

Hybrid model for urban air pollution forecasting: A stochastic spatio-temporal approach

Ana Russo ⁽¹⁾ and Amílcar O. Soares ⁽²⁾

⁽¹⁾ Instituto Dom Luiz, Universidade de Lisboa, Portugal

⁽²⁾ CERENA, Instituto Superior Técnico, Universidade Técnica de Lisboa, Portugal

(asoares@ist.utl.pt)

ABSTRACT: Air pollution is usually driven by a complex combination of factors in which meteorology, physical obstacles, and interactions between pollutants play significant roles. Considering the characteristics of urban atmospheric pollution and its consequent impacts on human health and quality of life, forecasting models emerge as an effective tool to identify and forecast AP episodes.

The overall objective of the present work was to produce forecasts of pollutant concentrations with high spatio-temporal resolution and to quantify the uncertainty in those forecasts. Therefore, a new approach was developed based on a two-step methodology. Firstly, neural network models were used to generate short-term temporal forecasts based on air pollution and meteorology data. The accuracy of those forecasts was then evaluated against an independent set of historical data. Secondly, local conditional distributions of the observed values with respect to the predicted values were used to perform spatial stochastic simulations with local distributions for the entire geographic area of interest.

Keywords: Air quality, Neural networks, Stochastic simulation, PM₁₀, Uncertainty.

1 Introduction

Urban air pollution (AP) is a complex mixture of toxic components that may induce acute and chronic pathologic responses in vulnerable individuals, especially children and people with cardiac and respiratory insufficiencies (Kolehmainen et al. 2001). In 2009, about 20% of the urban population in the EU was exposed to PM₁₀ above the limit value (EEA 2011).

Since the pioneer work of Bilonick (1983, 1985) and continuing to the most recent hybrid models coupled to determinist models (Russo et al. 2008, Russo et al 2014), geostatistical modeling has evolved tremendously. These models are able to mimic the dynamic behavior of physical phenomena, and specifically AP behavior. The mapping of pollutant concentrations has been treated by classical estimators of kriging (Atkinson and Lloyd 2001), stochastic simulation (Pereira 1999; Russo et al. 2008), Markov random fields (Cressie et al. 1999), and the co- simulation of contaminants (Franco et al. 2006). In addition, secondary information and time components have been

introduced into the frameworks of most geostatistical models (Kyriakidis and Journel 1999). These models have been widely used to characterize AQ in urban areas, where pollutant sources are considered diffuse, and in industrial areas, with localized emission sources (Nunes and Soares 2005; Soares and Pereira 2007; Russo et al. 2008). Considering the complexity of dispersion phenomena, non-stationary situations have been focused in several spatial models (Monestiez 1997) and space-time simulations (Kyriakidis and Journel 1999). Nonetheless, mainstream geostatistical models are basically interpolation exercises (estimation, simulation) using a set of variables in a spatial or spatio-temporal domain. The temporal prediction of a pollutant, i.e., extrapolation in the time domain, usually requires knowledge of both the main trends and the complex patterns of physical dispersion phenomena; that is, the dispersion of the main pollutants, the meteorological conditions, and the relevant chemical reactions. However, the types of information and knowledge needed to produce AQ predictions with the required rigor for urban areas are usually not compatible with the stationary assumptions of most geostatistical models.

Geostatistical space-time simulation models allow the characterization of uncertainty by supplying equiprobable images that reproduce patterns of spatial continuity quantified by the observations available. These space-time models can incorporate complex combinations of factors—which are usually non-linear—and thereby play a significant role in AQ forecasting (Kyriakidis and Journel 1999; Pereira 1999; Nunes and Soares 2005; Soares and Pereira 2007).

The objective of this work is to present an hybrid approach based on a two-step methodology which combines neural network short-term temporal forecast at the monitoring stations and a new spatial stochastic simulation, with local conditional probability distributions, to access the spatial uncertainty. To illustrate the proposed methodology a case study of the PM10 concentration in the city of Lisbon is presented.

2 Hybrid model for urban air pollution forecasting: a stochastic spatial-temporal approach

2.1 Short-term forecasts of the local conditional probability distributions at the monitoring stations.

Let us consider $Z(x,t)$, the pollutant's concentration at spatial location x and time period t , and $M(x,t)$ the general notation for the meteorological conditioning data.

