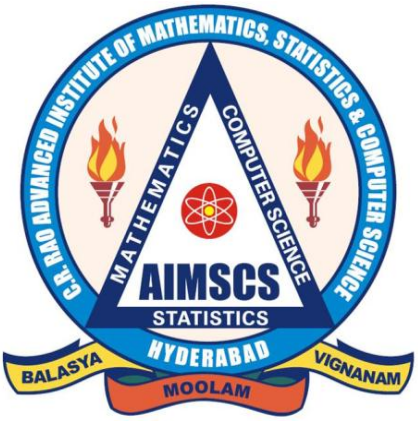
CRRAO Advanced Institute of Mathematics, Statistics and Computer Science (AIMSCS)

Research Report



Author (s): **Nataly A. Jimenez Monroy, Valderio A. Reisen, Tata Subba Rao**

Title of the Report: **Modeling and forecasting of sulfur dioxide using Space-Time Series models. A case study.**

Research Report No.: **RR2014-06**

Date: **March 13, 2014**

Prof. C R Rao Road, University of Hyderabad Campus,
Gachibowli, Hyderabad-500046, INDIA.
www.crraoaimscs.org

Modeling and forecasting of sulfur dioxide using Space-Time Series models. A case study.

Nátaly A. Jiménez Monroy[a,b], Valdério A. Reisen[a,b], Tata Subba Rao[c,d]

[a] *PPGEA – Universidade Federal do Espírito Santo, Brazil.*
[b] *DEST – Universidade Federal do Espírito Santo, Brazil.*
[c] *School of Mathematics, University of Manchester, UK.*
[d] *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 have larger percentage of observation covered than the intervals obtained under the Gaussian distribution assumption.

The fitted STARMA model indicated that the permanence 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, space-time models, STARMA, sulfur dioxide.

Email address: nataly.monroy@ufes.br (Nátaly A. Jiménez Monroy)

1. Introduction

The GVR is located on the Brazilian South Atlantic coast in the state of Espírito Santo (ES) and comprises of seven main cities, including the capital Vitória. Its population has grown significantly in the last decades as a consequence of rapid industrialization. The increase of the industrial activities, as well as the constat grown 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 such as nitrogen oxide 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 and 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, 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 of forecasting statistical models for SO_2 is based on univariate time series approaches. For example, Cheng and Lam (2000), Hassanzadeh et al. (2009), Kumar and Goyal (2011), Lalas et al. (1982), McCollister and Wilson (1975), Schlink et al. (1997). As explained by Turalioglu and Bayraktar (2005), such models are incapable of providing regional information on the spatial variations of air pollutants.

Some other researches 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 and 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 and Bayraktar (2005), Yu and 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, which can be used to direct new standards for air quality management policies and emission control at specific locations.

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 continu-

ously 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.

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

Font: IEMA

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).

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 and 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 and Deutsch (1980b). The processes which can be represented by STARMA models are characterized by a single 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 and 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, with diagonal entries 0 and off-diagonal entries related to the distances between the sites. By definition, $\mathbf{W}^{(0)} = \mathbf{I}_N$ and 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

$$E[\boldsymbol{\varepsilon}(t)] = \mathbf{0}, \tag{2}$$

$$E[\boldsymbol{\varepsilon}(t)\boldsymbol{\varepsilon}'(t+s)] = \begin{cases} \mathbf{G}, & \text{if } s = 0 \\ \mathbf{0}, & \text{otherwise} , \end{cases}$$

$$E[\mathbf{z}(t)\boldsymbol{\varepsilon}'(t+s)] = 0, \quad \text{for } s > 0,$$

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 and 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

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

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 and Deutsch (1980b) and Subba Rao and 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

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

Pfeifer and Deutsch (1980b) 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

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

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 and Deutsch (1980b), 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

$$\boldsymbol{\Phi} = [\phi_{10}, \ldots, \phi_{1\lambda_1}, \ldots, \phi_{p0}, \ldots, \phi_{p\lambda_p}]'$$

$$\boldsymbol{\Theta} = [\theta_{10}, \ldots, \theta_{1\lambda_1}, \ldots, \theta_{q0}, \ldots, \theta_{qm_q}]',$$

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

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

Finding the values of the parameters that minimize the log-likelihood function is equivalent to finding the values $\hat{\boldsymbol{\Phi}}$ and $\hat{\boldsymbol{\Theta}}$ 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 and Antunes (2003) proposed a procedure for initial estimation of the parameters of $S(\boldsymbol{\Phi}, \boldsymbol{\Theta})$ 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 and Deutsch (1980a) and Pfeifer and

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 and 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 and 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 $H_0 : \boldsymbol{\Phi} = \boldsymbol{\Theta} = \mathbf{0}$, or test the hypothesis that a particular ϕ_{kl} or θ_{kl} is zero with the remaining parameters unrestricted.

