ESTUDIOS DE ECONOMÍA APLICADA VOL. 33-1 2015 PÁGS. 145–174

Functional Kriging Prediction of Pollution Series: The Geostatistical Alternative for Spatially-Fixed Data

JOSÉ-MARÍA MONTERO [a], GEMA FERNÁNDEZ-AVILÉS [a]

[a] *University of Castile-La Mancha, Facultad de Ciencias Jurídicas y Sociales, Cobertizo de San Pedro Mártir, s/n, 45071 Toledo, España. E-mail:* jose.mlorenzo@uclm.es; gema.faviles@uclm.es

ABSTRACT

Proper assessment of air quality is of paramount importance. Accordingly, authorities in large cities have established air pollution monitoring networks that register levels of the most dangerous pollutants in a number of city locations on an hourly basis. Thus, the dataset including such measurements can be considered as a panel dataset where data are spatially fixed. Geostatistics deals with such panel or longitudinal data differently to econometrics: geostatistics takes advantage of the spatio-temporal dependencies to make kriging or cokriging predictions at non-observed sites. However spatio-temporal kriging implies a prohibitive computational burden and, as a consequence, functional kriging has emerged as an alternative strategy for dealing with this type of data since it deals with functional data representing the observations recorded at each location observed. Functional kriging could be said to reproduce the history of the phenomenon under study at non-observed sites. This novel approach has been applied in the city of Madrid (Spain) to particulate matter registers. We predict functional data at some of the most polluted sites of the city.

Keywords: Panel data, functional data, functional kriging, spatio-temporal dependencies, air pollution, PM_{10}.

Predicción de series de contaminación mediante kriging funcional. La alternativa geoestadística cuando los datos están anclados al espacio

RESUMEN

La calidad del aire es una cuestión de suma importancia. Por ello en las grandes ciudades se han implantado redes de seguimiento de la calidad del aire que proporcionan horariamente el nivel de los contaminantes más peligrosos en las ubicaciones donde se han instalado las estaciones de seguimiento. Por tanto, el conjunto de datos relativo a tales mediciones constituye un conjunto de datos panel en el que los datos están anclados al espacio. El tratamiento geoestadístico de dichos datos es muy diferente al tratamiento econométrico de los mismos; la Geoestadística se aprovecha de las dependencias espacio-temporales existentes en ellos para llevar a cabo predicciones krigeadas o cokrigeadas en localizaciones no observadas. Sin embargo, el krigeado espacio-temporal implica una enorme carga computacional, lo que ha favorecido la aparición del kriging funcional, una estrategia geoestadística para tratar este tipo de datos ya que trabaja con datos funcionales que representan las observaciones registradas en cada localización observada. Podría decirse que el kriging funcional reproduce la historia del fenómeno de interés en tales localizaciones. Este novedoso enfoque se aplica en Madrid (España) para predecir el nivel de PM_{10} en algunos de los puntos más contaminados de la ciudad.

Palabras clave: Datos panel, datos funcionales, kriging funcional, dependencias espacio-temporales, calidad del aire, PM_{10}.

JEL Classification*:* C21, C23, C55, O13, Q53

Artículo recibido en octubre de 2014 y aceptado en noviembre de 2014
Artículo disponible en versión electrónica en la página www.revista-eea.net, ref. ə-33104

ISSN 1697-5731 (online) – ISSN 1133-3197 (print)

1. INTRODUCTION

Air pollution is at the top of the list of citizen's environmental concerns (Montero-Lorenzo *et al*. 2013), this being particularly true in large cities, where more than half the world's population (3.3 billion people) lives. The link between air quality and human health worries health experts, policy-makers and inhabitants alike. Many health problems (e.g., respiratory and cardiovascular) can be caused or worsened by exposure to air pollution on a day-to-day basis. The World Health Organization (WHO) states that almost 2.5 million people die each year from causes directly attributable to air pollution, having ranked (one decade ago) urban air pollution as the 13th greatest contributor to global deaths (World Health Organization 2002). Therefore, studies concerning air pollution are increasing in number, while environmental issues have brought atmospheric science to the centre of both science and technology, where it now plays a key role in shaping national and international policy. Currently, environmental prediction plays a significant role in the planning of human affairs.

This article focuses on suspended particulate matter (PM), a mixture of solids and liquid droplets, and more specifically on PM_{10}, an important constituent of the atmosphere, contributing substantially to air pollution. Particles come in a wide range of sizes, some emitted directly and other formed in the atmosphere when other pollutants react. Those less than 10 micrometers (smaller than the width of a single human hair)[1] in diameter (PM_{10}) are so small that they can get into the lungs, potentially causing serious health problems. While air quality has improved substantially in European and American cities over the past decades PM, and especially PM_{10}, is of major current concern. Limit values are very often exceeded in European cities and this can have a critical impact not only on natural chemical processes but also on human health (Turoczi *et al*. 2012, Nguyen *et al*. 2013, Pope and Dockery 2013).Consequently PM levels, as well as many other pollutants, are monitored continuously. PM is considered as the air pollutant that most commonly affects people's health. Its effects on human health include coughing, wheezing, shortness of breath, aggravated asthma, lung damage (including decreased lung function and life-long respiratory disease) and premature death in individuals with existing heart or lung diseases. By way of example, epidemiological studies in the USA, continental Europe and the UK suggest that that every 10 $\mu g/m^3$ increase in PM_{10} leads to approximately 1% extra deaths (see Montero Lorenzo *et al*. 2011, and the references therein); Ayres (2002) estimated that particles contribute to around 8,100 deaths per year in urban areas of Great Britain; and Chay and Greenstone

[1] $PM_{2.5}$ particles (<2.5 µm) are especially dangerous, as they are small enough to penetrate deep into the lungs. Particles larger than 10 µm meanwhile, are not readily inhaled, and are removed relatively efficiently from the air by sedimentation.

(2003) estimated that a one-percent reduction in total suspended particulate lowered the infant mortality rate by 0.35% during the period 1970-80.

The sources of airborne PM_{10} can differ from site to site, but in European and American cities the principal source is road traffic emissions (e.g. exhaust emissions from vehicles and non-exhaust emissions from mechanical abrasion, such as brake-, tyre-, and road-wear by re-suspension), particularly from diesel vehicles. Thus, high PM_{10} levels are most often found in urban areas along busy roads, where the population density and, consequently, the number of people exposed is among the highest, exacerbating the health impact of PM_{10} emissions. As a consequence, policies to mitigate the impact of particulate matter usually focus on road transportation in cities. Natural emissions (e.g. crustal minerals originating from wind-eroded bare soils or transported from arid areas by episodic dust storms) and industrial emissions (fossil fuel combustion and industrial metallurgical processes) are still non-negligible sources of PM_{10} in some cities. It is of note that, unfortunately, PM_{10} remains in the atmosphere for longer periods due to its low intrinsic settling velocity.

To the extent that in large cities (especially in megacities) communications infrastructure plays a core role in inhabitants' social and economic development, industrial air pollution tends to be replaced by traffic air pollution. As stated in Montero *et al.* (2013), the challenge is to ascertain an optimal trade-off between advantages and disadvantages.

Therefore, it is easily understandable that the prediction of air pollution levels in general, and of PM_{10} levels in particular, and the detection of future extreme episodes (violations of the PM_{10} standards) are of particular interest in the field of air pollution control.

Following Montero Lorenzo *et al.* (2011), the most widely used procedure to predict or detect a future extreme air pollution episode is the extreme value models (Roberts 1979a,b, Surman *et al.* 1987, Sharma *et al.* 1999, Ercelebi and Kirmanli 1999, Lu and Fang 2003, Kan and Chen 2004, Hurairah *et al.* 2005, Sfetsos *et al.* 2006, Achcar *et al.* 2008, and Erceleby and Toros 2009, are good examples). Linear and non-linear regression models have also been widely used to predict future extreme air pollution episodes, particularly in relation to ozone (Robeson and Steyn 1990, Hubbard and Cobourn 1998, and Chaloulakou *et al.* 1999, are some interesting references in the past). Classification and regression trees models (Burrows *et al.* 1995) and neural networks (Baxt and White 1995, van Aalst and de Leeuw 1997, Comrie 1997, and Gardner and Dorling 2000, are good early references; more recently Sharma *et al.* 2003, Wang and Lu 2006, Rost *et al.* 2009, and Barai *et al.* 2009) are statistical procedures which have also been used, although they are not currently prominent research lines. Canonical analysis and other techniques related with linear models have also been sporadically used with some success.

The most promising approach to predict, detect or alert to a future extreme air pollution episode is time series analysis and, specifically, stochastic volatility models. These utilise two processes to model the series: a process to model the observations and a process to model the latent volatility. Montero Lorenzo *et al*. (2011) propose a threshold autoregressive asymmetric stochastic volatility (T-ARSV) strategy to alert to an immediate violation of the PM_{10} quality standards as an alternative to traditional ARCH, GARCH and ARSV models. This strategy takes into account the asymmetric response of volatility to a positive or negative relative variation of the level of the pollutant in the previous period. Thresholds were included in stochastic volatility models by So *et al.* (2002), and they have been recently developed by García and Mínguez (2009, 2011).

