg	Unidade de medida - Microgramas
ϕ_{kl}	Parâmetros autorregressivos nas defasagens temporal k e espacial l
$\Phi(B)$	Polinômio Autorregressivo
$\rho_{lk}(s)$	Função de autocorrelação espaço-temporal
$S(\boldsymbol{\Phi}, \boldsymbol{\Theta})$	Soma dos quadrados dos erros do modelo
θ_{kl}	Parâmetros de média móvel nas defasagens temporal k e espacial l
$\Theta(B)$	Polinômio de Média Móvel
$\boldsymbol{z}(t) = [z_1(t), \ldots, z_N(t)]'$	Vetor $N \times 1$ de observações no tempo $t = 1, \ldots, T$
$\boldsymbol{W}^{(l)}$	Matriz de ponderações $N \times N$ para a ordem espacial l

Resumo

Nos estudos de poluição atmosférica é comum observar dados medidos em diferentes posições no espaço e no tempo, como é o caso da medição de concentrações de poluentes em uma coleção de estações de monitoramento. A dinâmica desse tipo de observações pode ser representada por meio de modelos estatísticos que consideram a dependência entre as observações em cada localização ou região e as observações nas regiões vizinhas, assim como a dependência entre as observações medidas sequencialmente. Nesse contexto, a classe de Modelos Espaço-Temporais Autorregressivos e de Médias Móveis (STARMA) é de grande utilidade, pois permite explicar a incerteza em sistemas que apresentam uma complexa variabilidade nas escalas temporal e espacial. O processo com representação STARMA é uma extensão dos modelos ARMA para séries temporais univariadas, sendo que além de modelar uma série simples através do tempo, considera-se também sua evolução em uma grade espacial.

A aplicação dos modelos STARMA em estudos de poluição atmosférica é ainda pouco explorada. Nessa direção, propomos nesta Tese uma classe de modelos espaço-temporais que considera as características de longa dependência comumente observadas em séries temporais de concentrações de poluentes atmosféricos. Este modelo é aplicado a séries reais provenientes de observações diárias de concentração média de PM_{10} e SO_2 na Região da Grande Vitória, ES, Brasil. Os resultados evidenciaram que a dinâmica de dispersão dos poluentes estudados pode ser bem descrita usando modelos STARMA e STARFIMA, propostos nesta Tese. Essas classes de modelos permitiram estimar a influência dos poluentes sobre os níveis de poluição nas regiões vizinhas. O processo STARFIMA mostrou-se apropriado nas séries sob estudo, pois essas apresentaram características de longa memória no tempo. A consideração dessa propriedade no modelo conduziu a uma melhora significativa do ajuste e das previsões, no tempo e no espaço.

Abstract

In air pollution studies is frequent to observe data measured on time over several spatial locations. This is the case of measures of air pollutant concentrations obtained from monitoring networks. The dynamics of these kind of observations can be represented by statistical models, which consider the dependence between observations at each location or region and their neighbor locations, as well as the dependence between the observations sequentially measured. In this context, the class of the Space-Time Autoregressive Moving Average (STARMA) models is very useful since it explains the underlying uncertainty in systems with a complex variability on time and space scales. The process with STARMA representation is an extension of the univariate ARMA time series. In this case, besides the modeling of the single series on time, their evolution over a spatial grid is also considered.

The application of the STARMA models in air pollution studies is not much explored. This thesis proposes a class of space-time models which consider the long memory dependence usually observed in time series of air pollutant concentrations. This model is applied to real series of daily average concentrations of PM_{10} and SO_2 at Greater Vitória Region, ES, Brazil. The results obtained showed that the dispersion dynamics of the studied pollutants can be well described using the STARMA and STARFIMA models, here proposed. These class of models allowed to estimate the influence of the pollutants on the pollution levels over the neighbor regions. The STARFIMA process showed to be appropriate for the series under study since they have long memory characteristics. Taking into account the long memory properties lead to a significant improvement of the forecasts, both on time and space.

1 Introdução

O controle dos níveis de poluição atmosférica é necessário devido ao fato dos poluentes causarem problemas de saúde, deteriorarem materiais, danificarem a vegetação, entre outros. O tipo de controle pode ser fundamentado na investigação e na análise da dispersão de poluentes, assim como em metodologias de previsão de eventos de poluição que permitam, por exemplo, proporcionar alertas oportunos de saúde pública.

Nos estudos de poluição atmosférica é comum observar dados medidos em diferentes posições no espaço e no tempo, como por exemplo, a medição de concentrações de poluentes em uma coleção de estações de monitoramento ou a contagem de ocorrências de eventos hospitalares associados a problemas respiratórios em uma coleção de regiões geográficas. A dinâmica desse tipo de observações pode ser representada por meio de modelos estatísticos que consideram a dependência entre as observações em cada localização ou região e as observações nas regiões vizinhas, assim como dependência entre as observações medidas sequencialmente.

Nesse contexto, a classe geral dos modelos espaço-temporais é amplamente usada pois permite introduzir explicitamente a incerteza inerente aos dados, produzir previsões acuradas dos eventos de poluição em períodos de tempo futuros e realizar interpolação sobre regiões espaciais de interesse.

Nas últimas décadas, o interesse de pesquisadores pelas diversas metodologias de modelagem espaço-temporal tem aumentado consideravelmente. Essas metodologias têm sido aplicadas em diversas áreas como Ecologia, Epidemiologia, Geofísica, Hidrologia, Ciências Ambientais e em problemas de transporte, de processamento de imagens e de sistemas climáticos, entre outros. Como exemplos de aplicação nessas areas pode-se citar Haas (1995), Carroll et al. (1997), Epperson (2000), Shaddick & Wakefield (2002), Ma (2005) e Fernandez-Cortés et al. (2006), entre outros.

Recentemente, pesquisadores desenvolveram abordagens bayesianas hierárquicas para previsão de eventos de poluição do ar. De-Iaco et al. (2003) usaram dados da concentração média horária de NO_2 e CO (μ/m^3) em 18 estações de monitoramento em Milão. Paez & Gamerman (2003) estudaram a poluição atmosférica no Rio de Janeiro avaliando as concentrações diárias de PM_{10}. Huerta et al. (2004) introduziram um modelo espaço-temporal para concentrações horárias de ozônio na Cidade de México. Sahu & Mardia (2005) apresentaram uma análise de previsão de curto prazo para dados de $PM_{2,5}$ na cidade de Nova York no ano 2002.

No contexto dos modelos clássicos de probabilidade, diversas técnicas de modelagem têm sido desenvolvidas. Em geral, elas são extensões de modelos geoestatísticos que introduzem componentes temporais ou extensões de modelos de séries temporais que incorporam componentes espaciais. Höst et al. (1995) propuseram um modelo geoestatístico com componente temporal nos resíduos. Kyriakidis & Journel (1999) mostraram que esse modelo não consegue prever observações em tempos não amostrados e sugeriram um procedimento alternativo para estimar as componentes do modelo.

A classe de *Modelos Espaço-Temporais Autorregressivos e de Médias Móveis (STARMA)* é uma das classes de modelos espaço-temporais que têm mostrado maior utilidade para explicar

a incerteza em sistemas que apresentam uma complexa variabilidade nas escalas temporal e espacial. O processo com representação STARMA é uma extensão multivariada dos modelos ARMA para séries temporais univariadas (para detalhes sobre o modelo ARMA ver, e.g. Brockwell & Davis 2002), sendo que além de modelar a evolução de uma série simples através do tempo, considera-se a evolução temporal da série em uma grade espacial.

Em análise de séries temporais é fundamental estudar a estrutura de dependência das variáveis, pois o tipo de dependência das observações caracteriza o modelo que gera o processo. Uma classe de modelos que tem sido amplamente utilizada, devido a sua capacidade para captar os diferentes tipos de memórias, é o processo ARFIMA(p, d, q) (Autorregressivo Integrado Fracionário e de Média Móvel), sugerido por Granger & Joyeux (1980*a*) e Hosking (1981). No modelo, o parâmetro d assume valores reais e governa a memória do processo: curta ($d = 0$), intermediária ($d < 0$) e longa ($d > 0$).

Em particular, os modelos ARMA são de memória curta. Hosking (1981) mostrou que as séries que apresentam propriedade de memória longa são caracterizadas por correlações estatisticamente significativas entre observações distantes; equivalentemente, a função de densidade espectral tem singularidade na frequência zero.

A aplicação dos modelos STARMA em estudos de poluição atmosférica é ainda pouco explorada. Glasbey & Allcroft (2008) desenvolveram um modelo Espaço-Temporal Autorregressivo (STAR) para dados de radiação solar e mostraram sua utilidade para descrever outros conjuntos de dados que apresentam características similares ás dos dados de radiacç ao solar. Antunes & Subba Rao (2006) propuseram testes estatísticos para discriminação entre modelos STARMA e Multivariados Autorregressivos. A metodologia proposta foi ilustrada com uma aplicação em dados de concentrações horárias de CO para quatro estações de monitoramento em Londres.

A escassez de literatura sobre os modelos STARMA, relacionada à metodologia para diferentes estruturas de dependência, assim como à abordagem específica em estudos atmosféricos, estimula o interesse para o desenvolvimento desta Tese, tornando-se um tópico desafiador com amplo universo de investigação teórica e empírica.

Nessa direção, o objetivo principal desta Tese é estudar o processo STARMA no contexto de diferentes estruturas de dependência estocástica, com ênfase na longa dependência, isto é, o modelo ARFIMA Espaço-Temporal ou STARFIMA com $d > 0$. O modelo é justificado de forma teórica e empírica e sua aplicação é corroborada pela qualidade no ajuste e na previsão de dados de concentração de SO_2 e PM_{10} da Rede Automática de Monitoramento da Qualidade do Ar (RAMQAr) da Região da Grande Vitória, ES (RGV).

Esta Tese está organizada em forma de artigos. O Artigo 1 (vide p. 23), intitulado "**Daily average sulfur dioxide in Greater Vitória Region: a space-time analysis**", apresenta análise de ajuste e previsão de concentrações diárias de SO_2 medidas na RGV, por meio do modelo STARMA.

O modelo STARFIMA, as suas propriedades teóricas, o procedimento de estimação, os estudos empíricos e a aplicação nas séries do poluente PM_{10} medido na RAMQAr são os motivos de pesquisa do Artigo 2, intitulado "**Modeling and Forecasting PM_{10} concen-**

trations using the Space-Time ARFIMA Model" apresentado na p. 50 desta Tese. O estudo aplicado mostra que as séries de PM_{10} podem ser caracterizadas por processos de memória longa. Como é bem discutido na literatura sobre séries temporais, a flutuação média da série pode ser removida por meio do uso de parâmetros fracionários sem causar problemas de sobre-diferenciação. Em adição, se o processo realmente apresentar carcaterística de memória longa, o uso de modelos usuais ARMA pode levar a previsões pouco acuradas. Essas questões foram observadas na aplicação do modelo STARFIMA na análise espaço-temporal do poluente.

A Tese está dividida da seguinte forma: A Seção 2 apresenta os objetivos que motivaram esta pesquisa. Na Seção 3 apresenta-se uma síntese geral de trabalhos realizados na área da poluição atmosférica usando metodologias de modelos de séries temporais, análise espacial e modelos espaço-temporais.

Conceitos básicos usados na análise de séries temporais e no desenvolvimento desta Tese são abordados na Seção 4. Posteriormente, os resultados desta pesquisa se apresentam no Seção 5 em forma de dois artigos. As contribuições desta pesquisa são discutidas na Seção 6. Finalmente, as conclusões e algumas recomendações para pesquisas futuras são apresentadas nas Seções 7 e 8, respectivamente.

2 Objetivos

2.1 Objetivo Geral

Modelar processos espaço-temporais no contexto de estruturas de dependência estocástica curta e longa. Investigar as propriedades de estimação e identificação de Modelos Espaço-Temporais Autorregressivos e de Médias Móveis (STARMA) com estrutura de longa dependência (modelo STARFIMA) e aplicar o modelo em dados de concentração diária de SO_2 e PM_{10} da Região da Grande Vitória.

2.2 Objetivos Específicos

- Investigar e propor novas metodologias de análise de processos espaço-temporais com estruturas de dependência curta e longa.
- Aplicar a metodologia desenvolvida em dados de concentração diária de SO_2 e PM_{10}, obtidos da rede de monitoramento da qualidade do ar da Região da Grande Vitória, para obter previsões em tempos futuros.
- Implementar a metodologia estudada em software estatístico e disponibilizar para os potenciais usuários da técnica.

3 Revisão Bibliográfica

Uma ampla variedade de modelos estatísticos tem sido proposta para modelagem de fenômenos de poluição do ar, especialmente nas últimas décadas. No contexto dos mode-

los espaço-temporais, Cliff & Ord (1975) foram os primeiros pesquisadores a propor modelos estatísticos que relacionam variáveis no espaço e no tempo. Na mesma direção, Ali (1979) desenvolveu um método para o cálculo da função de verossimilhança dos parâmetros em Modelos Espaço-Temporais Autorregressivos (STAR), e discutiu o problema de previsão.

Pfeifer & Deutsch (1980*d*) extenderam as idéias de Cliff & Ord (1975) e propuseram os modelos Espaço-Temporais Autorregressivos e de Médias Móveis (STARMA), que são uma generalização dos modelos Autorregressivos e de Médias Móveis (ARMA) comumente estudados em séries temporais (ver Box et al. (1994)). Os autores apresentaram um procedimento iterativo para construir modelos STARMA diferenciados, denotados como STARIMA. Adicionalmente, desenvolveram as propriedades teóricas do modelo usando estimação por mínimos quadrados condicionais. Outras propriedades do modelo foram estudadas em Pfeifer & Deutsch (1980*b*), Pfeifer & Deutsch (1980*a*), Pfeifer & Deutsch (1980*c*), Deutsch & Pfeifer (1981), Pfeifer & Deutsch (1981) e Abraham (1983).

Reynolds & Madden (1988), Reynolds et al. (1988) e Madden et al. (1988) aplicaram o modelo STARMA em estudos de dispersão de doenças produzidas por fungos nas plantas de tabaco e de morango em seis campos dos Estados Unidos.

Haslett & Raftery (1989) estimaram a produção potencial de energia eólica a longo prazo na Irlanda usando dados de velocidade e direção do vento em 12 estações meteorológicas distribuídas no território do país. O enfoque dos autores foi orientado à verificação da estrutura de correlação espacial dos dados. Adicionalmente, eles propuseram um método para estimar a força do vento em um ponto não amostrado no espaço.

Epperson (1993) estudou as interações entre processos ecológicos e a estrutura espacial em sistemas de sub-populações com migração. Analisou a correlação de frequências de genes sobre o espaço e o tempo através de modelos STAR. Posteriormente, Epperson (1994) investigou a migração estocástica de populações por meio dos modelos STARMA para determinar correlações no espaço-tempo em sistemas com taxas de migração e número de dimensões espaciais gerais.

Niu & Tiao (1995) desenvolveram uma classe de modelos de regressão espaço-temporal para a análise de dados satelitais em uma latitude fixa e aplicaram os modelos a dados de mapeamento de ozônio total para verificação de tendências. Embora o modelo proposto por Niu & Tiao seja parsimonioso, isto é, com poucos parâmetros estruturais, não admite dependência estrutural devido a que o procedimento de estimação foi planejado especificamente para um processo espacial circular em uma latitude fixa e não aplica para sistemas gerais de lattices.

Dai & Billard (1998) propuseram a classe dos modelos Espaço-Temporais Bilineares (STBL) como uma extensão dos modelos STARMA para o caso de processos espaço-temporais que apresentam certo comportamento não-linear.

Epperson (2000) estudou correlações espaço-temporais para analizar dados ecológicos discretos no tempo e no espaço usando modelos STARMA. O autor defendeu a utilidade dessa classe de modelos nos estudos ecológicos devido à sua capacidade de incorporar características reais dos sistemas populacionais naturais, incluindo diversas formas de migração estocástica. Argumentou também que as correlações espaço-temporais são particularmente importantes

pois elas permitem ligar dados reais com processos teóricos e podem ser usadas para estimar taxas de migração, ajuste de modelos, testes e previsão de comportamento futuro de sistemas reais.

LaValle et al. (2001) utilizaram modelos STAR para identificar o comportamento de dados de praias e zonas costeiras coletados na praia nordeste do Lago Erie, Canada, nos anos 1978 a 1994. Os resultados obtidos pelos autores demostraram a influência dos processos estocásticos localizados nos fluxos de sedimentos na praia e nas variações da linha costeira. O modelo reforçou a hipótese dos pesquisadores sobre a interdependência do fluxo de sedimentos nas praias em lugares adjacentes.

Niu et al. (2003) propuseram uma classe de modelos espaço-temporais sazonais para sistemas gerais de lattices, sendo estes uma extensão do modelo proposto por Niu & Tiao (1995). Estes modelos foram aplicados a campos com altura geopotencial média mensal de 500 mb sobre um lattice de 10×10 cobrindo uma grande porção do hemisfério norte. Segundo os autores, o entendimento da estrutura estatística dos campos de altura geopontencial troposférica e a melhora na precisão das previsões desses campos são fatores muito importantes para previsão do clima no médio (de 6 dias até 2 semanas) e longo (mensal ou sazonal) prazos.

Dai & Billard (2003) consideraram o problema da estimação dos parâmetros do modelo STBL através do procedimento de estimação da máxima verossimilhança condicional. A metodologia proposta foi ilustrada com os dados de velocidade do vento estudados por Haslett & Raftery (1989) e comparada com o ajuste de um modelo STARMA. Os resultados do modelo mostraram que, para este conjunto particular de dados, o modelo STBL apresentou um melhor ajuste.

Giacomini & Granger (2004) compararam a eficiência relativa de diferentes métodos para previsão de variáveis espacialmente correlacionadas. Os resultados dos autores mostraram que as previsões podem ser melhoradas quando o modelo STAR é ajustado. Soni et al. (2004) usaram análise de intervenção em modelos STARMA para estudar dados de magnetoencefalografia fetal (fMEG) e determinar a influência de fatores como movimentos, respiração e outros, nos sinais resultantes.

Allcroft & Glasbey (2005) desenvolveram modelos STARMA para a radiação solar em Edinburgo. Embora esses modelos sejam computacionalmente custosos, os autores mostraram que a dimensão dos cálculos pode ser reduzida trabalhando em um espaço apropriado.

Motivados pela modelagem e previsão da atividade de furacões no Atlântico Norte, Jagger & Niu (2005) introduziram a classe dos modelos Espaço-Temporais Autorregressivos Exponenciais (ESTAR). Eles desenvolveram as propriedades assintóticas do estimador para os parâmetros e provaram a consistência e normalidade assintótica dos estimadores.

Antunes & Subba Rao (2006) propuseram testes estatísticos para discriminação entre modelos STARMA e modelos Multivariados Autorregressivos. A metodologia proposta foi ilustrada com uma aplicação em dados de variação de concentrações horárias de CO para quatro estações de monitoramento em Londres. Giacinto (2006) desenvolveu uma generalização dos modelos STARMA, denominada GSTARMA. Apresentou a metodologia para obtenção dos estimadores do modelo.

Finalmente, Borovkova et al. (2008) estudaram as propriedades assintóticas do Modelo Autorregressivo Espaço-Temporal Generalizado (GSTAR), que é um caso particular dos modelos GSTARMA.

Ao nosso conhecimento, até agora só existem desenvolvimentos teóricos ou empíricos de modelos STARMA com características de memória curta e não foram exploradas ainda as características de memória longa das séries envolvidas em aplicações. A partir desta revisão bibliográfica, pode-se perceber também, que os modelos STARMA têm sido pouco explorados no contexto dos estudos ambientais, especificamente na área da poluição do ar. Esses fatos motivam o interesse desta pesquisa para o desenvolvimento teórico e aplicação da metodologia nessa área da ciência.

4 Conceitos Básicos em Séries Temporais

Nesta seção são introduzidos conceitos básicos utilizados na análise de séries temporais. Em particular, é importante destacar o conceito de *estacionariedade*, no qual se encontram baseadas todas as técnicas de estimação e modelagem de séries temporais no domínio do tempo, através da função de autocovariância, e no domínio da frequência, através da função de densidade espectral. Para detalhes, ver, e.g., Brockwell & Davis (2006) e Priestley (1981)

4.1 Processos estacionários

A seguir são apresentadas as condições de estacionariedade para um processo estocástico linear geral. Adicionalmente, são definidas as funções que caracterizam a dinâmica do processo nos domínios do tempo e da frequência.

Definition 1. (*Processo estocástico*) Seja T um conjunto arbitrário. Um processo estocástico é uma família de variáveis aleatórias $\{y_t\}_{t\in T}$ $(:= \{y_t\})$, definidas no mesmo espaço de probabilidade, indexadas no tempo $t \in T$.

O conjunto T é comumente tomado como um subconjunto dos números inteiros $\mathbb{Z} = \{0, \pm 1, \pm 2, \ldots\}$. Seguindo a definição anterior, uma *série temporal* é uma realização de um certo processo estocástico. Os dois primeiros momentos de $\{y_t\}_{t\in\mathbb{Z}}$ (ou $\{y_t\}$) são definidos como

$$E[y_t] = \mu_t \text{ e } E(y_t - \mu_t)^2 = \sigma_t^2,$$

enquanto que a função de autocovariância do processo $\{y_t\}$ é

$$R_t(h) = Cov(y_t, y_{t+h}) = E[(y_t - \mu_t)(y_{t+h} - \mu_{t+h})] \text{ para } h \in \mathbb{Z},$$

e a função de autocorrelação é dada por

$$\rho_t(h) = \frac{R_t(h)}{\sqrt{\sigma_t^2 \sigma_{t+h}^2}} \text{ para } h \in \mathbb{Z}.$$

Definition 2. (estacionariedade) Um processo estocástico $\{y_t\}$ é dito ser (fracamente) estacionário se e somente se:

1. $E[y_t] = \mu$, para todo $t \in \mathbb{Z}$,
2. $E(y_t - \mu)^2 = \sigma^2$, $0 < \sigma^2 < \infty$, para todo $t \in \mathbb{Z}$,
3. $R(h) = Cov(y_t, y_{t+h})$ depende apenas de h, para todo $t \in \mathbb{Z}$.

As autocorrelações são obtidas normalizando as autocovariâncias através da sua divisão pelo produto dos respectivos desvios padrão, i.e., $\rho(h) = \frac{R(h)}{R(0)}$. O exemplo mais simples de um processo estacionário é o processo de ruído branco (RB), definido como uma sequência de variáveis aleatórias não-correlacionadas com média e variância constantes (sendo a variância estritamente positiva e finita) ao longo do tempo.

Definition 3. (Processo linear geral) $\{y_t\}$ é um processo linear se pode ser representado como

$$y_t = \sum_{j=-\infty}^{\infty} \psi_j \epsilon_{t-j}, \quad t \in \mathbb{Z},$$

onde $\{\epsilon_t\}$ é um RB com média 0 e variância σ_ϵ^2 (denotado por $\{\epsilon_t\} \sim RB(0, \sigma_\epsilon^2)$) e $\{\psi_j\}$ é uma sequência de constantes com $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$.

Definition 4. (Função geratriz de autocovariâncias) Seja $\{y_t\}$ um processo estacionário com função de autocovariâncias $R(h)$. A função geratriz de autocovariâncias de $\{y_t\}$ é definida como

$$g(z) = \sum_{h=-\infty}^{\infty} R(h) z^h,$$

onde z é um escalar complexo.

Em particular, a função de densidade espectral (ou espectro) de $\{y_t\}$ é a função dada por

$$\begin{aligned} f(\lambda) &= \frac{1}{2\pi} g(e^{-i\lambda}) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} e^{-ih\lambda} R(h) \\ &= \frac{1}{2\pi} \left[R(0) + 2 \sum_{h=1}^{\infty} R(h) \cos(\lambda h) \right], \quad \lambda \in [-\pi, \pi], \end{aligned} \tag{1}$$

onde $e^{-i\lambda} = \cos(\lambda) - i \sin(\lambda)$ e $i = \sqrt{-1}$. Neste caso, note que a somabilidade de $|R(\cdot)|$ implica que $f(\lambda)$ converge absolutamente.

Avaliando a Eq. 1 em $\lambda = 0$, o processo $\{y_t\}$ apresenta a propriedade de memória longa se

$$f(0) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} R(h) = \infty,$$

assim $f(\lambda)$ tem uma singularidade na frequência zero. Quando

$$f(0) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} R(h) = 0,$$

o processo apresenta dependencia negativa ou anti-persistencia; e $\{y_t\}$ apresenta propriedade de memória curta se $0 < f(0) < \infty$, como o caso dos processos ARMA definidos na Seção 4.2.1.

4.1.1 Estimação da média, autocovariâncias e espectro de um processo estacionário

Sejam $y_1, y_2, \ldots, y_n$ observações de um processo $\{y_t\}$ estacionário. Estimadores para $E[y_t] = \mu$ e $E(y_t - \mu)^2 = \sigma_Y^2$ são dados por $\bar{y} = \frac{1}{n}\sum_{t=1}^{n} y_t$ e $\widehat{R}(0) = \frac{1}{n}\sum_{t=1}^{n}(y_t - \bar{y})^2$, respectivamente. Um estimador da função de autocovariâncias é dado por

$$\widehat{R}(h) = \frac{1}{n}\sum_{t=1}^{n-h}(y_t - \bar{y})(y_{t+h} - \bar{y}), \qquad h = 0, \pm 1, \pm 2, \ldots, \pm(n-1),$$

e um estimador natural para $\rho(h)$ é $\widehat{\rho}(h) = \frac{\widehat{R}(h)}{\widehat{R}(0)}$.

No domínio da frequência, um estimador assintoticamente não-viesado para a função de densidade espectral $f(\lambda)$ é o *periodograma*, dado por $I(\lambda) = |w(\lambda)|^2$, onde $w(\lambda) = \frac{1}{\sqrt{2\pi n}}\sum_{t=1}^{n} y_t e^{i\lambda t}$. A função $w(\cdot)$ é chamada de transformada discreta de Fourier (TDF).