Denoting the present time as t_0 , and the spatial locations of the N_m monitoring stations as x_α , the objective of the first step was to calculate at any location x_α the conditional probability

$$p(Z(x_\alpha, t_0) | Z(x_\alpha, t_{0-i}), M(x_\alpha, t_{0-i}), i = 1, N_T). \quad (1)$$

Equation (1) expresses the probability of forecasting the concentration of a pollutant $Z(x_\alpha, t_0)$ at instant t_0 , taking into account both its concentration $Z(x_\alpha, t_{0-i})$ and the meteorological conditions $M(x_\alpha, t_{0-i})$ of previous time periods t_{0-i} ($i=1, N_T$). In order to calculate Eq. (1), our approach replaces the conditioning data $[Z(x_\alpha, t_{0-i}), M(x_\alpha, t_{0-i})]$ with a function φ that summarizes all the conditioning data in a “predicted” value at (x_α, t_0)

$$Z^*(x_\alpha, t_0) = \varphi(Z(x_\alpha, t_{0-i}), M(x_\alpha, t_{0-i}), i = 1, N_T). \quad (2)$$

The function φ can be any predictor and its construction is achieved using NN modeling. For this study a multi-layer NN was chosen. The NN model used was trained and implemented based on the historical time series $Z(x_\alpha, t_{0-i})$ and on the meteorological conditions $M(x_\alpha, t_{0-i})$ ($i=1, N_T$) of the local monitoring stations x_α (Sect. 3) (Russo et al, 2014).

Based on (2), local bivariate distributions of predicted and observed values are estimated, based on independent historical data that were not used in the construction of the NN prediction model at each monitoring station :

$$F[Z(x_\alpha, t_{0-i}), Z^*(x_\alpha, t_{0-i})], \quad (3)$$

For the present time period t_0 , $z^*(x_\alpha, t_0)$ are predicted using the local model of NN at the monitoring stations x_α , after which the local conditional distribution functions were estimated based on the historical bivariate distribution obtained from Eq. (4)

$$F[Z(x_\alpha, t_0) | Z^*(x_\alpha, t_0) = z^*]. \quad (5)$$

Details of the calculation of Eq. (5) are presented in Russo et al, 2014. For the sake of simplicity, in the following $Z(x_\alpha, t_0)$ is denoted as $Z(x)$, $Z^*(x_\alpha, t_0)$ as $Z^*(x)$ and z^* as m .

After the concentration of a pollutant has been forecasted and the conditional probability distributions of the observed values estimated for $x_\alpha = 1, N_m$ and $t = 1, N_T$, the

objective of the second step of the proposed methodology is to use the local conditional distributions (5) for that period to perform spatial simulations for the entire area of interest.

2.2 Stochastic simulation with local conditional distributions

A set of realizations of $Z(.,.)$ for instant t_0 was obtained based on the local distributions $F [Z|Z^*= m]$ determined at monitoring stations location x_α . Among the existing algorithms that are based on the spatial “re-sampling” of conditional distributions, P-field simulation (Srivastava 1992) implies knowledge of the local cumulative distribution functions (cdfs) on the entire grid of nodes to be simulated, and not only those at the locations of the experimental samples. One option is to adapt sequential Gaussian simulation (Gomez-Hernandez and Journel 1993; Goovaerts 1997) by transforming the local distributions; however, that would imply the spatial stationarity of the Gaussian transformation (in practice, the same Gaussian transform function is applied at any location x_α , which is not a valid assumption in this case). Hence, for our model we adapted the basics of the model of direct co-DSS with bi-distributions (Horta and Soares 2010) to the direct simulation with local cdfs.

Let us set the first nodes to be visited in the path of the sequential algorithm as precisely the N_m monitoring stations. At each “experimental” node (monitoring station location), a local mean and its variance is calculated at a given time and a simulated value $Z^{(l)}(x_\alpha, t_0)$ is obtained from the local distributions $F [Z|Z^*= m]$ following the algorithm of Horta and Soares (2010). The remaining $N-N_m$ nodes are then sequentially simulated following the traditional DSS algorithm (Soares 2001). The DSS algorithm with local cdfs can be summarized as follows:

- i) The spatial locations of local cdfs (in this case, the locations of the monitoring stations) are visited in a random path and at each one the simple kriging (sk) mean and its variance is calculated based on the previous simulated values. A simulated value $Z^{(l)}(x_\alpha, t_0)$ is obtained from the local distributions $F [Z|Z^*= m]$ according to the DSS algorithm with bi-distributions (Horta and Soares 2010).
- ii) The remaining $N-N_m$ nodes are then visited sequentially and $Z^{(l)}(x_i, t_0)$, $i=1, N-N_m$ is obtained from the global cdf by following the traditional approach (DSS algorithm).