As pollution data have been recorded by monitoring stations in large cities on a (usually) hourly basis since the 1990s, they constitute a simple but interesting one-factor panel data set. As they are available in a spatio-temporal format, they can be incorporated in traditional panel data sets including data on explanatory variables of the level of pollution, or data on explanatory variables of another variable of interest, the pollution variable being one additional explanatory variable. The availability of panel or longitudinal data sets including pollution information, together with the fact that (i) while it is possible to use ordinary multiple regression techniques on panel data, they may not be optimal, and (ii) their greater capacity for modeling complex phenomena than a single cross-section or times series data (see Hsiao 2006, for details on the advantages of panel data modeling) have led the econometric techniques for panel data to become more and more popular in the air pollution field over the two last decades.

This has led to a proliferation of empirical panel data studies using air pollution data. Although this is not an exhaustive list, panel data have been employed (i) to evaluate the short-term health effects of air pollution (e.g. Janes *et al.* 2008); (ii) to estimate the willingness to pay for an improved air quality (e.g. Gupta 2011); (iii) to validate the Environmental Kuznets Curve (EKC) hypothesis (Akbostanci *et al.* 2009, Selden and Song 2013, Omay and Canpolat 2013, Rahman and Porna 2014, and Acar and Tekce 2014, are some recent recommended references focusing on PM_{10}); (iv) to investigate the relationship between life expectancy rate and air pollution (PM_{10}) (Chay and Greenstone 2003); (v) to evaluate the efectiveness of environmental measures on reducing pollution levels: Auffhammer *et al.* (2009), and Malina and Fischer (2012), focus on the effects of regulation on PM_{10} concentrations (in USA and Germany respectively); Auffhammer and Kellogg (2011) illustrate the effects of gasoline content regulation on air quality in USA; Davis (2008) studies the effect of driving restrictions on air quality in Mexico City (Mexico); Atkinson (2009) focuses on policy regulation in the transport market in London (UK) and ambient air quality (including PM_{10} levels); (vi) to investigate the relationship between environmental regulation and innovation (Jaffe and Palme 1996 is a pioneer study in this topic); (vii) to examine the relationship between socio-economic factors, environment and policy responses: in a recent article by Laureti *et al.* (2014), in the framework of the 'urban development-transport-air pollution' connection, it is used an interesting and comprehensive framework, inspired by the DPSIR (Driving forces - Pressures - State - Impact - Responses) model, to identify the different variables indicating pressure of transportation, socio-economic conditions and policy response which may mitigate or worsen the effect of road transportation on air quality. The authors employ an augmented STIRPAT (Stochastic Impacts by Regression on Population, Affluence, and Technology) model to examine the relationship between socio-economic factors and the environment, which is estimated using several models for panel data; (viii) and finally, although many more topics could be listed, in the field of environmental degradation and happiness, which is a relatively new area of research. A well-known paper in this topic is Welsch (2006) uses a panel data approach to avoid unobserved heterogeneity and the danger of spurious correlations in previous study (Welsch 2002), where he used cross-sectional data from 54 countries to examine the relation between subjective wellbeing and pollution. More specifically, the 2006 study uses data on nitrogen dioxide, PM_{10} and lead for ten countries, from 1990 to 1997, and employs fixed effect estimation.

Nevertheless, as stated in Janes *et al.* (2008), many published panel studies could be improved if practitioners had a better understanding of the statistical issues pertaining to the analysis of longitudinal data and appropriate statistical analysis techniques.

The above points represent only a very small sample of issues (and literature) related to air quality that can be studied using econometrics for panel data. But in this article we focus on how geostatistics deals with one-factor or multi-factor panel data, an approach which is significantly different from the econometrics approach.

The econometric approach focuses on estimating the model parameters -a more accurate inference than in the case of cross-sectional data or time series data as panel data usually contains more degrees of freedom and less multicolinearity- being possible (i) consider relationships that are explicitly or implicitly dynamic, (ii) control for some types of omitted variables even without observing them, by observing changes in the dependent variable over time (this controls for omitted variables that differ between cases but are constant over time) or to control for omitted variables that vary over time but are constant between cases, (iii) and construct and test complicated behavioural hypotheses, etc. (more advantages can be seen in Hsiao 2006).

Geostatistics also deals with one-factor (the most frequent situation) or multi-factor panel data, but focuses directly on prediction rather than on parameter estimation and, to make predictions, takes advantage of the spatio-temporal dependencies existing in the data. Such predictions are called spatio-temporal kriging predictions when the panel data set includes only one variable (the variable of interest), and spatio-temporal cokriging[2] when it consists of one main variable and some auxiliary variables strongly correlated with the main variable.

Geostatistics can also use explanatory variables but, unlike in the econometric approach, the objective of such explanatory variables is usually to estimate the drift of the phenomenon under study in order to perform a spatio-temporal kriging on the residuals (see Cressie and Wikle 2011, Sherman 2011, and Montero *et al.* 2015 for details about spatio-temporal geostatistics). The keystone in the geostatistical process is the search for the semivariogram or covariance function that best captures the above mentioned spatio-temporal dependencies.

Literature regarding air quality prediction using cross-sectional data (that is, employing merely spatial geostatistics) with the focus on the kriging prediction of PM_{10} is lacking, generally emanating from the fields of environment health and toxicology (Lim *et al.* 2014, Feng *et al.* 2014, and Zhang *et al.* 2014, are three very recent examples). Similarly, the number of papers using panel data (that is, using spatio-temporal geostatistics to perform kriging or cokriging predictions on the level of the most dangerous pollutants, depending on whether we have a one-factor or a multi-factor panel data) is small, as performing spatio-temporal kriging and cokriging predictions implies a considerable, and sometimes prohibitive, computational burden. Studies which do use spatio-temporal kriging to investigate PM_{10} concentrations include: Yanosky *et al.* (2008), which predicts monthly outdoor PM_{10} concentrations in the northeastern and midwestern United States. Pollice and Lasinio (2010), where PM_{10} concentrations (among other pollutants) were analyzed in the municipal area of Taranto (southern Italy), an intensely industrialized area with elevated environmental impact activities. Their spatio-temporal modelling was addressed using a Bayesian kriging-based model characterized by the use of time varying covariates and a semiparametric covariance structure. Gräler *et al.* (2011), which studied different potential spatio-temporal kriging approaches for daily mean

[2] Kriging is a univariate procedure which interpolates the values of the target random function at unobserved (spatial or spatio-temporal) locations using the available observations of the same random functions (and other auxiliary random functions in the case of cokriging). This interpolation procedure uses the spatial or (spatio-temporal) covariance or semivariogram function to account for the correlation structure in making interpolative estimates. In the merely spatial case, the correlation function is the spatial equivalent of the autocorrelation function in time series analysis.

PM_{10} concentrations and applied them to daily mean rural background PM_{10} concentrations across Europe for the year 2005. And Hussain *et al.* (2013), which modelled the spatio-temporal structure of PM_{10} in northern USA (1998-2002) with and without a spatial trend madeof environmental covariates.

The number of studies with prediction of PM_{10} concentration using spatio-temporal cokriging (or multivariate geostatistics, in general) as a focal theme is very small, but it is worth mentioning the papers by Fasso and Finazzia (2011) developed the maximum likelihood estimation of the heterotopic spatio-temporal model with spatial linear coregionalization model components; De Iaco *et al.* (2012) proposed a thorough multivariate geostatistical analysis they apply to the south of the Apulian region (Italy) in 2009, including different tools for testing the symmetry assumption of the spatio-temporal linear coregionalization model, as well as a fitting procedure of such a model based on the simultaneous diagonalization of symmetric real-valued matrix variograms; and Campalani *et al.* (2014), which presented an application of spatio-temporal multivariate kriging for daily PM_{10} estimation, in Austria (2008 to 2010).

Following Liang and Kumar (2013), the computational challenge of spatio-temporal kriging (and cokriging) due to *the big n problem* (Banerjee *et al.* 2004) has been addressed in the literature by (i) approximating a large process by its realizations on a small set of knots (or subset of data) (Banerjee *et al.* 2008), (ii) fitting a Gaussian Markov random field to the original process on a lattice in order to utilize sparse matrix algorithms (Hartman and Hossjer 2008), although this approach is of little use for irregular spatio-temporal data, (iii) approximating the covariance function using linear combinations of a small number of low-dimensional deterministic (or basis) functions (Cressie and Johannesson 2008), andusing a flexible and computationally efficient Markov spatio-temporal Cube (or voxel) kriging, a Bayesian hierarchical spatio-temporal method of interpolation (Liang and Kumar 2013).

In this article we address the *big n problem* of spatio-temporal kriging from the functional perspective of geostatistics. Functional kriging is the alternative we propose when the available information is in the form of a spatio-temporal panel data set. Thus, not only the spatial-only or temporal-only dependencies but also the spatio-temporal dependencies existing in the data can be used to make spatio-temporal predictions, or even to reproduce the history of the phenomenon under study in a non-observed location.