Uma outra representação do periodograma, em função do estimador da autocovariância, pode ser escrita como

$$I(\lambda) = \frac{1}{2\pi}\left[\widehat{R}(0) + 2\sum_{h=1}^{n-1}\widehat{R}(h)\cos(\lambda h)\right]. \tag{2}$$

Um estimador consistente para o espectro de um processo estacionário é o *periodograma suavizado* dado por

$$I_s(\lambda) = \frac{1}{2\pi}\sum_{h=-(n-1)}^{n-1}\kappa(h)\widehat{R}(h)\cos(\lambda h), \quad \lambda \in [-\pi, \pi], \tag{3}$$

onde $\kappa(\cdot)$ é uma função contínua e par. Na literatura, essa função é conhecida como "janela" e é útil para reduzir a contribuição de covariâncias provenientes de defasagens (h) elevadas. A "janela" mais simples é a chamada *janela periodograma truncado*:

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

onde M $(< n-1)$ é o parâmetro de truncamento. Existem outras propostas para a função $\kappa(\cdot)$ considerando diferentes ponderações; para detalhes ver Priestley (1981, p. 437).

4.2 Modelos de séries temporais

O estudo das séries temporais pode ser motivado pelo interesse em investigar o mecanismo gerador de um conjunto de dados observados ao longo do tempo, para descrever sua dinâmica com o objetivo de gerar previsões acerca do seu comportamento futuro. Para tanto, são construídos modelos probabilísticos que pertencem a um domínio temporal previamente estabelecido. Tais modelos devem respeitar o princípio da parcimônia, ou seja, devem envolver o menor número possível de parâmetros.

A seguir, são descritos de forma geral alguns desses modelos e algumas de suas propriedades são apresentadas.

4.2.1 Processos autorregressivos e de médias móveis ARMA(p, q)

Seja $\{y_t\}$ um processo que satisfaz a equação de diferenças dada por

$$\Phi(B)y_t = \Theta(B)\epsilon_t, \tag{4}$$

onde $\{\epsilon_t\}$ é ruído branco, i.e., $\{\epsilon_t\} \sim RB(0, \sigma_\epsilon^2)$, B é o operador de defasagem definido como $B^k X_t = X_{t-k}$, $k = 1, \ldots, p$, $\Phi(z) = 1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p$ e $\Theta(z) = 1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_q z^q$. O processo $\{y_t\}$ definido na Eq. 4 é chamado de processo autorregressivo e de médias móveis, ARMA(p, q).

Definition 5. (Invertibilidade) Um processo $\{y_t\}$ com representação ARMA(p, q) é *invertível* se existem constantes $\{\pi_j\}$ tais que $\sum_{j=0}^{\infty} |\pi_j| < \infty$ e $\epsilon_t = \sum_{j=0}^{\infty} \pi_j y_{t-j}$, para todo $t \in \mathbb{Z}$.

Seguindo as Definições 2 e 5, o processo representado na Eq. 4 é estacionário e invertível se as raízes de $\Phi(z) = 0$ e $\Theta(z) = 0$ são não comuns e encontram-se fora do círculo unitário.

Definition 6. (Causalidade) Um processo $\{y_t\}$ com representação ARMA(p, q) é *causal*, ou função causal de $\{\epsilon_t\}$, se existem constantes $\{\psi_j\}$ tais que $\sum_{j=0}^{\infty} |\psi_j| < \infty$ e $y_t = \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j}$, para todo $t \in \mathbb{Z}$.

Note que as propriedades de invertibilidade e causalidade não são apenas do processo $\{y_t\}$, mas também da relação entre os processos $\{y_t\}$ e $\{\epsilon_t\}$ da definição da equação ARMA apresentada na Eq. 4. Invertibilidade e causalidade garantem que há uma solução única estacionária para a equação ARMA, quase certamente.

4.2.2 Função de autocovariâncias e espectro de um processo ARMA(p, q)

O cálculo da função de autocovariâncias para um processo $\{y_t\}$ com representação ARMA(p, q) causal é realizado através das equações

$$R(k) - \phi_1 R(k-1) - \cdots - \phi_p R(k-p) = \sigma_\epsilon^2 \sum_{j=0}^{\infty} \theta_{k+j}\psi_j, \quad 0 \neq k < m,$$

$$R(k) - \phi_1 R(k-1) - \cdots - \phi_p R(k-p) = 0, \quad k \geq m,$$

onde $m = \max(p, q+1)$, $\psi_j - \sum_{k=1}^{p} \phi_k \psi_{j-k} = \theta_j$, $j = 0, 1, 2, \ldots$. $\psi_j = 0$ para $j < 0$, $\theta_0 = 1$ e $\theta_j = 0$ para $j \notin \{0, 1, \ldots, q\}$; ver, e.g., Brockwell & Davis (2002, p. 88).

O espectro de $\{y_t\}$ é dado por

$$f_{ARMA}(\lambda) = \frac{\sigma_\epsilon^2}{2\pi} \frac{|\Theta(e^{-i\lambda})|^2}{|\Phi(e^{-i\lambda})|^2}, \quad \lambda \in [-\pi, \pi]. \tag{5}$$

4.2.3 Processos ARIMA(p, d, q) fracionários (ARFIMA(p, d, q))

No início da década de 80, Granger & Joyeux (1980*b*) e Hosking (1981) propuseram os modelos ARFIMA, utilizados na modelagem de séries que possuem memória longa ou longa dependência. A propriedade de memória longa ocorre em séries que apresentam correlações estatisticamente significativas mesmo para observações distantes; equivalentemente, o espectro apresenta singularidade para frequências próximas de 0.

Em particular, se o parâmetro de integração assume apenas valores inteiros positivos, i.e. $d \in \mathbb{Z}^+$, o modelo é conhecido como ARIMA(p, d, q). De maneira formal, o processo ARFIMA(p, d, q) é definido como a seguir:

Definition 7. Seja $d \in \mathbb{R}$. $\{y_t\}$ segue um processo ARFIMA(p, d, q) se satisfaz a equação em diferenças da forma

$$\Phi(B) y_t = \Theta(B)(1 - B)^{-d} \epsilon_t, \tag{6}$$

com $\Phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$ e $\Theta(z) = 1 - \theta_1 z - \cdots - \theta_p z^p$, $\{\epsilon_t\}$ sendo um processo ruído branco com média 0 e variância σ_ϵ^2. O filtro de diferenciação fracionária $(1 - B)^{-d}$ é definido pela expansão binomial

$$(1 - B)^{-d} = \sum_{j=0}^{\infty} \pi_j B^j,$$

onde $\pi_j = \frac{\Gamma(j+d)}{\Gamma(j+1)\Gamma(d)}$, $j = 0, 1, 2, \ldots$, e $\Gamma(\cdot)$ é a função gama dada por $\Gamma(x) = \int_0^\infty s^{x-1} e^{-s}\, ds$ se $x > 0$. Se $x < 0$ e não-inteiro, $\Gamma(\cdot)$ é definida em termos da fórmula $x\Gamma(x) = \Gamma(x + 1)$ para qualquer valor de x.

Quando $d \in (-0.5, 0.5)$ e as raízes dos polinômios $\Phi(z) = 0$ e $\Theta(z) = 0$ são não-comuns e estão fora do círculo unitário, o processo definido em (6) é estacionário e invertível e com função de densidade espectral dada por

$$f(\lambda) = \frac{\sigma^2}{2\pi} \left|1 - e^{i\lambda}\right|^{-2d} \left|\frac{\Theta(e^{i\lambda})}{\Phi(e^{i\lambda})}\right|^2, \quad \lambda \in [-\pi, \pi], \tag{7}$$

Nota 1. Observe-se que a Eq. 7 é da forma

$$f(\lambda) \sim G\lambda^{-2d}, \quad \text{quando } \lambda \to 0^+, \tag{8}$$

onde "$\sim$" significa que o quociente entre o lado esquerdo e o lado direito tende a 1. O valor

G é tal que $0 < G < \infty$ para todo λ e $-\frac{1}{2} < d < \frac{1}{2}$, porque para $d \geq \frac{1}{2}$ a função $f(\cdot)$ não é integrável. Para $d > 0$, o processo $\{y_t\}$ apresenta a propriedade de memória longa (e.g. Hosking (1981)).

4.3 Métodos de estimação do parâmetro de diferenciação fracionária

Existem vários estimadores do parâmetro de diferenciação fracionária d propostos na literatura que podem ser classificados em paramétricos e semi-paramétricos. Os primeiros envolvem a estimação simultânea dos parâmetros do modelo, em geral utilizando o método de máxima verossimilhança; ver, e.g., Fox & Taqqu (1986), entre outros. Nos procedimentos semi-paramétricos, a estimação dos parâmetros do modelo é realizada em dois passos: primeiro estima-se o parâmetro de memória longa d e, posteriormente, estimam-se os parâmetros autorregressivos e de médias móveis. O estimador mais popular dentro dessa classe é o estimador proposto por Geweke & Porter-Hudak (1983); variantes foram desenvolvidas por Chen et al. (1994), Reisen (1994), Robinson (1995*a*,*b*), entre outros.

4.3.1 Estimador Log-periodograma (LP)

Seja $f(\lambda_j)$ a função definida na Eq. 7 para $\lambda_j = \frac{2\pi j}{n}$, $j = 0, 1, \ldots, \lfloor \frac{n}{2} \rfloor$, onde n é o tamanho amostral e $\lfloor \cdot \rfloor$ denota a função parte inteira. Sejam $f(\lambda_j) := f_j$ e $f_0(\lambda_j) := f_{0j}$.

Supondo que a função f_j pode ser representada por $f_j = f_{0j} \left| 2 \sin\left(\frac{\lambda_j}{2}\right) \right|^{-2d}$, o logaritmo de f_j pode ser escrito como:

$$\ln f_j = \ln f_0(0) - d \ln \left\{ 2 \sin\left(\frac{\lambda_j}{2}\right) \right\}^2 + \ln \frac{f_{0j}}{f_0(0)}. \tag{9}$$

Adicionando $\ln I_j = \ln \frac{I_j}{f_j} + \ln f_j$ na Eq. 9, obtem-se a equação:

$$\ln I_j = \ln f_0(0) - d \ln \left\{ 2 \sin\left(\frac{\lambda_j}{2}\right) \right\}^2 + \ln \frac{I_j \left\{ 2 \sin\left(\frac{\lambda_j}{2}\right) \right\}^{2d}}{f_0(0)}, \tag{10}$$

que sugere a equação de regressão dada por

$$\ln I_j = \beta_0 + \beta_1 \ln \left\{ 2 \sin\left(\frac{\lambda_j}{2}\right) \right\}^2 + e_j, \quad j = 1, 2, \ldots, g(n),$$

onde $\beta_0 = \ln f_0(0)$ e $\beta_1 = -d$. Note que, para frequências próximas de zero e assumindo $g(n) = o(n)$, então

$$f_j \sim f_{u0} \left\{ 2 \sin\left(\frac{\lambda_j}{2}\right) \right\}^{-2d},$$

assim, $e_j \sim \ln \frac{I_j}{f_j}$, para $j = 1, 2, \ldots, g(n)$.

Geweke & Porter-Hudak (1983) sugerem um estimador semiparamétrico para d, dado por

$$d_{LP} = -\frac{\sum_{i=1}^{g(n)}(\upsilon_i - \bar{\upsilon})\ln I_i}{\sum_{i=1}^{g(n)}(\upsilon_i - \bar{\upsilon})^2}, \tag{11}$$

onde $\upsilon_j = \ln\left\{2\sin\left(\frac{\lambda_j}{2}\right)\right\}^2$, $\bar{\upsilon} = \frac{1}{g(n)}\sum \upsilon_j$ e $g(n)$ é chamado de *bandwidth* e corresponde ao número de frequências utilizadas na regressão.

Nota 2. Hurvich et al. (1998), sob algumas condições de regularidade, calculam um valor ótimo do *bandwidth* tal que $g(n) = O(n^{4/5})$.

As propriedades assintóticas do estimador LP foram derivadas por Robinson (1995*b*) e Hurvich et al. (1998), para o caso estacionário. No contexto não-estacionário, Velasco (1999*b*) estende os resultados obtidos por Robinson (1995*b*) e mostra a consistência do estimador LP para $d \in (0.5, 1]$. Kim & Phillips (2006) mostram que para valores $d > 1$ o estimador LP converge em probabilidade para 1. Phillips (1999) prova a normalidade assintótica do estimador para $d \in (0.5, 1)$, i.e.

$$\sqrt{g(n)}(d_{LP} - d) \xrightarrow{\mathcal{D}} N\left(0, \frac{\pi^2}{24}\right),$$

onde $\xrightarrow{\mathcal{D}}$ denota convergência em distribuição. No caso da presença de raíz unitária, Phillips (2007) mostra que o estimador LP assintoticamente apresenta distribuição normal mista com $\text{var}(d_{LP}) = 0.3948$, a qual resulta menor que $\frac{\pi^2}{24} = 0.4112$.

4.3.2 Estimador Whittle local (WL)

Seja $\{y_t\}$ um processo estacionário com espectro que satifaz a Eq. 8. Defina-se a função objetivo $\mathcal{Q}(G, d_0)$ dada por

$$\mathcal{Q}(G, d_0) = \frac{1}{g(n)}\sum_{j=1}^{g(n)}\left\{\ln G\lambda_j^{-2d_0} + \frac{\lambda_j^{2d_0}}{G}I_j\right\}, \tag{12}$$

onde $g(n)$ é um valor inteiro tal que $g(n) < \frac{n}{2}$ e $\frac{1}{g(n)} + \frac{g(n)}{n} \to 0$ quando $n \to \infty$. A estimativa para d resulta do valor $(\widehat{G}, d_{WL})$ que minimiza a Eq. 12, i.e.

$$(\widehat{G}, d_{WL}) = \arg\min \mathcal{Q}(G, d_0).$$

Substituindo G pela sua estimativa $\widehat{G} = \frac{1}{g(n)}\sum_{j=1}^{g(n)}\frac{I_j}{\lambda_j^{-2d_0}}$, obtem-se

$$\mathcal{R}(d_0) := \mathcal{Q}(\widehat{G}, d_0) - 1 = \ln\widehat{G} - 2d_0\frac{1}{g(n)}\sum_{j=1}^{g(n)}\lambda_j.$$

Robinson (1995a) mostra que o valor de d_0 que minimiza $\mathcal{R}(d_0)$, i.e.

$$d_{WL} = \arg\min \mathcal{R}(d_0),$$

é consistente e

$$\sqrt{g(n)}(d_{WL} - d) \xrightarrow{\mathcal{D}} N\left(0, \frac{1}{4}\right),$$

Nota 3. O cálculo das estimativas através do estimador WL requer o uso de métodos de aproximação numérica, mas como mostrado por Robinson (1995a), o estimador WL resulta estatisticamente mais eficiente que o estimador LP.

As propriedades assintóticas do estimador WL, para o caso não-estacionário, foram desenvolvidas por Velasco (1999a) e Phillips & Shimotsu (2004). Os autores mostram que o estimador WL é consistente para $d \in (\frac{1}{2}, 1]$ e assintoticamente normal para $d \in (\frac{1}{2}, \frac{3}{4})$. Para $d = 1$, o estimador apresenta distribuição normal mista com variância $var(d_{WL}) = 0.2028$, menor que no caso $d < 1$. Da mesma forma que o estimador LP, o WL resulta inconsistente para valores $d > 1$.

Variantes do estimador WL, considerando valores $d > 1$, foram propostas por Shimotsu & Phillips (2005) e Abadir et al. (2007). Os autores sugerem uma modificação do periodograma através de um termo de correção na TDF do processo.

Originally submitted to *Atmospheric Environment*, 2013

Daily average sulfur dioxide in Greater Vitória Region: a space-time analysis

Nátaly A. Jiménez Monroy[1,2*] *Valdério A. Reisen*[1,2] *and Tata Subba Rao*[3,4]

[1]Programa de Pós-Graduação em Engenharia Ambiental - UFES, Vitória, ES.

[2]Departamento de Estatística, UFES, Vitória, ES.

[3]School of Mathematics, University of Manchester, UK.

[4]CRRAO AIMSCS, University of Hyderabad Campus, India.

Abstract

This study explores the class of Space-Time Autoregressive Moving Average (STARMA) models in order to describe and identify the behavior of SO_2 daily average concentrations observed in the Greater Vitória Region (GVR), Brazil. These models are particularly useful in modeling atmospheric pollution data owing to the complex pollutant dispersion dynamics at temporal and spatial scales.

The data were obtained at the air quality monitoring network of GVR, recorded from January 2005 to December 2009. Our findings indicate that SO_2 daily averages tended to be higher than the guidelines suggested by the World Health Organization (daily average of 20 $\mu g/m^3$), for almost all the analyzed sites. The time series obtained for each monitoring station show high variability, mostly caused by some atypical values observed during the period. The main fluctuations in the data are caused by cyclical components, which change from one to another station. On the whole, the cycles are not only weekly (as expected, due to the daily measurements) but also monthly and seasonal.

Resampling bootstrap techniques were used in order to handle the lack of the distributional assumptions made for fitting the model. The obtained bootstrap prediction intervals showed to be much larger than the intervals obtained under the Gaussian distribution assumption.

The fitted STARMA model indicated that the influence time of SO_2 in GVR atmosphere is around 3-4 days. During the period observed, the pollutants released in a site disperse over a large expanse of the region, influencing SO_2 concentrations observed in the vicinity. The quality of the adjusted model suggests that the model is able to predict in-sample values, as well as to forecast average concentrations for one day in advance with good reliability.

Keywords: Air pollution, bootstrap, forecasting, STARMA models.

1 Introduction

The GVR is located on the Brazilian South Atlantic coast in the state of Espírito Santo (ES) and comprises seven main cities, including the capital, Vitória. Its population has grown

*Email: nataly.monroy@ufes.br

significantly in the last four decades as a consequence of rapid industrialization. The increase of the industrial activities, as well as the constant growth of traffic (almost 50% increases from 2001 to 2011), has caused a large impact on the atmospheric quality in the area.

Particularly, sulfur dioxide (SO_2) is considered to be the major indicator of the industrial activities in the area, where the mining and iron, as well as the steel industries, contribute with almost 76% of SO_2 released to the atmosphere (Instituto Estadual de Meio Ambiente e Recursos Hídricos [IEMA] 2011). An overall view of the air quality parameters in GVR shows that SO_2 levels do not exceed the standard levels established by the Brazilian law and there have not been any reported air pollution alerts due to this pollutant. However, according to the Instituto Brasileiro de Geografia e Estatística [IBGE] (2012), in 2010, Vitória was the city with the highest annual SO_2 average in Brazil.

Sulfur dioxide is the main precursor of acid rain and sulfuric acid smog pollution. At the same time, it can be oxidized in the atmosphere to form sulfate aerosol, which is an important component of fine particles suspended in the urban atmosphere. Its reaction with other major atmospheric pollutants can also affect the atmospheric concentrations of these pollutants. Therefore, SO_2 is a significant contributor to the quality of the environment (Yang et al. 2009).

In view of this pollution problem, it is important to develop statistical models for diagnosis and short-term prediction in order to provide accurate early warnings for the air quality control. As pointed out by McCollister & Wilson (1975), there is also the possibility that foreknowledge of high pollution potential could be used to reduce future atmospheric pollutant concentrations through timely reduction of emissions by traffic control or industrial shut-down.

Several statistical modeling approaches have been proposed to describe trends and forecasting SO_2 levels (Brunelli et al. (2007), Brunelli et al. (2008), Castro et al. (2003), Chelani et al. (2002), Lalas et al. (1982), Nunnari et al. (2004), Perez (2001), Roca Pardiñas et al. (2004), Tecer (2007), among others). The most used forecasting statistical models for SO_2 are based on univariate time series approaches. For example, Cheng & Lam (2000), Hassanzadeh et al. (2009), Kumar & Goyal (2011), Lalas et al. (1982), McCollister & Wilson (1975), Schlink et al. (1997). As explained by Turalioglu & Bayraktar (2005), such models are incapable of providing regional information on the spatial variations of air pollutants.

Some other researchers have modeled the spatial scale and used data reduction methods like principal component analysis to summarize the regional variation of SO_2 (Ashbaugh et al. (1984), Beelen et al. (2009), Ibarra Berástegui et al. (2009), de Kluizenaar et al. (2001), Kurt & Oktay (2010), Zou et al. (2009)). However, many of these spatial approaches do not account for the serial autocorrelation latent in data measured over time.

Considering that the data used in the majority of the air pollution studies are obtained from air quality monitoring networks, where the concentrations are observed over various spatial locations along time, it is reasonable to model time and space scales simultaneously aiming to capture explicitly the inherent uncertainty of the air pollution type data. Particularly, for SO_2 studies see Fan et al. (2010), Rouhani et al. (1992), Turalioglu & Bayraktar (2005), Yu & Chang (2006) and Zeri et al. (2011) among others.

In this context, the class of the space-time models is quite effective, allowing the practician to obtain accurate forecasts of the pollution events and to interpolate the spatial regions of interest. One of the most useful approaches of this kind of models, yet less explored in air pollution studies, is the class of STARMA models. This approach is an extension of the classic univariate ARMA time series models into the spatial domain, where the observations at each location at a fixed time are modeled as a weighted combination of past observations at different locations.

Our aim here is to explore the class of STARMA models as an alternative methodology to describe the dynamics of sulfur dioxide dispersion and to obtain short-term forecasts of SO_2 daily average in GVR.

This paper is outlined as follows: Section 2 presents the main characteristics of the region under the study as well as the description of the analyzed data. The three-stage procedure for STARMA modeling is also introduced in this section. Section 3 describes the data processing and the results obtained for the fitted STARMA model. Section 4 closes with a brief summary of the results obtained from the application of the model.

2 Data and methodology

2.1 Study area

The GVR is located in the Brazilian South Atlantic coast (latitude 20°19S, longitude 40°20W). The climate is tropical humid with average temperatures ranging from 23°C to 30°C. The rainfall occurs mainly from October to January, with annual precipitation volume higher to 1400 mm.

Its topography varies from plains to mountain range interspersed with small and medium size rocky massif, which favors the flowing of the humid winds from the sea (Instituto Jones dos Santos Neves [IJSN] 2012). Therefore, the dispersion of the pollutants is also favored over a large area of the region. Its main atmospherical flowing systems are the South Atlantic subtropical anticyclone, which causes the predominant eastern and northeastern winds, and the moving polar anticyclone, responsible for the cold fronts from the southern region of the continent, characterized by low temperatures, mist and strong winds (Instituto Estadual de Meio Ambiente e Recursos Hídricos [IEMA] 2007).

The region is constituted by seven main cities: Vitória (capital city of ES), Serra, Vila Velha, Cariacica, Viana, Guarapari and Fundão. These cities take almost half of total population of Espírito Santo State (48%) and 57% of the urban population in the State (Instituto Brasileiro de Geografia e Estatística [IBGE] 2012). According to the IJSN, the region occupies only 5% of ES territory, however its population density is nine times higher to the overall mean of State. Besides, it produces 58% of the wealth and consumes 55% of the total electric power produced in the State.

The GVR has two of the major seaports in Brazil: Vitória Port (located in downtown) and Tubarão Port (located at the North region of Vitória). The main industrial activities of GVR are related to iron and steel industry, stone quarry, cement and food industries, among

others. These activities represent nearly 55% to 65% of the total potentially pollutant fonts in the State (IEMA, 2011).

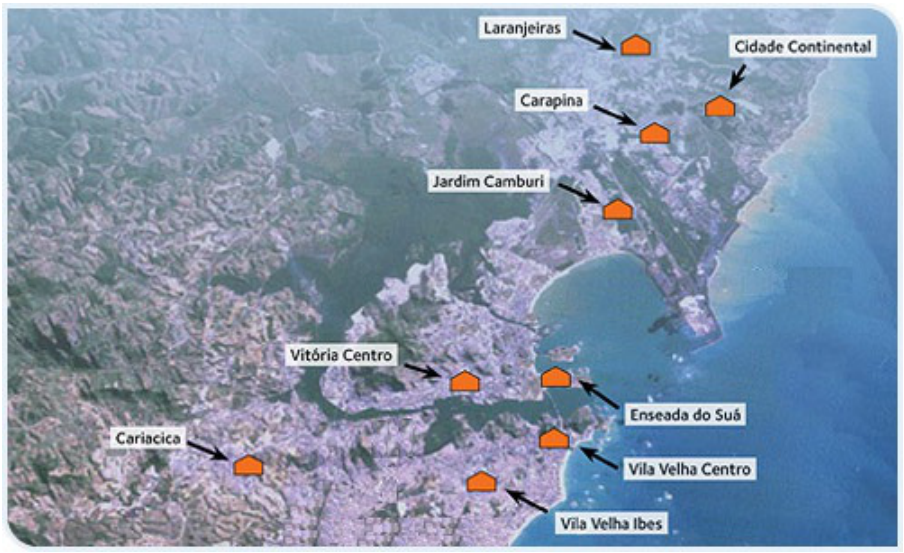


Figure 1: Map of the AAQMN monitoring stations in Greater Vitória Region.

In view of the increasing deterioration of the air quality, the IEMA installed the Automatic Air Quality Monitoring Network (AAQMN) of GVR in 2000. Currently, the network is composed of nine monitoring stations (the last one started operations in September 2012), all of them located in strategic urban areas (see Figure 1). The network measures continuously some meteorological variables as well as the concentration of the pollutants: particular matter, fine particles $< 10\mu$m (PM_{10}), sulfur dioxide (SO_2), carbon monoxide (CO), nitrogen oxides (NO_x), ozone (O_3) and hydrocarbons (HC).

2.2 Data

We analyzed daily average SO_2 concentration ($\mu g/m^3$) data from January 1 2005 to December 31 2009, obtained from seven AAQMN monitoring stations. The main sources of pollutants of each monitoring station are summarized in Table 1. Aiming to ensure the reliability of our study, the monitoring stations having more than 30% missing values for the full analyzed period were discarded. Except for Jardim Camburi station (36% missing values), all the stations met the criterion for inclusion in the study.

