Modelling spatio-temporal data with multiple seasonalities: the NO_2 Portuguese case*

Andreia Monteiro†
andreiaforte50@gmail.com

Raquel Menezes‡
rmenezes@math.uminho.pt

Maria Eduarda Silva§
mesilva@fep.up.pt

May 2017

Abstract

This study aims at characterizing the spatial and temporal dynamics of spatio-temporal data sets, characterized by high resolution in the temporal dimension which are becoming the norm rather than the exception in many application areas, namely environmental modeling. In particular, air pollution data, such as NO_2 concentration levels, often incorporate also multiple recurring patterns in time imposed by social habits, anthropogenic activities and meteorological conditions. A two-stage modelling approach is proposed which combined with a block bootstrap procedure correctly assesses uncertainty in parameters estimates and produces reliable confidence regions for the space-time phenomenon under study. The methodology provides a model that is satisfactory in terms of goodness of fit, interpretability, parsimony, prediction and forecasting capability and computational costs. The proposed framework is potentially useful for scenario drawing in many areas, including assessment of environmental impact and environmental policies, and in a myriad applications to other research fields.

Keywords: Geostatistics; Spatio-temporal modelling; Hourly air pollution data; Multiple seasonalities

1 Introduction

It is acknowledged that air pollution is a social as well as an environmental problem, leading to a multitude of adverse effects on human health, ecosys-

*Accepted author's manuscript (AAM) to be published in [Spatial Statistics]. [DOI:10.1016/j.spasta.2017.04.005]

†CIDMA & Centre of Mathematics, Departamento de Matemática, Universidade do Minho

‡Centre of Mathematics, Departamento de Matemática, Universidade do Minho

§CIDMA & Faculdade de Economia, Universidade do Porto

tems and the built environment. Several research studies, systematic reviews and meta-analysis have been carried out to analyse health effects of air pollutants: Shin et al. (2008) considered these issues by monitoring the risk of death associated with outdoor air pollution; McCarthy et al. (2009) used ambient monitoring data to determine the relative importance of individual air toxics for chronic cancer and noncancer exposures; Lai et al. (2013) analysed the risk estimates for mortality and morbidity outcomes due to air pollutants; Keramatinia et al. (2016) studied the relationship between exposure to NO_2 and breast cancer incidence; and Song et al. (2016) conducted a systematic review to provide an association between air pollution and cardiac arrhythmia. In fact, the European Environment Agency, EEA (2015) considers air pollution the single largest environmental health risk in Europe. Thus the need for accurate assessment of air pollution arises not only to investigate the linkage between ambient exposure and health effects but also with regard to compliance with legislated regulatory standards to control levels of environmental exposure. The above considerations advance the need for statistical models aimed at characterizing and predicting air quality events and assessing policies over specified areas.

In Portugal, estimation of the index of air quality involves measurements of the following chemical elements: carbon monoxide (CO), nitrogen dioxide (NO_2), sulphur dioxide (SO_2), ozone (O_3) and fine particulate matter as PM_{10}. The index is based on the pollutant with the highest concentration relative to the Portuguese annual limit values for the protection of human health. This work focus on NO_2 concentrations, which is considered a primary pollutant, formed naturally in the atmosphere by lightning and produced by plants, soil and water Carslaw (2005). However, the major sources are the fossil fuel combustion processes, the emissions from electricity generating stations and road traffic. Furthermore, NO_2 concentration levels closely follow vehicle emissions, in many situations, thus providing a reasonable marker exposure to traffic. Nitrogen dioxide is toxic by inhalation and there is evidence that long-term exposure to NO_2 at high concentrations has adverse health effects, namely in respiratory and cardiovascular systems, Ricciardolo et al. (2004). NO_2 and other nitrogen oxides are also precursor of ozone and particulate matter, whose effects on human health and the environment are well documented. Concentrations of NO_2 have been analysed extensively in many urban areas Carslaw (2005); Grice et al. (2009); Roberts-Semple et al. (2012) as well as in background sites Donnelly et al. (2011); Menezes et al. (2016). Moreover, these studies acknowledge that meteorological conditions influence NO_2 levels Shi and Harrison (1997); Donnelly et al. (2011); Russo and Soares (2014). Thus the overall results indicate recurrent multiple seasonal patterns resulting from anthropogenic activity and the influence of meteorological variables. Fassò and Negri (2002) propose a non-linear statistical model to deal with the problem of high frequency and multiple frequency periodicities underlying environmental data

dynamics. De Livera et al. (2011) also consider complex seasonal patterns into their modelling proposals, using exponential smoothing. The former works restrict their applications to one geographical location.

This work purposes a methodology to characterize the spatial and high resolution temporal evolution of spatio-temporal data using geostatistical approaches. The approach takes into account that environmental data often incorporate distinct recurring patterns in time and considering the influence of meteorological variables. The suggested framework is applied to hourly NO_2 concentration levels in Portugal. Spatio-temporal statistical modelling aims at revealing dependencies and spatio-temporal dynamics e.g. Cameletti et al. (2011) and, in our particular case, at obtaining hourly concentration predictions over the country. To this end the model by Menezes et al. (2016) is extended to hourly data and meteorological variables are included. A block bootstrap procedure is proposed to correctly assess uncertainty of parameters estimates, as well as to produce reliable confidence regions for (space-time) NO_2 concentrations. The model is potentially useful in many areas including assessment of environmental impact and environmental policies.

The paper is organized as follows. Section 2 describes the data chosen as a motivating example, as well as the results of the preliminary study. In Section 3, we present a brief review of the spatio-temporal methodology, highlighting the proposed approach based on a geostatistical framework. In Section 4, we show the application of the previously described methodology to the characterization of NO_2 concentrations in the Portugal case. Section 5 is devoted to make space time prediction and forecasting for NO_2 concentrations, as well as to analyse scenarios. Section 6 ends up with some discussion and main concluding remarks.

2 The Portuguese dataset

This study analyses hourly measurements of NO_2 obtained from the online database on air quality Qualar (2015) of the Portuguese Environment Agency, whose mission is to propose, develop and monitor the public policies for the environment and sustainable development. The database on air quality provides hourly measurements, resulting from monitoring activities, for various pollutants, including NO_2. The available data include information about the type of the site where the station is placed (background, industrial or traffic) and the environment of the zone (urban, suburban or rural). The most serious drawback of QualAr is that validated data are only available in October of the following year.

The hourly NO_2 concentrations under analysis concern 49 stations stations located over Portugal (mainland) from October 1st to December 31st in 2014, in a total of 108192 observations. From the 49 stations, 33 are

classified as background, 10 as traffic and 6 as industrial, 29 are located in urban areas, 11 in rural areas and 9 in suburban areas. The selected period corresponds to the highest NO_2 levels along the year, according Menezes et al. (2016), who analysed NO_2 data during 8 years. This study has about 18% of missing data in the hourly levels of NO_2.

The NO_2 concentrations have a mean of 20.6 $\mu g/m^3$, standard deviation of 21.9 and median of $13\mu g/m^3$. The histogram of NO_2 concentrations, represented in Figure 1, reveals asymmetry indicating departure from Gaussianity.

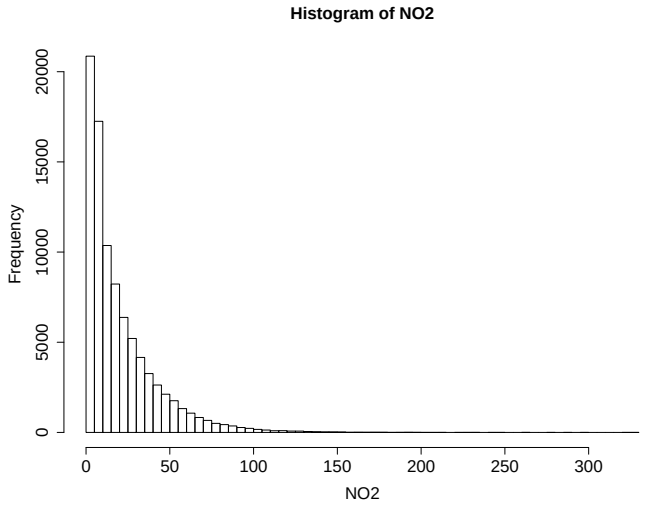

Figure 1: Histogram of NO_2 concentrations.