Specifically, we implement the above methodology to predict smooth curves representing the series of PM_{10} at four non-monitored sites in Madrid, a city daily transited by thousands of people. The main reason behind this research is the claim of ecologists that at such non-monitored sites there are higher PM_{10} values than in the locations where the monitoring stations are actually placed. The main idea is to compare the curves predicted at those non-observed sites with the curves obtained by a cross-validation process at the monitored sites. Assuming that the volatility of daily PM_{10} concentrations is similar throughout a small neighborhood, ecologists will be correct if the series of PM_{10} predicted at the four non-monitored sites show higher values than the curves obtained at the places where measurements are currently registered.

After this introductory section, the paper is organized as follows: Section 2 presents the basics ideas underlying spatio-temporal and functional kriging. Section 3 puts functional kriging to work in order to obtain PM_{10} functional predictions at four of the most transited places of the city of Madrid. Finally, Section 4 concludes and presents some challenging future research lines.

2. FUNCTIONAL KRIGING FOR SPATIO-TEMPORAL DATABASES. AN ALTERNATIVE TO SPATIO-TEMPORAL KRIGING

As stated in the introductory section, spatio-temporal geostatistics uses series of geo-referenced data sampled or recorded at certain locations of the area under study. Unlike the usual panel data employed in econometrics, the data normally refer to one and only one variable (in fact, a spatio-temporal random function or stochastic process), although this approach can be extended to the multivariate case if a number of auxiliary variables closely related to the main variable are also sampled (in the same spatio-temporal locations as the main variable or in different sites). When studying only one phenomenon, we are interested in taking advantage of the spatio-temporal dependencies existing in such a phenomenon (represented by one variable). Such spatio-temporal dependencies are vitally important for making predictions for the variable at non-observed sites (whatever the temporal instant). When using auxiliary variables closely correlated with the main variable, these auxiliary variables are not considered explanatory variables in the econometric sense, but they act to provide additional data on the main variable. The procedure geostatistics uses for prediction is called kriging, in honor to the mining engineer Daniel Gerardus Krige.

In formal terms, let $\left\{Z\left(\mathbf{s},t\right),\mathbf{s}\in D,t\in T\right\}$ with $D\subset\mathbb{R}^2$ and $t\subset\mathbb{R}$, be a spatio-temporal random function (r.f.) whose value has been observed on a set of n spatio-temporal locations $\left\{Z\left(\mathbf{s}_1,t_1\right),\ldots,Z\left(\mathbf{s}_n,t_n\right)\right\}$. Every spatio-temporal location can be seen as a point on $\mathbb{R}^d\times\mathbb{R}$, $\mathbb{R}^d$ being the d-dimensional Euclidean space and $\mathbb{R}$ the time dimension. Let $C\left(\left(\mathbf{s}_i,t_i\right),\left(\mathbf{s}_j,t_j\right)\right)$ and the time dimension. Let $C\left(\left(\mathbf{s}_i,t_i\right),\left(\mathbf{s}_j,t_j\right)\right)$ and $\gamma\left(\left(\mathbf{s}_i,t_i\right),\left(\mathbf{s}_j,t_j\right)\right)$ be the functions (covariogram and semivariogram, respectively) capturing the structure of the spatio-temporal dependencies existing in the phenomenon under study $C\left(\mathbf{s}_i-\mathbf{s}_j,t_i-t_j\right)=C\left(\mathbf{h},u\right)$, $\left(\mathbf{h},u\right)$ representing a spatio-temporal distance, when the r.f. is second-order stationary, and $\gamma\left(\mathbf{s}_i-\mathbf{s}_j,t_i-t_j\right)=\gamma\left(\mathbf{h},u\right)$ when it is either second order stationary or intrinsically stationary). Point spatio-temporal kriging (block spatio-temporal kriging when the prediction is made over a segment, area or volume) is aimed at predicting an unknown point value $Z\left(\mathbf{s}_0,t_0\right)$ at a non-observed point $\left(\mathbf{s}_0,t_0\right)$. In order to do so, all the information available about the regionalized variable is used, either at the points in the entire domain or in a subset of the domain called the neighborhood. In order to predict the value of a spatio-temporal r.f. at a non-observed point $\left(\mathbf{s}_0,t_0\right)$, geostatistics uses the following linear predictor (the BLUP):

$$Z^*\left(\mathbf{s}_0,t_0\right)=\sum_{i=1}^{n}\lambda_i Z\left(\mathbf{s}_i,t_i\right), \qquad (1)$$

which is a weighted average of the spatio-temporal observed values. The weights, λ_i, are obtained from the kriging equations, which depend on the degree of stationarity attributed to the r.f. that supposedly generates the observed realization. When $Z\left(\mathbf{s},t\right)$ is (i) second-order stationary with unknown constant mean μ and known covariance function $C\left(\mathbf{h},\mu\right)$[3], $\left(\mathbf{h},\mu\right)$ being a spatio-temporal lag distance, or (ii) intrinsically stationary, also with constant but unknown mean μ, but with unbounded variance (this is a common situation in reality), the kriging equations turn to be:

$$\begin{cases} \sum_{i=1}^{n}\lambda_j C\left(\mathbf{s}_i-\mathbf{s}_j,t_i-t_j\right)-\alpha=C\left(\mathbf{s}_i-\mathbf{s}_0,t_i-t_0\right), & \forall i=1,\ldots,n \\ \sum_{i=1}^{n}\lambda_i=1 \end{cases}, \qquad (2)$$

where α is a Lagrange multiplier associated with the condition of unbiasedness. They are known as spatio-temporal ordinary kriging (STOK) equations. The STOK prediction variance is given by:

$$V\left[Z^*\left(\mathbf{s}_0,t_0\right)-Z\left(\mathbf{s}_0,t_0\right)\right]=C\left(\mathbf{0},0\right)-\sum_{i=1}^{n}\lambda_i C\left(\mathbf{s}_i-\mathbf{s}_0,t_i-t_0\right)+\alpha, \qquad (3)$$

[3] In practice the covariance function is not known and should be deduced from the realization observed.

The STOK equations for the specific case where the mean of the r.f. is known are called simple spatio-temporal kriging (STSK) equations, and those for the case where the mean is not constant but there is a drift (the mean of the r.f. depends on the spatio-temporal locations $(\mathbf{s},t)$ are named universal spatio-temporal kriging (STUK) equations. Both of these, as well as details on STSK and STUK prediction variance, can be seen see in Montero *et al.* (2015).

Equations (2) and (3) can be also expressed in semivariogram terms as follows:

$$\begin{cases} \sum_{i=1}^{n} \lambda_j \gamma\left(\mathbf{s}_i - \mathbf{s}_j, t_i - t_j\right) + \alpha = \gamma\left(\mathbf{s}_i - \mathbf{s}_0, t_i - t_0\right), \quad \forall i = 1,\ldots,n \\ \sum_{i=1}^{n} \lambda_i = 1 \end{cases}, \qquad (4)$$

with

$$V\left[Z^*\left(\mathbf{s}_0,t_0\right) - Z\left(\mathbf{s}_0,t_0\right)\right] = \sum_{i=1}^{n} \lambda_i \gamma\left(\mathbf{s}_i - \mathbf{s}_0, t_i - t_0\right) + \alpha \,. \qquad (5)$$

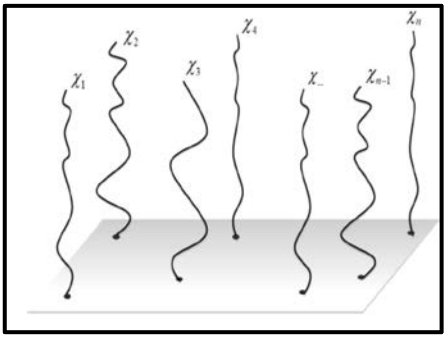
When the r.f. is second-order stationary both types of equations can be used. However, if the r.f. is intrinsically stationary, only the equations in semivariogram terms are operational as the variance of the r.f. does not appear in them.

As can be appreciated, the spatio-temporal kriging equations are expressed in terms of the covariance function -or the semivariogram- that best captures the spatio-temporal dependencies existing in the phenomenon under study. However, in the spatio-temporal case, (i) determining such a covariance function or semivariogram from their respective empirical counterparts is not an easy task. It is especially difficult in the case of non-separable covariance functions (or semivariograms), that is, where the covariance function (or semivariogram) is not constructed by combining two spatial-only and temporal-only covariance functions; and (ii) where the number of spatial locations and/or instants of time is relatively large, solving the spatio-temporal kriging equations becomes a prohibitively difficult task due to the computational burden it implies. In the case of spatio-temporal cokriging the computational burden is even greater, as the spatio-temporal cokriging equations involve spatio-temporal cross-covariograms or cross-semivariograms (in case of second-order stationary vectors); therefore, in addition to *the big n problem***,** there is the *large number of auxiliary variables problem*. For these reasons, we use functional kriging herein as an alternative to spatio-temporal kriging (functional cokriging would be the alternative to spatio-temporal cokriging). The functional alternative proposes a predictor, similar to the spatial-only kriging predictor, but based on functional data (curves) instead of unidimensional data. More specifically, we focus on functional ordinary kriging (FOK), that is, on either second-order or intrinsically stationary spatio-temporal r.f.s.