The missing values were filled using the Gibbs sampling for multiple imputations of the incomplete multivariate data suggested by Aerts et al. (2002). This algorithm imputes an incomplete column (in our case, each column corresponds to a monitoring station) by generating plausible synthetic values given the other columns in the data. Each incomplete column must act as a target column, and has its own specific set of predictors. The default set of predictors for a given target consists of all other columns in the data set. All these computations were made using the language and environment for statistical computing `R 2.15.2` (R Core Team 2012).

Table 1: Description of the AAQMN monitoring stations in GVR.

Monitoring station	Main pollution sources	Longitude	Latitude
Laranjeiras	Industrial and traffic	40°15'24.74"W	20°11'26.88"S
Jardim Camburi	Industrial and traffic	40°16'06.49"W	20°15'15.03"S
Enseada do Suá	Port of Tubarão and traffic	40°17'26.92"W	20°18'43.29"S
Vitória Centro	Traffic, seaports, Industrial	40°20'13.87"W	20°19'09.42"S
Ibes	Traffic and industrial	40°19'04.38"W	20°20'53.47"S
Vila Velha Centro	Traffic and industrial	40°17'37.77"W	20°20'04.81"S
Cariacica	Traffic and industrial	40°24'01.59"W	20°20'29.92"S

Source: IEMA

Once the database was filled, we calculated the 24-hour average concentrations. Therefore, the analyzed database contains 1826 observations for the six monitoring stations (sites) considered here. The first 1811 observations were used for modeling purposes and the last 15, corresponding to the last two weeks of the full period, were used for forecasting purposes.

2.3 The STARMA Model

Spatial time series can be viewed as time series collected simultaneously in a number of fixed sites with fixed distances between them. As pointed out by Subba Rao & Antunes (2003), the space-time models are used to explain the dependence along time in situations that present systematic dependence between observations in several sites.

The class of STARMA models was developed by Pfeifer & Deutsch (1980*b*). The processes which can be represented by STARMA models are characterized by a random variable $Z_i(t)$, observed at N fixed spatial locations ($i = 1, 2, \ldots, N$) on T time periods ($t = 1, 2, \ldots, T$). The N spatial locations can represent several situations, like states of a country or regions with monitoring stations inside a city, for example.

The dependence between the N time series is incorporated into the model through hierarchical weighting $N \times N$ matrices, specified before the data analysis. These matrices must include the relevant physical characteristics of the system into the model, as for example, the distance between the center of several cities or the distance between monitoring stations from a monitoring network (Kamarianakis & Prastacos 2005).

As in the case of univariate time series, observations $z_i(t)$ from the process $\{Z_i(t)\}$, are expressed in terms of a linear combination of previous observations and errors at the site $i = 1, 2, \ldots, N$. In this case, due to the spatial dependence of the system, the model must incorporate also past observations and errors from the neighboring spatial orders. In this paper, the first order neighbors are those sites which are closer to the location of interest, the second order neighbors are those more distant than the first ones, even less distant than the third order neighbors, and so on.

The STARMA model, denoted by STARMA($p_{\lambda_1,\lambda_2,\ldots,\lambda_p}, q_{m_1,m_2,\ldots,m_q}$), can be represented by the matrix equation:

$$\mathbf{z}(t) = -\sum_{k=1}^{p}\sum_{l=0}^{\lambda_k}\phi_{kl}\mathbf{W}^{(l)}\mathbf{z}(t-k) + \sum_{k=1}^{q}\sum_{l=0}^{m_k}\theta_{kl}\mathbf{W}^{(l)}\boldsymbol{\varepsilon}(t-k) + \boldsymbol{\varepsilon}(t), \tag{1}$$

where $\mathbf{z}(t) = [z_1(t), \ldots, z_N(t)]'$ is a $N \times 1$ vector of observations at time $t = 1, \ldots, T$, p represents the autoregressive order (AR), q represents the moving average order (MA), λ_k is the spatial order of the $k-$th AR term, m_k is the spatial order of the $k-$th MA term, ϕ_{kl} and θ_{kl} are the parameters at temporal lag k and spatial lag l, $\mathbf{W}^{(l)}$ is the $N \times N$ weighting matrix for the spatial order $l > 0$, with diagonal entries 0 and off-diagonal entries related to the distances between the sites. If $l = 0$, then $\mathbf{W}^{(0)} = \mathbf{I}_N$. Each row of $\mathbf{W}^{(l)}$ must add up to 1. It is assumed that $\boldsymbol{\varepsilon}(t) = [\epsilon_1(t), \ldots, \epsilon_N(t)]'$, the random error vector at time t, is a weakly stationary Gaussian process, with

$$\begin{aligned} E[\boldsymbol{\varepsilon}(t)] &= \mathbf{0}, \\ E[\boldsymbol{\varepsilon}(t)\boldsymbol{\varepsilon}'(t+s)] &= \begin{cases} \mathbf{G}, & \text{if } s = 0 \\ \mathbf{0}, & \text{otherwise} \end{cases}, \\ E[\mathbf{z}(t)\boldsymbol{\varepsilon}'(t+s)] &= 0, \quad \text{for } s > 0, \end{aligned} \tag{2}$$

where $E(\cdot)$ is the expected value of the variable.

There are two subclasses of the model in Equation 1: STAR($p_{\lambda_1,\lambda_2,\ldots,\lambda_p}$) when $q = 0$ and STMA($q_{m_1,m_2,\ldots,m_q}$) when $p = 0$. The *stationarity* condition is based on:

$$\det\left(\mathbf{I_N} + \sum_{k=1}^{p}\sum_{l=0}^{\lambda_k}\phi_{kl}\mathbf{W}^{(l)}x^k\right) \neq 0,$$

for $|x| \leq 1$. This condition determines the region of ϕ_{kl} values for which the process is weakly stationary.

As explained by Deutsch & Pfeifer (1981), the proper approach to estimation is highly dependent upon the nature of the variance-covariance matrix of the errors. If $\mathbf{G}$ is assumed to be diagonal, the model estimation should proceed using weighted least squares method. In particular, when the processes for all the N sites have the same variance ($\mathbf{G} = \sigma^2\mathbf{I_N}$, where $\mathbf{I_N}$ is the $N \times N$ identity matrix), the estimation technique reduces to ordinary least squares.

Lastly, when $\mathbf{G}$ is not diagonal, estimation should be performed using generalized least squares. The authors develop procedures for testing hypotheses about $\mathbf{G}$ and provide tables of the critical values for the proposed tests.

The covariance between the l and k order neighbors at the time lag s is defined as *space-*

time covariance function (STCOV). Let $E[Z(t)] = 0$, the STCOV can be expressed as

$$\gamma_{lk}(s) = E\left\{\frac{[\mathbf{W}^{(l)}\mathbf{z}(t)]'[\mathbf{W}^{(k)}\mathbf{z}(t+s)]}{N}\right\} = tr\left\{\frac{\mathbf{W}^{(k)'}\mathbf{W}^{(l)}\mathbf{\Gamma}(s)}{N}\right\}, \tag{3}$$

where $tr[\mathbf{A}]$ is the trace of the square matrix $\mathbf{A}$ and $\mathbf{\Gamma}(s) = E[\mathbf{z}(t)\mathbf{z}(t+s)']$. More details, see for example Pfeifer & Deutsch (1980*b*) and Subba Rao & Antunes (2003).

2.3.1 Model identification

The identification of the STARMA model is carried out by using the *space-time autocorrelation function* (STACF). The STACF between the l and k order neighbors, at the time lag s, is defined as

$$\rho_{lk}(s) = \frac{\gamma_{lk}(s)}{[\gamma_{ll}(0)\gamma_{kk}(0)]^{1/2}}.$$

Given the vector $\mathbf{z}(t) = [z_1(t), \ldots, z_N(t)]'$ of observations at time $t = 1, \ldots, T$, the estimator of $\mathbf{\Gamma}(s)$ is given by

$$\hat{\mathbf{\Gamma}}(s) = \sum_{l=1}^{T-s} \frac{\mathbf{z}(t)\mathbf{z}(t+s)'}{T-s}, \quad s \geq 0.$$

$\hat{\mathbf{\Gamma}}(s)$ can be substituted in Equation 3 in order to obtain the sample estimates $\hat{\gamma}_{lk}$ of the STCOV. Therefore, the sample estimator of the STACF is

$$\hat{\rho}_{lk}(s) = \frac{\hat{\gamma}_{lk}(s)}{[\hat{\gamma}_{ll}(0)\hat{\gamma}_{kk}(0)]^{1/2}}. \tag{4}$$

Pfeifer & Deutsch (1980*b*) demonstrated that identification can usually proceed strictly on the basis of $\hat{\rho}_{l0}$ for $l = 1, \ldots, \lambda$.

Each particular model of the STARMA family has a unique space-time autocorrelation function (see Table 2). However, if the model is autoregressive but with unknown order, is not easy to determine its correct order using $\hat{\rho}_{lk}(s)$. This difficulty can be handled using the *space-time partial autocorrelation function* (STPACF), which can be expressed as

$$\rho_{h0} = \sum_{j=1}^{k}\sum_{l=0}^{\lambda} \phi_{jl}\rho_{hl}(s-j), \quad s = 1, \ldots, k;\ h = 0, 1, \ldots, \lambda. \tag{5}$$

The last coefficient, $\phi_{k\lambda}$, obtained from solving the system in Equation 5 for $\lambda = 0, 1, \ldots$ and $k = 1, 2, \ldots$, is called space-time partial correlation of spatial order λ. The selection of the spatial order is established by the researcher. As suggested by Pfeifer & Deutsch (1980*b*), the value of λ must be at least the maximum spatial order of any hypothetic model.

Table 2: Characteristics of the theoretical STACF and STPACF for STAR, STMA and STARMA models.

Process	STACF	STPACF
STAR	Tails off with both space and time	Cuts off after p lags in time and λ_p lags in space
STMA	Cuts off after q lags in time and m_q lags in space	Tails off with both space and time
STARMA	Tails off	Tails off

2.3.2 Parameter estimation

Assuming that the $\boldsymbol{\varepsilon}(t)$, $t = 1, \ldots, T$, are independent with distinct variances for each of the N sites, that is, the variance-covariance matrix $\boldsymbol{G}$ is a $N \times N$ diagonal matrix, the maximum likelihood estimates of

$$\begin{aligned}\boldsymbol{\Phi} &= [\phi_{10}, \ldots, \phi_{1\lambda_1}, \ldots, \phi_{p0}, \ldots, \phi_{p\lambda_p}]' \\ \boldsymbol{\Theta} &= [\theta_{10}, \ldots, \theta_{1\lambda_1}, \ldots, \theta_{q0}, \ldots, \theta_{qm_q}]',\end{aligned}$$

the parameter vectors of the STARMA model defined in Equation 1, are obtained by maximizing the log-likelihood function

$$\begin{aligned}l(\boldsymbol{\varepsilon}|\boldsymbol{\Phi}, \boldsymbol{\Theta}, \boldsymbol{G}) &= -\frac{TN}{2}\log|2\pi\,\boldsymbol{G}| - \frac{1}{2}\sum_{t=1}^{T}\boldsymbol{\varepsilon}(t)'\boldsymbol{G}^{-1}\boldsymbol{\varepsilon}(t), \\ &= -\frac{TN}{2}\log|2\pi\,\boldsymbol{G}| - \frac{1}{2}S(\boldsymbol{\Phi}, \boldsymbol{\Theta})\end{aligned}$$

where

$$S(\boldsymbol{\Phi}, \boldsymbol{\Theta}) = \sum_{t=1}^{T}\boldsymbol{\varepsilon}(t)'\boldsymbol{G}^{-1}\boldsymbol{\varepsilon}(t), \tag{6}$$

is the weighted sum of squares of the errors and

$$\begin{aligned}\boldsymbol{\varepsilon}(t) = \mathbf{z}(t) &+ \sum_{k=1}^{p}\sum_{l=0}^{\lambda_k}\phi_{kl}\mathbf{W}^{(l)}\mathbf{z}(t-k) \\ &- \sum_{k=1}^{q}\sum_{l=0}^{m_k}\theta_{kl}\mathbf{W}^{(l)}\boldsymbol{\varepsilon}(t-k).\end{aligned}$$

Finding the values of the parameters that maximize the log-likelihood function is equiva-

lent to finding the values $\hat{\mathbf{\Phi}}$ and $\hat{\mathbf{\Theta}}$ that minimize the sum of squares in Equation 6. Therefore, the problem is reduced to finding the weighted least squares estimates of the parameters.

Numerical techniques must be used to minimize the sum of squares in Equation 6. Subba Rao & Antunes (2003) proposed a procedure for initial estimation of the parameters of $S(\mathbf{\Phi}, \mathbf{\Theta})$ as well as an efficient criterion for order determination.

2.3.3 Model Adequacy

If the fitted model represents adequately the data, the residuals should have gaussian distribution with mean zero and variance-covariance matrix equal to $\boldsymbol{G}$. There are several tests to verify these conditions in the residuals. Particularly, Pfeifer & Deutsch (1980a) and Pfeifer & Deutsch (1981) suggested to calculate the sample space-time autocorrelations of the residuals and to compare them with their theoretical variance. The authors proved that, if the model is adequate,

$$\text{var}(\hat{\rho}_{l0}(s)) \approx \frac{1}{N(T-s)},$$

where $\approx$ means approximately and $\hat{\rho}_{l0}(s)$ is the space-time autocorrelation function of the fitted model residuals. Since the space-time autocorrelations of the residuals should be approximately gaussian, they can be standardized for, subsequently, testing their significance.

Pfeifer & Deutsch (1980a) pointed out that if the residuals have spatial correlation they can be represented by a STARMA model. Usually, identifying the model and incorporating into the candidate model that generated the residuals, is the best form of updating the model.

According to Subba Rao & Antunes (2003), the estimated parameters can be tested for statistical significance in two ways: use the confidence regions for the parameters to test the hypothesis that $H_0 : \mathbf{\Phi} = \mathbf{\Theta} = \mathbf{0}$, or test the hypothesis that a particular ϕ_{kl} or θ_{kl} is zero with the remaining parameters unrestricted.

Let $\hat{\boldsymbol{\delta}} = (\hat{\mathbf{\Phi}}, \hat{\mathbf{\Theta}})' = (\delta_1, \ldots, \delta_K)'$ be the least squares estimate of the full parameter vector, and let $\hat{\boldsymbol{\delta}}^* = (\delta_1, \ldots, \delta_i, \ldots, \delta_K)'$ be the least squares estimate of the parameter vector with δ_i, $i = 1, \ldots, K$, constrained to be zero. The test for the hypothesis $H_0 : \delta_i = 0$ is based on the statistic:

$$\Upsilon = \frac{(TN-K)[S(\hat{\boldsymbol{\delta}}^*) - S(\hat{\boldsymbol{\delta}})]}{S(\hat{\boldsymbol{\delta}})}.$$

Under H_0, Υ is approximately distributed as an $F_{1,TN-K}$. Any parameter that is statistically insignificant must be removed from the model to obtain a simpler model which must be considered as candidate and the estimation stage must be repeated.

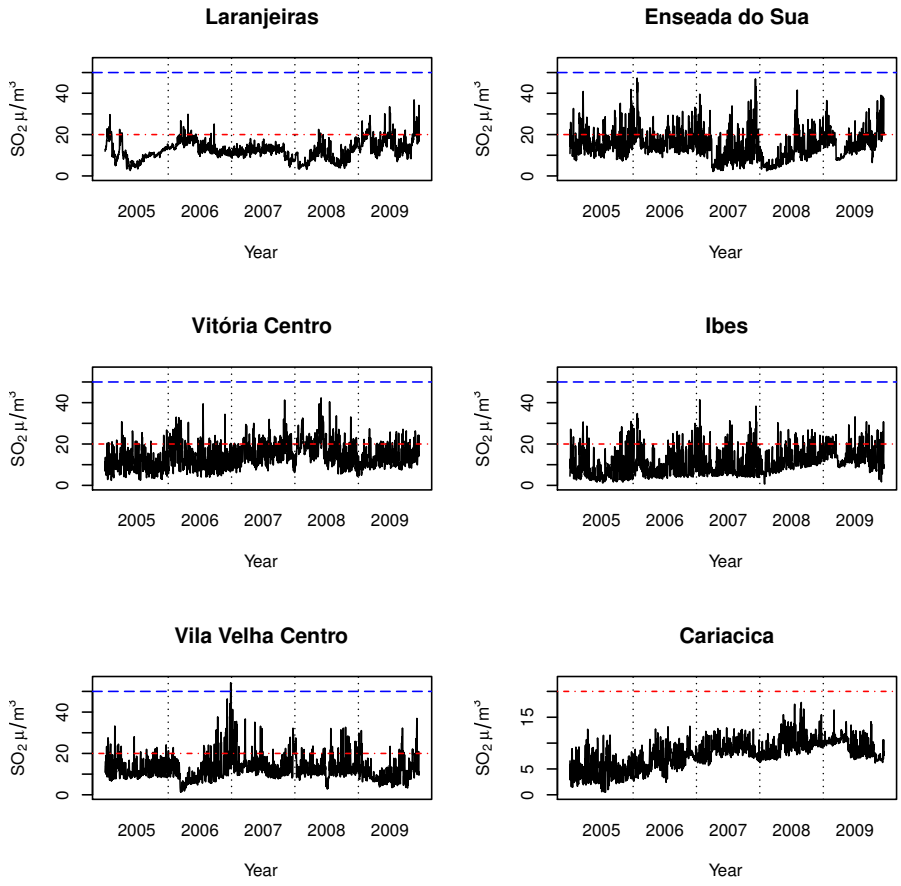


Figure 2: SO_2 daily average concentrations at the AAQMN monitoring stations (- · - 2005 WHO guideline –– 2005 WHO interim guideline).

3 Results and discussion

3.1 Data preparation

Outliers detection

Figure 2 shows the time series plots of the six monitoring stations considered in this study. Some sites (like Laranjeiras at the beginning of the year 2009, for example) show outliers that can affect the modeling and forecasting model performance.

In this context, Fox (1972) suggested four classes of outliers: additive outliers (AO), level shift (LS), temporal change (TC) and innovational outliers (IO). According to (Peña 2001), the effect of AO, TC and LS outliers is limited and independent of the model, AO and TC have transitory effects while LS have permanent effects. However, the effect of an IO depends on the kind of model and its statistical characteristic.

We used the methodology proposed by Gomez & Maravall (1998), which is implemented on the software `TRAMO` (http://www.bde.es/), for outliers detection and correction of the time series obtained from each monitoring station. Table 3 shows the number of the observation

detected as outlier as well as its type.

There were not any IO outliers and the only LS outlier was detected in Cariacica corresponding to observation 568 (July 22, 2006). This level shift can be observed in Figure 2, there is a sudden fall of concentrations observed from this date on, maybe because of a measuring equipment change or any calibration adjusting of the equipment.

Almost all time series observed have outliers with immediate effects, like observation 1536 in Laranjeiras, recorded on March 16th, 2009 (AO outlier); or short-time effects (TC outliers), like the observation 848 in Enseada do Suá, corresponding to April 28th, 2007, where there is a temporary fall in the concentrations, but rapidly they back to the mean levels.

Considering the high quantity of outliers detected by the previous analysis, we decided to transform all the time series in order to correct the distortions caused by the atypical values.

Table 3: List of detected outliers at each AAQMN monitoring station.

	Outlier type		
Station	**AO**	**LS**	**TC**
Laranjeiras	1536, 1335, 1367, 1755, 1224, 1680, 1719, 1378, 1170, 1340, 1290, 1082, 127, 1331, 1402, 1397, 627		57, 123, 52, 1673, 1409, 1344, 1156
Enseada do Suá	1029, 897, 882, 889, 343, 178, 171, 350, 140, 268		848, 970
Vitória Centro	1301, 538, 406, 568, 247, 506, 302, 365, 188, 1739, 688, 553, 898, 532		184, 199, 35, 527, 510
Ibes	301, 1800		
Vila Velha Centro	447, 629		451, 455, 1725, 1700
Cariacica	412, 133, 171, 1240, 1246, 203, 92, 68, 1601, 763, 564, 1600, 515, 1376, 1235, 97, 196, 636, 812, 817, 415, 952, 140	568	

Cycles determination

It is well known that air pollution and meteorological data are influenced by cycles and seasons. In order to determine the cycles affecting SO_2 daily average concentrations, we estimated the periodogram for the time series from each monitoring station. The plots of the

periodograms are not shown due to space constraints, however, the most significant periods are given in Table 4.

Table 4: Significant cycles by monitoring station.

Station	Cycle (days)
Laranjeiras	None
Enseada do Suá	$16.5, 17.5, 18.5, 82$
Vitória Centro	$32, 7, 3.5, 19$
Ibes	$18.5, 16.5, 57, 25$
Vila Velha Centro	$82, 56.5, 18.5, 75$
Cariacica	$7, 3.5, 32$

The expected period of 7 days (since the time series are daily measurements) is significant only in Vitória Centro and Cariacica stations, both sites also present significant periods of 3.5 and 32 days. The remaining monitoring stations have significant periods of approximately 19, 57 and 82 days. These findings indicate that SO_2 concentration levels are affected not only by weekly cycles, but also by monthly and seasonal periods. Following Antunes & Subba Rao (2006), we removed the cyclical component in each time series. Denoting by $\mathbf{Y}(t)$ the outliers-corrected time series, the transformed series to be used for STARMA modeling can be written as

$$\mathbf{Z}(t) = \mathbf{Y}(t) - \mathbf{X}(t),$$

where $\mathbf{X}(t) = [X_1(t), \ldots, X_6(t)]'$ is a periodic function that can be represented as an harmonic series, i.e.

$$X_i(t) = \sum_{j=1}^{s} \left[\xi_{i,j} \cos\left(\frac{2\pi jt}{C_j}\right) + \xi_{i,j}^{\dagger} \sin\left(\frac{2\pi jt}{C_j}\right) \right],$$
$$i = 1, \ldots, 6, \quad t = 1, \ldots, T$$

where $\xi_{i,j}$ and $\xi_{i,j}^{\dagger}$ are unknown parameters which are estimated by least squares, s is the number of significant cycles and C_j represents the period (or cycle) of the time series.

3.2 Descriptive analysis

As observed on Figure 2, for every year the average concentrations are lower than the standard level established by the Brazilian law (CONAMA Nº. 03 of 28/06/90) which are: average of $365\mu g/m^3$ for a 24-hour period (cannot be exceeded more than once a year) and annual arithmetic average of $80\mu g/m^3$. Nevertheless, the concentrations are quite higher than the guideline suggested by the World Health Organization (World Health Organization [WHO] 2006), which is 24-hour average concentration of $20\mu g/m^3$, or even the interim guideline of $50\mu g/m^3$ average suggested for developing countries like Brazil.

Particularly, Vila Velha Centro station exceed the interim limit only once in 2006. Cariacica station does not exceed any limit and shows the lowest values and variability.

These assertions can be confirmed from the results displayed in Table 5. Besides, it can be observed that some stations show a high variability and maximum values much larger than the most of observed concentrations, for example, while 75% of concentrations from Ibes station is lower than $14.48\mu g/m^3$, the maximum concentration observed is $41.385\mu g/m^3$ (more than four times the mean value).

Table 5: Summary statistics of daily average SO_2 concentrations in GVR (2005-2009).

Station	Minimum	1st. Quartil	Median	Mean	3rd. Quartil	Maximum
Laranjeiras	2.630	9.675	12.100	12.478	14.861	36.770
Enseada do Suá	2.159	10.349	14.195	14.942	18.452	47.288
Vitória Centro	2.417	9.651	13.233	14.165	17.915	42.295
Ibes	0.623	5.738	9.694	10.898	14.476	41.385
Vila Velha Centro	1.288	8.914	11.195	12.422	14.918	54.165
Cariacica	0.479	6.316	7.927	7.872	9.797	17.852

The highest SO_2 mean concentrations were observed at Enseada do Suá and Vitória Centro stations. This situation can be explained by the direct influence of industrial and port activities for both monitoring stations, as showed in Table 1.

The boxplots shown in Figure 3 show that the mean concentrations and variability are different for all stations. Higher concentrations are observed in regions influenced by the main industrial activities of GVR, and lower values are observed in regions far away from that influence (like Laranjeiras and Cariacica stations). This behavior suggests there is an influence of the location, which reinforces the importance of including spatial characteristics into the model.

Figure 4 displays the boxplots of the average concentrations by day of the week. As observed in Section 3.1, there is a weekly cycle in Vitória Centro and Cariacica monitoring stations because the median is slightly lower on weekends and the concentration rises along the week. The remaining stations do not show any obvious trend along the week.

The sample autocorrelation functions (ACF) of the outliers-corrected SO_2 time series obtained for each monitoring station are shown in Figure 5. The slow decay of the correlations suggest non-stationarity of the time series in all the stations, however, the Augmented Dickey-Fuller test, proposed by Dickey & Fuller (1979), was used to examine the hypothesis of stationarity of SO_2 average concentrations at each monitoring station. Results indicate that there is not enough evidence to consider the series as non-stationary ($p\_value < 0.02$ for all stations).