A periodogram analysis of the data reveals periodicities at 12, 24 and 168 hrs, which corresponds to intra-daily, daily and weekly periods. These recurring patterns are clearly observed in Figure 2, which represents mean hourly values for both weekdays and weekends. NO_2 levels show two daily peaks, one in the morning (8:00) and one in the afternoon (18:00) which coincide with rush-hour traffic, with the second peak being more pronounced than the first. Moreover, the mean NO_2 concentrations are much lower on weekends (particularly on Sunday) than on weekdays, displaying, also, smaller variations on weekends, which reflect reduced levels of vehicular emissions on non-working days. Thus, the two main seasonal effects in the data: intra-day as well as intra-week periodicities, may be, at least partially, explained by characteristics of the station. In fact, Figure 3 illustrates the influence of the location and the environment of the station in values of NO_2. It is clear that the stations located in traffic areas and urban zones present higher values for their NO_2 quartiles as well higher variability. This

analysis indicates that the type of site and the environment zone must be considered as explanatory variables.

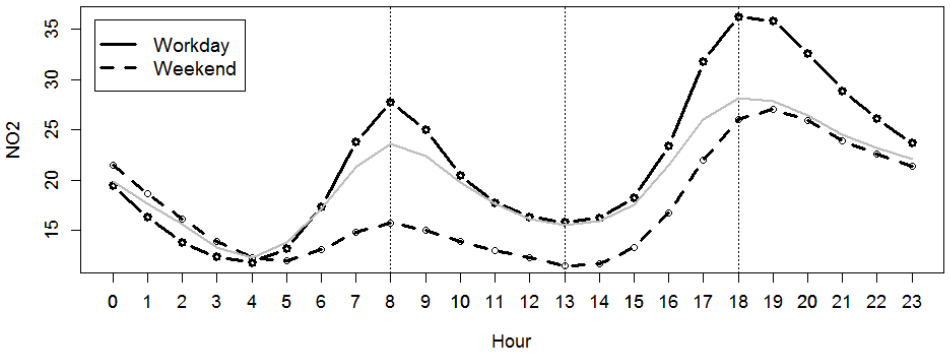

Figure 2: Mean NO_2 concentrations, for workdays and weekends. The gray line identifies the trigonometric representation based on Fourier series of the cyclical component, section 3.

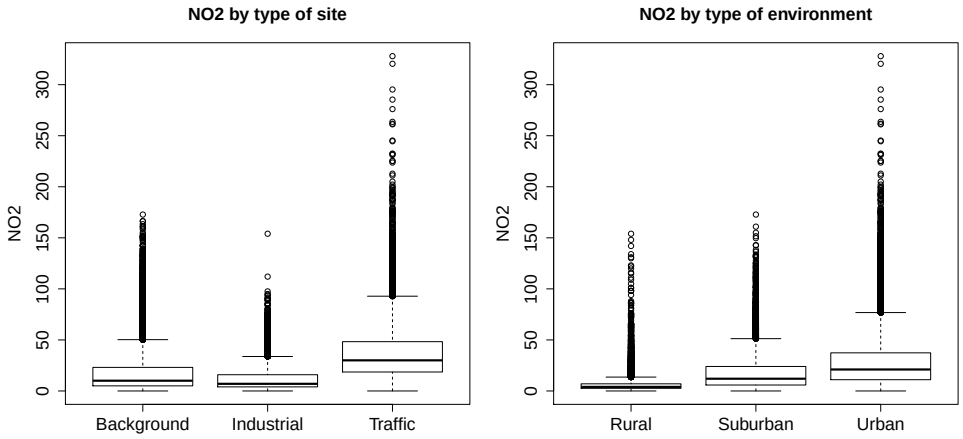

Figure 3: Boxplots of NO_2 concentrations, by type of site and type of environment.

Since it is acknowledge that meteorological variables influence NO_2 levels, hourly data from the following meteorological variables were obtained

from Weather Underground (2015), which provides weather data collected hourly from around the world: wind speed (km/h); air temperature (^{0}C) and relative humidity (%). The analysis of the correlation between these meteorological variables and NO_2 levels identified the well known negative associations among them. High NO_2 concentrations are favored by cold and drier weather; on the other hand, an increase of wind-speed, generally, promotes dilution and dissipation of the pollutants, thus yielding lower levels of NO_2, in accordance with Shi and Harrison (1997). Additionally Spearmans rank correlation coefficient between NO_2 and the meteorological variables for several lags, represented in Figure 4 indicates that the strongest correlations occur at 6-hour lag with air temperature, 1-hour lag with wind-speed and 5-hour lag with relative humidity. Therefore, these meteorological variables at the identified lags are considered as explanatory variables for NO_2 levels.

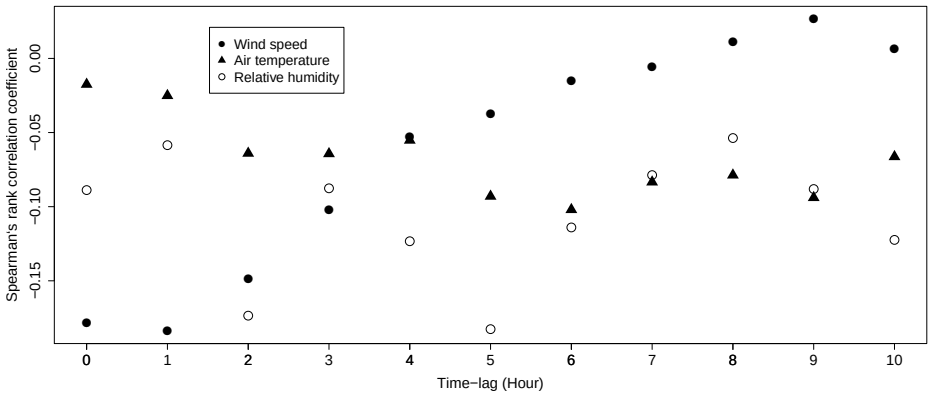

Figure 4: Spearmans rank correlation coefficient between hourly NO_2 and the meteorological variables for several lags.

The above exploratory analysis indicates two main seasonal effects in the temporal dynamics of NO_2 levels: daily and weekly. This preliminary study also shows that the variables type of site (background, industrial or traffic) and type of environment (urban, suburban or rural), together with the meteorological variables air temperature (6-hour lag), wind speed (1-hour lag) and relative humidity (5-hour lag) are possible explanatory variables for the NO_2 levels. Further analysis, not reported here, shows the presence of strong spatial dependence in the NO_2 dataset as widely reported in environmental pollution data literature.

These remarks evidence the importance of using a spatio-temporal model incorporating multiple seasonalities for describing the complex structure and

dynamics of the phenomenon.

3 Methodology

Consider a spatio-temporal stochastic process $Z(\mathbf{s}, t)$ indexed in space by $s \in \mathbb{R}^d$ and in time by $t \in \mathbb{N}$. The process can be represented as

$$Z(\mathbf{s}, t) = \mu(\mathbf{s}, t) + \delta(\mathbf{s}, t) \tag{1}$$

where $\mu(\mathbf{s}, t) = \mathrm{E}(Z(\mathbf{s}, t))$ represents a spatio-temporal mean field modelling the trend, usually refered to as the large-scale variation component and $\delta(\mathbf{s}, t)$ is a zero-mean smooth stationary spatio-temporal process that models the small-scale variation (hereafter referred to as stationary residual).

3.1 Large-scale variation

The mean component $\mu(\cdot)$ in the above model may be a deterministic function when the physics of the underlying phenomenon is known. However, in the large majority of problems and spatio-temporal data sets such knowledge is unavailable and we must resort to stochastic specifications which aim at representing the patterns of the observed variability. Accordingly, in the specification of the mean component we include regression variables observed jointly with the response variables and incorporate, also, complex nested or non nested seasonal and cyclic effects. In fact, many time series exhibit multiple seasonal patterns: hourly pollution levels reveal a daily pattern with period of 12 or 24 as well as a weekly pattern with period $24 \times 7 = 168$ and a long series might also exhibit an annual seasonal pattern with period 24×365, resulting from the natural cycles and anthropogenic activity. Thus, a flexible approach to model (1) consists on considering the generalized linear model (GLM) which combines three components:

- A random component specifying the conditional distribution of the response variable $Z(\mathbf{s}, t)$, given the values of explanatory variables. This conditional distribution may be any from the exponential family thus avoiding transformations of the response variable.