FOK is an adaptation by Giraldo (2009) of the functional analysis recently developed by Ramsay and Silverman (2005) that tackles the problem of spatial prediction of functional data.

According to Ferraty and Vieu (2006), a random variable, χ , is called a functional variable (f.v.) if it takes values in a dimensional infinite space (or functional space). An observation of that variable, χ, is called a functional observation. Thus, as stated in Levitin *et al.* (2007), a functional datum (or observation) is not a single observation but rather a set of measurements along a continuum that, taken together, are to be regarded as a single entity, curve or image. Functional datum can be thought to represent such a set of measurements (observations). To do this, a smooth curve is fitted to the discrete observations, which approximates the continuous underlying process. Then the discrete points are set aside and the functional objects retained for subsequent analyses, as shown in Figure 1.

Figure 1
Functional data in a set of locations



Source: Own elaboration.

In the FOK strategy, the predicted curve is a linear combination of the functional data (smooth curves) at the observed locations, with the coefficients being real numbers. The statistical statements are as follows: Let us consider a functional random process $\boldsymbol{\chi}_{\mathbf{s}}$: $\mathbf{s} \in D \subseteq \mathbb{R}^d$, usually d=2, such that χ_{s} is a functional variable for any $\mathbf{s} \in D$. Let $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n$ be arbitrary locations in D, and assume that we can observe a realization of the functional random process $\boldsymbol{\chi}_{\mathrm{s}}$ at these sites, $\chi_{\mathrm{s}_1}, \chi_{\mathrm{s}_2}, \ldots, \chi_{\mathrm{s}_n}$. The predictor is the simple ordinary kriging predictor (Cressie 1993) but using curves instead of variables:

$$\chi^*_{\mathbf{s}_0} = \sum_{i=1}^{n} \lambda_i \chi_{\mathbf{s}_i} \ . \qquad (6)$$

That is to say, this approach treats the whole curve as a single entity, and the weights in the predictor give more influence to the curves of locations closer to the prediction point, $\mathbf{s}_0$, than to more distant curves.

In order to find the best linear unbiased predictor (BLUP), we extend the criterion given in Myers (1982) to the functional case, and the *n* weights in the kriging predictor of $\chi_{\mathbf{s}_0}$ are given by the solution of the following optimization problem:

$$\min_{\lambda_1,\lambda_2,\ldots,\lambda_n} \int_T V\left(\chi^*_{\mathbf{s}_0}(t) - \chi_{\mathbf{s}_0}(t)\right) dt \quad \text{with} \quad \sum_{\mathrm{i}=1}^{\mathrm{n}} \lambda_i = 1 \quad \text{(unbiasedness constraint)}. \qquad (7)$$

Observe that the unbiasedness constraint and Fubini theorem imply that:

$$\min_{\lambda_1,\lambda_2,\ldots,\lambda_n} \int_T V\left(\chi^*_{\mathbf{s}_0}(t) - \chi_{\mathbf{s}_0}(t)\right) dt = E\left[\int_T \left(\chi^*_{\mathbf{s}_0}(t) - \chi_{\mathbf{s}_0}(t)\right)^2 dt\right]. \qquad (8)$$

Furthermore, under the second-order stationarity assumption, the integral in the above equation can be written as:

$$\int_T V\left(\chi^*_{s_0}(t) - \chi_{s_0}(t)\right) dt = \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j \int_T C_{ij}(t)\,dt + \int_T \sigma^2(t)\,dt - 2\sum_{i=1}^{n} \lambda \int_T C_{i0}(t)\,dt, \qquad (9)$$

where $C_{ij}(t)$ is the value of the spatial covariance function for the observed locations $\mathbf{s}_i$ and $\mathbf{s}_j$, $C_{i0}(t)$ is the analogous for the observed location $\mathbf{s}_i$ and the unobserved site $\mathbf{s}_0$, and $\sigma^2(t)$ is the variance of the random process.

As a consequence, the objective function can be expressed as:

$$\begin{aligned}\int_T V\left(\chi^*_{\mathbf{s}_0}(t) - \chi_{\mathbf{s}_0}(t)\right) dt = \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j \int_T C_{ij}(t)\,dt + \int_T \sigma^2(t)\,dt - 2\sum_{i=1}^{n} \lambda \int_T C_{i0}(t)\,dt \\ -2\mu\left(\sum_{i=1}^{n} \lambda_i - 1\right)\end{aligned} \ \cdot \qquad (10)$$

The result of the optimization process, in matrix notation and in terms of the trace-covariance (a global measure of space-time dependence), is the following:

$$\begin{pmatrix} \int_T C_{11}(t)\,dt & \cdots & \int_T C_{1n}(t)\,dt & 1 \\ \vdots & \ddots & \vdots & \vdots \\ \int_T C_{n0}(t)\,dt & \cdots & \int_T C_{n0}(t)\,dt & 1 \\ 1 & \cdots & 1 & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_1 \\ \mu \end{pmatrix} = \begin{pmatrix} \int_T C_{10}(t)\,dt \\ \vdots \\ \int_T C_{n0}(t)\,dt \\ 1 \end{pmatrix}, \qquad (11)$$

where $\int_T C_{ij}\left(t\right)dt$ denotes the trace-covariance function of the process evaluated at $h = \left\| \mathbf{s}_i - \mathbf{s}_j \right\|$ (Menafoglio *et al.* 2013).

The FOK prediction trace-variance, which is considered as a global measure of uncertainty, can be easily obtained from the first n equations of the above system of equations:

$$\sigma_{FOK}^2 = \int_T \sigma^2\left(t\right)dt - \sum_{i=1}^{n} \lambda_i \int_T C_{i0}\left(t\right)dt - \mu \,. \tag{12}$$

Since in the stationary case the relationship $\gamma_{\mathbf{s}_i,\mathbf{s}_j}\left(t\right) = \sigma^2\left(t\right) - \gamma C_{\mathbf{s}_i,\mathbf{s}_j}\left(t\right)$ is fulfilled, then, $\int_T C_{ij}\left(t\right)dt = \int_T \sigma^2\left(t\right)dt - \int_T \gamma_{ij}\left(t\right)dt$, and the FOK equations can be written in trace-semivariogram terms, $\left(\int_T \gamma_{ij}\left(t\right)dt\right)$, as follows:

$$\begin{pmatrix} \int_T \gamma_{11}\left(t\right)dt & \cdots & \int_T \gamma_{1n}\left(t\right)dt & 1 \\ \vdots & \ddots & \vdots & \vdots \\ \int_T \gamma_{n1}\left(t\right)dt & \cdots & \int_T \gamma_{nn}\left(t\right)dt & 1 \\ 1 & \cdots & 1 & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_1 \\ -\mu \end{pmatrix} = \begin{pmatrix} \int_T \gamma_{10}\left(t\right)dt \\ \vdots \\ \int_T \gamma_{n0}\left(t\right)dt \\ 1 \end{pmatrix}, \tag{13}$$

so that the FOK prediction trace-variance in terms of the trace-semivariogram can be written as:

$$\sum_{i=1}^{n} \lambda_i \int_T \gamma_{i0}\left(t\right)dt - \mu \,. \tag{14}$$

As in the STOK equations, in the case of second-stationarity both systems of equations can be used. However, if the spatio-temporal r.f. being studied is intrinsically stationary only the FOK equations in trace-semivariogram terms are operative.

In order to solve the system of FOK equations in trace-semivariogram terms, the estimator of the valid trace-semivariogram FOK uses is the following generalization of the classical *Method of Moments* estimator:

$$\hat{\gamma}(h) = \frac{1}{2\#N(h)} \sum_{i,j \in N(h)} \int_T \left(\boldsymbol{\chi}_{\mathbf{s}_i}(t) - \boldsymbol{\chi}_{\mathbf{s}_j}(t)\right)^2 dt \,, \tag{15}$$

where $N(h) = \left\{(\mathbf{s}_i,\mathbf{s}_j) : h = \| \mathbf{s}_i - \mathbf{s}_j \|\right\}$, and $\#N(h)$ is the number of distinct elements in $N(h)$. In the case of data irregularly distributed in space, $N(h) = \left\{(\mathbf{s}_i,\mathbf{s}_j) : h = \| \mathbf{s}_i - \mathbf{s}_j \|\right\}$ is replaced with $N(h) = \left\{(\mathbf{s}_i,\mathbf{s}_j) : \| \mathbf{s}_i - \mathbf{s}_j \| \in (h-\varepsilon, h+\varepsilon)\right\}$, $\varepsilon > 0$ being a small value.