3.3 Weighting matrix

As indicated by Pfeifer & Deutsch (1980*b*), the weighting matrix $\mathbf{W}^{(l)}$ must be defined prior to modeling. Since the GVR has a small number of stations irregularly distributed over a relatively small area, it is reasonable to consider each site as first order neighbor of every other site. Therefore, the maximum spatial order of the STARMA model is one. So we have

$$\mathbf{W}^{(0)} = \mathbf{I}_N \qquad \text{and} \qquad \mathbf{W}^{(1)} = \mathbf{W}.$$

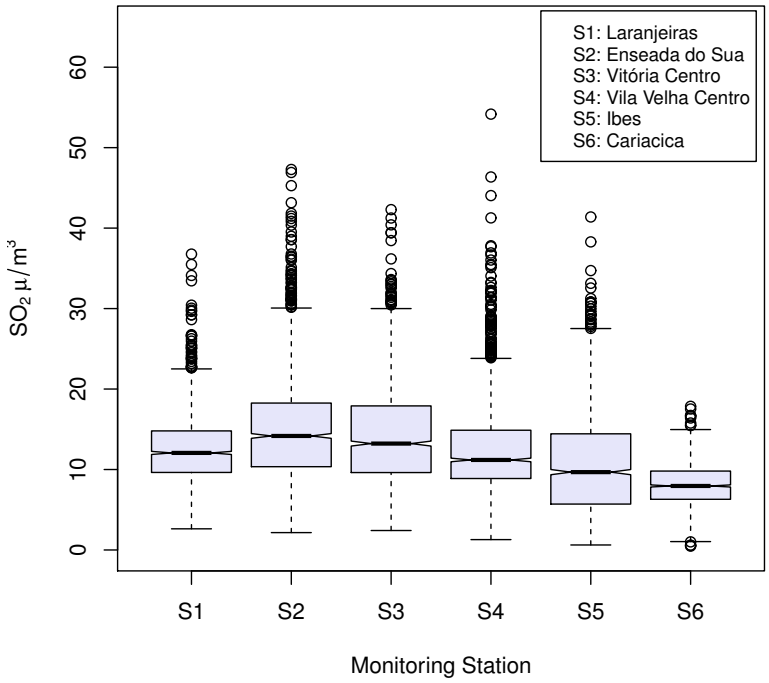


Figure 3: Boxplots of SO_2 daily average by monitoring station.

There are several ways to define the weighting matrix, see Cliff & Ord (1981) and Anselin & Smirnov (1996). In particular, we chose **W** formed by weights inversely proportional to the Euclidean distance between the monitoring stations since this is the most widely used and simplest approach.

The distance (Km) between the stations was calculated using the expression:

$$\begin{aligned} d_{ij} =& 6378.7 \times \text{acos}(\sin(\text{lat}_i/57.296) \times \sin(\text{lat}_j/57.296) \\ &+ \cos(\text{lat}_i/57.296)\cos(\text{lat}_j/57.296) \\ &\times \cos(\text{lon}_j/57.296 - \text{lon}_i/57.296)), \end{aligned}$$

for $i, j = 1, 2, \ldots, 6$, where lat_i and lon_i represent the latitude and longitude of the station i, respectively (www.meridianworlddata.com/Distance-Calculation.asp). Therefore, the weighting matrix **W** was defined considering weights (w_{ij}) as,

$$w_{ij} = \begin{cases} 1/d_{ij}, & \text{for } i \neq j \\ 0, & \text{for } i = j. \end{cases}$$

The weights were scaled so that the sum of the elements at each line equals one. The resulting **W** matrix is:

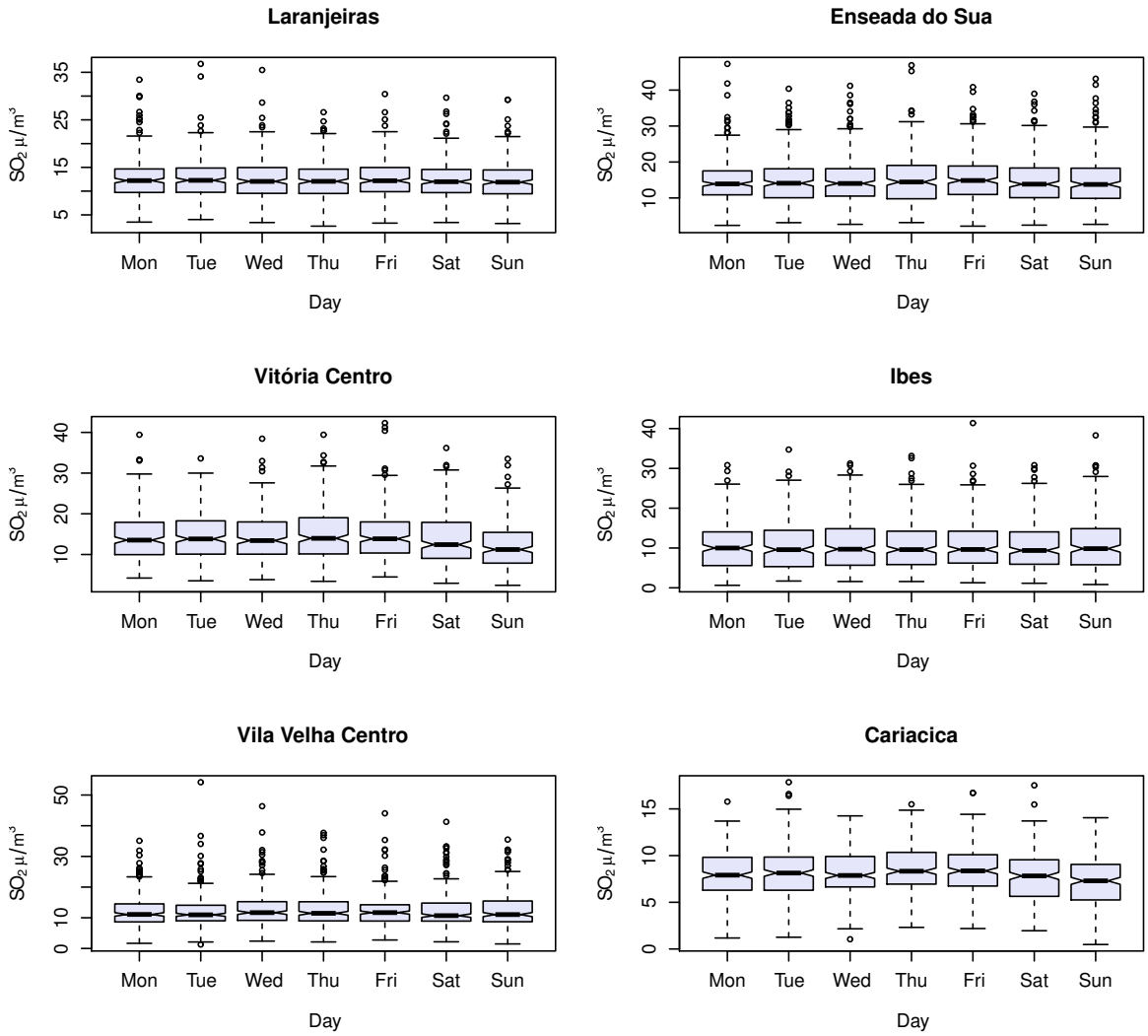


Figure 4: Boxplots of SO_2 daily average by day of the week.

$$\mathbf{W} = \begin{bmatrix} 0.000 & 0.252 & 0.206 & 0.184 & 0.211 & 0.148 \\ 0.081 & 0.000 & 0.212 & 0.211 & 0.409 & 0.087 \\ 0.073 & 0.232 & 0.000 & 0.299 & 0.235 & 0.161 \\ 0.058 & 0.208 & 0.269 & 0.000 & 0.348 & 0.118 \\ 0.060 & 0.359 & 0.188 & 0.311 & 0.000 & 0.082 \\ 0.096 & 0.176 & 0.297 & 0.242 & 0.188 & 0.000 \end{bmatrix}$$

3.4 Fitted model

From Figures 6 and 7 we can observe that there is no remaining seasonality or cycles in the data. According to the characteristics described on Table 2, the slow decaying of the STFAC and the cutting-off in the STPACF after the first 6 time lags in the spatial lag zero indicates that a suitable model is a STAR with maximum autoregressive order 6.

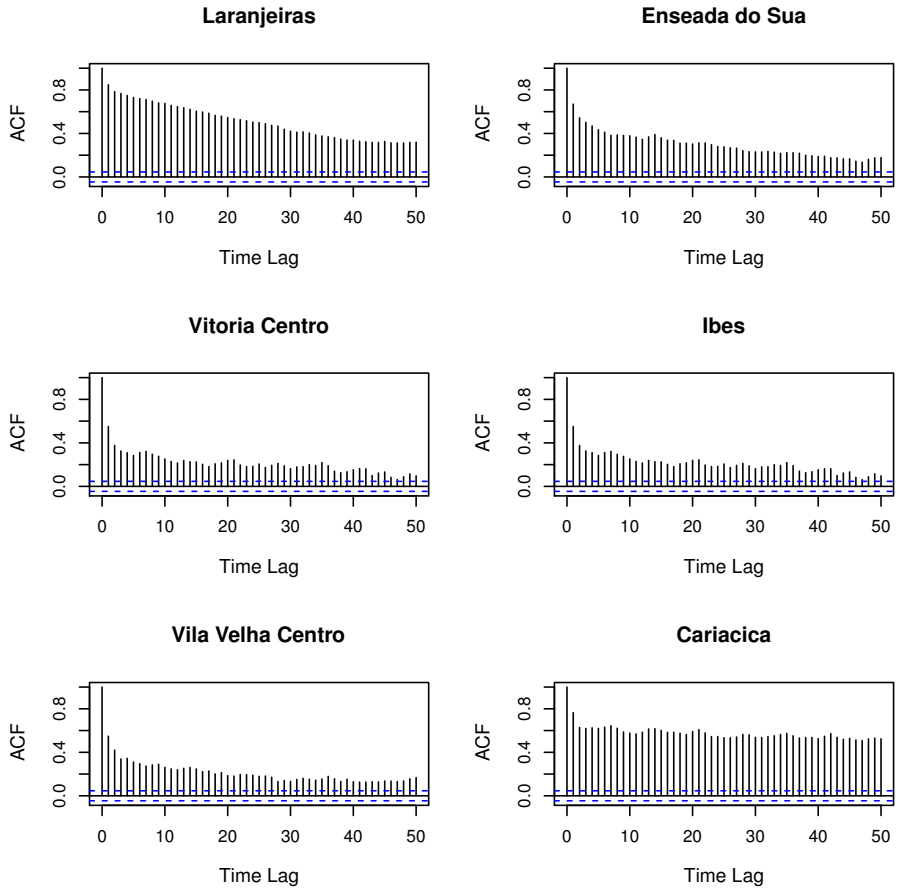


Figure 5: Autocorrelation Functions for SO_2 daily average by monitoring station.

The partial space-time autocorrelations are not significant for the spatial order 1 after the first time lag, indicating that a spatial order one could be enough. The STACF and STPACF were calculated based on the assumption that the errors $\boldsymbol{\varepsilon}$ have a diagonal variance-covariance matrix $\boldsymbol{G}$, estimated from the data.

The model with the best performance is the STAR($4_{1,0,0,0}$) with parameters (the standard errors are shown in brackets):

$$\begin{aligned} \phi_{10} &= -0.475\,(0.0109) & \phi_{11} &= -0.066\,(0.0306) \\ \phi_{20} &= -0.066\,(0.0121) & \phi_{21} &= 0.058\,(0.0335) \\ \phi_{30} &= -0.108\,(0.0121) & \phi_{31} &= -0.004\,(0.0335) \\ \phi_{40} &= -0.156\,(0.0109) & \phi_{41} &= -0.019\,(0.0306) \end{aligned}$$

The parameters ϕ_{21}, ϕ_{31} and ϕ_{41} were not significant at a 5% level of significance. There-

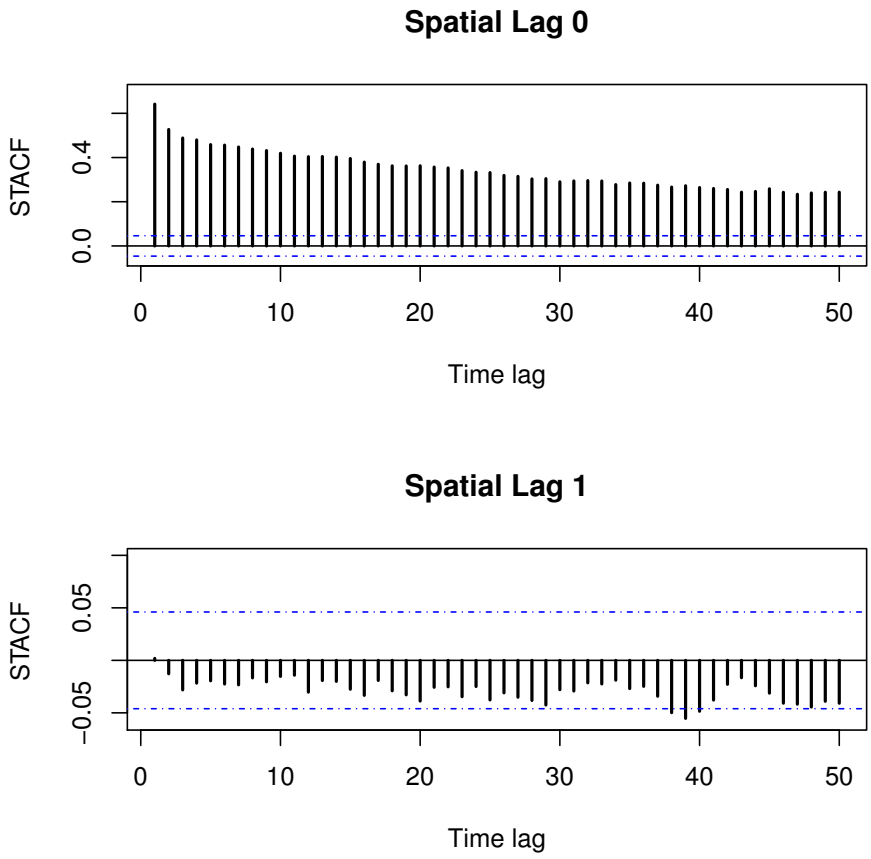


Figure 6: Space-time Autocorrelation Function (STACF) for SO_2 daily average time series.

fore, the final fitted model is:

$$\begin{aligned}\widehat{\mathbf{z}}(t) = {} & 0.475\mathbf{z}(t-1) + 0.066\mathbf{W}\mathbf{z}(t-1) + 0.066\mathbf{z}(t-2) \\ & + 0.108\mathbf{z}(t-3) + 0.156\mathbf{z}(t-4).\end{aligned} \tag{7}$$

The sample STACF of the residuals, displayed in Figure 8, shows very small autocorrelation values, suggesting that the assumption of uncorrelated errors is satisfied by the fitted model.

Normality tests and quantile-quantile plots of the residuals (Figure 9) show that the errors are not normally distributed. The lack of Gaussian distribution affects the inferential process, that is, the significance tests as well as the confidence and prediction intervals.

In order to guarantee the reliability of the model, bootstrap resampling techniques were used to obtain confidence intervals for the estimated parameters as well as the prediction intervals. The bootstrap approach here adopted was resampling from the residuals $\boldsymbol{\varepsilon}(t)$ of the fitted model as follows,

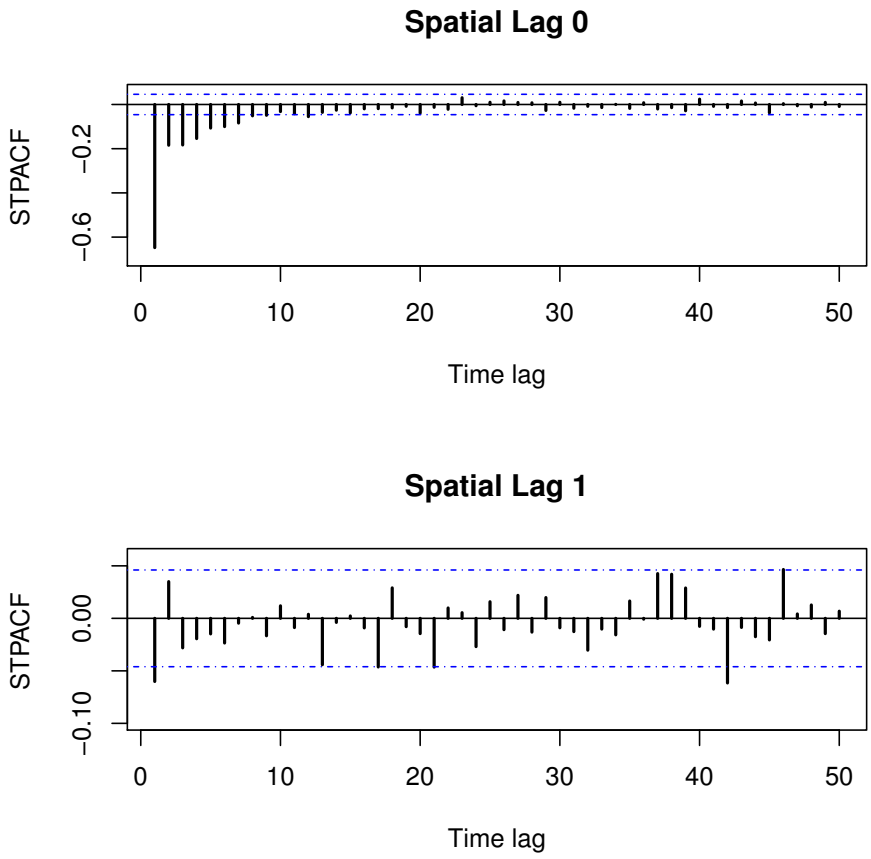


Figure 7: Partial Space-time Autocorrelation Function (STPACF) for SO_2 daily average time series.

a. Calculate the residual for each observation:

$$\widehat{\boldsymbol{\varepsilon}}(t) = \mathbf{z}(t) - \widehat{\mathbf{z}}(t) \qquad t = 1, \ldots, T.$$

b. Select bootstrap samples of the residuals, $\mathbf{e}_b^\star = [\boldsymbol{\varepsilon}_b^\star(1), \ldots, \boldsymbol{\varepsilon}_b^\star(T)]'$, and from these, calculate bootstrapped $\mathbf{z}$ values $\overline{\mathbf{z}}_b^\star = [\mathbf{z}_b^\star(1), \ldots, \mathbf{z}_b^\star(T)]'$, where $\mathbf{z}_b^\star(t) = \widehat{\mathbf{z}}(t) + \boldsymbol{\varepsilon}_b^\star(t)$, for $t = 1, \ldots, T$.

c. Fit the model using $\mathbf{z}$ values to obtain the bootstrap coefficients

$$\boldsymbol{\delta}_b^\star = (\phi_{10,b}^\star, \phi_{11,b}^\star, \phi_{20,b}^\star, \phi_{21,b}^\star, \phi_{30,b}^\star, \phi_{31,b}^\star, \phi_{40,b}^\star, \phi_{41,b}^\star)',$$

for $b = 1, \ldots, r$, where r is the number or bootstrap replicates.

d. The resampled $\boldsymbol{\delta}_b^\star$ can be used to construct bootstrap standard errors and confidence intervals for the coefficients.

As is well known, the bootstrap samples have the property of mimic the original sample.

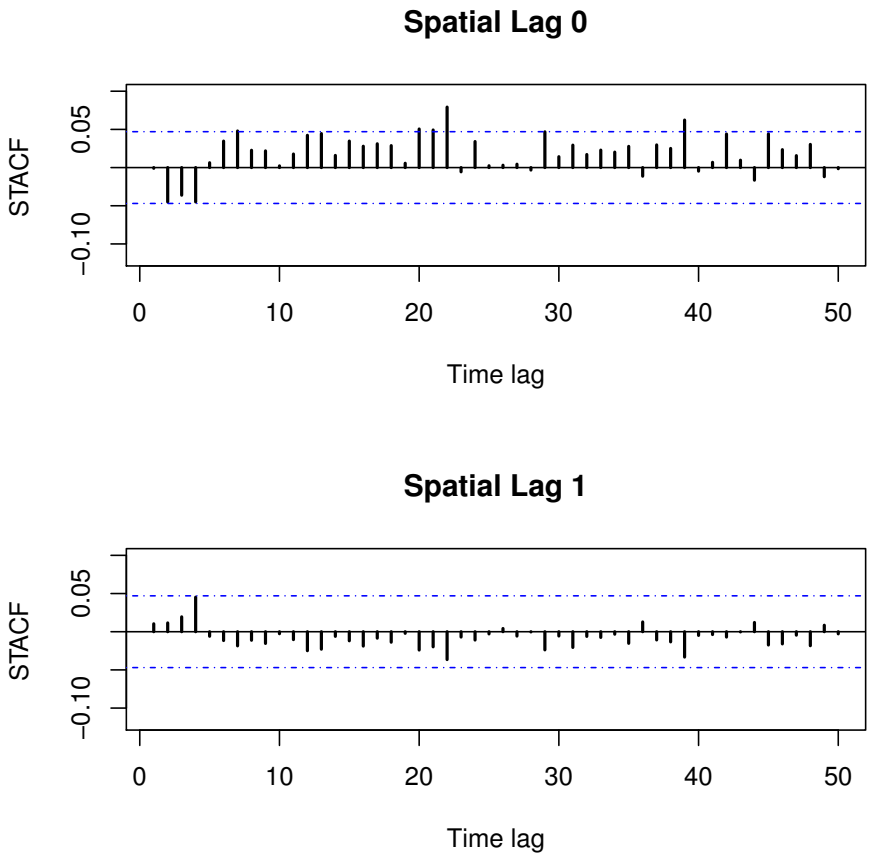


Figure 8: Space-time Autocorrelation Function (STACF) of the residuals from the fitted STARMA$(4_{1,0,0,0}, 0)$ model.

More details about bootstrap techniques can be obtained in Wu (1986), Efron & Tibshrani (1993) and Lam & Veall (2002) among others.

Figure 10 displays the predicted values of the observed time series by using the fitted model. This figure suggests a reasonably good performance of the model. It well captures the variability, tendency and the periods of the data.

The model indicates that SO_2 concentrations in a site are highly influenced by the levels presented in the previous day ($\phi_{10} = -0.475$). Moreover, the influence of SO_2 over the region is around 3-4 days and the concentration level in a site is influenced by the concentration observed at its neighbors in the day before. Based on the good in-sample performance of the model, it is reasonable to consider it as an alternative method for estimating missing data.

3.5 Forecasting

The fitted model shown in Equation 7 was used in order to determine one-step-ahead forecasts for a 15-days period, that is, we obtained forecasts for the last two weeks of the full period. The forecasts were calculated using the Minimum Mean Square Error (MMSE)

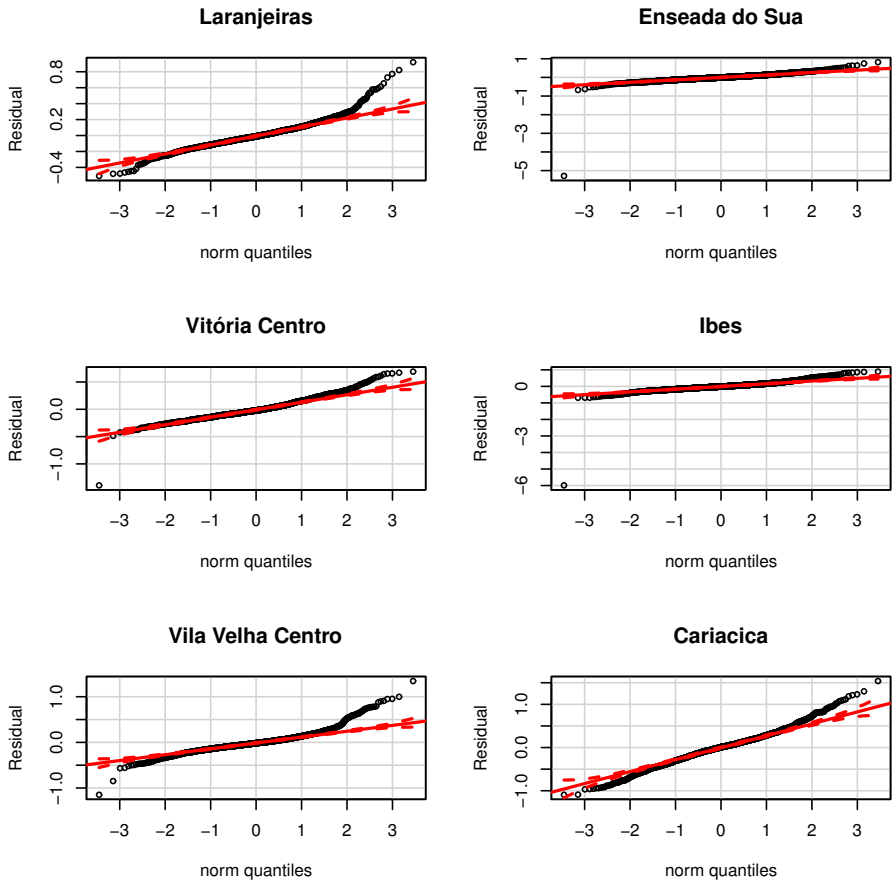


Figure 9: Quantile-quantile plot of the residuals from the fitted STARMA$(4_{1,0,0,0}, 0)$ model.

criterion as

$$\widehat{\mathbf{z}}_{(1)}(t) = \mathrm{E}[\mathbf{z}(t+1)|\mathbf{z}(s),\ s \leq t].$$

The forecasts and their 95% prediction intervals are displayed in Figure 11. It can be observed that forecasts describe well the time series behavior and trend for all the stations. Even knowing that Gaussian distribution assumption is not met, the prediction intervals under this supposition were calculated only for comparative purposes. It becomes clear that the errors were underestimated for the most of stations and, therefore, the reliability of the inferences based on the Gaussian assumption was strongly compromised. This fact reinforces the usefulness of the resampling techniques in order to perform efficient inferences.

In particular, for the time series which have the lower variability (Laranjeiras and Cariacica stations), almost all the real data falls within the prediction intervals and their forecasts are more accurate than those for the sites which have observations very distant from the mean, as is the case of Enseada do Suá station, for example. For the remaining series, it can be observed that even for the model capturing the high variability in the data, the discrepant values are not covered by the prediction intervals.