- A systematic component which specifies a linear predictor that is a function of a set explanatory variables $\mathbf{X}$

  $$\eta(\mathbf{s}, t) = \mathbf{AX} \tag{2}$$

  where $\mathbf{A}$ is a matrix of real coefficients and $\mathbf{X}$ a matrix of regressors.

  A smooth and invertible linearising link function $g(\cdot)$ which transforms the expectation of the response variable $\mathrm{E}(Z(\mathbf{s}, t)) = \mu(\mathbf{s}, t)$ into the linear predictor $\eta(\mathbf{s}, t) = g(\mu(\mathbf{s}, t))$.

Matrix $\mathbf{X}$ contains the K regression variables $X_i(\mathbf{s},t)$, $i=1,\ldots,K$ observed jointly with the response $Z(\mathbf{s},t)$, and the periodic regressors that capture the periodicities in the time series. Assume that there are L identified periods $(m_1,\ldots,m_L)$ and assume for each cyclic component at time t, $S_{t,l}$ a trigonometric representation based on Fourier series with the form $S_{t,l}=\sum_{j=1}^{k_l}\left[\phi_{j,1}\cos\left(\frac{2\pi jt}{m_l}\right)+\phi_{j,2}\sin\left(\frac{2\pi jt}{m_l}\right)\right]$, where k_l represents the number of harmonics required for the lth cyclic component. The number of periodic regressors L depends on the data under study and may be determined by frequency analysis of the time series. Thus we can write

$$\begin{aligned}\eta(\mathbf{s},t) &= \alpha+\sum_{i=1}^{K}\beta_i X_i(\mathbf{s},t)+\sum_{l=1}^{L}S_{t,l}\\ &= \alpha+\sum_{i=1}^{K}\beta_i X_i(\mathbf{s},t)+\sum_{l=1}^{L}\sum_{j=1}^{k_l}\left[\phi_{j,1}\cos\left(\frac{2\pi jt}{m_l}\right)+\phi_{j,2}\sin\left(\frac{2\pi jt}{m_l}\right)\right]\end{aligned} \tag{3}$$

where α, β_i, $\phi_{j,1}$, $\phi_{j,2}\in\mathbb{R}$ are regression parameters.

3.2 Small-scale variation

It is now necessary to consider the space-time dependence structure underlying the stationary spatio-temporal residual $\delta(\mathbf{s},t)$. Many methods have been proposed in the literature to define valid models for the spatio-temporal dependence structures e.g. Gneiting (2002); De Cesare et al. (2001). For a comparative review of the characteristics of many of these currently accepted and implemented models see De Iaco (2010). One of the main distinctions between these models is based on the notion of separability. A separable space-time covariance function can be written as the product of a purely spatial component and a purely temporal component. This allows for efficient estimation (especially computationally), and inference but the separability is restrictive and often require unrealistic assumptions Bruno et al. (2003), and a major disadvantage of these models is that they can not incorporate the space-time interaction. Thus, in our study, the attention has shifted to non-separable covariance structures, namely the product-sum and sum-metric models, which are widely used in the literature. Other parametric families of non-separable models are discussed in Cressie and Huang (1999), Ma (2008) and Rodrigues and Diggle (2010). For more general classes of non-separable covariance functions see Fonseca and Steel (2011), Ip and Li (2015).

The product-sum model can be defined in terms of the semivariogram as

$$\gamma_{\mathbf{s}t}(\mathbf{h}_s,h_t)=\gamma_{\mathbf{s}}(\mathbf{h}_s)+\gamma_t(h_t)-k\gamma_{\mathbf{s}}(\mathbf{h}_s)\gamma_t(h_t) \tag{4}$$

where γ_s and γ_t are the corresponding valid semivariogram functions in space and time, $(\mathbf{h}_s, h_t) \in \mathbb{R}$ and

$$k = \frac{sill_s + sill_t - sill_{st}}{sill_s \, sill_t}$$

where $sill_s$ and $sill_t$ represent the sill of the marginal semivariograms in space and time, repectively, and $sill_{st}$ is the global sill.
The sum-metric model can be defined:

$$\gamma_{\mathbf{st}} (\mathbf{h}_s, h_t) = \gamma_s (\mathbf{h}_s) + \gamma_t (h_t) + \gamma (|\mathbf{h}_s| + \alpha |h_t|) \tag{5}$$

with $\alpha \in \mathbb{R}$ and γ_s e γ_t the semivariograms.

3.3 Parameter estimation and inference by block bootstrap

The estimation of model (1) is accomplished in a 2-step approach which estimates separately the trend (large-scale variation) and the spatio-temporal dependence structure (small-scale variation) components. First obtain point estimates for the regression parameters using maximum likelihood (ML) and relaxing the assumption of non-correlated errors, underlying ML estimation in GLM. Then fit a valid non-separable space-time variogram to the residuals resulting from the previous step, fully accomplishing the estimation of the spatio-temporal correlation in the data.

An important issue arising in the first step as a consequence of relaxing the assumption of uncorrelated residuals is that of assessing the statistical significance of the estimated parameters. To handle this issue we resort a bootstrap procedure for serially correlated data. We consider a modification of the so called block bootstrap Kreiss and Paparoditis (2011), based on moving and overlapping blocks in the time dimension, when taking fixed data in the space dimension. The main idea consists of dividing the temporal data, $(X_1, \ldots, X_T)$ say, into blocks of consecutive observations of length l, $(X_t, \ldots, X_{t+l-1})$. The first block corresponds to $(X_1, \ldots, X_l)$ and each new block slides M time units, becoming $(X_{1+k\times M}, \ldots, X_{l+k\times M})$ with $k = 1, \ldots, K$, $M << l$ and $l + K \times M \leq T$, allowing for a total of $K + 1$ blocks. This bootstrap approach is particularly appropriate when one has long time series, as it is usually the case with hourly data, collected at a small number of geographical locations. This bootstrap approach further allows to obtain a confidence band for large-scale variation predictions.

The estimation of the parameters in the semivariograms (4) and (5) relies in a least-squares approach over a space-time empirical variogram. At this stage the sample marginal variograms in space and time, defined in De Iaco and Posa (2012) are important to give some guidance for the selection of the one-dimensional variogram components in (4) and (5). In fact, the selection of adequate models in (4) and (5) is crucial to guarantee that the resulting function is valid for prediction using kriging tools. Myers (2004)

provide some guidelines that may be useful for model selection. To evaluate the final variograms, a cross-validation approach originally introduced in Stone (1974), and meanwhile adapted to the context of dependent data, is used. This procedure consists on eliminating one observation from the whole set and then predicting its value from the remaining data through a kriging methodology. Repeating the procedure for all the observations, the Mean Square Error (MSE) of the resulting errors can be used to choose between several (variogram) models. Following an adequate choice of a spatio-temporal variogram, a block bootstrap procedure is once more resorted to correctly assess uncertainty in its parameters estimates.

4 Results

The preliminary data analysis of NO_2 concentrations in Portugal carried out in Section 2, indicates that the underlying process presents several characteristics such as non Gaussianity, multiple periodicities and spatial dependence, for which model (1) introduced in Section 3 may be particularly useful. The 2-step estimation procedure proposed is carried out leading to the characterization of the mean or large-scale variation component in Section 4.1, and that of the stationary residual or small-scale variation component in Section 4.2. The estimation procedure is implemented in R environment R Core Team (2015) and the following packages are used: gstat Pebesma (2004), sp and space-time Bivand et al. (2013).