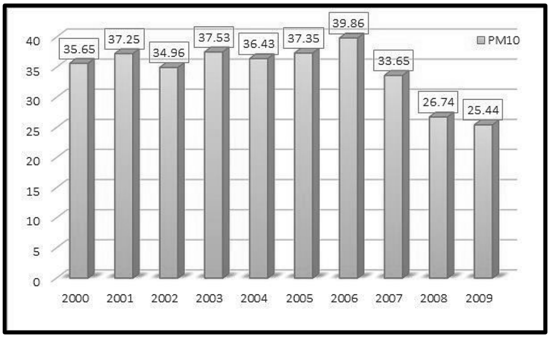
3. CASE STUDY: PREDICTING FUNCTIONAL DATA ON PM_{10} AT HIGHLY TRANSITED SITES IN MADRID, 2000-2009

As stated in the introductory section, particulate matter, together with nitrogen dioxide and, on occasion, tropospheric ozone, are the most problematic pollutants in the city of Madrid. A high level of economic activity translates into a high level of pollutant emissions in an area with low wind speed and a high solar radiation. As stated by the WHO (2011), based on the limits set by the EU Directive in 2008, over 6 million people (14% of the Spanish population) are exposed to polluted air, many of them in Madrid. According to the Spanish Ministry of the Environment, Rural and Marine Resources, around 16,000 people die prematurely each year from air pollution in Spain's large cities (Madrid is the largest city in Spain). This represents a figure eight times more than the death toll caused by traffic accidents. However, based on stricter WHO recommendations, the above figures rise to six times as high[4].

Figure 2

PM_{10} concentrations in the city of Madrid, 2000-2009

(µg/m^3, annual average)



Source: Own elaboration.

Since 1990 the city of Madrid has made great efforts to reduce the emissions of the aforementioned pollutants, although concentrations of PM_{10} are still a serious problem (see Figure 2), and new efforts (including green cars, and easy measures such as promotion of public transport, walking and cycling) must be made to meet the expectations of the European Commission for 2020 on PM -to reduce the loss of life expectancy as a result of exposure to particulate matter by 47%. Figure 2 demonstrates that concentrations of PM_{10} decreased significantly in the city in 2008 and 2009. However, this reduction has been linked to the economic crisis afflicting Spain at that time, meaning that concentrations of PM_{10} will increase again with a recovery in the economic activity.

[4] The annual limit approved in Spain for PM_{10} is twice as high as WHO's guideline.

However, a serious problem arises when measuring real reduction of PM_{10} levels and understanding whether Madrid exceeds the limits set by the EU Directive in 2008 (or the WHO recommendations). Ecologist groups (especially the Ecologist in Action organization) claim that the monitoring stations registering levels of PM_{10} are not located at sites with intensive road traffic but rather (especially since 2009) that they are “deliberately” placed in non-polluted sites. (The official reason given is that they aim to measure PM_{10} in the new areas of the city, although PM_{10} levels are much lower there than in the city centre).

For these reasons, we aim to apply functional geostatistics to predict the functional data (2000-2009) on PM_{10} at four of the most transited sites of the city that, in addition, are severely affected by road traffic: Plaza de Cibeles, Plaza de Callao, Plaza Carlos V (also known as Atocha), and Puerta del Sol. The predicted functional data will be compared with the functional data representative of the series of PM_{10} values at the locations where the operative monitoring stations are located, in order to verify whether the functional data at the four chosen sites are above the functional data of the monitoring stations. If the claims of the ecologists are correct, new locations for the monitoring stations should be considered.

In this section, the area under study is briefly introduced, the raw data on PM_{10} are described, as is their conversion into functional data, and finally, the functional data at the prediction sites are presented and compared with the functional data at the monitored locations.

3.1. The study area

Madrid (the capital of Spain) is the third most populous city in the European Union after London and Berlin with a population of 3,346,441 in 2013 (6,492,468 in the region). As for other capital cities, government institutions, the Parliament, embassies, museums, and central offices of the most relevant companies are located in Madrid. This has made Madrid a large city covering 60,430.76 ha, together with a closely-linked, large peripheral metropolitan area with more than five million inhabitants. Obviously, this implies a great deal of transit of both population and also goods, etc., which has necessitated a complex transportation system.

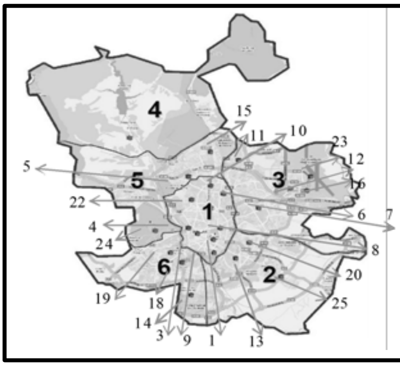
Specifically, Madrid has both a dense ring road network (M-30, M-40, M-45 and M-50) and a dense radial highway network, both of which have enormously improved accessibility to emerging industrial and high economic activity areas, resulting in competitiveness and dynamism. These factors contributed to a 5.6% increase in the number of vehicles in Madrid over the last decade, amounting in 2013 to a total of 1,960,112. This implies 1,213.4 vehicles per km. and 585.7 vehicles per 1,000 inhabitants, with two million drivers transiting though the city on a daily basis. This increase in traffic has an obvious negative environmental impact, as road traffic is the main emission source for PM_{10} (and also other pollutants such as NO_2). In addition, Madrid has the fourth largest European airport and is a centre for train connections - 500 trains enter Madrid from the 10 most important Spanish cities, as well as from Paris and Lisbon. Freight transportation by train is also vitally important in Madrid, with 400 trains entering and leaving the city each day, transporting 150,000 tons of commodities. In fact, Madrid has the largest inland maritime customs centre in Europe. However, as a negative consequence of these factors, transport and specifically road traffic has become the main source of particulate matter. Therefore, in spite of the decrease in PM_{10} levels experienced in recent years, this pollutant still continues to be of particular concern for the inhabitants of Madrid.

3.2. Raw data and functional data

The data used in this article have been provided by the Atmosphere Pollution Monitoring System of Madrid, Spain. Data consists of hourly measurements at 21 fixed operative monitoring stations from January 2000 to December 2009, normalized to a temperature of 293K and a pressure of 101.3 $k\rho_2$. Subsequently, we have computed the daily means, so that the database can be used to compare the concentration of the pollutant with the PM_{10} daily standard for Madrid. Figure 3 shows the location of the air quality monitoring stations, as well as the statistical areas where they are located and the population of such areas. Table 1 reports, for each of the 21 operative monitoring stations (*E*), the main descriptive statistics for the daily mean of the pollutant in the period under study.

Figure 3

Location of the air quality monitoring stations in the city of Madrid



Zone	Population	Area in km^2
1	1,071,003	41
2	593,498	120
3	604,034	100
4	13,498	170
5	300,544	80
6	534,702	89

(*) The monitoring station located in Zone 4 has not been used because it was installed in 2010.

Source: Own elaboration.

As can be seen in Figure 3, most monitoring stations are located in the urban centre and relatively few in peripheral sites. Note that this does represent a reasonable coverage of the study area by the monitoring stations since most of the population is concentred in the urban centre (zone 1) and zones 2, 3 and 6, which also represent areas with the highest intensity of road traffic.

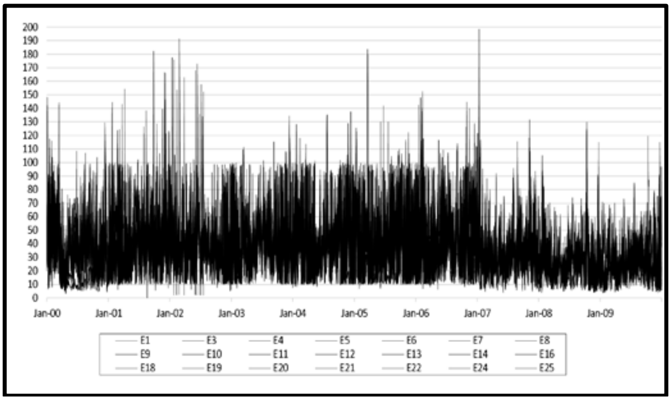
Table 1
PM_{10}: Main descriptive statistics
(µg/m^3)

Monitoring station	Mean	Stand. deviat.	Median	Min.	Max.	Skew.	Kurt.
E1	38.56	19.35	34.59	8.17	163.22	1.31	2.56
E3	32.92	17.72	29.34	6.21	179.91	1.48	4.07
E4	30.47	17.53	27.13	4.14	122.85	1.25	1.98
E5	37.11	20.93	33.02	4.96	157.16	0.95	0.81
E6	39.41	18.90	35.81	6.08	156.54	1.13	1.88
E7	32.93	18.09	29.0	4.68	158.11	1.51	3.64
E8	37.29	21.23	32.72	2.03	91.50	1.84	6.07
E9	41.10	21.91	37.31	4.90	198.46	1.24	2.54
E10	35.73	19.95	31.45	6.24	183.86	1.49	3.46
E11	30.23	17.23	26.69	4.00	154.28	1.43	3.00
E12	32.59	17.30	29.14	3.89	120.42	1.18	1.66
E13	31.73	18.04	27.73	3.28	165.95	1.41	2.81
E14	38.53	21.02	34.57	3.93	151.42	1.08	1.31
E16	30.52	17.97	26.32	3.79	128.71	1.47	2.47
E18	34.34	19.78	29.79	4.85	139.69	1.25	1.68
E19	33.00	18.29	29.02	3.57	130.04	1.27	1.82
E20	31.05	17.08	27.57	5.32	155.52	1.39	2.96
E21	29.93	17.56	25.52	4.72	139.08	1.49	2.79
E22	36.23	19.42	32.12	4.34	157.54	1.24	1.92
E24	30.27	18.65	25.98	3.75	163.49	1.53	3.03
E25	39.77	21.52	36.45	3.80	153.41	0.85	0.56

Source: Own elaboration.