In order to quantify the forecasting ability of the fitted model for each monitoring station

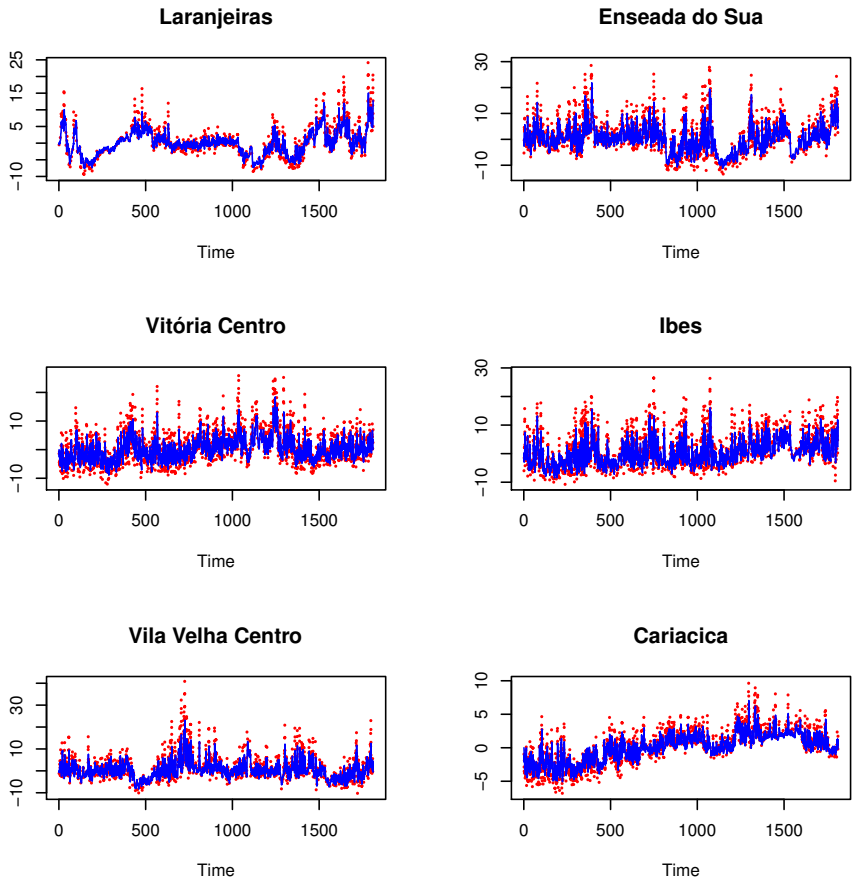


Figure 10: Within-sample prediction for the transformed SO_2 time series (· · · Observed concentrations — Predicted concentrations).

we used the criterions: root mean squared error (RMSE) and mean absolute error (MAE), defined as

$$RMSE_i = \sqrt{\frac{1}{H}\sum_{t=T+1}^{T+H} \boldsymbol{\epsilon}_i(t)^2},$$

$$MAE_i = \frac{1}{H}\sum_{t=T+1}^{T+H} |\boldsymbol{\epsilon}_i(t)|,$$

where $i = 1, 2, \ldots, 6$ and $H = 1, \ldots, 15$. The MAE measures the average magnitude of errors considering their absolute magnitude. The RMSE is also known as the standard error of the forecast and it is more sensitive to outliers than MAE (Hyndman & Koehler 2006).

As observed in Table 6, Laranjeiras and Cariacica stations have the most accurate forecasts (MAE of about 1.71 and 0.25, respectively). The highest values for the MAE criterion were obtained for Ibes, Enseada do Suá and Vitória Centro stations (about 2.64, 2.59 and 2.11, respectively), which means that the average absolute difference between the forecasts and the observed concentrations was approximately 2 $\mu g/m^3$.

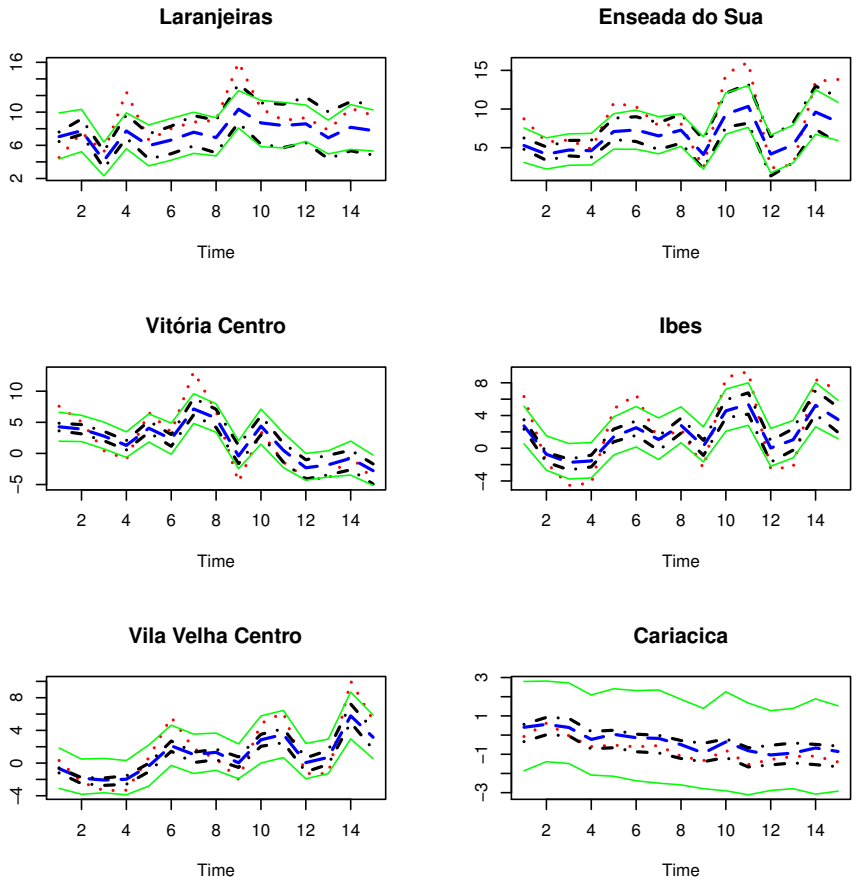


Figure 11: Out-of-sample one-step-ahead forecasts for the transformed SO_2 time series ($\cdots$ Observed data $-\,-$ Forecasted data $-\cdot-$ 95% confidence limits for Gaussian interval — 95% confidence limits for bootstrap interval).

The most imprecise forecasts were obtained for Enseada do Suá with a residual standard deviation of 3.04 $\mu g/m^3$, followed by Ibes station which has a RMSE of 2.91 $\mu g/m^3$.

4 Final Remarks

This study applies a STARMA model to daily average SO_2 concentrations in order to describe the dynamics of the pollutant at GVR, as well as to forecast future concentrations. The analysis of the individual time series at the monitoring stations reveals that there are some significant cycles affecting the behavior of the dispersion over the region.

Based on the fitted model, the persistence of SO_2 in the region is about four days and its concentration levels are influenced by the levels observed at nearby sites. The residual analysis indicated a good fit for in-sample observations, so that it can be used for imputation of missing values. Regarding the out-of-sample performance, the model can be a reasonable tool for predicting future values with a certain reliability. The higher values of the accuracy measures for the series with more discrepant values indicate that the forecasting capability of

Table 6: Model accuracy measures.

Station	RMSE	MAE
Laranjeiras	2.1409	1.7090
Enseada do Suá	3.0442	2.5917
Vitória Centro	2.5027	2.1073
Ibes	2.9062	2.6408
Vila Velha Centro	2.0422	1.7597
Cariacica	0.2770	0.2503

the model is highly influenced by outliers.

Acknowledgements

This work was performed under the CAPES financial support.

Prof. Tata Subba Rao thanks the University of Manchester, UK and CRRAO AIMSCS, University of Hyderabad Campus,India.

Prof. Valderio Reisen thanks FAPES and CNPq for the financial support.

The authors would like to thank the Instituto Estadual de Meio Ambiente e Recursos Hídricos (IEMA) of Espírito Santo State for providing the data.

References

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

Anselin, L. & Smirnov, O. (1996), 'Efficient algorithms for constructing proper higher order spatial lag operators', *Journal of Regional Science* **36**(1), 67–89.

Antunes, A. & Subba Rao, T. (2006), 'On hypotheses testing for the selection of spatio-temporal models', *Journal of Time Series Analysis* **27**(5), 767–791.

Ashbaugh, L., Myrup, L. & Flocchini, R. (1984), 'A principal component analysis of sulfur concentrations in the Western United States', *Atmospheric Environment* **18**, 783–791.

Beelen, R., Hoek, G., Pebesma, E., Vienneau, D., de Hoogh, K. & Briggs, D. J. (2009), 'Mapping of background air pollution at a fine spatial scale across the European Union', *Science of The Total Environment* **407**(6), 1852 – 1867.

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.

Brunelli, U., Piazza, V., Pignato, L., Sorbello, F. & Vitabile, S. (2008), 'Three hours ahead prevision of SO_2 pollutant concentration using an Elman neural based forecaster', *Building and Environment* **43**, 304–314.

Castro, F. B., Prada, J., Gonzalez, W. & Febrero, M. (2003), 'Prediction of SO_2 levels using neural networks', *Journal of the Air and Waste Management Association* **53**, 532–539.

Chelani, A., Rao, C., Phadke, K. & Hasan, M. (2002), 'Prediction of sulphur dioxide concentrations using artificial neural networks', *Environmental Modelling and Software* **17**(2), 161–168.

Cheng, S. & Lam, K. (2000), 'Synoptic typing and its application to the assesment of climatic impact on concentrations of sulfur dioxide and nitrogen oxides in Hong Kong', *Atmospheric Environment* **34**, 585–594.

Cliff, A. & Ord, J. (1981), *Spatial Processes: Models and Applications*, London: Pion.

de Kluizenaar, Y., Aherne, J. & Farrell, E. (2001), 'Modelling the spatial distribution of SO_2 and NO_x emissions in Ireland', *Environmental Pollution* **112**, 171–182.

Deutsch, S. & Pfeifer, P. (1981), 'Space-time ARMA modeling with contemporaneously correlated innovations', *Technometrics* **23**(4), 401–409.

Dickey, D. & Fuller, W. (1979), 'Distribution of estimators for autoregressive time series with a unit root', *Journal of the American Statistical Association* **74**, 427–431.

Efron, B. & Tibshrani, R. (1993), *An Introduction to the Bootstrap*, New York: Chapman & Hall.

Fan, S., Burstyn, I. & Senthilselvan, A. (2010), 'Spatiotemporal modeling of ambient sulfur dioxide concentrations in Rural Western Canada', *Environmental Modeling and Assessment* **15**, 137–146.

Fox, A. (1972), 'Outliers in time series', *Journal of the Royal Statistical Society* **34**(3), 350–363.

Gomez, V. & Maravall, A. (1998), Guide for using the program TRAMO and SEATS, Technical report, Research Department, Banco de España.

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.

Hyndman, R. J. & Koehler, A. B. (2006), 'Another look at measures of forecast accuracy', *International Journal of Forecasting* **22**(4), 679 – 688.

Ibarra Berástegui, G., Sáenz, J., Ezcurra, A., Ganzedo, U., Díaz de Argadoña, J., Errasti, I., Fernandez Ferrero, A. & Polanco Martínez, J. (2009), 'Assessing spatial variability of SO_2

field as detected by an air quality network using self-orginized maps, cluster and principal component analysis', *Atmospheric Environment* **43**, 3829–2826.

Instituto Brasileiro de Geografia e Estatística [IBGE] (2012), Indicadores de desenvolvimento sustentável, Technical report.

Instituto Estadual de Meio Ambiente e Recursos Hídricos [IEMA] (2007), Relatório da qualidade do ar na Região da Grande Vitória 2006., Technical report.

Instituto Estadual de Meio Ambiente e Recursos Hídricos [IEMA] (2011), Inventário de emissões atmosféricas da Região da Grande Vitória, Technical report.

Instituto Jones dos Santos Neves [IJSN] (2012), Perfil do Espírito Santo. Dados gerais. Vitória − ES, 2012., Technical report.

Kamarianakis, Y. & Prastacos, P. (2005), 'Space-time modeling of traffic flow', *Computers and Geosciences* **31**, 119–133.

Kumar, A. & Goyal, P. (2011), 'Forecasting of daily air quality index in Delhi', *Science of the total environment* **409**, 5517–5523.

Kurt, A. & Oktay, A. B. (2010), 'Forecasting air pollutant indicator levels with geographic models 3 days in advance using neural networks', *Expert Systems with Application* **37**, 7986–7992.

Lalas, D., Veirs, V., Karras, G. & Kallos, G. (1982), 'An analysis of the SO_2 concentration levels in Athens, Greece', *Atmospheric Environment* **16**(3), 531–544.

Lam, J. & Veall, M. (2002), 'Bootstrap prediction intervals for single period regression forecasts', *International Journal of Forecasting* **18**(1), 125–130.

McCollister, G. & Wilson, K. (1975), 'Linear stochastic models for forecasting daily maxima and hourly concentrations of air pollutants', *Atmospheric Environment* **9**, 417–423.

Nunnari, G., Dorling S., Schlink, U., Cawley, G., Foxall, R. & Chatterton, T. (2004), 'Modelling SO_2 concentration at a point with statistical approaches', *Environmental Modelling and Software* **10**(10), 887–905.

Peña, D. (2001), Outliers, influential observations, and missing data, *in* D. Peña, G. Tiao & R. Tsay, eds, 'A course in advanced time series analysis', J. Wiley and Sons, chapter 6.

Perez, P. (2001), 'Prediction of sulfur dioxide concentrations at a site near downtown Santiago, Chile', *Atmospheric Environment* **35**(29), 4929–4935.

Pfeifer, P. & Deutsch, S. (1980*a*), 'Identification and interpretation of first order space-time ARMA models', *Technometrics* **22**(3), 397–408.

Pfeifer, P. & Deutsch, S. (1980*b*), 'A three-stage iterative procedure for space-time modeling', *Technometrics* **22**(1), 35–47.

Pfeifer, P. E. & Deutsch, S. (1981), 'Variance of the sample space-time autocorrelation function', *Journal of the Royal Statistical Society* **43**(1), 28–33.

R Core Team (2012), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
URL: *http://www.R-project.org/*

Roca Pardiñas, J., Gonzalez Manteiga, W., Febrero Bande, M., Prada Sánchez, J. & Cadarso Suárez, C. (2004), 'Predicting binary time series of SO_2 using generalized additive models with unknown link function', *Environmetrics* **15**, 729–742.

Rouhani, S., Ebrahimpour, M., Yaqub, I. & Gianella, E. (1992), 'Multivariate geostatistical trend detection and network evaluation of space-time acid deposition data – I. Methodology', *Atmospheric Environment. Part A. General Topics* **26**(14), 2603 – 2614.

Schlink, U., Herbarth, O. & Tetzlaff, G. (1997), 'A component time-series model for SO_2 data: forecasting, interpretation and modification', *Atmospheric Environment* **31**(9), 1285–1295.

Subba Rao, T. & Antunes, A. (2003), Spatio-temporal modelling of temperature time series: a comparative study, *in* 'Time Series Analysis and Applications to Geophysical Systems', The IMA volumes in Mathematics and its Applications, pp. 123–150.

Tecer, L. (2007), 'Prediction of SO_2 and PM concentrations in a coastal mining area (Zonguldak, Turkey) using an artificial neural network', *Polish Journal of Environmental Studies* **16**(4), 633–638.

Turalioglu, F. S. & Bayraktar, H. (2005), 'Assessment of regional air pollution distribution by point cumulative semivariogram method at Erzurum urban center, Turkey', *Stochastic Environmental Research and Risk Assessment* **19**, 41–47.

World Health Organization [WHO] (2006), Who air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide - Global update 2005, Technical report.

Wu, C. (1986), 'Jackknife, bootstrap and other resampling methods in regression analysis', *The Annals of Statistics* **14**(4), 1261–1295.

Yang, S., Yuesi, W. & Changchun, Z. (2009), 'Measurements of the vertical profile of atmospheric SO_2 during the heating period in Beijing on days of high air pollution', *Atmospheric Environment* **43**, 468–472.

Yu, T.-Y. & Chang, I.-C. (2006), 'Spatiotemporal features of severe air pollution in Northern Taiwan', *Environmental science and pollution research international* **13**(4), 268–275.

Zeri, M., Oliveira-Júnior, J. & Lyra, G. (2011), 'Spatiotemporal analysis of particulate matter, sulfur dioxide and carbon monoxide concentrations over the city of Rio de Janeiro, Brazil', *Meteorology and Atmospheric Physics* **113**, 139–152.

Zou, B., Gaines Wilson, J., Benjamin Zhan, F. & Zeng, Y. (2009), 'An emission-weighted proximity model for air pollution exposure assessment', *Science of The Total Environment* **407**(17), 4939 – 4945.

Originally submitted to *Environmetrics*, 2013

Modeling and Forecasting PM_{10} concentrations using the Space-Time ARFIMA Model

Nátaly A. Jiménez Monroy[1,2*] *Valdério A. Reisen*[1,2] *and Tata Subba Rao*[3,4]

[1]Programa de Pós-Graduação em Engenharia Ambiental - UFES, Vitória, ES.

[2]Departamento de Estatística, UFES, Vitória, ES.

[3]School of Mathematics, University of Manchester, UK.

[4]CRRAO AIMSCS, University of Hyderabad Campus, India.

Abstract

This paper proposes the Space-Time ARFIMA model (STARFIMA) as an extension of the STARMA class models in order to account for time series with long-memory behavior, a phenomenon that is quite common in the atmospheric pollutant variables. The model is introduced and the semiparametric estimation procedure given in Shimotsu (2007) is suggested to estimate the fractional parameters of the STARFIMA processes. Empirical results from Monte Carlo simulations show the importance of considering not only the spatial dependence between the processes, but also the long memory characteristics of the time series involved. The proposed methodology is applied to PM_{10} daily average concentrations. The comparison of the results obtained using STARFIMA and STARMA models reinforces the usefulness of considering the long-memory characteristics to this particular data set in order to improve the forecasting ability.

Keywords: Atmospheric pollution, ARFIMA, forecasting, long-memory, STARMA models, particulate matter.

1 Introduction

The space-time models have shown their usefulness in situations where the data are observed simultaneously in time and space scales. This is the case of the air quality monitoring networks, where the concentration of various pollutants are measured over several spatial locations (monitoring stations) along time (usually at each minute or hour). See, for example Rouhani et al. (1992), De-Iaco et al. (2003), Huerta et al. (2004), Yu & Chang (2006) and Zeri et al. (2011), among others.

In particular, the class of STARMA (space-time autoregressive moving average) models has been used successfully in several research areas as meteorology (Glasbey & Allcroft (2008)), oceanography (Stoffer (1986), LaValle et al. (2001)), ecology (Reynolds & Madden (1988), Reynolds et al. (1988), Epperson (1994), Epperson (2000)), spatial econometrics (Terzi (1995),

*Email: nataly.monroy@ufes.br

Pace et al. (1998), Giacinto (2006)), hydrology (Deutsch & Ramos (1986)), transportation research (Garrido (2000), Kamarianakis & Prastacos (2005)) and imaging (Soni et al. (2004), Crespo et al. (2007)). Nevertheless, its application to atmospheric pollution studies is rare (Antunes & Subba Rao (2006), Glasbey & Allcroft (2008)).

In time series modeling is fundamental to analyze the stochastic dependence structure of the series. The class of dependence between the observations determines the model underlying the process. In general, the dependence (or memory) classes are characterized in three forms: short, intermediate and long.

The ARFIMA(p, d, q) (Fractionally Integrated Autoregressive Moving Average) class models, suggested by Granger & Joyeux (1980) and Hosking (1981), has been broadly used due to its capability for capturing the three memory classes previously described in univariate time series. The parameter d assumes real values and characterizes the memory of the process as follows: short ($d = 0$), intermediate ($d < 0$) and long-memory ($d > 0$). The ARMA(p, q) model is a particular case of the ARFIMA(p, d, q) that has the short memory property.

The aim of this work is to propose the STARFIMA model, as an extension of the STARMA class models, taking into account the long memory of the processes under analysis, a phenomenon which is usually observed in the dispersion dynamics of some atmospheric pollutants. The paper also suggests a two-step procedure to estimate the model. The model and the estimation procedure are the motivations of Section 2. In Section 3 a simulation study is presented in order to show the performance of the model estimates for small sample sizes and other considerations. Section 4 shows an application of the proposed model for forecasting PM_{10} concentrations at the Greater Vitória Region (GVR), Brazil. In addition, the comparison of the fitting and forecasting ability of the proposed model with respect to the STARMA approach is studied. Some final remarks and recommendations are presented in Section 5.

2 The space-time ARFIMA model

Let $\boldsymbol{Z}_t = (Z_{1,t}, Z_{2,t}, \ldots, Z_{N,t})'$ be a vector of observations at N fixed spatial locations on time t. The space-time autoregressive fractionally integrated moving average (STARFIMA) model, denoted as STARFIMA$(p_{\lambda_1,\lambda_2,\ldots,\lambda_p}; \boldsymbol{d}; q_{m_1,m_2,\ldots,m_q})$, is defined as

$$\boldsymbol{\Phi}_{p,\lambda}(B)\boldsymbol{Z}_t = \boldsymbol{\Theta}_{q,m}(B)\mathcal{D}(B)^{-1}\boldsymbol{\varepsilon}_t, \quad t = 1, 2, \ldots, n, \tag{1}$$

where $\boldsymbol{\Phi}_{p,\lambda}(x) = \mathbb{I}_N - \sum_{k=1}^{p}\sum_{l=0}^{\lambda_k}\phi_{kl}\boldsymbol{W}_l x^k$, $x \in \mathbb{C}$, represents the autoregressive polynomial with temporal order p and spatial order λ_k, $\boldsymbol{\Theta}_{q,m}(x) = \mathbb{I}_N - \sum_{k=1}^{q}\sum_{l=0}^{m_k}\theta_{kl}\boldsymbol{W}_l x^k$, $x \in \mathbb{C}$, represents the moving averaged polynomial with temporal order q and spatial order m_k, $\mathbb{I}_N$ is the $N \times N$ identity matrix and $\boldsymbol{W}_l$ is a nonzero $N \times N$ matrix of weights for the spatial order l with diagonal entries 0 and off-diagonal entries related to the distances between the sites. By definition, $\boldsymbol{W}_0 = \mathbb{I}_N$. Each row of $\boldsymbol{W}_l$ adds up to 1. $\boldsymbol{d} = (d_1, \ldots, d_N)$ is the fractional difference vector, $\mathcal{D}(B)$ is the $N \times N$ fractional difference operator matrix such that

$\mathcal{D}(B) = \text{diag}\left\{(1-B)^{d_1}, (1-B)^{d_2}, \ldots, (1-B)^{d_N}\right\}$, where

$$(1-B)^{-d_i} = \sum_{k=0}^{\infty} \frac{\Gamma(d_i+k)}{\Gamma(d_i)\Gamma(k+1)} B^k, \quad i = 1, 2, \ldots, N, \tag{2}$$

with $d_i \in \mathbb{R}$, B is the backward shift operator and $\Gamma(\cdot)$ represents the Gamma function. The N-dimensional vectors $\boldsymbol{\varepsilon}_t = [\varepsilon_{1,t}, ..., \varepsilon_{N,t}]^{'}; t = 1, 2, \ldots, n$ are weakly stationary processes, such that $\mathbb{E}[\boldsymbol{\varepsilon}_t] = \mathbb{E}[\boldsymbol{\varepsilon}_t | \mathcal{F}_{t-1}] = 0$ and

$$\mathbb{E}\left[\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}'_{t+s} | \mathcal{F}_{t-1}\right] = \begin{cases} \boldsymbol{\Sigma}_\varepsilon, & \text{for } s = 0; \\ 0, & \text{otherwise.} \end{cases}$$

$\mathcal{F}_{t-1}$ represents the past information available at time t.

A special class of the STARFIMA model defined in Eq. 1 is the space-time autoregressive moving averaged (STARMA), obtained when $\boldsymbol{d} = \boldsymbol{0}$. It was proposed by Cliff & Ord (1975) and broadly studied by Pfeifer & Deutsch (1980*a*,*b*,*c*), Stoffer (1986) and Antunes & Subba Rao (2006) among others.

The representation given in Eq. 1 is akin to that used in multivariate ARFIMA models, also known as VARFIMA. Both models consider the intrinsic relationships between the processes under study and have $N \times N$ coefficient matrices. The fundamental difference between them is the fact that in STARFIMA models, the spatial dependencies are imposed a priori by the model builder using a weighting matrix. Therefore, the coefficient matrices are simpler since they are products of scalars and known weighting matrices. However, as pointed out by Antunes & Subba Rao (2006), the parameters of the STARFIMA model cannot be obtained from the parameters of a VARFIMA model. Therefore, the STARFIMA is not a special case of the multivariate ARFIMA models except for the particular case when both models have the same orders.

2.1 The spatial weighting matrix

The definition of the weighting matrix $\boldsymbol{W}_l$ is non-trivial and can be rather arbitrary. There are several suggestions to define the weights of $\boldsymbol{W}_l$, all of them depend on the regularity of the grid. For example, when the grid is regularly spaced, uniform weights can be used, i.e., $w_{ij}^{(k)} = \frac{1}{n_{ki}}$ if the sites i and j are $k-$th order neighbors, and zero otherwise. The value n_{ki} represents the number of k-th order neighbors at the $i-$th site (Besag 1974).

One widely used approach is based on the definition of the weights as the inverse of the Euclidean distances between sites (Cliff & Ord 1981). It is specially useful when the sampled sites are not on a regular grid. In this case, defining weighting matrices of higher spatial order is not an easy task. As pointed out by Gao & Subba Rao (2011), to avoid these difficulties, all sites may be considered as the first order neighbors of each other site. That is, it can be assumed the spatial orders $\lambda_k = 1$ and $m_k = 1$ for all k. In such a case, there are only two weighting matrices: $\boldsymbol{W}_0 = \mathbb{I}_N$ and $\boldsymbol{W}_1 = \boldsymbol{W}$. Other ways to define the weighting matrix can be found in Bennet (1979), Anselin & Smirnov (1996) and Garrido (2000) among others.