The image representing the previous instant (t_{0-1}) of a pollutant's concentrations is used as the local spatial trend. As we have an average image of simulated realizations $Z^{(l)}(x, t_{0-1})$, $l=1, N_s$ for the period t_{0-1} , calculated with the observations at t_{0-1}

$$O(x, t_{0-1}) = \frac{1}{N_s} \sum_{l=1}^{N_s} Z^{(l)}(x, t_{0-1}), \quad (6)$$

then it is possible to use the $O(x, t_{0-1})$ as the local trend of the sk estimates for the next period t_0

$$Z(x_\alpha, t_0)^* - O(x, t_{0-1}) = \sum_{\alpha} \lambda_{\alpha}(x, t_0) Z(x_{\alpha}, t_0) - O(x, t_{0-1}). \quad (7)$$

This approach is repeated for all the N_T time instants. Thus, this approach allows forecasting of the space-time distribution of AP, accounting for spatial as well as temporal uncertainties. An application to PM₁₀ concentration forecasting is presented in the next section.

3 Space-time application of PM₁₀ concentration forecasts for the greater Lisbon area

This illustrative example focuses on the city of Lisbon (Fig. 1), a large coastal city in central Portugal; The area is covered by a conventional AQ monitoring network comprising urban, traffic, industrial and suburban monitoring stations that record the atmospheric concentrations of major pollutants.

3.1 Data

The datasets consisted of the daily concentrations of several pollutants' (NO, NO₂, CO, PM₁₀) measured at 12 monitoring stations within the greater Lisbon area for a period of 5 years (from 1/1/2002 to 31/12/2006). Data from the first 4 years were used to construct the models whereas data from the last year (2006) were used for independent evaluation. Meteorological hourly observational data were also available. Daily values of boundary layer height (BLH) and SLP (sea-level pressure), obtained from the ECMWF (European Centre for Medium Weather Forecast), were used as well.