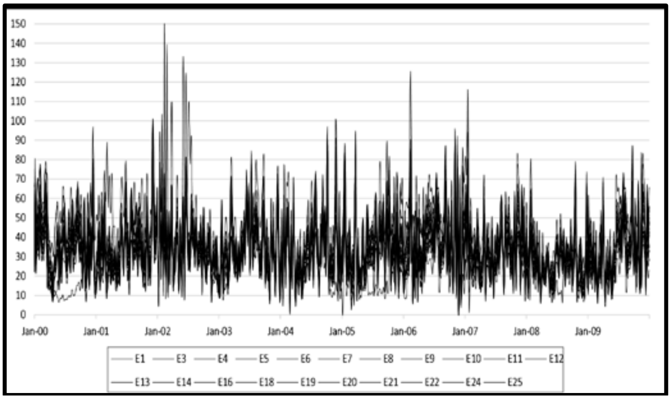
Figure 4 displays the raw data used in the paper, that is, the mean level of PM_{10} at each operative monitoring station on the 3,650 days considered in the analysis. It suggests a strong seasonality, as well as a marked cycle-tendency component (as expected where environmental variables are concerned). These temporal features make it difficult to find parametric models to represent this type of data with a smooth curve (functional data). As a result, in order to construct functional data, we have opted for a non-parametric approach and chose *B*-splines as basis functions to obtain smooth curves from the original series. The cross-validation process was used to find the optimum number of inner knots (L) and the value of the parameter that controls the trade-trade off between the fit of the data observed and the smoothness of the approximating spline (η), also known as a roughness penalty. On the basis of a preliminary exploration of the data, 250, 251, 252, 253… 350, and 0, 1, 10, 10^2, 10^3, were considered as the possible values of L and η, respectively. The values $L^* = 275$ and $\eta^* = 0$ were found as optimal. The resulting functional data are depicted in Figure 5. It can be seen that the functional data obtained for the 21 operative monitoring stations in the city of Madrid in the period 2000-2009 display less variability than the raw data, due to the smoothing process. They capture the trend and seasonality of the corresponding raw series but do not take into account the most extreme variability (the peaks).

Figure 4
Raw data. PM_{10} mean level, 2000 - 2009



Source: Own elaboration.

Figure 5
Functional data. PM_{10} mean level, 2000 - 2009



Source: Own elaboration.

3.3. Results

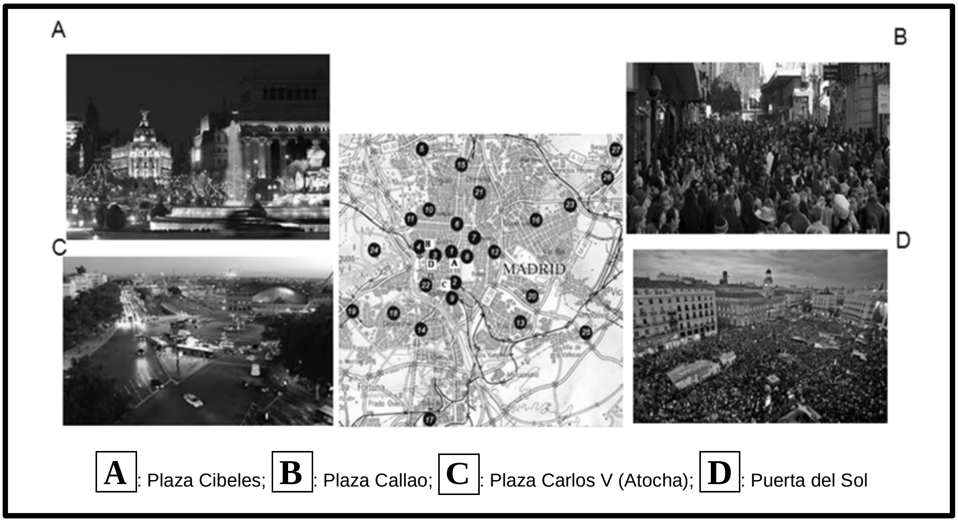
Once the functional data have been constructed, the spatial correlations of such functional data must be represented by a valid semivariogram. In light of the empirical trace-semivariogram for the functional data (which is not presented because it appears in form of a "movie"), and taking into account the results of the cross-validation process, the isotropic exponential semivariogram model in (16) was used for prediction tasks:

$$\gamma(|\mathbf{h}|) = m\left(1-\exp\left(-\frac{|\mathbf{h}|}{\alpha}\right)\right), \tag{16}$$

where m is the sill, which is only asymptotically reached, and 3α is the effective range, which is the distance above which the semivariogram value is $0.95m$.

On the basis of both the functional data constructed from the raw data on PM_{10} and also an exponential semivariogram, we proceeded to the prediction of the PM_{10} functional data (2000-2009) at four non-monitored sites corresponding to the most problematic locations in Madrid, with regard to PM_{10} pollution. Those sites are (A) Plaza de Cibeles, (B) Plaza Callao, (C) Plaza Carlos V, and (D) Puerta del Sol (see Figure 6). Figure 6 demonstrates the relative activity at these sites, so that, at high levels of PM_{10}, the exposed population is also very large.

Figure 6
Prediction sites (A) Plaza de Cibeles, (B) Plaza Callao, (C) Plaza Carlos V, and (D) Puerta del Sol

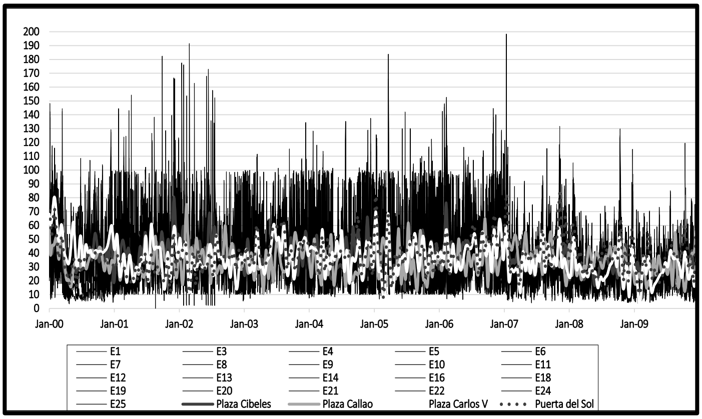

A
B
C
D
MADRID
A: Plaza Cibeles; B: Plaza Callao; C: Plaza Carlos V (Atocha); D: Puerta del Sol

Source: Own elaboration.

Figures 7 and 8 show the predicted (kriging) curves along with the original functional data (Figure 7) and the kriging functional data (Figure 8). Figure 7 suggests that the PM_{10} levels at the prediction sites is lower than in the sites where the monitoring stations are located, except for the economic crisis period

(2007-2009) when the series of the pollutant at monitored sites experienced a strong decrease.

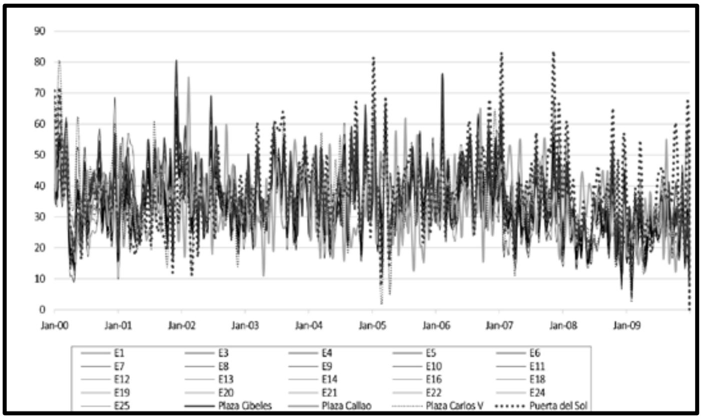
Figure 7
Comparison of the observed and predicted functional data. PM_{10} mean level, 2000 – 2009



Source: Own elaboration.