Let $\hat{\boldsymbol{\delta}} = (\hat{\boldsymbol{\Phi}}, \hat{\boldsymbol{\Theta}})' = (\delta_1, \ldots, \delta_K)'$ 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.

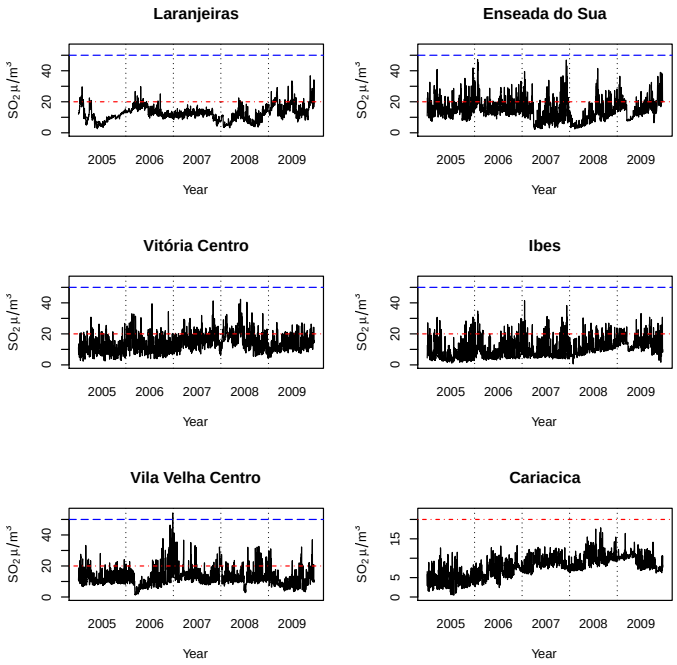


Figure 2: SO_2 daily average concentrations at the AAQMN monitoring stations (- · - 2005 WHO guideline −− 2005 WHO interim guideline).

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 and 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 and 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 a harmonic series, i.e.

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

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.

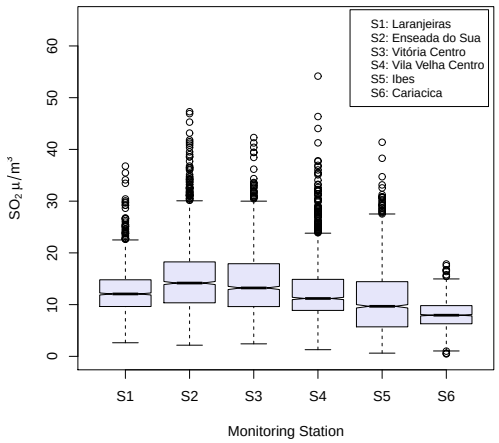


Figure 3: Boxplots of SO_2 daily average by monitoring station.

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 and 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 and Deutsch (1980b), 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

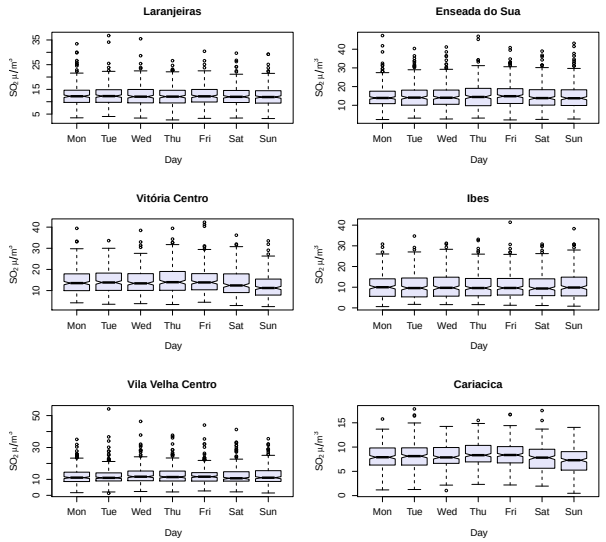


Figure 4: Boxplots of SO_2 daily average by day of the week.