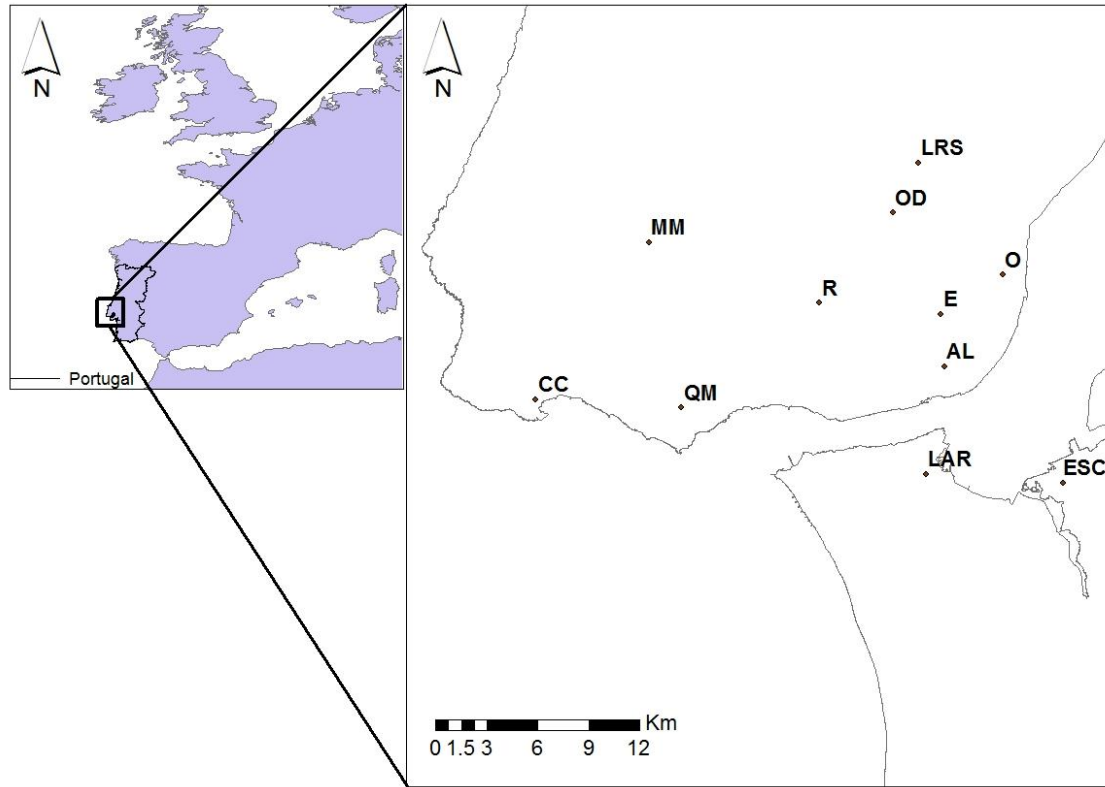


Fig 1 Case study area

After a pre-processing, one conclude that for the majority of the monitoring stations, the most significant variables in the prediction of PM_{10} were the previous-day PM_{10} concentration, followed by the 0 a.m. and previous-day maximum PM_{10} concentrations, and the previous-day values of NO_2 concentration, wind direction and humidity.

3.2 *Forecasting daily contaminants with neural networks*

The first step in the proposed approach consists of forecasting the present day (t_0) PM_{10} concentration using NN models, producing for each one of the monitoring stations an individual temporal forecast. In their construction, an AP system is considered as one that receives information from various sets of inputs and responds by producing a specific output. This model assumes no prior knowledge about the relationship between input and output variables. The NN models were trained with data from the first 4 years (01/01/2002 to 31/12/2005).

A cross-validation exercise was performed, dividing the available period into four sets and completing the calibration-validation procedure four times independently, i.e., from the 4 years of data available, data from 3 years were used to build the model and data from 1 year for validation. The cross-validation approach was applied such that the equivalent period of one year was taken out, for validation purposes, in all runs.

After calibration and validation with data from 2002–2005, the models were used to produce daily average PM_{10} forecasts for a period of 1 year. For this purpose an independent 1-year sample (1/1/2006 to 31/12/2006) was omitted in order to evaluate the performance of the models for the respective monitoring stations during individual daily average predictions. The PM_{10} forecasts were then compared with the actual values of the pollutants at the monitoring stations and with persistence.

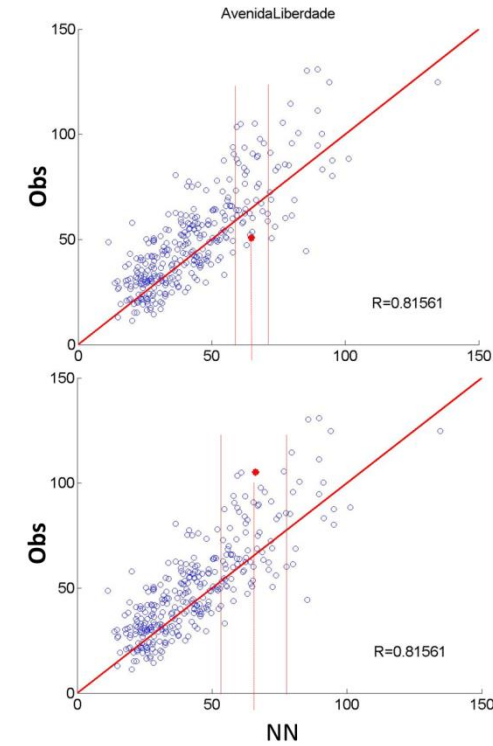


Fig 2 Scatter plots of the NN results versus the actual observed PM_{10} values for the Avenida Liberdade monitoring station. The filled dot represents the NN forecast for (a) 29/4/2006 and (b) 7/8/2006