4.1 Large-scale variation

Firstly, we model the trend of NO_2 data using a Generalized Linear Model as given in equation (3). In the case of NO_2 concentrations, exploratory analysis revealed that it is a continuous variable with an asymmetric distribution, in particular, we assume that the response variable is gamma distributed with log-link. As the gamma distribution is only defined for strictly positive values, we make a translation of the data set by 0.0001. We consider six explanatory variables: type of site, type of environment, if weekend, air temperature (6-hour lag effect), wind speed (1-hour lag effect) and relative humidity (5-hour lag effect). Other factors were also considered, like the distinction between the days of the weekend (week, Saturday and Sunday), but this did not result in significant improvements. Furthermore, we consider other hour lag effects for meteorological variables, however, the best model is selected under Akaike information criterion and by graphical observation of NO_2 fitted values vs. NO_2 levels.

For modelling the seasonal effects in the data set, we proceed as represented in (3), assuming a trigonometric representation for each cyclic component. The dominant frequencies of the data were estimated, based on those

stations without missing values, which made it possible to identify two important periodicities equal to 12 and 24 hours. Consequently, although we have tested distinct periodic regressors, including one for the weekly cycles, the simpler model restricted to the daily (or half-daily) cycles proved to be preferable.

The results of the gamma regression of the hourly NO_2 concentrations are summarized in Table 1. The standard errors were obtained using a moving block bootstrap in the time dimension, each block with 5 weeks sliding 3 hours, generating 456 replicates. All 49 monitoring stations were kept as fixed. According to the notation presented in Section 3, the block length $l = 5 \times 7 \times 24 = 840$ hours, $\delta = 3$ hours and $K = 455$. Two weeks blocks were also considered, however, these were not able to capture patterns of intra- and inter-day variability, meaning that the seasonal components became no significant in the trend model. From the results in Table 1, we conclude that the values of NO_2 concentrations are greater during the week and in monitoring stations where the environment is urban or suburban and the type of site is traffic. Besides that, NO_2 levels increase by a factor of 3.64 from rural to urban, by a factor of 1.64 from background to traffic, and by a factor of 1.22 during the week. In respect of meteorological variables, these variables have significant negative associations with NO_2 levels. These conclusions confirm the results from the preliminary data analysis. Wind speed has a stronger influence on NO_2 concentrations than humidity and air temperature. Furthermore, NO_2 level decrease 3% by an increase of 1 km/h in wind speed and decrease 1% by an increase of 1% in humidity. In the case of air temperature, the lack of significance in its coefficient was confirmed under the proposed block bootstrap approach, which can be explained by the fact that only months with low temperature are selected (in October to December, mean is 14.6^0C and standard deviation is 5.4^0C). The acquired coefficient of determination shows that 41% of the large-scale variation of NO_2 concentrations is explained under this trend model.

4.2 Small-scale variation

Having estimated the large-scale variation $\mu(\mathbf{s}, t)$ as $g^{-1}(\eta(\mathbf{s}, t))$ in (3), we now aim at estimating the dependence structure of the stationary residual $\delta(\mathbf{s}, t)$, resulting from $Z(\mathbf{s}, t) - \mu(\mathbf{s}, t)$ in (1). This issue is addressed through the approximation of the spatio-temporal variogram. The fit of the empirical variogram demands estimation of the unknown parameters of the theoretical model, namely, the nugget τ^2, the partial variance σ^2 and the range ϕ. We start by analyzing the marginal spatial and the marginal temporal correlation structures, defined in De Iaco and Posa (2012). The Gaussian model is selected for the approximation of the spatial variogram, suggesting the parameters estimates $\hat{\tau}_s^2 = 0.19$, $\hat{\sigma}_s^2 = 0.59$ and $\hat{\phi}_s = 35.47$km. For the temporal variogram, it is selected the Exponential model, and the resulting

Parameter	Estimate	Over-optimistic Std. Error$^{(*)}$	Bootstrap Std. Error
Intercept	2.452	0.019	0.259
Type of site (baseline: Background)			
Industrial	-0.517	0.009	0.026
Traffic	0.489	0.008	0.031
Day of the week (baseline: Weekend)			
Week	0.202	0.007	0.024
Environment (baseline: Rural)			
Suburban	1.147	0.010	0.128
Urban	1.310	0.008	0.153
Air Temperature	-0.008	0.0006	0.015
Wind Speed	-0.029	0.0004	0.002
Relative Humidity	-0.006	0.0002	0.002
$\sin(\frac{2\pi t}{12})$	-0.228	0.004	0.019
$\cos(\frac{2\times 2\pi t}{12})$	-0.015	0.004	0.007
$\sin(\frac{2\times 2\pi t}{12})$	0.033	0.004	0.011
$\cos(\frac{4\times 2\pi t}{12})$	0.008	0.004	0.002
$\cos(\frac{2\pi t}{24})$	0.093	0.005	0.039
$\sin(\frac{2\pi t}{24})$	-0.167	0.005	0.018
$\cos(\frac{3\times 2\pi t}{24})$	0.018	0.004	0.008
$\sin(\frac{3\times 2\pi t}{24})$	0.102	0.004	0.012
$\cos(\frac{5\times 2\pi t}{24})$	0.016	0.004	0.005
$\sin(\frac{5\times 2\pi t}{24})$	-0.021	0.004	0.004

Table 1: Estimates of the gamma regression coefficients for hourly NO_2 concentrations, together with the corresponding standard errors obtained by bootstrap. The standard errors given in (*) were obtained by GLM when relaxing the assumption of non-correlated residuals.

parameters estimates are $\hat{\tau}_t^2 = 0.60$, $\hat{\sigma}_t^2 = 0.06$ and $\hat{\phi}_t = 47.47$ hours. We examined other models, however, the results for the parameter estimates were similar.

To decide whether to adopt the product-sum model in (4) or sum-metric model in (5), we proceed with a cross-validation study to compare both models, according to which the eliminated observations are predicted through the kriging tools. For each model, we estimate the mean error (ME) and the mean square error (MSE) based on all resulting prediction errors. The results in Table 2 are very similar, however, the model sum-metric has an extra parameter for anisotropy which allows dealing with spatial and temporal distances in the same term. Besides that, the sum-metric model makes it possible to use specific variogram for space, time, and space-time. There-

fore, we decide to choose the sum-metric model with a Exponential function for the temporal component and Gaussian functions for the spatial and the spatio-temporal components.

Model	joint	temporal	space	ME	MSE
Product-sum model	-	Exp	Gau	-0.016	0.214
Sum-metric model	Gau	Exp	Gau	-0.007	0.219

Table 2: ME and MSE estimates of the cross-validation study.

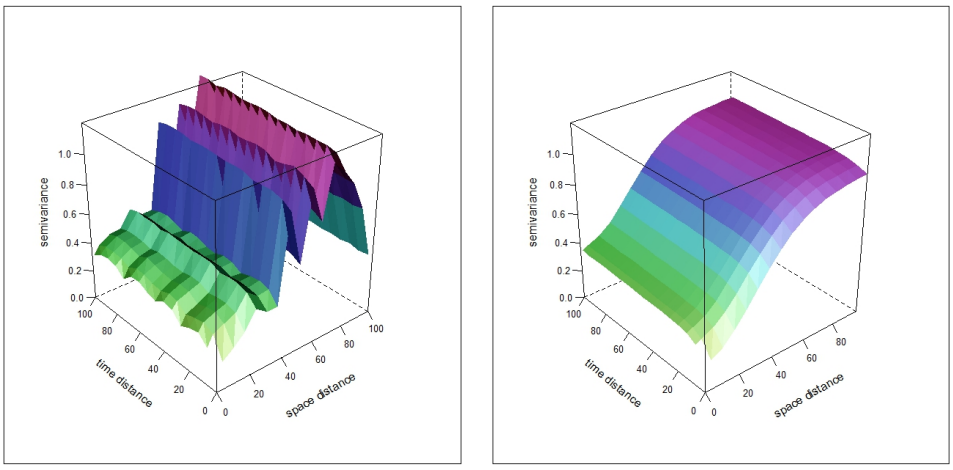

Figure 5: Plots of the experimental estimator (left) and the fitted model (right) for the space-time variogram.

Variogram	Model	τ^2	σ^2	ϕ	α
Spatial	Gau	0.015 (0.025)	0.662 (0.128)	40km (1.348)	
Temporal	Exp	0.010 (0.020)	0.071 (0.022)	100h (0.003)	
Joint	Gau	0.172 (0.018)	0.132 (0.030)	70 (0.024)	13.007 (0.074)

Table 3: Parameters estimates, and corresponding bootstrap standard errors obtained by moving block bootstrap with blocks of 5 weeks sliding 8 hours, generating 171 replicates, for the spatial, temporal and spatio-temporal variograms.