Since in most of practical applications the sites are not scattered on a regular grid, here we consider the weighting matrices based on irregularly spaced sites, i.e. we only consider weighting matrices up to first order neighborhood. In this case, the STARFIMA process in Eq. 1 is simplified to the STARFIMA$(p_1; \boldsymbol{d}; q_1)$ given by

$$\boldsymbol{\Phi}_{p,1}(B)\boldsymbol{Z}_t = \boldsymbol{\Theta}_{q,1}(B)\mathcal{D}(B)^{-1}\boldsymbol{\varepsilon}_t, \quad t = 1, 2, \ldots, n, \tag{3}$$

where $\boldsymbol{\Phi}_{p,1}(z) = \mathbb{I}_N - \sum_{k=1}^{p}(\phi_{k0}\mathbb{I}_N + \phi_{k1}\boldsymbol{W})z^k$, $z \in \mathbb{C}$ and $\boldsymbol{\Theta}_{q,1}(z) = \mathbb{I}_N - \sum_{k=1}^{q}(\theta_{k0}\mathbb{I}_N + \theta_{k1}\boldsymbol{W})z^k$, $z \in \mathbb{C}$.

2.2 Properties of the STARFIMA$(p_1; \boldsymbol{d}; q_1)$ process

The values of ϕ_{k0}, ϕ_{k1} and $\boldsymbol{W}$ must keep stationary and causal conditions in order to assure the existence of a unique solution of he difference equations representing the process. The space-time ARFIMA process is said to be *causal* (or stable) if there is an equivalent infinite moving average representation. Additionally, the process is *invertible* if it can be expressed as an infinite order autoregressive process. The conditions for stationarity and invertibility of the STARFIMA$(p_1; \boldsymbol{d}; q_1)$ process are given by the Theorem 1. Then, the Theorem 2 defines the functions for analyzing the space-time dependence structure of the process in time and frequency domains, respectively.

Theorem 1. *Let $\boldsymbol{Z}_t$ the STARFIMA process defined in Eq. 3 with $d_i \in (-1, 0.5)$, $i = 1, 2, \ldots, N$. Then,*

a. if $\det\left\{\mathbb{I}_N - \sum_{k=1}^{p}\left(\phi_{k0}\mathbb{I}_N + \phi_{k1}\boldsymbol{W}\right)z^k\right\} \neq 0$, *for* $|z| \leq 1$ *with* $z \in \mathbb{C}$, *there is a unique stationary condition solution of Eq. 3 given by*

$$\boldsymbol{Z}_t = \sum_{j=0}^{\infty}\boldsymbol{\Psi}_j\boldsymbol{\varepsilon}_{t-j} \tag{4}$$

where $\boldsymbol{\Psi}(z) = \boldsymbol{\Phi}_{p,1}(z)^{-1}\boldsymbol{\Theta}_{q,1}(z)\mathcal{D}(z)^{-1}$.

b. if $\det\left\{\mathbb{I}_N - \sum_{k=1}^{p}\left(\phi_{k0}\mathbb{I}_N + \phi_{k1}\boldsymbol{W}\right)z^k\right\} \neq 0$, *for* $|z| \leq 1$ *with* $z \in \mathbb{C}$, *the process is said to be causal.*

c. if $\det\left\{\mathbb{I}_N - \sum_{k=1}^{q}(\theta_{k0}\mathbb{I}_N + \theta_{k1}\boldsymbol{W})z^k\right\} \neq 0$, *for* $|z| \leq 1$ *with* $z \in \mathbb{C}$, *the process is said to be invertible.*

Theorem 2. *Let $\boldsymbol{Z}_t$ a causal and invertible STARFIMA process with representation in Eq. 3 and $d_i \in (-1, 0.5)$, $i = 1, 2, \ldots, N$.*

a. The space-time covariance function is

$$\gamma_{lk}(s) = tr\left[\frac{\boldsymbol{W}_k'\boldsymbol{W}_l\boldsymbol{\Gamma}(s)}{N}\right], \quad k, l = 0, 1, \tag{5}$$

where $tr[A]$ is the trace of the square matrix A. The function $\gamma_{lk}(s)$ represents the covariance between the l and k order neighbors at the time lag s and the $\mathbf{\Gamma}(s)$ matrix is such that

$$\mathbf{\Gamma}(s) \sim diag\{s^{d_1-0.5}, s^{d_2-0.5}, \ldots, s^{d_N-0.5}\}\boldsymbol{A}\ diag\{s^{d_1-0.5}, s^{d_2-0.5}, \ldots, s^{d_N-0.5}\}, \quad s \to \infty,$$

where the $(i,j)th$ element of the $N \times N$ matrix $\boldsymbol{A}$ is

$$\frac{\Gamma(1-d_i-d_j)}{\Gamma(d_j)\Gamma(1-d_j)}\boldsymbol{\pi}_i'\mathbf{\Sigma}_\varepsilon\boldsymbol{\pi}_j,$$

$\Gamma(\cdot)$ the gamma function, $\boldsymbol{\pi}_j$ the jth row of $\mathbf{\Phi}_{p,1}(1)^{-1}\mathbf{\Theta}_{q,1}(1)$ and the symbol "$\sim$" means that the ratio of left- and right-hand sides tends to 1.

b. The spectral matrix density function $\boldsymbol{f}(\omega)$ at ω frequency, is given by

$$\boldsymbol{f}(\omega) = \mathcal{D}\left(e^{i\omega}\right)^{-1}\boldsymbol{f}_{ST}(\omega)\left[\mathcal{D}\left(e^{i\omega}\right)^{-1}\right]^*, \tag{6}$$

with $\boldsymbol{f}_{ST}(\omega) = \frac{1}{2\pi}\mathbf{\Phi}_{p,1}\left(e^{i\omega}\right)^{-1}\mathbf{\Sigma}_\varepsilon\left[\mathbf{\Phi}_{p,1}\left(e^{i\omega}\right)^{-1}\right]^*$ and $\boldsymbol{M}^*$ represents the conjugate transpose of the complex matrix $\boldsymbol{M}$. The matrix function $\boldsymbol{f}_{ST}(\omega)$ represents the space-time (ST) spectral density of the vector.

It can be seen that the dependence structure of the process is influenced by the memory parameter. Furthermore, as $s \to \infty$, the autocovariances die out as a hyperbolic rate.

Note that, as $\omega \to 0^+$ we have

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

where G is a symmetric and positive definite matrix. Hence, the espectral density defined in Eq. 6 is such that $f(\omega) \sim \Lambda(\omega;d)G\Lambda^*(\omega;d)$ where $\Lambda(\omega;d) = \mathcal{D}\left(1-\omega e^{i\frac{(\omega-\pi)}{2}}\right)^{-1}$ and the symbol "$\sim$" means that the ratio of left- and right-hand sides tends to 1. In this case, to estimate the vector of parameters $\boldsymbol{d} = (d_1, d_2, \ldots, d_N)$ we may apply the existing results for vector ARFIMA models.

2.3 Parameter estimation

The procedure of parameter estimation is carried out in two steps. In the first step, we consider the semiparametric estimation of the vector $\boldsymbol{d} = (d_1, d_2, \ldots, d_N)'$ in a neighborhood of the origin, based on the local Whittle estimator suggested by Kunsch (1987) and widely studied in a series of papers by Robinson(1995a, 1995b, 2008). Having estimated the memory parameters, the data must be filtered in order to obtain the data that will be analyzed.

In the second step, we estimate the vector of parameters of the STARMA model for the transformed data from the first step by using the methodology developed by Pfeifer & Deutsch (1980*a*).

2.3.1 Memory estimates

Let $\boldsymbol{I}(\omega_j)$ be the periodogram matrix function of $\boldsymbol{Z}_t$ evaluated at Fourier frequencies $\omega_j = \frac{2\pi j}{n}$ and given by

$$\boldsymbol{I}(\omega_j) = \frac{1}{2\pi n}\left(\sum_{t=1}^{n} \boldsymbol{Z}_t e^{it\omega_j}\right)\left(\sum_{t=1}^{n} \boldsymbol{Z}_t e^{it\omega_j}\right)^{*}, \tag{7}$$

where $j = 1, 2, \ldots, \lfloor \frac{n}{2} \rfloor$ and $\lfloor \cdot \rfloor$ denotes the integer part. The periodogram function is an estimator of the espectral density function of the process $\boldsymbol{Z}_t$ and it can be rapidly computed by fast Fourier transform, even when n is quite large.

The local approximation of the Gaussian log-likelihood function at the origin is given by

$$\mathcal{Q}(G, \boldsymbol{d}) = \frac{1}{m}\sum_{j=1}^{m}\left\{\log\det\left\{\Lambda(\omega_j;\boldsymbol{d})G\Lambda^{*}(\omega_j;\boldsymbol{d})\right\} + \mathrm{tr}\left[\Lambda(\omega_j;\boldsymbol{d})G\Lambda^{*}(\omega_j;\boldsymbol{d})\boldsymbol{I}(\omega_j)^{-1}\right]\right\},$$

where $\boldsymbol{I}(\omega_j)$ is defined in Eq. 7, $m \in [1, n/2]$ is a bandwidth number which satisfies at least $\frac{1}{m} + \frac{m}{n} \to 0$ as $n \to \infty$ (e.g., $m = o(n)$ and tends to infinity as $n \to \infty$, but at a slower rate than n) and "tr" denotes the trace of a matrix. The local Whittle estimator of the vector $\boldsymbol{d}$ is defined as

$$\widehat{\boldsymbol{d}} = \arg\min_{\boldsymbol{d}} \mathcal{R}(\boldsymbol{d}), \tag{8}$$

where $\mathcal{R}(\boldsymbol{d}) = \mathcal{Q}(\widehat{G}; \boldsymbol{d}) = \log\det\widehat{G}(\boldsymbol{d}) - \frac{2}{m}\sum_{i=1}^{N}\sum_{j=1}^{m} d_i \log\omega_j$, and

$$\widehat{G} = \frac{1}{m}\sum_{j=1}^{m}\mathbf{Re}\left\{\Lambda(\omega_j;\boldsymbol{d})^{-1}\boldsymbol{I}(\omega_j)\Lambda^{*}(\omega_j;\boldsymbol{d})^{-1}\right\}$$

and **Re** denotes the real part of a complex number.

Lobato (1999) derived the semi-parametric two-step estimator in a multivariate long memory model, by extending the work by Robinson (1995a) on the univariate local Whittle (LW) estimator, initially proposed by Kunsch (1987). Shimotsu (2007) shows that the estimator of Lobato (1999) is consistent since the spectral density representation is more precise, and the limiting distribution is more evolved. Therefore, it follows that the estimator of Shimotsu (2007) has a smaller limiting distribution than the two-step estimator of Lobato (1999). Under some regularity conditions, Shimotsu (2007) established the asymptotic normality of the Gaussian semi-parametric estimator of multivariate stationary fractionally integrated processes in Eq. 8, i.e.,

$$m^{1/2}(\widehat{\boldsymbol{d}} - \boldsymbol{d}_0) \xrightarrow{\mathcal{D}} N(0, \Omega^{-1}),\ \Omega = 2\left[G^0 \odot (G^0)^{-1} + \mathbb{I}_N + \frac{\pi^2}{4}(G^0 \odot (G^0)^{-1} - \mathbb{I}_N)\right],$$

$\widehat{G}(\widehat{\boldsymbol{d}}) \xrightarrow{p} G^0$, where $\odot$ denotes the Hadamard product and the true parameter values are denoted by $\boldsymbol{d}_0$ and G^0. Nielsen (2011) extend the results, presented by Shimotsu (2007), to cover non-stationary values of $\boldsymbol{d}$ by using the notion of the extended discrete Fourier transform. The author established the central limit theorem under the same argument as in the stationary case $|d_i| < \frac{1}{2}$, $i = 1, \ldots, N$, derived by Robinson (1995a), for the univariate case, and Shimotsu (2007), for the multivariate case,

for $d_i \in \left(-\frac{1}{2}, \infty\right)$, $i = 1, \ldots, N$.

3 Empirical Results

We conducted a simulation study aiming to explore the behavior of the proposed estimation methodology for different values of the parameters and weighting matrices.

We assume a STARFIMA$(1_1, \boldsymbol{d}, 0)$ process with four variables. The considered weighting matrix is based on the real data matrix obtained for the monitoring stations analyzed in Section 4. It is given by:

$$\boldsymbol{W}^{(1)} = \begin{pmatrix} 0.00 & 0.40 & 0.25 & 0.35 \\ 0.40 & 0.00 & 0.30 & 0.30 \\ 0.30 & 0.55 & 0.00 & 0.15 \\ 0.08 & 0.20 & 0.78 & 0.00 \end{pmatrix}.$$

The data were generated assuming combinations of the parameters $\phi_{10} = 0.1, 0.12$; $\phi_{11} = 0.1, 0.51$ and $\boldsymbol{d} = \{(0,0,0,0), (0.0, 0.1, 0.1, 0.2), (0.1, 0.1, 0.3, 0.3), (0.45, 0.45, 0.45, 0.45)\}$, in order to reflect different assumptions about them. These values of the parameters jointly with the specifications of the matrix $\boldsymbol{W}$ are such that the causality condition is satisfied. The combinations $(\phi_{01}, \phi_{11}) = (0.1, 0.1)$, $(0.12, 0.1)$ lead to the maximal absolute eigenvalue of the matrix $(\phi_{10}\mathbb{I}_N + \phi_{11}\boldsymbol{W})$[1] equal to 0.58, whilst the combinations $(\phi_{01}, \phi_{11}) = (0.1, 0.51)$, $(0.12, 0.51)$, lead to the maximal absolute eigenvalue 0.99. Sample sizes were set to $n = \{300, 1000\}$ and bandwidth $m = \lfloor n^{\alpha} \rfloor$, were $\alpha \in \{0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$. The mean and MSE were computed using 1000 replications. Due to space issues, we present the results for $m = n^{0.5}$ since this value lead to the least bias of the estimates. The remaining results are available upon request.

Here we concentrate on the performance of the memory parameter estimates, since the behavior of the parameter estimates from the second step of the estimation procedure are highly influenced by the estimates of $\boldsymbol{d}$. Studies on the performance of the parameter estimates for the STARMA processes (second step) have been conducted by Subba Rao & Antunes (2003), Giacomini & Granger (2004) and Borovkova et al. (2008) among others.

Table 1 shows the estimates of the memory parameter when there is no long-range dependence ($\boldsymbol{d} = \boldsymbol{0}$), i.e., the classic STARMA case. It can be observed that the estimates are close to the real value when the maximal eigenvalues of the matrix $(\phi_{01} + \phi_{11}\boldsymbol{W})$ are within the unit circle, even for the smaller sample size. Nevertheless, when the eigenvalues are close to 1, the bias increases significantly for small sample sizes. In this case, even a small raise of the ϕ_{01} parameter causes an increase of the bias. The MSE stays stable for all combinations of the parameters.

When there is long-range dependence and the processes are stationary (Tables 2 and 3), the simulation results show that, as n increases, the bias of the $\boldsymbol{d}$ estimates tends to decrease. For those models which the maximal eigenvalues are close to 1, the bias is large even for larger sample sizes. As in the case of the STARMA process, a small increase of the ϕ_{10} parameter leads to a significant increasing of the bias at a slower rate if the sample size is greater. The MSE remains stable for all the cases.

Table 4 displays the performance of the estimates when the memory parameter is close to the non stationary region. In this case, the bias is significantly large even for the larger sample sizes. The performance of the estimates get poorer when the maximal eigenvalues are close to 1.

[1] This condition is analogous to the causality condition in Theorem 1.

Table 1: Memory parameter values and estimates for the STARMA$(1_1, 0)$ process ($\boldsymbol{d} = \mathbf{0}$).

n	300				1000			
ϕ_{01}	0.10		0.12		0.10		0.12	
ϕ_{11}	0.10	0.51	0.10	0.51	0.10	0.51	0.10	0.51
Mean	0.0309	0.1133	0.0314	0.1267	-0.0162	0.0115	-0.0161	0.0161
MSE	0.0394	0.0484	0.0393	0.0495	0.0245	0.0208	0.0245	0.0208
Mean	-0.0084	0.1176	-0.0075	0.1310	-0.0176	0.0188	-0.0172	0.0245
MSE	0.0342	0.0325	0.0341	0.0321	0.0234	0.0179	0.0233	0.0174
Mean	0.0250	0.1220	0.0257	0.1347	0.0028	0.0363	0.0031	0.0431
MSE	0.0251	0.0314	0.0251	0.0323	0.0196	0.0179	0.0196	0.0172
Mean	-0.0134	0.0928	-0.0128	0.1046	0.0029	0.0387	0.0034	0.0427
MSE	0.0407	0.0577	0.0408	0.0585	0.0197	0.0202	0.0195	0.0207

Table 2: Memory parameter values and estimates for the STARFIMA$(1_1, \boldsymbol{d}, 0)$ process

$\boldsymbol{d}$	n	300				1000			
	ϕ_{01}	0.10		0.12		0.10		0.12	
	ϕ_{11}	0.10	0.51	0.10	0.51	0.10	0.51	0.10	0.51
0.0	**Mean**	0.0295	0.1174	0.0300	0.1327	-0.0169	0.0120	-0.0168	0.0183
	MSE	0.0396	0.0406	0.0395	0.0383	0.0248	0.0186	0.0248	0.0181
0.1	**Mean**	0.0950	0.1986	0.0959	0.2277	0.0837	0.1246	0.0842	0.1297
	MSE	0.0337	0.0269	0.0336	0.0315	0.0228	0.0182	0.0227	0.0179
0.1	**Mean**	0.1266	0.1981	0.1274	0.2237	0.1028	0.1446	0.1031	0.1495
	MSE	0.0242	0.0377	0.0242	0.0380	0.0191	0.0162	0.0191	0.0163
0.2	**Mean**	0.1851	0.2880	0.1855	0.3026	0.2045	0.2477	0.2048	0.2516
	MSE	0.0420	0.0392	0.0423	0.0430	0.0175	0.0172	0.0174	0.0169

4 Application: daily average PM_{10} in GVR

In this section, we apply the developed methodology to daily average PM_{10} concentrations ($\mu g/m^3$). We compare the fitting and forecasting ability of the proposed STARFIMA model with the performance of the STARMA model with no consideration about the memory properties of the PM_{10} time series.

The raw series consists of observations from June 15, 2008 to December 31, 2009, obtained from six monitoring stations of the Automatic Air Quality Monitoring Network (AAQMN) in the Greater Vitória Region, Brazil. Thus, we have $N = 6$ sites and $n = 560$ observations in time. Figures 1 and 2 show the locations of the sites and the time series obtained from each one of them, respectively.

We estimated the missing values using the Gibbs sampling for multiple imputations of the incomplete multivariate data suggested by Aerts et al. (2002). The first 546 observations were used for modeling purposes and the last 14, corresponding to the last two weeks of the full period, were used for forecasting purposes.

Since the region has a small number of stations distributed irregularly over a relatively small area,

Table 3: Memory parameter values and estimates for the STARFIMA($1_1, \boldsymbol{d}, 0$) process

$\boldsymbol{d}$	n	300				1000			
	ϕ_{01}	0.10		0.12		0.10		0.12	
	ϕ_{11}	0.10	0.51	0.10	0.51	0.10	0.51	0.10	0.51
0.1	**Mean**	0.1348	0.2249	0.1355	0.2383	0.0859	0.1307	0.0860	0.1376
	MSE	0.0377	0.0429	0.0376	0.0428	0.0238	0.0215	0.0238	0.0208
0.1	**Mean**	0.0976	0.2146	0.0985	0.2295	0.0869	0.1299	0.0874	0.1344
	MSE	0.0341	0.0347	0.0341	0.0336	0.0226	0.0196	0.0225	0.0197
0.3	**Mean**	0.3280	0.4048	0.3285	0.4169	0.3046	0.3514	0.3050	0.3561
	MSE	0.0242	0.0363	0.0244	0.0374	0.0183	0.0159	0.0183	0.0159
0.3	**Mean**	0.2895	0.4064	0.2901	0.4187	0.3048	0.3415	0.3050	0.3467
	MSE	0.0445	0.0468	0.0445	0.0474	0.0176	0.0184	0.0176	0.0183

Table 4: Memory parameter values and estimates for the STARFIMA($1_1, \boldsymbol{d}, 0$) process

$\boldsymbol{d}$	n	300				1000			
	ϕ_{01}	0.10		0.12		0.10		0.12	
	ϕ_{11}	0.10	0.51	0.10	0.51	0.10	0.51	0.10	0.51
0.45	**Mean**	0.4895	0.5893	0.4903	0.6021	0.4550	0.4749	0.4551	0.4809
	MSE	0.0421	0.0437	0.0420	0.0460	0.0246	0.0212	0.0246	0.0211
0.45	**Mean**	0.4697	0.5764	0.4706	0.5898	0.4463	0.4744	0.4464	0.4785
	MSE	0.0309	0.0382	0.0308	0.0376	0.0233	0.0165	0.0233	0.0162
0.45	**Mean**	0.4849	0.5535	0.4855	0.5681	0.4568	0.4959	0.4570	0.4999
	MSE	0.0227	0.0412	0.0228	0.0400	0.0185	0.0164	0.0186	0.0163
0.45	**Mean**	0.4515	0.5476	0.4524	0.5585	0.4635	0.4945	0.4638	0.4992
	MSE	0.0527	0.0510	0.0524	0.0526	0.0170	0.0184	0.0171	0.0185

we consider the weighting matrix $\boldsymbol{W}$ as suggested by Gao & Subba Rao (2011). Then we obtain

$$\boldsymbol{W} = \begin{bmatrix} 0.0000 & 0.4879 & 0.2292 & 0.1066 & 0.0872 & 0.0891 \\ 0.3887 & 0.0000 & 0.3355 & 0.1076 & 0.0818 & 0.0864 \\ 0.2031 & 0.3732 & 0.0000 & 0.1762 & 0.1183 & 0.1292 \\ 0.0850 & 0.1077 & 0.1586 & 0.0000 & 0.2212 & 0.4275 \\ 0.0989 & 0.1164 & 0.1513 & 0.3145 & 0.0000 & 0.3189 \\ 0.0768 & 0.0934 & 0.1256 & 0.4618 & 0.2424 & 0.0000 \end{bmatrix}.$$

The analysis of the periodograms of the series from each station (Figure 3) reveals that there are some significant periods at each site. Following Antunes & Subba Rao (2006), we subtracted the cyclical component in each time series individually. Denoting by $\boldsymbol{Y}_t$ the original time series, the transformed series can be written as $\boldsymbol{Z}_t = \boldsymbol{Y}_t - \boldsymbol{X}_t$, where $\boldsymbol{X}_t = [X_{1,t}, \ldots, X_{6,t}]'$ is a periodic function that can be represented as harmonic series, that is

$$X_{i,t} = \sum_{k=1}^{s} \left[\xi_{i,k} \cos\left(\frac{2\pi k t}{p_k}\right) + \xi_{i,k}^{\dagger} \sin\left(\frac{2\pi k t}{p_k}\right) \right], \; i = 1, \ldots, 6, \; t = 1, \ldots, n$$

where $\xi_{i,k}$ and $\xi_{i,k}^{\dagger}$ are unknown parameters which have to be estimated by least squares and p_k represents the periods of the time series.

Once the transformed series Z_t were obtained, we proceed to differentiate them by using the

Figure 1: Map of the studied AAQMN monitoring stations in the Greater Vitória Region.

approach presented in Section 2.3.1. These filtered series are the time series to be used for modeling. The estimates of the memory parameters were obtained using different bandwidth values $m = \lfloor n^{\alpha} \rfloor, \alpha \in \{0.4, 0.5, 0.6\}$. The estimates showed to be stable across the bandwidth values, inspired on the results showed by the simulation procedures, we decided to chose the estimates for $\alpha = 0.5$. Here we only present the results for this bandwidth, however the results for the other m values are available upon request. Thus, the estimates are $\widehat{\boldsymbol{d}} = (0.47, 0.40, 0.31, 0.38, 0.35, 0.49)$. From the estimates, it can be observed that the series in all the monitoring stations have long memory behavior and are stationary.

The temporal order is chosen by analyzing the space-time autocorrelation (STACF) and partial autocorrelation (STPACF) functions (Figures 4a and 4b). The cutting-off in the STFAC and STPACF after the second time lag suggest that a suitable model is a STARFIMA with maximum order 2 for the AR and MA components. There are some significant partial correlations at the first spatial lag, which indicates that this spatial order in the autoregressive component should be included.

The model with the best performance for the filtered series is the STARFIMA$(2_{10}, \widehat{\boldsymbol{d}}, 0)$ with estimates of the parameters given by[2]: $\phi_{10} = 0.1060$ (0.01978), $\phi_{20} = 0.1101$ (0.02697) and $\phi_{11} = -0.0980$ (0.01981). The STACF of the residuals, displayed in Figure 5, shows very small autocorrelation values, suggesting that the assumption of uncorrelated errors is satisfied by the fitted model.

According to the model, the influence of the PM_{10} over the region is around 1-2 days. The concentrations of the pollutant are highly influenced by the concentrations observed in the site and its neighbors the day before ($\phi_{10} = 0.1060$ and $\phi_{11} = -0.0980$).

STARMA Modeling

Considering the STARMA modeling methodology, the model with the best performance is the STARMA$(2_{10}, 0)$ with estimated parameters $\phi_{10} = -0.3372$ (0.0198), $\phi_{20} = -0.1029$ (0.0269) and $\phi_{11} = -0.0987$ (0.0198). The STACF of the residuals (not shown here, but available upon request) indicate that the model is adequate for the data.

[2]The standard deviations are shown in parentheses.