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

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

There are several ways to define the weighting matrix, see Cliff and Ord (1981) and Anselin and Smirnov (1996). In particular, we chose $\mathbf{W}$ 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:

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

for $i, j = 1, 2, \ldots, 6$, where lat_i and lon_i represent the latitude and longitude of the station i, respec-

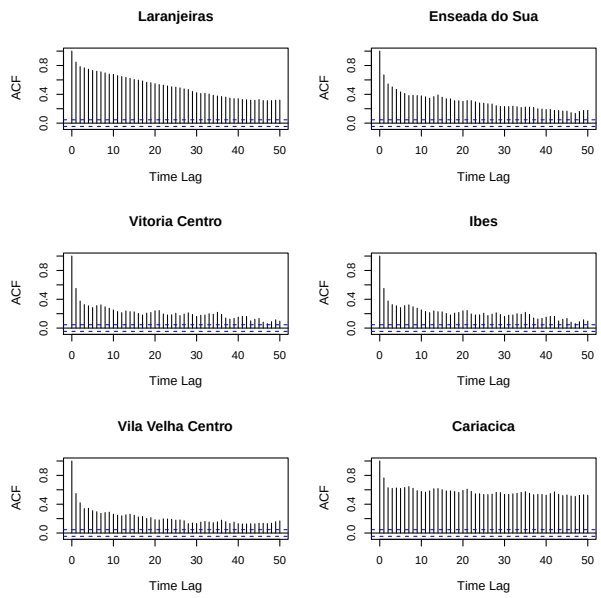


Figure 5: Autocorrelation Functions for SO_2 daily average by monitoring station.

tively (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:

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

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.

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.

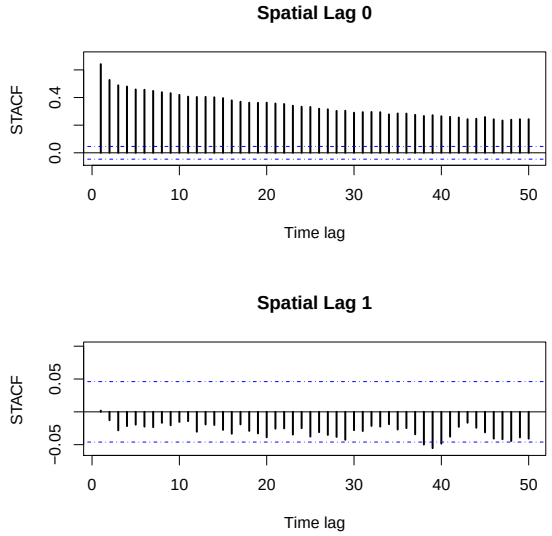


Figure 6: Space-time Autocorrelation Function (STACF) for SO_2 daily average time series.

The model with the best performance is the STAR($4_{1,0,0,0}$) with parameters (the standard

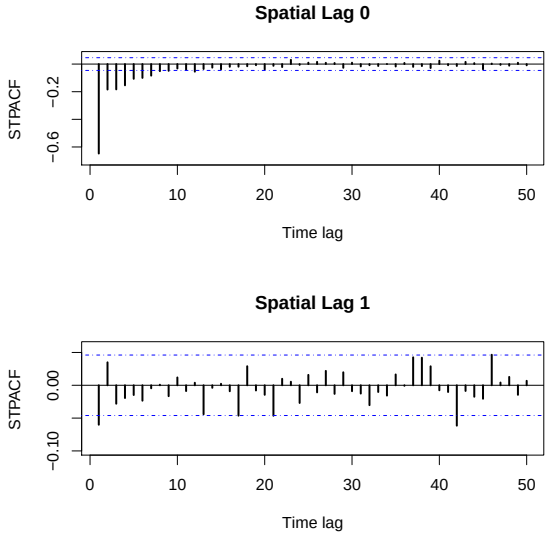


Figure 7: Partial Space-time Autocorrelation Function (STPACF) for SO_2 daily average time series.