Under this selection, the fitted final model is represented in Figure 5 (right), being the corresponding empirical variogram given in the left panel. The resulting parameters estimates and corresponding standard errors, obtained by moving block bootstrap, blocks of 5 weeks sliding 8 hours, generating 171 replicates, are given in Table 3. Initially, we tried the option

of sliding 3 hours instead of 8 hours, as done for the regression coefficients estimates in the trend, but the computational cost associated to the estimation of the variogram was not acceptable. According to the results, we conclude that the majority of the total variation is explained by the spatial component. The temporal and spatio-temporal components have a smaller contribution. Furthermore, NO_2 concentrations have a significative spatial correlation up to 40 km and a temporal correlation up to 100 hours (approximately 4 days).

4.3 Model assessment

To assess the goodness of fit of the model two measures are chosen: the Mean Absolute Percentual Error (MAPE) and the Mean Absolute Scaled Error (MASE). The MAPE, being a percentage error has the advantage of being scale-independent, and so is frequently used to compare model predictive performance between different data sets, in this case stations with different environment characteristics. On the other hand, the MAPE, being a measure based on percentage errors has the disadvantage of presenting large values for observations close to zero. Hyndman and Koehler (2006) proposed the MASE as an alternative measure based on scaled errors, which, in fact, compare the error in the value predicted by the model with that of a naive prediction. The naive prediction must take into account the data seasonality.

For model assessment the predictions are defined as $\hat{Z}(\mathbf{s},t) = \hat{\mu}(\mathbf{s},t) + \hat{\delta}(\mathbf{s},t)$, where: $\hat{\mu}(\mathbf{s},t)$ is the fitted large scale variation at location $\mathbf{s}$ and time t, given climate conditions; and $\hat{\delta}(\mathbf{s},t)$ is the predicted small scale-variation, obtained under a cross-validation approach. This means that data from station at location $\mathbf{s}$ is eliminated and $\delta(\mathbf{s},t)$ is predicted from the remaining data by kriging tools. Considering T observations for any particular station $\mathbf{s}$, one has

$$\text{MAPE} = \frac{1}{T}\sum_{t=1}^{T}\frac{|e_t|}{Z(\mathbf{s},t)}\times 100\% \qquad \text{MASE} = \sum_{t=1}^{T}\frac{|e_t|}{\sum_{t=1}^{T}|Z(\mathbf{s},t) - Z(\mathbf{s},t-168)|}$$

where $e_t = \hat{Z}(\mathbf{s},t) - Z(\mathbf{s},t)$.

Since our data set is of high dimensionality, model assessment is performed for a subset of seven monitoring stations (Loures, Beato, Entrecampos, Avenida da Liberdade, Matosinhos, Lourinhã, Sonega) representative of the different types of environments during five consecutive working days: from 2014-10-13 at 0:00 (Monday) to 2014-10-17 at 24:00 hrs (Friday). Goodness of fit measures, MAPE and MASE for the seven stations are presented in Table 4. For the computation of the MASE, a more adequate measure in our case, we considered a naive prediction of the NO_2 concentration at a location, the value of the concentration at that location, at the

Station	Environment	Type	MASE	MAPE(×100%)
Loures	urban	background	0.841	0.54
Beato	urban	background	0.618	0.40
Entrecampos	urban	traffic	0.898	0.48
Avenida da Liberdade	urban	traffic	0.443	0.47
Matosinhos	suburban	background	0.623	0.51
Lourinhã	rural	background	0.874	0.38
Sonega	rural	industrial	0.757	0.68

Table 4: MASE and MAPE errors for some stations, according environment of the zone and type of the site.

same day and same time of the day of the previous week, computed for mean climate conditions of that time of day. This procedure takes into account the multiple seasonalities present in NO_2 concentrations. The MASE values range from 0.44 to 0.90 and are all less than one indicating that the model predicts more accurately that the naive predictor. There is not a clear pattern on the errors with urban, traffic stations (Avenida da Liberdade and Entrecampos) presenting the lowest and highest MASE errors. The absence of such a pattern may be explained on one hand by the high variability that hourly concentrations present and on the other hand, the low number of stations classified as rural and industrial.

The predicted large and small scales variation and observed concentration in Loures, an urban and background station, represented in Figure 6 illustrates the high variability present in the data. Although the overall mean intra-day pattern of the NO_2 concentrations is well described by the model, see Figure 2, individual stations and days present particularities that remain unexplained by the model.

Even so, this assessment exercise allows to conclude that the model provides a good enough representation of the data and can be used for out of sample prediction and scenario generation.

5 Space-time prediction and forecasting

This section illustrates the potential of the proposed spatio-temporal modelling strategy for prediction and forecasting. The former is accomplished by interpolating in the observed space-time dimension, through the kriging tools. The latter is accomplished through the mean predictor given in (3), as it allows to obtain NO_2 forecasts as a function of the explanatory variables.

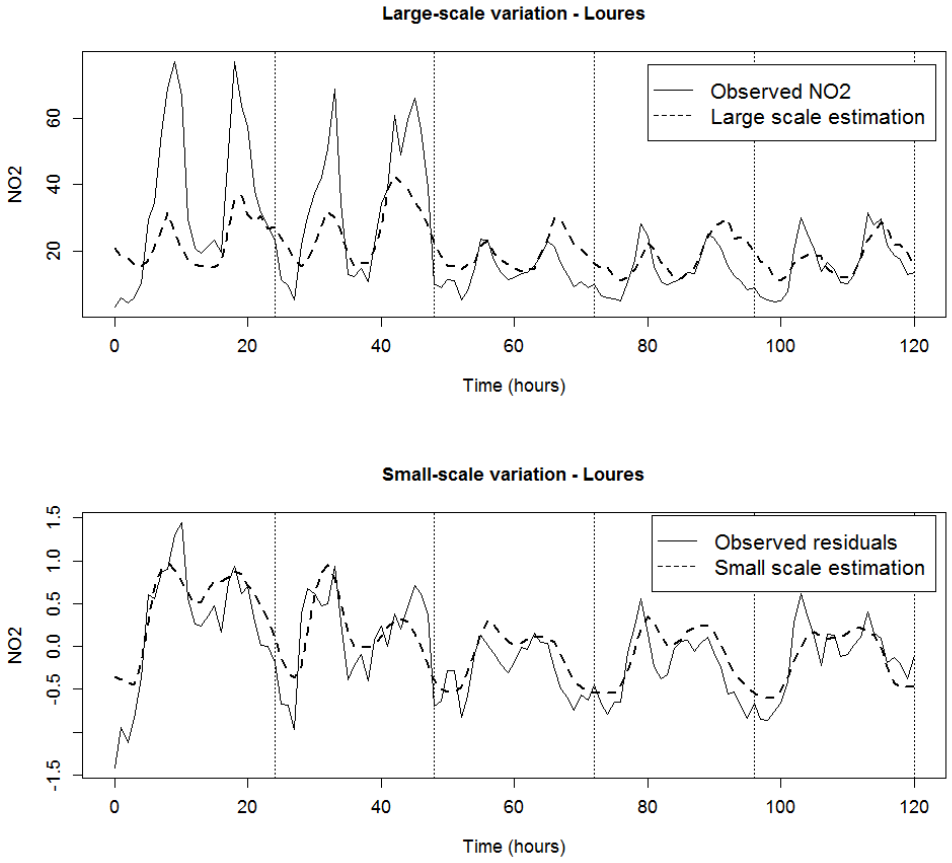

Figure 6: Estimation of the large-scale variation (top) and small-scale variation (bottom) of NO_2 concentrations in Loures Station from 2014-10-13 (Monday) to 2014-10-17 (Friday).