Scatter plots and the correlation coefficients for two days, 29/4/2006 and 7/8/2006, determined at the Av. Liberdade monitoring station, are presented (Fig. 2). These two days were intentionally chosen to illustrate the potential of the proposed method, because they represent two daily periods characterized by a relative sharp

change in AQ relative to the previous day, hence, with difficult forecasting properties. The scatter plot results for 29/4/2006 show generally lower PM_{10} values than on the previous day, i.e., 28/4/2006. By contrast, on 7/8/2006, the PM_{10} values were generally higher than on 6/8/2006. Consequently, the forecasts for 29/4/2006 were usually underestimated while those for 7/8/2006 were mostly overestimated. These results influenced the conditional spatial simulations. Also, the simulated NN results for 29/4/2006 were generally closer to the perfect correlation curve (the red line which represents that predictions (y) are equal to observations (x), $y=x$) whereas those for 7/8/2006 were usually farthest away from the $y=x$ curve. For more detailed discussion of these results see Russo et al, 2014. Based on each day's forecast (red dot on each scatter plot of Fig. 4), bivariate distributions were determined based on the methodology presented in Horta and Soares (2010).

3.3 *Conditional spatial simulations with local distributions*

Based on the bivariate distributions between the predicted values resulting from the independent NN forecasts and the previously determined equivalent real values, conditional distributions were estimated for each monitoring station according to the methodology described in Sect. 2.2. Then, and in order to reproduce the PM_{10} spatial dispersion, a conditional spatial simulation was performed by applying the methodology described in Sect. 2.2. A set of $Nr=50$ equiprobable simulated realizations was generated by the algorithm described in Sect. 2.2, using 100-m $\times$ 100-m grids for each day, with the trend being the simulated average image of the previous day. Figure 3 illustrates the trend image for the previous day (t_{0-1}) and the conditional distributions of real PM_{10} at the monitoring station locations for the present day (t_0). An average image was computed based on the 50 realizations (Figures 4(a) and 4(b)).

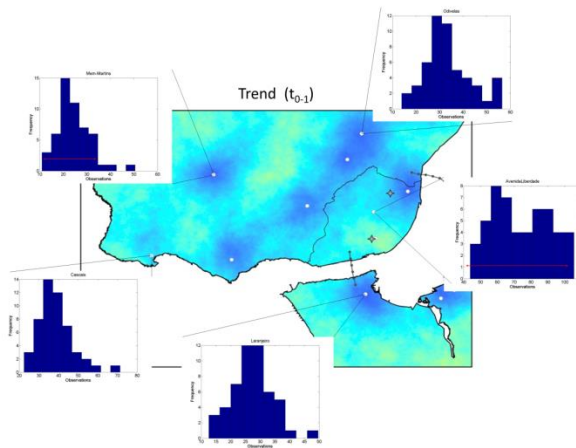


Fig 3 Schematic representation of the second stage of the proposed method, showing a trend image for the previous day (t_{0-1}) and the conditional distributions of real PM_{10} at the monitoring station locations for the present day (t_0)

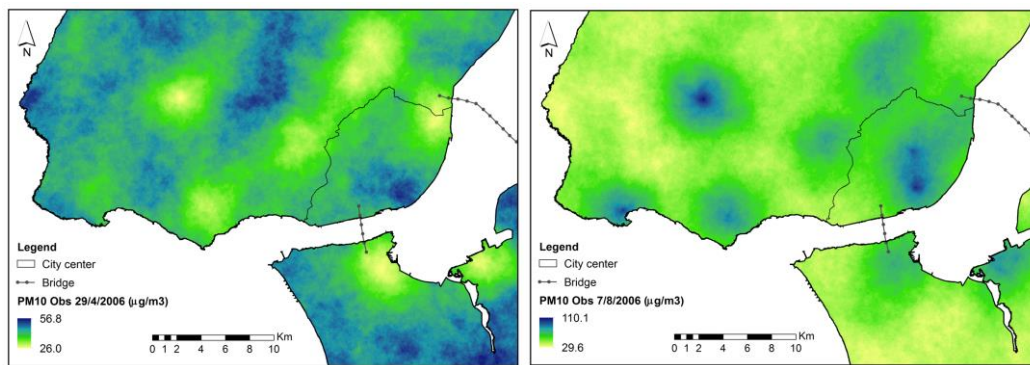


Fig 4 Average image (50 simulations) representing the spatial dispersion of PM_{10} on (a) 29/4/2006 and (b) 7/8/2006