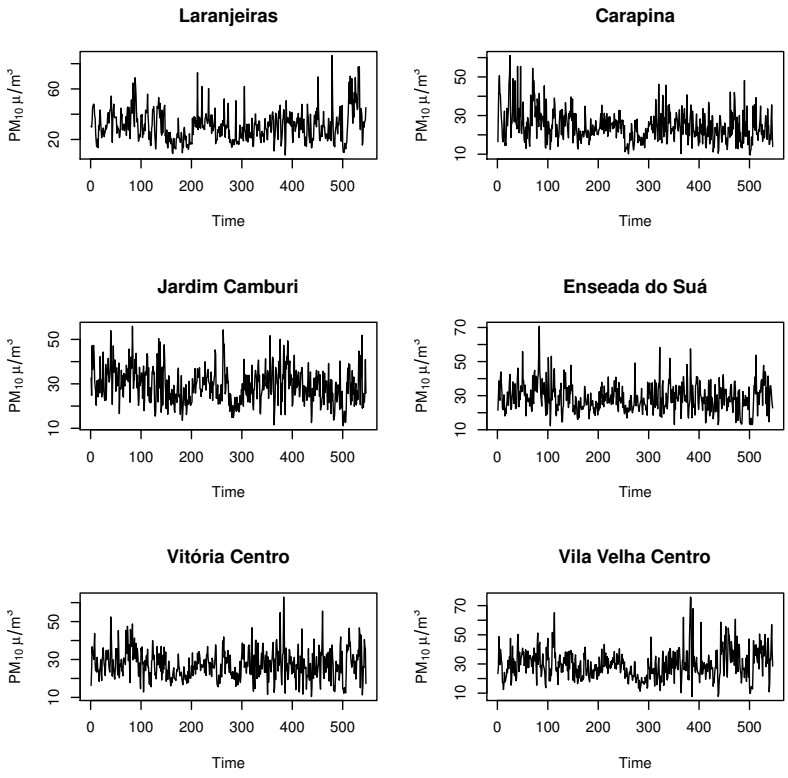


Figure 2: Time series obtained for each monitoring station.

Performance comparison

Figure 6 displays the predicted values of the observed time series by using the two fitted models. Figure 6b shows the superior in-sample performance of the STARFIMA model. It can be considered as a more suitable method for estimating missing data than the STARMA model (Figure 6a) because it can predict the larger values with more accuracy.

Regarding to the forecasting ability, we obtained one-step-ahead forecasts for a 14-days period using the Minimum Mean Square Error (MMSE) criterion. Figure 7 displays the forecasts and their 95% prediction intervals. The forecasts obtained using the STARMA model follow well the behavior of the time series (Figure 7a), nevertheless, the model cannot capture the variability with good reliability. In this sense, the results showed in Figure 7b show that the performance of the STARFIMA model is superior for all the sites.

Aiming to quantify the forecasting ability for each monitoring station, we calculated the root mean squared error (RMSE) for both models. As observed in Table 5, taking into account the memory characteristics in the model led to an improvement of the accuracy of, at least, 38%. For example, the RMSE of Vila Velha Centro obtained using the STARMA model is 1.39 times the RMSE obtained using the STARFIMA methodology. Similarly, the RMSE for Enseada do Suá station using the STARMA model is 1.78 times the RMSE obtained with the STARFIMA model, which means an approximately 78% improving of the forecasting performance.

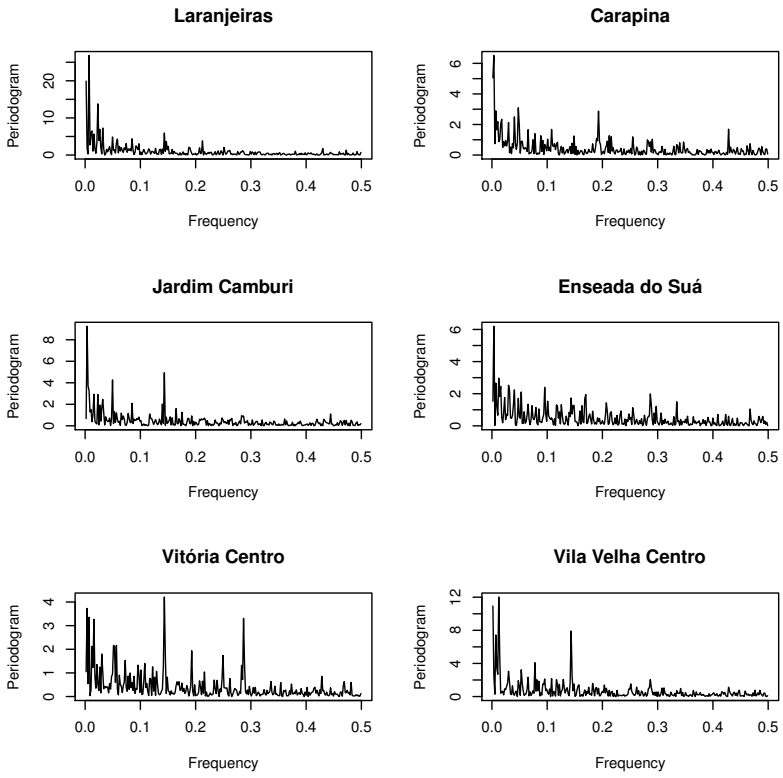


Figure 3: Periodograms for the time series at each monitoring station.

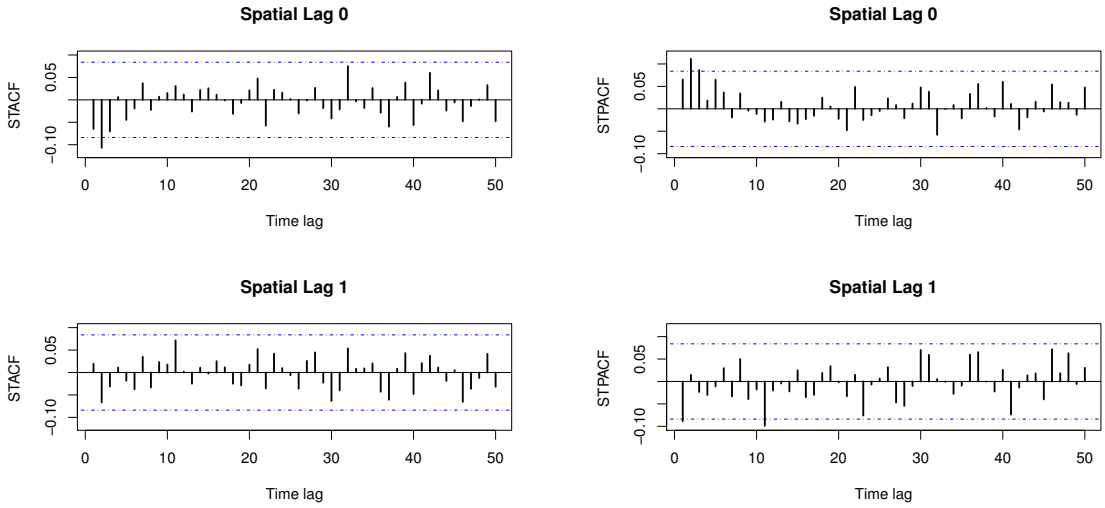
5 Final Remarks

This study presents the space-time ARFIMA model as a suitable alternative for modeling air pollution data. The developed methodology is applied to daily average PM_{10} concentrations in order to describe the dynamics of the pollutant at the Greater Vitória Region, as well as to forecast future concentrations.

According to the fitted model, the persistence of the PM_{10} in the region is about two days and its concentration levels are highly influenced by the levels observed at the closest sites the day before. The residual analysis indicated a good fit for in-sample observations, so that it can be used for imputation of missing values. Regarding the out-of-sample performance, the model showed to be a very good tool for predicting future values.

Table 5: Model accuracy measures for both fitted models.

Station	**STARMA**$(2_{10}, 0)$	**STARFIMA**$(2_{10}, \widehat{d}, 0)$
Laranjeiras	5.5767	3.2323
Carapina	2.6455	1.5250
Jardim Camburi	4.6144	2.9156
Enseada do Suá	5.9992	3.3684
Vitória Centro	4.8821	3.2148
Vila Velha Centro	3.3488	2.4147



(a) STACF (b) STPACF

Figure 4: Space-time Autocorrelation (STACF) and Partial Autocorrelation (STPACF) Functions for the differenced PM_{10} daily average.

Acknowledgements

This work was performed under the CAPES financial support.
Prof. Tata Subba Rao thanks the University of Manchester, UK and CRRAO AIMSCS, University of Hyderabad Campus,India.
Prof. Valdério Reisen thanks FAPES and CNPq for the financial support.
The authors would like to thank the Instituto Estadual de Meio Ambiente e Recursos Hídricos (IEMA) of Espírito Santo State for providing the data.

References

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

Anselin, L. & Smirnov, O. (1996), 'Efficient algorithms for constructing proper higher order spatial lag operators', *Journal of Regional Science* **36**(1), 67–89.

Antunes, A. & Subba Rao, T. (2006), 'On hypotheses testing for the selection of spatio-temporal models', *Journal of Time Series Analysis* **27**(5), 767–791.

Bennet, R. (1979), *Spatial time series, analysis-forecasting-control*, Holden-Day, Inc. San Francisco, CA.

Besag, J. S. (1974), 'Spatial interaction and the statistical analysis of lattice system', *Journal of the Royal Statistical Society, Series B* **36**, 197–242.

Borovkova, S., Lopuhaa, H. & Ruchjana, B. (2008), 'Consistency and asymptotic normality of least squares estimators in generalized star models', *Statistica Neerlandica* pp. 1–27.

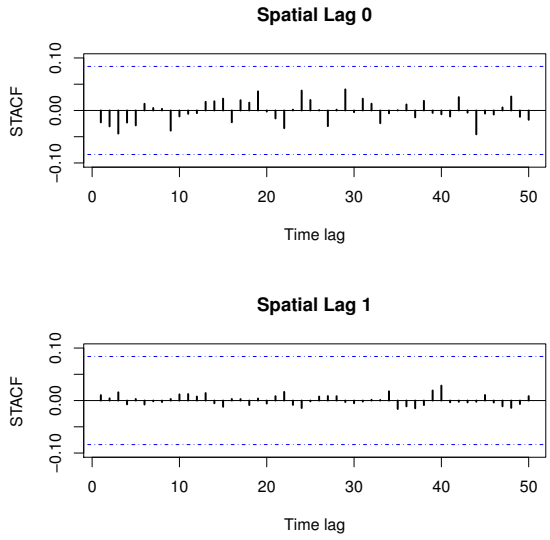


Figure 5: Space-time Autocorrelation Function (STACF) of the residuals from the fitted STARFIMA$(2_{10}, \widehat{\boldsymbol{d}}, 0)$ model.

Brockwell, P. & Davis, R. (2006), *Time Series: Theory and Methods*, second edn, Springer.

Cliff, A. & Ord, J. (1975), 'Space-time modeling with an application to regional forecasting', *Transactions of the Institute of British Geographers* **64**, 119–128.

Cliff, A. & Ord, J. (1981), *Spatial Processes: Models and Applications*, London: Pion.

Crespo, J., Zorrilla, M., Bernardos, P. & Mora, E. (2007), 'A new image prediction model based on spatio-temporal techniques', *Visual Computer* **23**(6), 419–431.

De-Iaco, S., Myers, D. & Posa, D. (2003), 'The linear coregionalization model and the product-sum space-time variogram', *Mathematical Geology* **35**(1), 25–38.

Deutsch, S. & Ramos, J. (1986), 'Space-time modeling of vector hydrologic sequences', *Water Resources Bulletin* **22**, 967–980.

Epperson, B. (1994), 'Spatial and space-time correlations in systems of subpopulations with stochastic migration', *Theoretical Population Biology* **46**, 106–197.

Epperson, B. (2000), 'Spatial and space-time correlations in ecological models', *Ecological Modelling* **132**, 63–76.

Gao, X. & Subba Rao, T. (2011), Regression models with starma errors: An application to the study of temperature variations in the antarctic peninsula, *in* M. T. Wells & A. SenGupta, eds, 'Advances in Directional and Linear Statistics', Physica-Verlag HD, pp. 27–50.

Garrido, R. (2000), 'Spatial interaction between the truck flows through the mexico-texas border', *Transportation Research Part A* **34**, 23–33.

Giacinto, V. D. (2006), 'A generalized space-time ARMA model with an application to regional unemployment analysis in italy', *International Regional Science Review* **29**(2), 159–198.

Giacomini, R. & Granger, C. (2004), 'Aggregation of space-time processes', *Journal of Econometrics* **118**, 7–26.

Glasbey, C. & Allcroft, D. (2008), 'A spatiotemporal auto-regressive moving average model for solar radiation', *Journal of the Royal Statistical Society: Applied Statistics* **57**(3), 343–355.

Granger, C. & Joyeux, R. (1980), 'An introduction to long memory time series models and fractional differencing', *Journal of Time Series Analysis* **1**, 15–30.

Hosking, J. (1981), 'Fractional differencing', *Biometrika* **68**, 165–167.

Huerta, G., Sansï¿½, B. & Stroud, J. (2004), 'A spatiotemporal model for Mexico City ozone levels', *Applied Statistics* **53**(2), 231–248.

Kamarianakis, Y. & Prastacos, P. (2005), 'Space-time modeling of traffic flow', *Computers and Geosciences* **31**, 119–133.

Kunsch, H. (1987), Statistical aspects of self-similar processes, *in* 'Proceedings of the First World Congress of the Bernoulli Society', Prokhorov, Yu., Sazanov, V.V., pp. 67–74.

LaValle, P., Lakhan, V. & Trenhaile, A. (2001), 'Space-time series modelling of beach and shoreline data', *Environmental Modelling & Software* **16**, 299–307.

Lobato, I. N. (1999), 'A semiparametric two-step estimator in a multivariate long-memory model', *Journal of Econometrics* **90**, 129–153.

Nielsen, F. (2011), 'Local whittle estimation of multivariate fractionally integrated processes', *Journal of Time Series Analysis* **32**, 317–335.

Pace, R., Barry, R., Clapp, J. & Rodriguez, M. (1998), 'Spatiotemporal autoregressive models of neighborhood effects', *Journal of Real Estate Finance and Economics* **17**(1), 15–33.

Palma, W. (2007), *Long-memory time series: theory and methods*, Vol. 662, Wiley. com.

Pfeifer, P. & Deutsch, S. (1980*a*), 'A comparison of estimation procedures for the parameters of the star model', *Communications in Statistics, Serie B - Simmulation and Computation* **9**(3), 255–270.

Pfeifer, P. & Deutsch, S. (1980*b*), 'Identification and interpretation of first order space-time ARMA models', *Technometrics* **22**(3), 397–408.

Pfeifer, P. & Deutsch, S. (1980*c*), 'A three-stage iterative procedure for space-time modeling', *Technometrics* **22**(1), 35–47.

Reynolds, K. & Madden, L. (1988), 'Analysis of epidemics using spatio-temporal autocorrelation', *Phytopathology* **78**(2), 240–246.

Reynolds, K., Madden, L. & Ellis, M. (1988), 'Spatio-temporal analysis of epidemic development of leather rot of strawberry', *Phytopathology* **78**(2), 246–252.

Robinson, P. (1995a), 'Gaussian semiparametric estimation of long range dependence', *The annals of statistics* **23**, 1630–1661.

Robinson, P. (1995b), 'Log-periodogram regression of time series with long range dependence', *The annals of statistics* **23**, 1048–1072.

Robinson, P. (2008), 'Multiple local whittle estimation in stationary systems', *The Annals of statistics* **36**, 2508–2530.

Rouhani, S., Ebrahimpour, M., Yaqub, I. & Gianella, E. (1992), 'Multivariate geostatistical trend detection and network evaluation of space-time acid deposition data – I. Methodology', *Atmospheric Environment. Part A. General Topics* **26**(14), 2603 – 2614.

Seber, G. A. (2008), *A matrix handbook for statisticians*, Vol. 15, Wiley. com.

Shimotsu, K. (2007), 'Gaussian semiparametric esimation of multivariate fractionally integrated processes', *Journal of Econometrics* **137**, 277–310.

Soni, P., Chan, Y., Preissl, H., Eswaran, H., Wilson, J., Murphy, P. & Lowery, C. (2004), 'Spatial-temporal analysis of non-stationary fMEG data', *Neurology and Clinical Neurophysiology* **100**, 1–6.

Stoffer, D. (1986), 'Estimation and identification of space-time ARMAX models in the presence of missing data', *Journal of the American Statistical Association* **81**(395), 762–772.

Subba Rao, T. & Antunes, A. (2003), Spatio-temporal modelling of temperature time series: a comparative study, *in* 'Time Series Analysis and Applications to Geophysical Systems', The IMA volumes in Mathematics and its Applications, pp. 123–150.

Terzi, S. (1995), 'Maximum likelihood estimation of a generalized star$(1, 1_p)$ model', *Statistical Methods and Applications* **4**(3), 377–393.

Yu, T.-Y. & Chang, I.-C. (2006), 'Spatiotemporal features of severe air pollution in Northern Taiwan', *Environmental science and pollution research international* **13**(4), 268–275.

Zeri, M., Oliveira-Júnior, J. & Lyra, G. (2011), 'Spatiotemporal analysis of particulate matter, sulfur dioxide and carbon monoxide concentrations over the city of Rio de Janeiro, Brazil', *Meteorology and Atmospheric Physics* **113**, 139–152.

A Appendix

To prove Theorem 1, the following results are used.

Definition A1. (Definition 19.3 in Seber (2008)). Let $\{\boldsymbol{A}_n\}$ $(n = 1, 2, \ldots)$ be a sequence of $N \times N$ matrices and let $a^n_{i,j}$ denote the (i, j)th element of $\{\boldsymbol{A}_n\}$. The sequence $\{\boldsymbol{A}_k\}$ converges to $\boldsymbol{A} = (a_{i,j})$, that is $\lim_{n\to\infty} \boldsymbol{A}_n = \boldsymbol{A}$, if $\lim_{n\to\infty} a^n_{i,j} = a_{i,j}$, $\forall\ i, j$, when $n \to \infty$.

Lemma A1. *Let $\boldsymbol{A}_n$ $(n = 1, 2, \ldots)$ be a sequence of $N \times N$ matrices. Furthermore, let a_n be a sequence of positive numbers. Then, $\boldsymbol{A}_n = O(a_n)$ if and only if $a^n_{i,j} = O(a_n)$ where $a^n_{i,j}$ denotes the (i, j)th element of $\{\boldsymbol{A}_n\}$.*

Proof of Theorem 1. a. For a fixed location $i = 1, 2, \ldots, N$, let $Y_{i,t} = \sum_{j=0}^{\infty} \eta_j \varepsilon_{i,t-j}$ be a random variable at the site i where η_j are the coefficients of the (i, i)-th entry of the diagonal matrix $[\mathcal{D}(B)]^{-1}$, that is, η_j are the coefficients of $(1 - B)^{-d_i}$ $(\eta^i_j = O(j^{d_i - 1}))$ and $\varepsilon_{i,t}$ is the white noise process of the N-dimensional vectors $\boldsymbol{\varepsilon}_t$ with $\mathbb{E}[\varepsilon_{i,t}] = 0$, $t = 1, ...T$ and $\mathbb{E}[\varepsilon^2_{i,t}] = \sigma^2_i$. For $d_i < 1/2$, $\forall i = 1, \ldots, N$, it follows that $\sum_{j=0}^{\infty} \eta^2_j < \infty$ and, therefore, $\sum_{j=0}^{T} \eta_j e^{\iota\omega j}$ converge to $(1 - e^{\iota\omega})^{-d_i}$ as $T \to \infty$ in the Hilbert space $\mathbf{L}^2(d\omega)$ and $d\omega$ denotes the Lebesgue measure. By Theorems 4.10.1 and 1.4 in Brockwell & Davis (2006) and Palma (2007), respectively, the process $Y_{i,t}$ is

well-defined. Therefore, by Definition A1 and the above results for $Y_{i,t}$, $\boldsymbol{Y}_t = \sum_{j=0}^{\infty} \boldsymbol{\eta}_j \boldsymbol{\varepsilon}_{t-j}$ where $\sum_{j=0}^{\infty} \|\boldsymbol{\eta}_j\|_2 < \infty$, that is, $\boldsymbol{Y}_t$ is also well-defined. $\|\boldsymbol{A}\|_2$ denotes the 2-norm for the matrix $\boldsymbol{A}$ such as $\|\boldsymbol{A}\|_2^2 = \text{tr}\{\boldsymbol{A}'\boldsymbol{A}\}$.

Note that, by Lemma A1, $\boldsymbol{\eta}_j = O(j^{\max_{i=1,..,N}\{d_i\}})$, the condition

$$\det\left\{\mathbb{I}_N - \sum_{k=1}^{p} \left(\phi_{k0}\mathbb{I}_N + \phi_{k1}\boldsymbol{W}\right) z^k\right\} \neq 0$$

for $|z| \leq 1$ with $z \in \mathbb{C}$ implies that $\exists\, \xi > 0$ such that $\boldsymbol{\Phi}_{p,1}^{-1}(z)$ exits for $|z| < 1 + \xi$. Since each of the N^2 elements of $\boldsymbol{\Phi}_{p,1}^{-1}(z)$ is a rational function of z with no singularities in $|z| < 1 + \xi$, consequently $\boldsymbol{\Phi}_{p,1}^{-1}(z)$ can be written as the absolutely convergent Laurent series, that is, it has the power expansion

$$\boldsymbol{\Phi}_{p,1}^{-1}(z)\boldsymbol{\Theta}_{q,1}(z) = \sum_{j=0}^{\infty} A_j z^j = A(z) \text{ for } |z| < 1 + \xi. \tag{9}$$

Thus, by Theorem 1.5(a) in Palma (2007), the process

$$\boldsymbol{Z_t} = \boldsymbol{\Phi}_{p,1}^{-1}(B)\boldsymbol{\Theta}_{q,1}(B)\boldsymbol{Y_t} = \sum_{j=0}^{\infty} \boldsymbol{A}_j B^j. \tag{10}$$

is a stationary vector process. Consequently $\boldsymbol{A}_j(1+\xi) \to 0$ as $j \to \infty$, so there exists $K \in (0, \infty)$, independent of j, such that all components of $\boldsymbol{A}_j$ are bounded in absolute value by $K(1+\xi/2)^{-j}$, $j = 0, 1, \ldots$. This implies absolute summability of the components of the matrices $\boldsymbol{A}_j$. Moreover, by Theorem 1.5(b) in Palma (2007), the vector $\boldsymbol{Z_t}$ can be written as Eq. (4), that is,

$$\boldsymbol{Z}_t = \sum_{j=0}^{\infty} \boldsymbol{\Psi}_j \boldsymbol{\varepsilon}_{t-j} \tag{11}$$

where $\boldsymbol{\Psi}(B) = \boldsymbol{\Phi}_{p,1}^{-1}(B)\boldsymbol{\Theta}_{q,1}(B)\boldsymbol{\eta}(B)$.

Now, premultiplying Eq. 11 by $\boldsymbol{\Phi}_{p,1}(B)$ and applying Theorem 1.5 in Palma (2007), then

$$\boldsymbol{\Phi}_{p,1}(B)\boldsymbol{Z}_t = \boldsymbol{\Theta}_{q,1}(B)\boldsymbol{\eta}(B)\boldsymbol{\varepsilon}_t, \tag{12}$$

which shows that $\boldsymbol{Z}_t$ is a stationary vector process that satisfies Eqs. 1 and 4.

b. The proof of the casual property follows the same lines of the univariate case as Theorem 3.4(b) given in Palma (2007).

c. The proof that $\boldsymbol{Z}_t$ is invertible can be obtained using similar arguments of the proof in (a), excepts that conditions are required on the convergence of $[\mathcal{D}(z)]\, \dfrac{\Phi_{p,1}(z)}{\boldsymbol{\Theta}_{q,m}(z)}$.

□

To prove Theorem 1, the following results are used.