However, comparing original and (kriging) predicted functional data is not a fair comparison. The functional data for the prediction sites display less variability than the functional data representing the data observed at the monitored sites, and this different variability is a consequence of kriging being a smoothing method (leading to an additional decrease in variance). A fairer comparison should not involve the original functional data for the sites where the monitoring stations are located but instead the kriging prediction of such curves, based on the functional data computed for the other 20 monitoring stations (that is, using a cross-validation or leave-one-out procedure). This comparison is shown in Figure 8 and Table 2, where it can be seen that the claims of the ecologist group are correct: the functional data predicted at Plaza Cibeles and Puerta del Sol are clearly above the curves obtained for the monitored sites. The curves for Plaza Callao and Plaza Carlos V are also among the highest. In addition, the variability underlying the functional data for Plaza Carlos V and Puerta del Sol is much higher than that of the kriging curves for the monitored sites, meaning that at such prediction sites, the most transited sites of the city, the probability of a violation of the legal standard for the pollutant is much higher than at the monitored sites. The volatility underlying the kriging curve for Plaza Callao and Plaza Cibeles is among the highest of the kriging curves obtained for the monitored sites, which implies that at these two sites the risk of violating the legal standard is also high.

Figure 8
Comparison of the kriging functional data at the locations observed and prediction sites.
PM10 mean level, 2000 – 2009



Source: Own elaboration.

Table 2
PM_{10} kriging functional data: Main descriptive statistics
($\mu g/m^3$)

Monitoring Station	Mean	Stand. deviat.	Median	Min.	Max.	Skew.	Kurt.
E1	36.07	9.51	35.24	12.51	72.93	0.41	0.20
E3	35.55	9.29	34.66	9.38	71.46	0.49	0.47
E4	32.37	8.36	31.62	7.81	59.46	0.26	-0.05
E5	33.88	8.86	33.52	8.02	61.72	0.14	-0.14
E6	33.42	8.44	32.84	9.65	65.46	0.41	0.26
E7	36.41	9.36	35.72	14.45	69	0.28	-0.22
E8	36.39	8.92	35.75	12.72	66.76	0.32	0.06
E9	36.95	9.63	35.84	11.13	67.3	0.28	-0.17
E10	33.83	9.02	33.28	7.06	61.51	0.30	-0.21
E11	36.62	9.23	36.16	6.37	66.65	0.14	-0.17
E12	32.58	8.75	31.96	10.88	60.86	0.26	-0.12
E13	34.60	9.33	33.83	7.80	63.74	0.17	-0.18
E14	36.02	10.28	35.40	4.30	68.65	0.31	-0.01
E16	34.34	9.14	33.87	6.59	61.42	0.12	-0.24
E18	36.05	9.83	35.26	7.43	66.26	0.21	-0.21
E19	31.68	9.29	31.43	4.75	59.56	0.22	-0.19
E20	32.74	9.26	32.4	7.92	62.65	0.11	-0.20
E21	33.85	9.71	33.03	5.98	66.6	0.43	0.31
E22	35.18	9.93	33.59	3.68	75.87	0.48	0.37
E24	33.19	9.05	33.07	7.19	60.12	0.05	0.05
E25	33.51	8.81	32.93	8.09	61.94	0.15	-0.14

Table 2 ***(continue)***
PM$_{10}$ kriging functional data: Main descriptive statistics
(µg/m^3)

Monitoring Station	Mean	Stand. deviat.	Median	Min.	Max.	Skew.	Kur.
Plaza Cibeles	37.59	9.76	36.78	14.36	80.58	0.55	0.60
Plaza Callao	34.19	9.61	33.68	11.00	75.15	0.41	0.24
Plaza Carlos V	36.17	11.73	35.83	1.70	80.50	0.26	0.49
Puerta del Sol	38.70	11.87	37.78	7.73	83.60	0.55	0.68

Source: Own elaboration.

4. CONCLUSIONS AND FUTURE RESEARCH LINES

As pollution data have been recorded in large cities by monitoring stations on a (usually) hourly basis since the 1990s, they constitute a simple but interesting one-factor panel data set. Since they are available in a spatio-temporal format, they can be incorporated in traditional panel data sets including data on explanatory variables of the level of pollution, or data on explanatory variables of another variable of interest, the pollution variable being one additional explanatory variable. This fact, together with the advantages of panel data modelling, has led to a proliferation of empirical panel data studies using air pollution data.

In this article we focussed on how geostatistics deals with one-factor (the most frequent situation in the geostatistical field) or multi-factor panel data, which is significantly different from how econometrics deals with such data. The econometric approach focuses on estimating the model parameters. Geostatistics focuses directly on prediction rather than on parameter estimation and, to make spatio-temporal kriging or cokriging predictions, takes advantage of the spatio-temporal dependencies existing in the data. However, spatio-temporal kriging or cokriging predictions imply a considerable computational burden when the number of spatio-temporal locations is large (*the big n problem*). In this article, we propose functional kriging as a method to overcome this problem from the geostatistical perspective when the available information is in the form of a panel data set and, thus, not just the spatial-only or temporal-only dependencies, but also the spatio-temporal ones existing in the data, can be used to make spatio-temporal predictions. Going even further, one can reproduce the history of the phenomenon under study in a non-observed location.

Given that there is generalized consensus that air quality control is of crucial importance in large cities, and any help to prevent a violation of the air quality standards is of particular interest for the authorities responsible for the environment, we address the PM$_{10}$ concentration problem in the city of Madrid. More specifically, we address this problem by implementing an ordinary version of functional kriging, the tool functional geostatistics uses to make predictions on the basis of a one-factor or multi-factor panel data (spatio-temporal databases in general). We focus on PM$_{10}$, as it is particularly detrimental to human health and its concentrations in large cities still are a significant problem for inhabitants. We chose Madrid as the study area because PM$_{10}$ continues to be one of the worst air pollution problems to affect the city. According to current legislation and limits set down within this legislation, levels of PM$_{10}$ are not satisfactory in the city, although due to the economic crisis in recent years, there has been a significant decrease.

We implement the ordinary version of functional kriging in order to predict smooth curves representing a PM$_{10}$ series at four non-monitored sites in Madrid that are daily transited by thousands of people and have high road traffic intensity. Ecologist groups claim that at such non-monitored sites PM$_{10}$ levels have higher values in comparison to the locations where the monitoring stations are currently placed. The main objective of this study was to compare the curves predicted at those non-observed sites with the curves obtained by a cross-validation process at the monitored sites. Assuming that the daily volatility of the PM$_{10}$ registers is similar at different points within a small neighbourhood, ecologists will be correct if the PM$_{10}$ series predicted at the four non-monitored sites show higher values than the curves obtained at the currently monitored sites. Following our analysis, we conclude that the ecologist groups are correct in their claim. At the four prediction sites, both the level and volatility of the PM$_{10}$ concentrations have been, and currently are, either significantly higher than at the monitored sites or among the highest readings. As a consequence, a new configuration of the monitoring network in the city of Madrid should be implemented.

Although functional kriging could be one interesting alternative to both spatio-temporal kriging and the econometric approach to panel data when dealing with spatio-temporal databases, there is much work still to do. For example, limiting ourselves to the geostatistical perspective and to the case under study, as the smoothing process that converts the raw data into functional data captures both the trend and the seasonality of the raw data but does not capture its extreme variability (the peaks), functional data are not optimal tools for predicting exceedances over the daily legal standard of the pollutant under study. For this reason, a promising and challenging avenue of research is the appropriate modelling of the volatility component in order to be properly combined with the functional data. We use the word “appropriate” because one stylized fact that environmental series share with financial series is the asymmetric response of the volatility. Consequently, the model used to account for the volatility must consider this stylized fact. In our opinion, the best option is the Threshold Autoregressive Asymmetric Stochastic Volatility (TA-ARSV) model by García-Centeno and Minguez-Salido (2009a,b), a strategy which includes two new parameters in the volatility equation of the ARSV model (ϕ_{11} and ϕ_{12}), used to capture the asymmetric behaviour of volatility. A more challenging option is a

TA-ARSV model including more than one threshold. The combination functional kriging-TA-ARSV model could perfectly explain the dynamics of the pollution series and serve as a warning system for violations of the legal standards of the pollutant under study. The only weak point of this combined strategy is the prediction of ϕ_{11} and ϕ_{12} at the non-observed locations; although this does not represent an unsolvable problem as these coefficients are spatially correlated, and can be predicted using kriging techniques.

Other interesting initiatives in the air pollution field from the functional geostatistics perspective are the implementation of two challenging proposals by Giraldo (2009): functional ordinary kriging and cokriging of air pollution data with functional weights. Of course, this proposal could be combined with a TA-ARSV model to warn of violations of the legal standard of the pollutants under control.

From the perspective of spatial and spatio-temporal econometrics there is also much work still left to do in the air pollution arena. Although recently the spatial versions of the traditional strategies have been applied to the environmental sciences in general and to air quality in particular, we find spatio-temporal econometric strategies which focus (and estimate) at the same time on the large- and short-scale spatio-temporal correlation especially interesting. For this reason our proposal from this perspective is to extend the non-parametric P-spline approach applied to SAR models with nonparametric spatial trends by Montero *et al.* (2012) to the spatio-temporal framework, which, in addition, allows the estimation of the smoothing parameter for the P-spline together with the parameters for the short and large scale correlation.

REFERENCES

ACAR, S. and TEKCE, M. (2014). “Economic Development and Industrial Pollution in the Mediterranean Region: A Panel Data Analysis”. *Topics in Middle Eastern and African Economies*, 16(1), pp. 65-95.