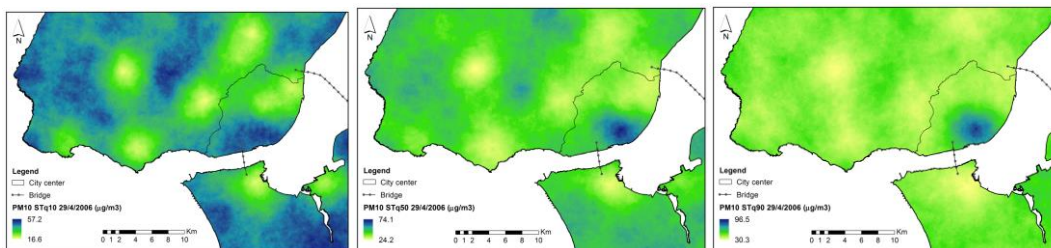


Fig 5 Average image representing the (a) 10th, (b) 50th, and (c) 90th percentile forecasts for 29/4/2006

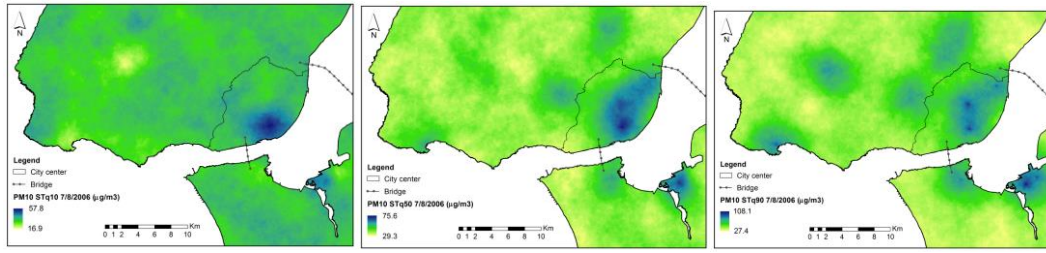


Fig 6 Average image representing the (a) 10th, (b) 50th, and (c) 90th percentile forecasts for 7/8/2006

Figures 5(a) and 6(a) show the 10th percentile average image for the 29/4/2006 and 7/8/2006 events, Figs. 5(b) and 6(b) the 50th percentile, and Figs. 5(c) and 6(c) the 90th percentile, respectively. This percentile representation allows an assessment of the spatial uncertainty of the predictions..

3.4 Discussion

Sharp transitions in pollutant levels from one day to the next are among the most difficult situations to forecast by means of machine learning techniques, which learn based on historical past events and thus require a certain amount of repetitiveness. As these situations are uncommon relative to the mean behavior, prediction becomes a difficult task. By contrast, the NN approach performed reasonably well on 29/4/2006 and 7/8/2006, i.e., both situations were well forecasted within the range of risk defined by the 10th and 90th percentiles.

Geographic areas repeatedly subjected to daily PM₁₀ values higher than 50µg/m³ or located close to major roads, intersections, or transport infrastructures are considered pollution hotspots. The health and quality of life of people living and working close to those areas may be at risk due to increased levels of AP and noise (EC 2006). Therefore, the correct prediction of high pollutant levels and the identification of pollution hotspots are vital in order to equip the responsible entities with the tools to produce alerts and to facilitate sustainable AQ management. The simulated images (Figs. 7 and 8) show the areas with higher and lower PM₁₀ values; some of those areas are hotspots. Critical risk areas are those with PM₁₀ values higher than the 50 µg/m³ threshold. These were also well forecasted within the risk intervals of the 10th and 90th percentiles. The choice among the percentile forecasts might be decided upon based on an evaluation of the main spatial trend of the previous day (12 hours ahead).

Through this hybrid approach, the uncertainty resulting from the prediction errors and from the local spatial variability of pollutants is quantified by the simulated maps, giving rise to critical risk areas. Thus, the resulting images provide a threshold of risk for use in risk assessment studies and by decision-makers. It is worth noting that the time prediction could be performed by using an alternative machine learning method or by deterministic simulation, as long as the chosen method is able to integrate as many factors into the prediction as the NN does. This means that the hybrid approach proposed in this study is, above all, a methodological framework, coupling a time predictor (in this case a NN) with a new geostatistical simulation methodology for the spatial characterization of a pollutant's concentration, in which the uncertainty in the time prediction is accounted for.

Acknowledgments: The authors acknowledge the Instituto de Meteorologia and **Agência Portuguesa do Ambiente** for the meteorological and environmental data, respectively. The authors also acknowledge the Fundação para a Ciência e Tecnologia from the Science, Technology and Superior Education Ministry, for supporting this research through grant SFRH/BD/27765/2006.