errors are shown in brackets):

$$\phi_{10} = -0.475\,(0.0109) \qquad \phi_{11} = -0.066\,(0.0306)$$

$$\phi_{20} = -0.066\,(0.0121) \qquad \phi_{21} = \quad 0.058\,(0.0335)$$

$$\phi_{30} = -0.108\,(0.0121) \qquad \phi_{31} = -0.004\,(0.0335)$$

$$\phi_{40} = -0.156\,(0.0109) \qquad \phi_{41} = -0.019\,(0.0306)$$

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

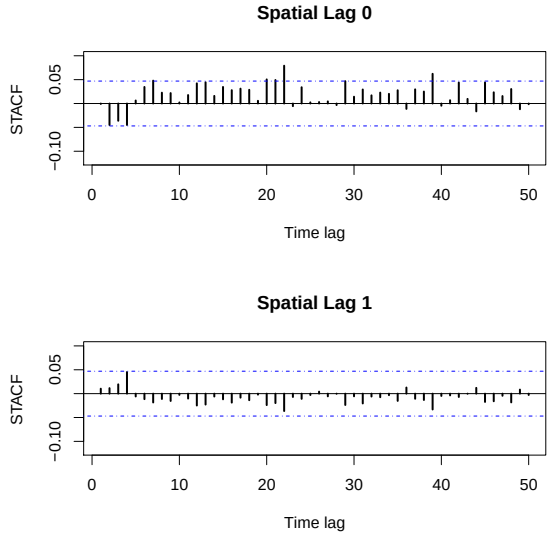


Figure 8: Space-time Autocorrelation Function (STACF) of the residuals from the fitted STARMA$(4_{1,0,0,0}, 0)$ model.

not normally distributed. The lack of Gaussian distribution affects only 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,

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

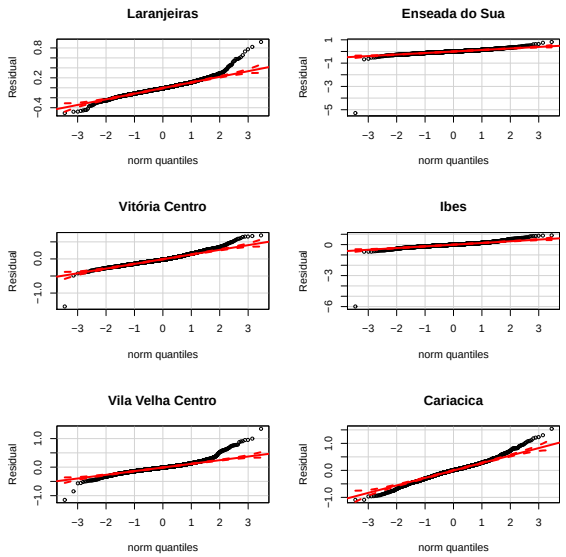


Figure 9: Quantile-quantile plot of the residuals from the fitted STARMA$(4_{1,0,0,0}, 0)$ model.

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 permanence of SO_2 in the atmosphere of 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.

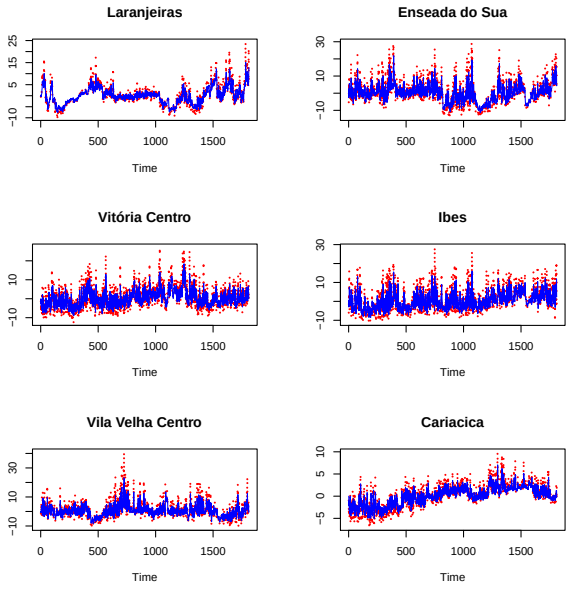


Figure 10: Within-sample prediction for the transformed SO_2 time series ($\cdots$ Observed concentrations — Predicted concentrations).