5.1 Space-time prediction

A major advantage of the proposed modelling methodology is the possibility of using space-time kriging techniques, namely ordinary kriging, to make predictions at any space-time point within the observation domain. Thus it allows to assess how pollution patterns change over space and time, as well as extending the current sampling design to locations without monitoring stations. This is illustrated in Figure 7 which represents the predicted spatio-temporal NO_2 concentrations process (small-scale variation) over Portugal on a Friday and a Sunday at 8:00, 13:00 and 18:00. We choose these days because Friday and Sunday are the days of the week with the highest and lowest concentration levels, respectively, while the choice of the times correspond to daily maxima, 8:00 and 18:00 and minimum, 13:00. Note that most of the temporal patterns in NO_2 concentrations result from anthropogenic

activities and are captured by the mean or large-scale variation. The first comment is that NO_2 concentrations present a strong spatial pattern that does not present much variation over time: along the day and overs different days. The space-time residual process achieves higher values on the coast where most of the urban and traffic monitoring station are located, corresponding to higher population density.

One may find slight differences on spatial patterns in interior zones of Portugal probably justified by the lack of monitoring stations, becoming harder to produce accurate estimations. Moreover, we can conclude that the estimated residuals slightly decrease, when comparing Friday and Sunday, mainly at 8:00 and 18:00. This should be explained by the lower traffic typical from weekends at these moments of the day.

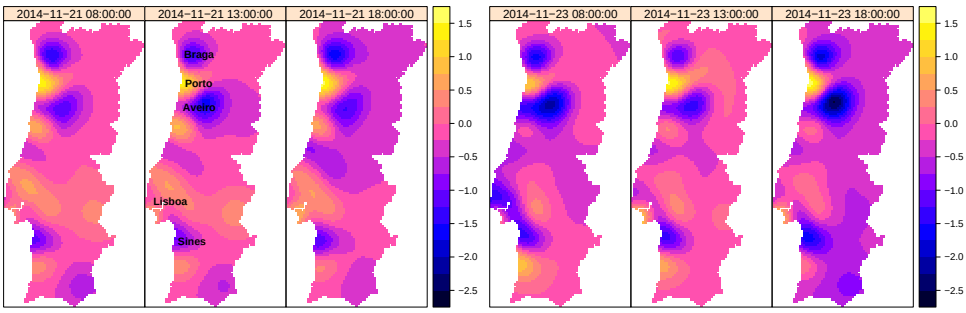

Figure 7: Space kriging maps for 2014-11-21 (Friday) and 2014-11-23 (Sunday), aiming to estimate the intra- and inter-day spatial patterns of NO_2 after removing the estimated trend.

A further application of space-time kriging allows to predict missing values in a specific station. These missing values may occur occasionally at some time points or when the station becomes inactive. Firstly, to illustrate this application, we proceed with the estimation of large and small-scale variation from Monday 2014-10-06 to Friday 2014-10-10 for Vila Nova da Telha, a suburban and background station from Maia county with no observations during this period. The results are presented at Figure 8, dashed lines, in the top panel, represent the 95% confidence bands for the estimated large-scale variation obtained by moving block bootstrap in time dimension, as explained in Section 3. The 95% confidence bands for the estimated small-scale variation, in the bottom panel, were obtained using kriging tools. We note that the estimated afternoon peak seems to occur 1 hour later which might be explained by the fact that Maia is a satellite town of Porto, leading

to a postponed rush hour traffic. Wednesday's NO_2 concentrations are lower with a somewhat different pattern from the remaining weekdays, which is also noted for other stations.

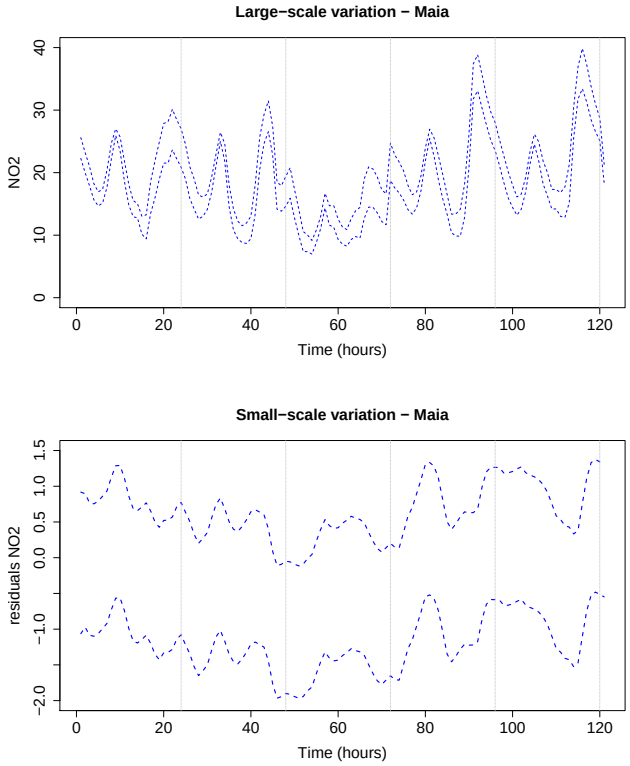

Figure 8: Estimation of the large-scale variation, top panel, and small-scale variation, bottom panel, of NO_2 concentrations in Maia station from Monday 2014-10-06 to Friday 2014-10-10. The dashed-lines identify the 95% confidence bands for: large-scale variation, obtained by a moving block bootstrap, each block with 5 weeks sliding 3 hours, generating 456 replicates (top panel); small-scale variation obtained by kriging tools (bottom panel).

5.2 Forecasting

The proposed model and associated modelling strategy enables to produce forecasts for NO_2 and quantify the associated uncertainty, as well as to analyse scenarios of possible future situations such as climate change and environmental policies. As explained before, the NO_2 forecasts are acquired through the mean predictor.

In Portugal, December 2015 was considered atypically warm with a mean temperature of 11.8^0C, the second warmest since 1931. Consider the 14th of December, a Monday, with mean valued for temperature, wind speed and relative humidity 16.1^0C, 15.6 km/h and 87%, respectively. The daily mean forecasts for NO_2 in the 39 stations are represented in the right panel of Figure 9. The point estimates are classified for easiness of representation. Since QualAr NO_2 levels for December 2015 are not available at the time of writing, we compare these forecasts with fitted NO_2 levels for Monday 15th December 2014, left panel of Figure 9, a day with somewhat different meteorological conditions: mean temperature of 11.2^0C, wind speed of 9.9 km/h and relative humidity of 78.5%.

As expected, due to the altered weather conditions in 2015, the predictions of NO_2 levels for this year are lower than for 2014, in particular in the north of the country.

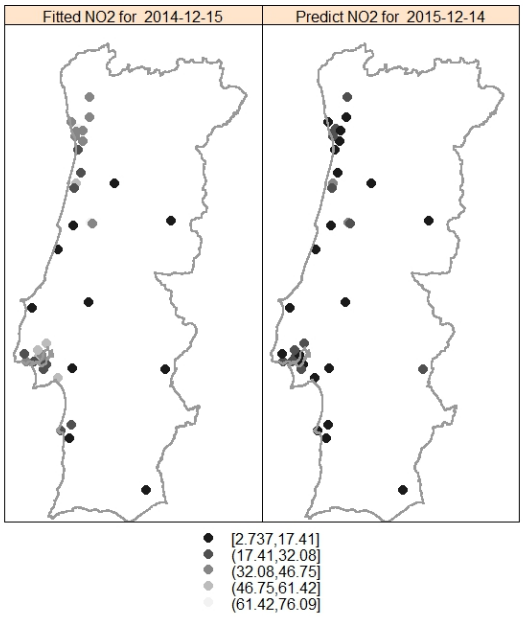

Figure 9: Daily mean of fitted NO_2 levels for 2014-12-15 (left). As meteorological data are available earlier than NO_2 levels, predictions for NO_2 levels for 2015-12-14 (right).

To further analyze the impact of meteorological variables (wind speed and relative humidity), we now compare hourly NO_2 concentrations observed during a week in 2014 with the corresponding 2015 forecasts for the same weekdays. The analysis is illustrated in Vila do Conde, a suburban and background station, between 15th and 21st December of 2014 (14th

to 20th December of 2015). In Figure 10, all the meteorological variables and NO_2 levels for 2014 represent observed values, while the bottom right panel represents NO_2 forecasts for 2015. Bearing in mind that in 2014 the values of wind speed ranged from 0 to 15 km/h and in 2015 ranged from 0 to 30 km/h (top panels), and the increased variability of relative humidity in 2015 (middle panels), we note a significant decrease in the forecasts of NO_2 concentrations for 2015. Furthermore, the maximum peaks in the wind speed correspond to the minimum peaks of NO_2 concentrations, showing a "mirror" alike effect.