References

- Atkinson PM, Lloyd CD (2001) Ordinary and Indicator kriging of monthly mean nitrogen dioxide concentrations in the United Kingdom. In: Monestiez et al. (eds.) *geoENV VIII-Geostatistics for Environmental Applications*. Kluwer, pp 33–44.
- Bilonick RA (1983) Risk qualified maps of hydrogen ion concentration for the New York state area for 1966-1978. *Atmos Environ* 17:2513-2524
- Bilonick RA (1985) The space-time distribution of sulfate deposition in the northeastern United States. *Atmos Environ* 19:1829-1845
- Cressie N, Kaiser MS, Daniels MJ, Aldworth J, Lee J, Lahiri SN, Cox LH (1999) Spatial analysis of particulate matter in an urban environment. In: Gomez-Hernandez, J, Soares, A., Froidevaux, R. (eds). *geoENV II - Geostatistics for Environmental Applications..* Kluwer, Dordrecht, pp 41-52.
- EEA-European Environment Agency (2011) Air quality in Europe. Technical report No 12/2011.

EEA-European Environment Agency (2010) The European environment – state and outlook 2010: synthesis. European Environment Agency, Copenhagen.

Franco C, Soares A, Delgado J (2006) Geostatistical modelling of heavy metal contamination in the topsoil of Guadiana river margins (Spain) using a stochastic simulation technique, *Geoderma*, Volume 136, Issues 3–4: 852-864.

Gardner M, Dorling S (2000) Statistical surface ozone models: an improved methodology to account for non-linear behaviour. *Atmos Environ* 34:21–34.

Gardner M, Dorling S (1999) Neural network modelling and prediction of hourly NO_x and NO₂ concentrations in urban air in London. *Atmos Environ* 33:709–719.

Gomez-Hernandez J, Journel AG (1993) Joint sequential simulation of multi-gaussian fields, in *Geostatistics TROIA'92*. In: Soares A (ed). Kluwer Academic Pub., Dordrecht, The Netherlands, pp 85–94.

Goovaerts P (1997) *Geostatistics for Natural Resources Evaluation*. Oxford Univ. Press, New-York.

Horta A, Soares A (2010) Direct sequential co-simulation with joint probability distributions. *Math Geosci* 42: 269–292.

Kolehmainen M, Martikainen H, Ruuskanen J (2001) Neural networks and periodic components used in air quality forecasting. *Atmos Environ* 35: 815-825.

Kyriakidis P, Journel A (1999) Geostatistical space-time models: a review. *Math Geol* 31(6): 651-685.

Monestiez P, Meiring W, Sampson PD, Guttorp P (1997) Modelling non-stationary spatial covariance structure from space-time monitoring data. *Ciba Foundation symposium*. 01/1997; 210:38-48; discussion 48-51, 68-78.

Nunes C, Soares A (2005) Geostatistical Space-Time Simulation Model, *Environmetrics*, 16:393-404.

Pereira MJ (1999) Air quality modelling and simulation. PhD Dissertation, Instituto Superior Técnico, Lisbon, Portugal.

Perez P (2001) Prediction of sulfur dioxide concentrations at a site near downtown Santiago, Chile. *Atmos Environ* 35:4929–4935.

Perez P, Trier A, Reyes J (2000) Prediction of PM_{2.5} concentrations several hours in advance using neural networks in Santiago, Chile. *Atmos Environ* 34:1189–1196.

Russo A, Trigo R, Soares A (2008) Stochastic modelling applied to air quality space-time characterization. In: Soares A, Pereira MJ, Dimitrakopoulos R (eds). *geoENV VI-Geostatistics for Environmental Applications*. Springer, pp 83–93.

Russo A, Soares A. (2014) Hybrid model for urban air pollution forecasting: A stochastic spatio-temporal approach. *Mathematical Geology*.

Soares A (2001) Sequential direct simulation and co-simulation. *Math Geol* 33(8): 911–926.

Soares A, Pereira MJ (2007) Space-time modelling of air quality for environmental-risk maps: a case study in south Portugal. *Comput & Geosci* 33(10): 1327-1336.

Srivastava RM (1992) Reservoir characterization with probability field simulation. *SPE Paper* 24753.