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) 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

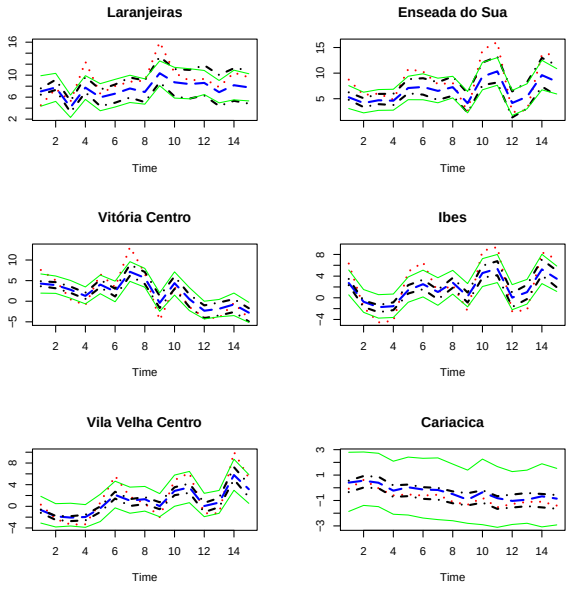


Figure 11: Out-of-sample one-step-ahead forecasts for the transformed SO_2 time series (··· Observed data – – Forecasted data · – · 95% confidence limits for Gaussian interval — 95% confidence limits for bootstrap interval).

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 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 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 and 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 $\mu g/m^3$.

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 $\mu g/m^3$.

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

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 the model is highly influenced by outliers.

Acknowledgements

This work was performed under the CAPES financial support.

Professor T. Subba Rao (Adjunct Professor, CRRAO AIMSCS) also wishes to thank the Department of Science and Technology, Government of India, for their financial support through their research grant to the institute No. SR/S4/516/07 which supported his visit to the Institute.

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. Tech. rep., 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. Tech. rep.

Instituto Estadual de Meio Ambiente e Recursos Hídricos [IEMA], 2007. Relatório da qualidade do ar na

Região da Grande Vitória 2006. Tech. rep.

Instituto Estadual de Meio Ambiente e Recursos Hídricos [IEMA], 2011. Inventário de emissões atmosféricas da Região da Grande Vitória. Tech. rep.

Instituto Jones dos Santos Neves [IJSN], 2012. Perfil do Espírito Santo. Dados gerais. Vitória − ES, 2012. Tech. rep.

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: Peña, D., Tiao, G., Tsay, R. (Eds.), A course in advanced time series analysis. J. Wiley and Sons, Ch. 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., 1980a. Identification and interpretation of first order space-time ARMA models. Technometrics 22 (3), 397–408.

Pfeifer, P., Deutsch, S., 1980b. 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. Tech. rep.

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.

Suggested Reviewer List (include up to 5 names and their contact details)

1) Maria Eduarda da Silva, Porto,
mesilva@fep.up.pt
2) Patricio Perez; Chile.
patricio.perez@usach.cl, pperez@fisica.usach.cl
3) José Francisco de Oliveira Júnior, Universidade Federal Rural do Rio de Janeiro (UFRRJ).
http://www.if.ufrrj.br/dca/dca.html
4) M. Zeri. University of Illinois, USA
mzeri@illinois.edu, mzelli@illinois.edu
5) (Yiannis) Kamarianakis
Assistant Professor
School of Mathematics & Statistical Sciences, yiannis76@asu.edu
Arizona State University
6) Yu Tai-Yi: yutaiyi@gmail.com
7) Chang I-Cheng: d2507002@hotmail.com
8) Prof. Wilfredo Palma Ph.D. Statistics ■ Carnegie Mellon University ■ 1995
Master in Statistics ■ CarnegieMellon University ■ 1992
Mathematical Civil Engineer ■ University Chile ■ 1990
Office ■ 106 Tel ■ 354-5465 Fax ■ 354-7229
Email ■ wilfredo@mat.puc.cl Department of Statistics Faculty of Mathematics
Pontificia Universidad Católica de Chile
9) Dr. Yulia Gel , ygl@math.uwaterloo.ca Department of Statistics and Actuarial Science 200 University Avenue West
Waterloo, ON, Canada N2L 3G1
(519) 888-4567 x33550