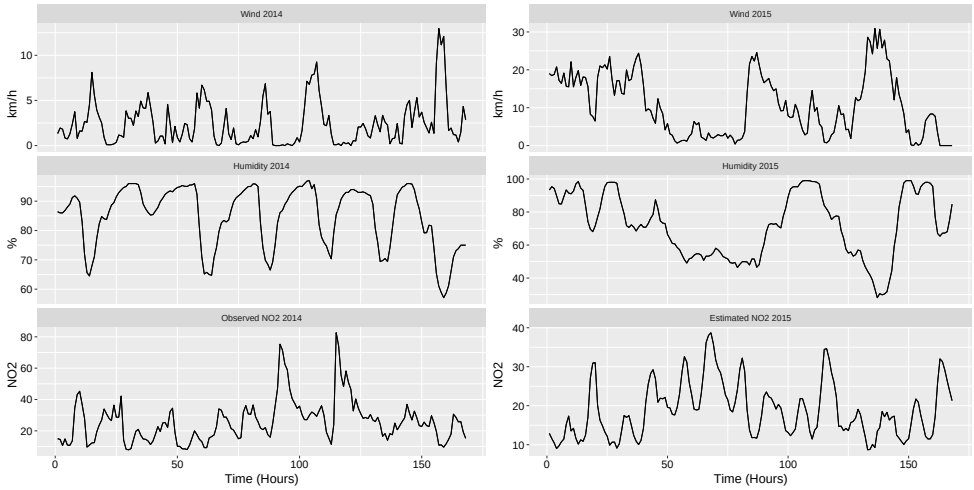

Figure 10: Observed NO_2 concentrations and meteorological variables in Vila do Conde, suburban and background station from 2014-12-15 (Monday) to 2014-12-21 (Sunday) (left). Meteorological variables in Vila do Conde from 2015-12-14 (Monday) to 2015-12-20 (Sunday) (right) and corresponding NO_2 forecasts.

5.3 Scenario analysis

Scenario analysis is achieved with conditional forecasting in which future (unknown) realizations of the explanatory variables are fixed at plausible values of interest. To illustrate the potential of the model in scenario generation, we obtain NO_2 forecasts under two distinct scenarios: if wind speed duplicates, and if relative humidity is reduced by half. In particular, we choose again Vila do Conde station, as being located in the north Portuguese coast, typically a windy and humid region. Figure 11 displays the observed NO_2 concentrations from 2014-12-12 (Monday) to 2014-12-18 (Sunday), against the NO_2 forecasts under the two scenarios which are being considered. The

results confirm that an increase in wind speed provokes, in general, a decrease in NO_2 concentrations and a decrease in relative humidity provokes, generally, an increase in NO_2 levels.

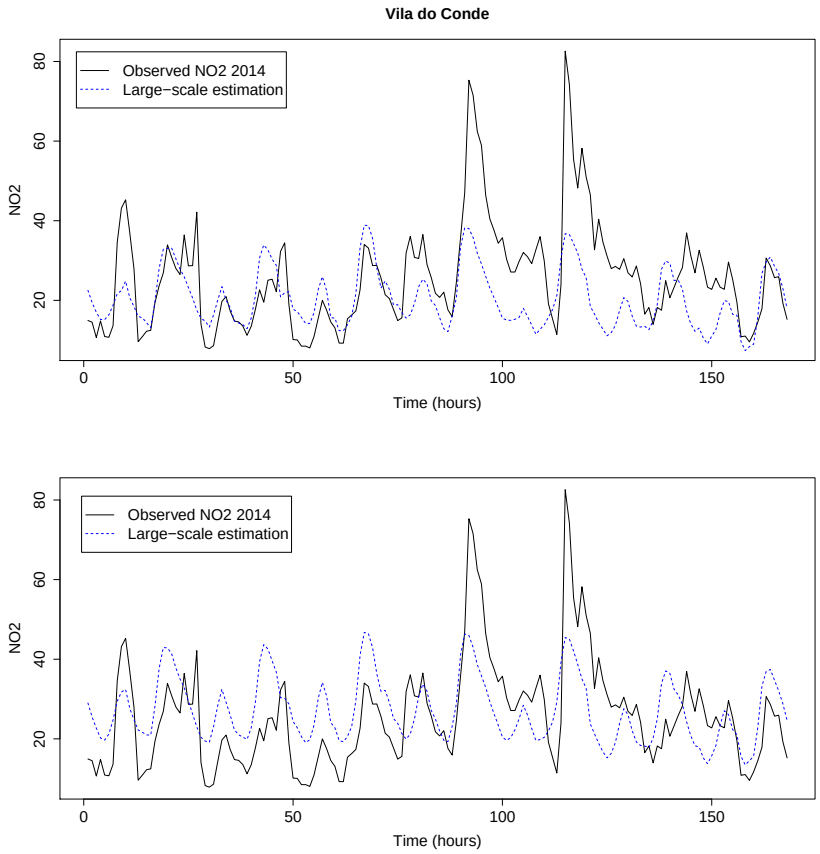

Figure 11: Observed NO_2 concentrations, in Vila do Conde station, from 2014-12-12 (Monday) to 2014-12-18 (Sunday). The dashed-lines represent NO_2 forecasts under the scenarios: wind speed duplicates (top panel) and relative humidity reduced by half (bottom panel).

A last example of scenario generation is the enforcement of environmental policies that many European cities are taking by pondering the permanent prohibition of vehicles in certain areas. This is equivalent to changing the type of site of a station located in a city from traffic to background. To illustrate this situation we consider Entrecampos which is an urban and traffic station located in Lisbon, where only vehicles registered after 1996 can circulate. Figure 12 displays the observed NO_2 levels together with NO_2 forecast if Entrecampos station becomes classified as background, assuming

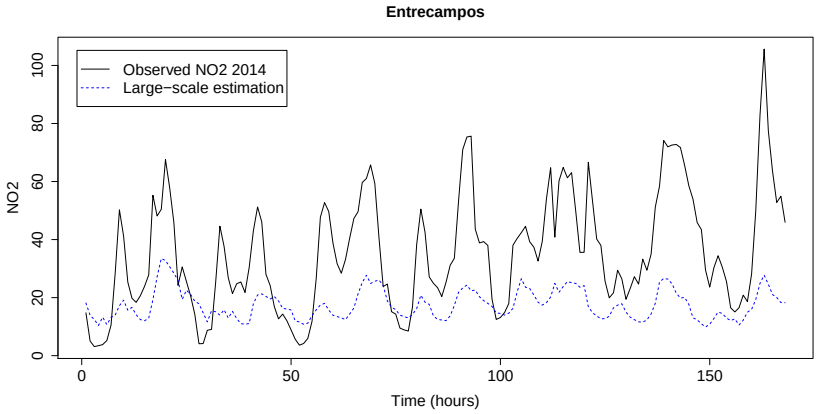

Figure 12: Observed NO_2 concentrations in Entrecampos station from 2014-12-12 (Monday) to 2014-12-18 (Sunday). The dashed-line represents NO_2 forecast under the scenario of changing this station from traffic to background classification.

that the meteorological variables are the same as in 2014. The decrease not only in mean but also in variability of NO_2 levels is noteworthy.

6 Discussion and concluding remarks

In this work, an easily implementable two-step approach is suggested to model spatial and high resolution temporal data, which exhibits multiple seasonal patterns imposed by social habits, anthropogenic activity and natural cycles explained by meteorological condition, simultaneously incorporating any additional information considered relevant to explain the phenomenon. The framework allows inference on the large-scale and small-scale variation components of the spatio-temporal stochastic process. Our proposal uses a block bootstrap procedure to correctly assess uncertainty in parameter estimates and produce reliable confidence regions for (space-time) unobserved values of the variable of interest. The suggested modelling approach, supported by well-known geostatistical tools such as kriging, is a methodology accessible to a wide range of practitioners, within the scope of spatial statistics.