Definition A2. (Section 3 in Pfeifer & Deutsch (1980*c*)). Assuming that $\mathbb{E}\{Z_{i,t}\} = 0$, $i = 1, \ldots, N$, $t = 1, \ldots, T$, the space-time covariance function can be expressed as

$$\gamma_{lk}(s) = \mathbb{E}\left\{\frac{[\boldsymbol{W}_l\boldsymbol{Z}_t]'[\boldsymbol{W}_k\boldsymbol{Z}_{t+s}]}{N}\right\}, \quad k, l = 0, 1, \ldots \tag{13}$$

Lemma A2. *Let $\boldsymbol{Z}_t$ an i-dimensional process that has an infinite-order moving average representation as:*

$$\boldsymbol{Z}_t = \sum_{j=0}^{\infty} \boldsymbol{\Psi}_j \boldsymbol{\varepsilon}_{t-j},$$

such that

$$\boldsymbol{\Psi}_s \sim diag\{\Gamma(d)^{-1} s^{d-1}\}\boldsymbol{\Pi}, \quad as\ s \to \infty,$$

where $\Gamma(\cdot)$ is the Gamma function and $\boldsymbol{\Pi}$ is a nonsingular $N \times N$ matrix of constants that are independent of s. The notation $diag\{s^{d-1}/\Gamma(d)\}$ represents a diagonal matrix $N \times N$ with $s^{d_1-1}/\Gamma(d_1), \ldots, s^{d_N-1}/\Gamma(d_N)$ on the diagonal. Then,

$$\sum_{k=0}^{s} \boldsymbol{\Psi}_k \sim diag\{\Gamma(d+1)^{-1} s^{d}\}\boldsymbol{\Pi}, \quad as\ s \to \infty.$$

Lemma A3. *Given the assumptions of Theorem 2,*

$$\begin{aligned}
&diag\{s^{0.5-d}\}\left(\sum_{k=0}^{\infty} \boldsymbol{\Psi}_k \boldsymbol{\Sigma}_\varepsilon \boldsymbol{\Psi}_{k+s}^{'}\right) diag\{s^{0.5-d}\} \\
&= diag\{s^{0.5-d}\}\left\{\sum_{k=1}^{\infty} diag\{\Gamma(d)^{-1}k^{d-1}\}\boldsymbol{\Pi}\boldsymbol{\sigma}_\varepsilon\boldsymbol{\Pi}^{'} diag\{\Gamma(d)^{-1}(k+s)^{d-1}\}\right\} diag\{s^{0.5-d}\} \\
&+ o(1),\ \ as\ s \to \infty
\end{aligned}$$

Proof of Theorem 2. a. By Definition A2,

$$\begin{aligned}
\gamma_{lk}(s) &= \frac{1}{N}\mathbb{E}\left\{[\boldsymbol{W}_l \boldsymbol{Z}_t]^{'}[\boldsymbol{W}_k \boldsymbol{Z}_{t+s}]\right\} \quad k, l = 0, 1, \\
&= \frac{1}{N}\mathbb{E}\left\{\boldsymbol{Z}_t^{'}\boldsymbol{W}_l^{'}\boldsymbol{W}_k \boldsymbol{Z}_{t+s}\right\} = \frac{1}{N}\mathbb{E}\left\{\mathrm{tr}(\boldsymbol{Z}_t^{'}\boldsymbol{W}_l^{'}\boldsymbol{W}_k \boldsymbol{Z}_{t+s})\right\} \\
&= \frac{1}{N}\mathbb{E}\left\{\mathrm{tr}(\boldsymbol{W}_k^{'}\boldsymbol{W}_l \boldsymbol{Z}_t \boldsymbol{Z}_{t+s}^{'})\right\} = \frac{1}{N}\mathrm{tr}\left\{\mathbb{E}(\boldsymbol{W}_k^{'}\boldsymbol{W}_l \boldsymbol{Z}_t \boldsymbol{Z}_{t+s}^{'})\right\}
\end{aligned}$$

since $\mathbb{E}[Z_t] = 0$,

$$\begin{aligned}
\gamma_{lk}(s) &= \frac{1}{N}\mathrm{tr}\left\{\boldsymbol{W}_k^{'}\boldsymbol{W}_l \mathbb{E}(\boldsymbol{Z}_t \boldsymbol{Z}_{t+s}^{'})\right\} = \frac{1}{N}\mathrm{tr}\left\{\boldsymbol{W}_k^{'}\boldsymbol{W}_l \boldsymbol{\Gamma}(s)\right\} \\
&= tr\left[\frac{\boldsymbol{W}_k^{'}\boldsymbol{W}_l \boldsymbol{\Gamma}(s)}{N}\right]
\end{aligned}$$

where $\boldsymbol{\Gamma}(s) = \mathbb{E}(\boldsymbol{Z}_t \boldsymbol{Z}_{t+s}^{'})$.

From Lemmas A2 and A3, we have for the covariances $\boldsymbol{\Gamma}(s) \equiv \mathrm{cov}(\boldsymbol{Z}_t, \boldsymbol{Z}_{t+s})$ that

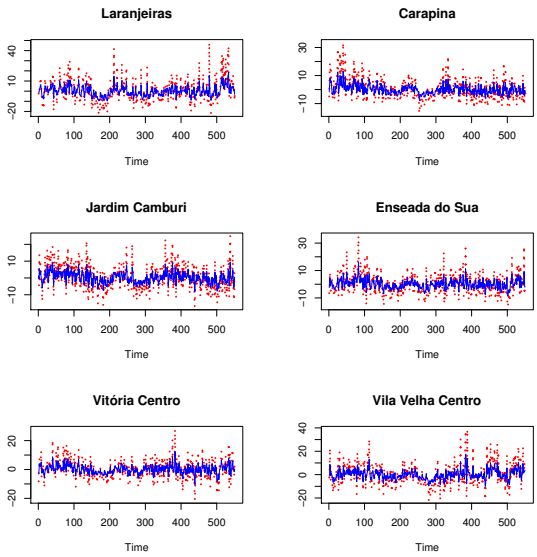
$$\begin{aligned}
&diag\{s^{0.5-d}\}\mathrm{cov}(\boldsymbol{Z}_t, \boldsymbol{Z}_{t+s}) diag\{s^{0.5-d}\} \\
&= diag\{s^{0.5-d}\}\mathrm{cov}\left(\sum_{k=0}^{\infty} \boldsymbol{\Psi}_k \boldsymbol{\varepsilon}_{t-k}, \sum_{k=-s}^{\infty} \boldsymbol{\Psi}_{k+s} \boldsymbol{\varepsilon}_{t-k}\right) diag\{s^{0.5-d}\} \\
&= diag\{s^{0.5-d}\}\left(\sum_{k=0}^{\infty} \boldsymbol{\Psi}_k \boldsymbol{\Sigma}_\varepsilon \boldsymbol{\Psi}_{k+s}^{'}\right) diag\{s^{0.5-d}\}.
\end{aligned}$$

From Lemma A3 it follows

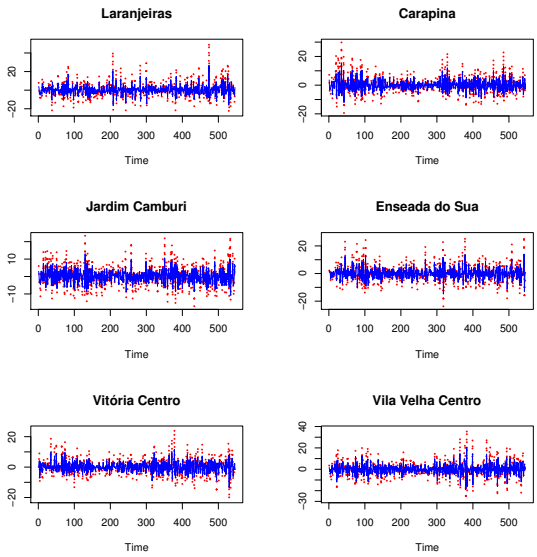
$$
\begin{aligned}
&diag\{s^{0.5-d}\}\mathrm{cov}(\boldsymbol{Z}_t, \boldsymbol{Z}_{t+s})diag\{s^{0.5-d}\} \\
&\to \int_0^\infty diag\{\Gamma(d)^{-1}z^{d-1}\}\boldsymbol{\Pi}\boldsymbol{\Sigma}_\varepsilon\boldsymbol{\Pi}^{'}diag\{\Gamma(d)^{-1}(z+1)^{d-1}\}dz \quad \text{as } s\to\infty \\
&= \left[(\pi_i^{'}\Sigma_\varepsilon\pi_k)\frac{1}{\Gamma d_i\Gamma d_k}\int_0^\infty z^{d_i-1}(z+1)^{d_k-1}dz\right], \quad i,k=1,\ldots,N.
\end{aligned}
$$

b. The proof is straightforward and is omitted here.

□

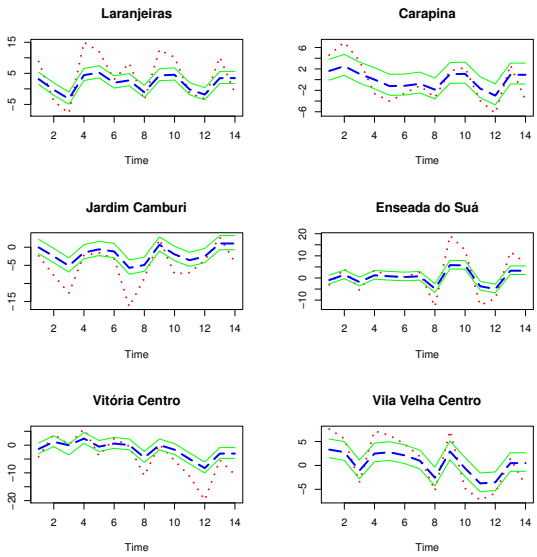


(a) STARMA$(2_{10}, 0)$

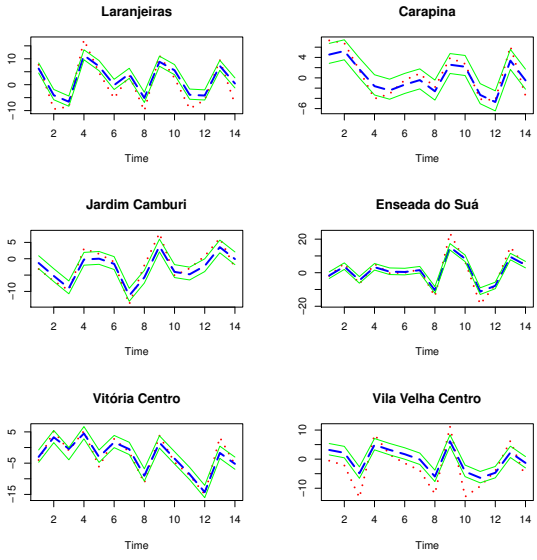


(b) STARFIMA$(2_{10}, \widehat{\boldsymbol{d}}, 0)$

Figure 6: Within-sample prediction (··· Observed concentrations — Predicted concentrations).



(a) STARMA$(2_{10}, 0)$



(b) STARFIMA$(2_{10}, \widehat{\boldsymbol{d}}, 0)$

Figure 7: Out-of-sample one-step-ahead forecasts for the transformed SO_2 time series ($\cdots$ Observed data $-\,-$ Forecasted data — 95% confidence limits for prediction interval).

6 Discussão Geral

Estudos teóricos e empíricos de modelos espaço-temporais com diferentes estruturas de dependência (curta e longa) e suas aplicações para a análise de dados de concentração de SO_2 e PM_{10} observados na Rede Automática de Monitoramento da Qualidade do Ar da RGV (RAMQAr), foram as motivações principais desta pesquisa. Os resultados evidenciaram que a dinâmica de dispersão dos poluentes estudados pode ser bem descrita usando os modelos espaço-temporais propostos, especificamente, os processos STARMA e STARFIMA. Essas classes de modelos permitiram estimar o tempo de permanência dos poluentes na atmosfera e sua influência sobre os níveis de poluição nas regiões vizinhas. O processo STARFIMA mostrou-se apropriado nas séries sob estudo, pois essas apresentaram características de longa memória no tempo. A consideração dessa propriedade no modelo conduziu a uma melhora significativa do ajuste e das previsões, no tempo e no espaço.

Os resultados principais estão apresentados em dois artigos e suas contribuições resumidas a seguir.

Pelo motivo da escassez de estudos de poluição atmosférica que envolve os modelos espaço-temporais autorregressivos de médias móveis (STARMA), pelas características da RGV e dada a distribuição espaço-temporal do poluente SO_2, o processo STARMA foi usado como aplicação de uma ferramenta alternativa na modelagem da dinâmica de dispersão de um dos poluentes que mais afeta a qualidade do ar da RGV. Os dados usados correspondem a observações de concentrações médias diárias de SO_2 obtidas de seis estações da RAMQAr. O modelo ajustado indicou que o tempo de influência do poluente na atmosfera da região é de aproximadamente 3 a 4 dias e que as concentrações observadas num local específico são afetadas não apenas pelos níveis observados em dias anteriores, mas também pelas concentrações observadas nos locais vizinhos. Por meio do modelo ajustado, foram obtidas previsões de concentrações para um dia à frente com boa precisão. Os resultados desse estudo estão no artigo Daily average sulfur dioxide in Greater Vitória Region: a space-time analysis, submetido a um periódico da área.

Com base na propriedade de memória longa, comumente encontrada em processos de dispersão atmosférica, at ese propôs a classe dos modelos espaço-temporais autorregressivos de médias móveis fracionalmente integrados (STARFIMA), uma extensão da classe de modelos STARMA. Essa vertente de pesquisa é o coração central deste trabalho com a apresentação do modelo STARFIMA e suas propriedades teóricas, do procedimento de estimação dos parâmetros e de estudos empíricos e aplicados.

O confronto entre as qualidades de ajustes dos modelos STARMA e STARFIMA nas séries de PM_{10} é a parte final desta pesquisa. Os resultados mostraram que para esse particular poluente, o modelo STARFIMA apresentou melhor performance tanto no ajuste quanto na capacidade preditiva. A comparação entre os modelos foi realizada por meio dos erros quadráticos médios EQM (estimados), da previsão de um passo à frente, calculados para cada estação de monitoramento, e o modelo SARFIMA apresentou uma redução de pelo menos 38% no valor do EQM. Esses resultados correspondem à parte aplicada do artigo Modeling and Forecasting

PM_{10} concentrations using the Space-Time ARFIMA Model, a ser submetido para o periódico *Environmetrics*.

7 Conclusões

Nesta Tese propomos a classe dos modelos ARFIMA espaço-temporais visando melhorar a precisão das previsões de concentrações médias de poluentes atmosféricos considerando não apenas a dinâmica espacial e temporal dos processos envolvidos, mas também sua estrutura de dependência temporal.

Nesse contexto, as propriedades da classe de modelos STARMA foram investigadas como um primeiro passo para o desenvolvimento da extensão do modelo para situações com comportamento de longa dependência no tempo. O modelo foi aplicado a dados de SO_2 obtidos da RAMQAr com o objetivo de descrever a dinâmica de dispersão do poluente na região assim como obter previsões um dia à frente. O modelo ajustado consegue descrever a tendência das séries temporais envolvidas no estudo, porém observa-se uma certa dificuldade para descrever adequadamente a variabilidade das mesmas.

Posteriormente, a classe dos modelos ARFIMA espaço-temporais foi proposta como uma extensão da classe dos modelos STARMA. Este modelo incorpora a estrutura de dependência dos processos sob estudo através dos parâmetros de memória definidos por Hosking (1981). Foi proposta uma metodologia de estimação semi-paramêtrica em duas etapas e as propriedades assintóticas dos estimadores foram estudadas teoricamente e através de simulações de Monte Carlo. O modelo desenvolvido foi aplicado a dados de concentrações diárias de PM_{10} na RGV. Os resultados obtidos indicam que o modelo descreve com boa precisão a dinâmica das séries temporais sob estudo, sendo que consegue descrever não apenas a tendência das séries mas também a variabilidade com maior precisão quando comparado com os resultados obtidos pelo modelo STARMA.

Os modelos STARMA e STARFIMA foram comparados empíricamente usando a aplicação aos dados de PM_{10} quanto ao ajuste e à capacidade preditiva. Observou-se que a consideração das características de longa dependência do poluente na região conduziram a um ganho signficativo na precisão das previsões para um dia à frente.

Destaca-se que todos os desenvolvimentos e simulações foram implementados nos softwares estatísticos R Core Team (2012) e `Ox`. Os programas estão disponibilizados para quem desejar consultá-los.

8 Recomendações para trabalhos futuros

Os modelos STARMA e STARFIMA assumem estrutura de correlação espacial isotrópica. Esta suposição implica que a correlação entre estações é igual em qualquer direção. Entretanto, em problemas de dispersão de poluentes atmosféricos esta suposição é pouco realista devido à influência de características da topografia local, às condições de trânsito e presença de algumas fontes pontuais de poluição próximas às estações de monitoramento. Adicionalmente, os

eventos meteorológicos como temperatura, pressão, velocidade e direção do vento influenciam diretamente no processo de dispersão dos poluentes. Por essas razões, outras especificações da matriz de ponderações W devem ser exploradas para permitir que as correlações entre estações sejam melhor descritas nas diferentes direções. Entre as opções que podem ser exploradas para a matriz W, pode-se citar:

- ⋆ Modelagem espacial *a priori* para obter a matriz de covariâncias e usá-la como matriz de ponderações no modelo STARFIMA.
- ⋆ Modelagem STARFIMA com variáveis meteorológicas exógenas, seguindo a metodologia STARMAX proposta por Stoffer (1986).
- ⋆ Inclusão das variáveis meteorológicas relevantes usando modelos de regressão com erros STARFIMA.

Finalmente, como foi observado nos resultados da aplicação dos modelos, mesmo que a dinâmica dos poluentes seja descrita com precisão, nenhum deles consegue estimar os pontos com valores mais extremos nas séries temporais. Sugere-se o estudo de extensões de modelos com erros GARCH visando melhorar a capacidade dos modelos para descrever a alta variabilidade mostrada nos processos de dispersão de poluentes.

Referências

Abadir, K., Distaso, W. & Giraitis, L. (2007), 'Nonstationarity-extended local whittle estimation', *Journal of econometrics* **141**, 1353–1384.

Abraham, B. (1983), 'The exact likelihood function for a space time model', *Metrika* **30**, 239–243.

Ali, M. (1979), 'Analysis of stationary spatial-temporal processes: estimation and prediction', *Biometrika* **66**, 513–518.

Allcroft, D. & Glasbey, C. (2005), Starma processes applied to solar radiation, Technical report, Biomathematics and Statistics Scotland.

Antunes, A. & Subba Rao, T. (2006), 'On hypotheses testing for the selection fo Spatio-Temporal models', *Journal of Time Series Analysis* **27**(5), 767–791.

Borovkova, S., Lopuhaa, H. & Ruchjana, B. (2008), 'Consistency and asymptotic normality of least squares estimators in generalized star models', *Statistica Neerlandica* pp. 1–27.

Box, G., Jenkins, G. & Reinsel, G. (1994), *Time Series Analysis: Forecasting and Control*, third edn, Prentice Hall.

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

Brockwell, P. J. & Davis, R. A. (2006), *Time Series: Theory and Methods*, 2nd edn, Springer Series in Statistics.

Carroll, R., Chen, R., George, E., Li, T., Newton, H., Schmiediche, H. & Wang, N. (1997), 'Ozone exposure and population density in Harris County, Texas', *Journal of the American Statistical Association* **92**, 392–404.

Chen, G., Abraham, B. & Peiris, S. (1994), 'Lag window estimation of the degree of differencing in fractionally integrated time series models', *Journal of Time Series Analysis* **15**, 473–487.

Cliff, A. & Ord, J. (1975), 'Space-time modeling with an application to regional forecasting', *Transactions of the Institute of British Geographers* **64**, 119–128.

Dai, Y. & Billard, L. (1998), 'A space-time bilinear model and its identification', *Journal of Time Series Analysis* **19**(6), 657–679.

Dai, Y. & Billard, L. (2003), 'Maximum likelihood estimation in space time bilinear models', *Journal of Time Series Analysis* **24**(1), 25–44.

De-Iaco, S., Myers, D. & Posa, D. (2003), 'The linear coregionalization model and the product-sum space-time variogram', *Mathematical Geology* **35**(1), 25–38.

Deutsch, S. & Pfeifer, P. (1981), 'Space-time ARMA modeling with contemporaneously correlated innovations', *Technometrics* **23**(4), 401–409.

Epperson, B. (1993), 'Spatial and space-time correlations in systems of subpopulations with genetic drift and migration', *Genetics* **133**, 711–727.

Epperson, B. (1994), 'Spatial and space-time correlations in systems of subpopulations with stochastic migration', *Theoretical Population Biology* **46**, 106–197.

Epperson, B. (2000), 'Spatial and space-time correlations in ecological models', *Ecological Modelling* **132**, 63–76.

Fernandez-Cortés, A., Calaforra, J., Jiménez-Espinoza, R. & Sánchez-Martos, F. (2006), 'Geostatistical spatiotemporal analysis of air temperature as an aid to delineating thermal zones in a potencial show cave: Implications for environmental management', *Journal of Environmental Management* **81**, 371–383.

Fox, R. & Taqqu, M. S. (1986), 'Large-sample properties of parameters estimates for strongly dependent stationary gaussian time series', *The Annals of Statistics* **14**, 517–532.

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

Giacinto, V. D. (2006), 'A generalized space-time ARMA model with an application to regional unemployment analysis in italy', *International Regional Science Review* **29**(2), 159–198.

Giacomini, R. & Granger, C. (2004), 'Aggregation of space-time processes', *Journal of Econometrics* **118**, 7–26.

Glasbey, C. & Allcroft, D. (2008), 'A spatiotemporal auto-regressive moving average model for solar radiation', *Journal of the Royal Statistical Society: Applied Statistics* **57**(3), 343–355.

Granger, C. & Joyeux, R. (1980*a*), 'An introduction to long memory time series models and fractional differencing', *Journal of Time Series Analysis* **1**, 15–30.

Granger, C. W. J. & Joyeux, R. (1980*b*), 'An introduction to long-memory time series models and fractional differencing', *Journal of Time Series Analysis* **1**, 15–30.

Haas, T. (1995), 'Local prediction of a Spatio-Temporal process with an application to wet sulfate deposition', *Journal of the American Statistical Association* **90**, 1189–1199.

Haslett, J. & Raftery, A. (1989), 'Space-time modelling with long-memory dependence: assessing Ireland's wind power resource (with discussion)', *Applied Statistics* **38**, 1–50.

Hosking, J. (1981), 'Fractional differencing', *Biometrika* **68**, 165–167.

Höst, G., Omre, H. & Switzer, P. (1995), 'Spatial interpolation errors for monitoring data', *Journal of the American Statistical Association* **90**(431), 853–861.

Huerta, G., Sansó, B. & Stroud, J. (2004), 'A spatiotemporal model for Mexico City ozone levels', *Applied Statistics* **53**(2), 231–248.

Hurvich, C. M., Deo, R. & Brodsky, J. (1998), 'The mean square error of Geweke and Porter-Hudak's estimator of the memory parameter of a long-memory time series', *Journal of Time Series Analysis* **19**(1), 19–46.

Jagger, T. & Niu, X.-F. (2005), 'Asymptotic properties of ESTAR models', *Statistica Sinica* **15**, 569–595.

Kim, C. S. & Phillips, P. (2006), Log periodogram regression: the nonstationary case, Technical report, Cowles Foundation Discussion Paper, Yale University.

Kyriakidis, P. & Journel, A. (1999), 'Geostatistical space-time models: a review', *Mathematical Geololgy* **31**(6), 651–684.

LaValle, P., Lakhan, V. & Trenhaile, A. (2001), 'Space-time series modelling of beach and shoreline data', *Environmental Modelling & Software* **16**, 299–307.

Ma, C. (2005), 'Semiparametric spatio-temporal covariance models with the ARMA temporal margin', *Annals of the Institute of Statistical Mathematics* **57**(2), 221–233.

Madden, L., Reynolds, K., Pirone, T. & Raccah, B. (1988), 'Modeling of tobacco virus epidemics as spatio-temporal autoregressive integrated moving-average process', *Phytopathology* **78**(10), 1361–1366.

Niu, X.-F., McKeague, I. & Elsner, J. (2003), 'Seasonal Space-Time Models for Climate Systems', *Statistical Inference for Stochastic Processes* **6**(2), 111–133.

Niu, X.-F. & Tiao, G. (1995), 'Modeling satellite ozone data', *Journal of the American Statistical Association* **90**, 969–983.

Paez, M. & Gamerman, D. (2003), 'Study of the space-time effects in the concentration of airborne pollutants in the Metropolitan Region of Rio de Janeiro', *Environmetrics* **14**, 387–408.

Pfeifer, P. & Deutsch, S. (1980*a*), 'A comparison of estimation procedures for the parameters of the STAR model', *Communications in Statistics, Serie B - Simmulation and Computation* **9**(3), 255–270.

Pfeifer, P. & Deutsch, S. (1980*b*), 'Identification and interpretation of first order space-time ARMA models', *Technometrics* **22**(3), 397–408.

Pfeifer, P. & Deutsch, S. (1980*c*), 'Independence and sphericity tests for the residuals of space-time ARIMA models', *Communications in Statistics, Serie B - Simmulation and Computation* **9**(5), 533–549.

Pfeifer, P. & Deutsch, S. (1980*d*), 'A three-stage iterative procedure for space-time modeling', *Technometrics* **22**(1), 35–47.

Pfeifer, P. E. & Deutsch, S. (1981), 'Variance of the sample space-time correlation function of contemporaneously correlated variables', *SIAM Journal on Applied Mathematics* **40**(1), 133–136.

Phillips, P. (1999), Discrete fourier transforms of fractional processes, Technical report, Cowles Foundation Discussion Paper, Yale University.

Phillips, P. (2007), 'Unit root log periodogram regression', *Journal of econometrics* **138**, 104–124.

Phillips, P. & Shimotsu, K. (2004), 'Local whittle estimation in nonstationary and unit root cases', *Annals of statistics* **32**, 656–692.

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

R Core Team (2012), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
URL: *http://www.R-project.org/*

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

Reynolds, K. & Madden, L. (1988), 'Analysis of epidemics using spatio-temporal autocorrelation', *Phytopathology* **78**(2), 240–246.

Reynolds, K., Madden, L. & Ellis, M. (1988), 'Spatio-temporal analysis of epidemic development of leather rot of strawberry', *Phytopathology* **78**(2), 246–252.

Robinson, P. M. (1995*a*), 'Gaussian semiparametric estimation of long range dependence', *The annals of statistics* **23**, 1630–1661.

Robinson, P. M. (1995*b*), 'Log-periodogram regression of time series with long range dependence', *The annals of statistics* **23**, 1048–1072.

Sahu, S. & Mardia, K. (2005), 'A Bayesian Kriged-Kalman model for short-term forecasting of air pollution levels', *Journal of the Royal Statistical Society, Series C* **54**, 223–244.

Shaddick, G. & Wakefield, J. (2002), 'Modelling daily multivariate pollutant data at multiple sites', *Journal of the Royal Statistical Society, Series C* **51**, 351–372.

Shimotsu, K. & Phillips, P. (2005), 'Exact local whittle estimation of fractional integration', *Annals of statistics* **33**, 1890–1933.

Soni, P., Chan, Y., Preissl, H., Eswaran, H., Wilson, J., Murphy, P. & Lowery, C. (2004), 'Spatial-temporal analysis of non-stationary fMEG data', *Neurology and Clinical Neurophysiology* **100**, 1–6.

Stoffer, D. (1986), 'Estimation and identification of space-time ARMAX models in the presence of missing data', *Journal of the American Statistical Association* **81**(395), 762–772.

Velasco, C. (1999*a*), 'Gaussian semiparametric estimation of non-stationary time series', *Journal of Time Series Analysis* **20**(1), 87–127.

Velasco, C. (1999*b*), 'Non-stationary log-periodogram regression', *Journal of Econometrics* **91**, 325–371.