Nonetheless, the discussed model presents some limitations, one of which is the difficulty in capturing temporal specificities intrinsic to a location. In fact, as discussed in section 4.3 in the illustrating example, although the overall mean intra-day pattern of the NO_2 concentrations is well described by the model, individual stations and days present particularities that re-

main unexplained. For example, stations located in the surroundings of major cities present anticipated and/or postponed rush-hour traffic leading to lagged peaks of NO_2 concentrations. To overcome this issue interactions between harmonic regression and type of station could be incorporated into the model, or time and space-varying model parameters could be allowed. Furthermore, this method, as a two-stage approach may introduce some extra-variance in the inferential procedures, which is expected to be negligible. A simulation study could be conducted to better assess its impact.

An alternative advocated approach is the Stochastic Partial Differential Equations (SPDE) approach implemented via the Integrated Nested Laplace Approximation (INLA) R package which is currently widely used in spatio-temporal modelling. For high resolution time series, such as the ones considered in the present work, Blangiardo and Cameletti (2015) point out that INLA becomes computationally expensive and advise lowering the temporal resolution by defining the model on a set of time knots, instead of on the set of all the time points. In our view, this could, however, mask high frequency variability, such as intra-day variability resulting from anthropogenic activities and meteorological conditions.

This work contributes to the characterization of the space-time dynamics, which can be used to complement the current sampling design by space-time prediction, to obtain forecasts and perform scenario analysis in environmental data as NO_2 concentrations, as well as in other data sets with similar characteristics, such as electrical demand.

Acknowledgments

The authors acknowledge Foundation FCT (Fundação para a Ciência e Tecnologia) for funding through Individual Scholarship PhD PD/BD/ 105743/2014, Centre of Mathematics of Minho University within project UID/MAT/00013/2013 and Center for Research & Development in Mathematics and Applications of Aveiro University within project UID/MAT/04106/2013.

References

References

Bivand RS, Pebesma E, Gomez-Rubio V. Applied spatial data analysis with R, Second edition. Springer, NY, 2013. URL: `http://www.asdar-book.org/`.

Blangiardo M, Cameletti M. Spacetime model lowering the time resolution. In: Spatial and spatio-temporal Bayesian models with R-INLA. John Wiley & Sons; 2015. p. 295–303.

Bruno F, Guttorp P, Sampson PD, Cocchi D. Non-separability of space-time covariance models in environmental studies. In: The ISI International Conference on Environmental Statistics and Health. Univ Santiago de Compostela; number 141; 2003. p. 153.

Cameletti M, Ignaccolo R, Bande S. Comparing spatio-temporal models for particulate matter in Piemonte. Environmetrics 2011;22(8):985–96.

Carslaw DC. Evidence of an increasing NO2/NOx emissions ratio from road traffic emissions. Atmospheric Environment 2005;39(26):4793–802.

Cressie N, Huang HC. Classes of nonseparable, spatio-temporal stationary covariance functions. Journal of the American Statistical Association 1999;94(448):1330–9. doi:10.1080/01621459.1999.10473885.

De Cesare L, Myers D, Posa D. Estimating and modelling space–time correlation structures. Statistics & Probability Letters 2001;51(1):9–14.

De Iaco S. Space–time correlation analysis: a comparative study. Journal of Applied Statistics 2010;37(6):1027–41.

De Iaco S, Posa D. Predicting spatio-temporal random fields: some computational aspects. Computers & Geosciences 2012;41:12–24.

De Livera AM, Hyndman RJ, Snyder RD. Forecasting time series with complex seasonal patterns using exponential smoothing. Journal of the American Statistical Association 2011;106(496):1513–27.

Donnelly A, Misstear B, Broderick B. Application of nonparametric regression methods to study the relationship between NO2 concentrations and local wind direction and speed at background sites. Science of the Total Environment 2011;409(6):1134–44.

European Environment Agency, EEA . Air quality in europe - 2015 report. 2015. URL: http://www.eea.europa.eu/publications/air-quality-in-europe-2015.

Fassò A, Negri I. Non-linear statistical modelling of high frequency ground ozone data. Environmetrics 2002;13(3):225–41.

Fonseca TC, Steel MF. A general class of nonseparable space–time covariance models. Environmetrics 2011;22(2):224–42.

Gneiting T. Nonseparable, stationary covariance functions for space–time data. Journal of the American Statistical Association 2002;97(458):590–600.

Grice S, Stedman J, Kent A, Hobson M, Norris J, Abbott J, Cooke S. Recent trends and projections of primary NO2 emissions in europe. Atmospheric Environment 2009;43(13):2154–67.

Hyndman RJ, Koehler AB. Another look at measures of forecast accuracy. International journal of forecasting 2006;22(4):679–88.

Ip RH, Li W. Time varying spatio-temporal covariance models. Spatial Statistics 2015;14:269–85.

Keramatinia A, Hassanipour S, Nazarzadeh M, Wurtz M, Monfared AB, Khayyamzadeh M, Bidel Z, Mhrvar N, Mosavi-Jarrahi A. Correlation between nitrogen dioxide as an air pollution indicator and breast cancer: a systematic review and meta-analysis. Asian Pacific Journal of Cancer Prevention 2016;17(1):419–24.

Kreiss JP, Paparoditis E. Rejoinder: Bootstrap methods for dependent data: A review. Journal of the Korean Statistical Society 2011;40(4):393–5.

Lai HK, Tsang H, Wong CM. Meta-analysis of adverse health effects due to air pollution in chinese populations. BMC Public Health 2013;13(1):360.

Ma C. Recent developments on the construction of spatio-temporal covariance models. Stochastic Environmental Research and Risk Assessment 2008;22(1):39–47.

McCarthy MC, O'Brien TE, Charrier JG, Hafner HR. Characterization of the chronic risk and hazard of hazardous air pollutants in the united states using ambient monitoring data. Environmental health perspectives 2009;117(5):790.

Menezes R, Piairo H, García-Soidán P, Sousa I. Spatial–temporal modellization of the NO2 concentration data through geostatistical tools. Statistical Methods & Applications 2016;25(1):107–24.

Myers DE. Estimating and modelling space-time variograms. In: Proceedings of the joint meeting of TIES-2004 and ACCURACY-2004. 2004. .

Pebesma EJ. Multivariable geostatistics in s: the gstat package. Computers & Geosciences 2004;30:683–91.

Qualar . Online database on air quality. 2015. URL: `http://qualar.apambiente.pt/`.

R Core Team . R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; Vienna, Austria; 2015. URL: `https://www.R-project.org/`.

Ricciardolo FL, Sterk PJ, Gaston B, Folkerts G. Nitric oxide in health and disease of the respiratory system. Physiological reviews 2004;84(3):731–65.

Roberts-Semple D, Song F, Gao Y. Seasonal characteristics of ambient nitrogen oxides and ground–level ozone in metropolitan northeastern new jersey. Atmospheric Pollution Research 2012;3(2):247–57.

Rodrigues A, Diggle PJ. A class of convolution-based models for spatio-temporal processes with non-separable covariance structure. Scandinavian Journal of Statistics 2010;37(4):553–67.

Russo A, Soares AO. Hybrid model for urban air pollution forecasting: A stochastic spatio-temporal approach. Mathematical Geosciences 2014;46(1):75–93.

Shi JP, Harrison RM. Regression modelling of hourly NOx and NO2 concentrations in urban air in london. Atmospheric Environment 1997;31(24):4081–94.

Shin HH, Stieb DM, Jessiman B, Goldberg MS, Brion O, Brook J, Ramsay T, Burnett RT. A temporal, multicity model to estimate the effects of short-term exposure to ambient air pollution on health. Environmental health perspectives 2008;116(9):1147.

Song X, Liu Y, Hu Y, Zhao X, Tian J, Ding G, Wang S. Short-term exposure to air pollution and cardiac arrhythmia: a meta-analysis and systematic review. International Journal of Environmental Research and Public Health 2016;13(7):642.

Stone M. Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society Series B (Methodological) 1974;:111–47.

Weather Underground . Site which provides weather data. 2015. URL: `https://www.wunderground.com/`.