ACHAR, J.A.; FERNÁNDEZ-BREAMAUNTZ, A.A.; RODRIGUES, E.R. and TZINTZUN, G. (2008). “Estimating the number of ozone peaks in Mexico City using a non-homogeneous Poisson model”. *Environmetrics*, 19(5), pp. 469-485.

AKBOSTANCI, E.S.; TURUT-ASIK, S. and TUNC, G.I. (2009). “The relationship between income and environment in Turkey: Is there an Environmental Kuznet Curve?”. *Energy Policy*, 37, pp. 861-867.

ATKINSON R.W.; BARRATT, B.; ARMSTRONG, B.; ANDERSON, H.R.; BEEVERS, S.D.; MUDWAY, I.S.; GREEN, D.; DERWENT, R.G.; WILKINSON, P.; TONNE, C. and KELLY, FJ. (2009): “The impact of the Congestion Charging Scheme on ambient air pollution concentrations in London”. *Atmospheric Environment*, 43, pp. 5493-5500.

AUFFHAMMER, M.; BENTO, A.M. and LOWE, S.E. (2009). "Measuring the effects of the Clean Air Act Amendments on ambient PM10 concentrations. The critical importance of a spatially disaggregated approach". *Journal of Environmental Economics and Management*, 58, pp. 15-26.

AUFFHAMMER, M. and KELLOGG, R. (2011). "Clearing the air. The effects of gasoline content regulation on air quality". *American Economic Review*, 101, pp. 2687-2722.

AYRES, J.G. (2002). "Chronic effects of air pollution". *Occupational and Environmental Medicine*, 59, pp. 147-148.

BANERJEE, S.; CARLIN, B.P. and GELFAND, A.E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Boca Raton, Florida: Chapman & Hall/CRC.

BANERJEE, S.; GELFAND, A.E.; FINLEY, A.O. and SANG, H. (2008). "Stationary process approximation for the analysis of large spatial datasets". *Journal of the Royal Statistical Society*, Series B-Statistical Methodology, 70, 825-848.

BARAI, S.V.; GUPTA, A.K. and KODALI, J. (2009). "Air Quality Forecaster: Moving Window Based Neuro Models". In E. AVINERI (Ed.): *Applications of Soft Computing*, pp. 137-146, ASF 52. Berlin: Springer-Verlag.

BAXT, W. and WHITE, H. (1995). "Bootstrapping confidence intervals for clinical input variable effects in a network trained to identify the presence of acute myocardial infarction". *Neural Computation*, 7, pp. 624-638.

BURROWS, W.R.; BENJAMIN, M.; BEAUCHAMP, S.; LORD, E.R.; McCOLLOR, D. and THOMSON, B. (1995). "CART decision tree statistical analysis and prediction of summer season maximum surface ozone for the Vancouver, Montreal, and Atlantic Regions of Canada". *Journal of Applied Meteorology*, 34, pp. 1848-1862.

CAMPALANI, P.; MANTOVANI, S. and BAUMANN, P. (2014). "Spatiotemporal Interactions for Daily Mapping of PM_{10} with MODIS and Meteorological Data". In E. Pardo-Igúzquiza *et al*. (eds): *Mathematics of Planet Earth* (pp. 239-243), Lecture Notes in Earth System Sciences. Berlin: Springer.

CHALOULAKOU, A., ASSIMACOPOULOS, D. and LEKKAS T. (1999). "Forecasting daily maximum O_3 concentrations in the Athens Basin". *Environmental Monitory and Assessment*, 56, pp. 97-112.

CHAY, K.Y. and GREENSTONE, M. (2003). "Air quality, infant mortality, and the Clean Air Act of 1970". *NBER Working Paper No.* W10053.

COMRIE, A. (1997). "Comparing neural networks and regression models for ozone forecasting". *Journal of the Air & Waste Management Association*, 47, pp. 653-663.

CRESSIE, N. (1993). *Statistics for spatial data*. New York: Wiley.

CRESSIE, N. and JOHANNESSON, G. (2008). "Fixed rank Kriging for very large spatial data sets". *Journal of the Royal Statistical Society*, Series B-Statistical Methodology,70, pp. 209-226.

CRESSIE, N. and WIKLE, C.K. (2011). *Statistics for Spatio-Temporal Data*. Hoboken (New Jersey): Wiley.

DAVIS, L.W. (2008). "The Effect of Driving Restrictions on Air Quality in Mexico City". *Journal of Political Economy*, 116(1), pp. 38-79.

DE IACO, S.; PALMA, M. and POSA, D. (2013). "Prediction of particle pollution through spatio-temporal multivariate geostatistical analysis: spatial special issue". *Advances in Statistical Analysis*, 97(2), pp.133-150.

ERCELEBI, S.G. and TOROS, H. (2009). "Extreme Value Analysis of Istanbul Air Pollution Data". *Clean*, 37(2), pp. 122-131.

FASSO, A. and FINAZZIA, F. (2011). "Maximum likelihood estimation of the dynamic coregionalization model with heterotopic data". *Environmetrics*, 22(6), pp.735-748.

FENG, X.; LI, Q; ZHU, Y.; WANG, J.; LIANG, H.; XU, R. (2014). "Formation and dominant factors of haze pollution over Beijing and its peripheral areas in winter". *Atmospheric Pollution Research*, 5, pp. 528-538.

FERRATY, F. and VIEU, P. (2006). *Non parametric functional data analysis. Theory and practice.* New York: Springer.

GARCÍA CENTENO, M.C. and MÍNGUEZ SALIDO, R. (2009a). "Estimation of Asymmetric Stochastic Volatility Models for Stock-Exchange Returns". *International Advances in Economic Research*, 15, pp. 71-87.

GARCÍA CENTENO, M.C. and MÍNGUEZ SALIDO, R. (2009b). "Asymmetric Stochastic Volatility Models and Multicriteria Decision Methods in Finance". *AESTIMATIO, the IEB International Journal of Finance*, 3, pp. 2-23.

GARDNER, M. and DORLING, S. (2000). "Statistical surface ozone models: an improved methodology to account for non-linear behaviour". *Atmospheric Environment*, 34, pp. 21-34.

GIRALDO, R. (2009). *Geostatistical Analysis of Functional Data*. Ph.D. thesis. Barcelona: Universidad Politécnica de Cataluña.

GRÄLER, B.; GERHARZ, L.; PEBESMA, E. (2011). "Spatio-temporal analysis and interpolation of PM10 measurements in Europe". European Topic Centre on Air Pollution and Climate Change Mitigation (ETC/ACM) Technical Paper 2011/10, The Netherlands.

GUPTA, U. (2011). "Estimation of Welfare Losses from Urban Air Pollution Using Panel Data from Household Health Diaries". Conference Paper presented in Sustaining Commons: Sustaining Our Future, the Thirteenth Biennial Conference of the International Association for the Study of the Commons, Hyderabad, India, January 10-14.

HARTMAN, L. and HOSSJER, O. (2008). "Fast kriging of large data sets with Gaussian Markov random fields". *Computational Statistics & Data Analysis*, 52, pp. 2331-2349.

HSIAO, C. and YANAN, W. (2006). "Panel Data Analysis - Advantages and Challenges". *WISE Working Paper* No. 602, Xiamen University, China.

HUBBARD, M.C. and COBOURN, W.G. (1998). "Development of a Regression Model to Forecast Ground Level Ozone in Louisville, Kentucky". *Atmospheric Environment*, 32, pp. 2637-2647.

HURAIRAH, A.; IBRAHIM, N.A.; DAUD, I.B. and HARON, K. (2005). "An application of a new extreme value distribution to air pollution data". *Management Environmental Quality: An International Journal,* 16(1), pp. 17-25.

HUSSAIN, I.; KAZIANKA, H.; PILZ, J. and FAISAL, M. (2013). "Spatio-Temporal Modelling of Particulate Matter Concentrations Including Covariates". *Science International*, 25(1), pp. 15-21.

JAFFE, A.B. and PALMER, K. (1996). "Environmental Regulation and Innovation; A Panel Data Study". NBER Working Paper Series, Cambridge M.A.

JANES, H.; SHEPPARD, L. and SHEPERD, K. (2008). "Statistical Analysis of Air Pollution Panel Studies: An Illustration". *Annals of Epidemiology*, 18, pp.792-802.

KAN, H.D. and CHEN, B.H. (2004). "Statistical distributions of ambient air pollutants in Shanghai, China". *Biomedical and Environmental Sciences*, 17(3), pp. 366-372.

LAURETI T.; MONTERO, J.-M. and FERNÁNDEZ-AVILES, G. (2014). “A local scale analysis on influencing factors of NO_x emissions: Evidence from the Community of Madrid, Spain”. *Energy Policy*, Available online 22 July 2014.

LEVITIN, D.J.; NUZZO, R.L.; BRADLEY, W.V. and RAMSAY, J.O. (2007). “Introduction to Functional Data Analysis”. *Canadian Psychological* 48(3), pp. 135-155.

LIANG, D. and KUMAR, N. (2013). “Time-space kriging to address the spatio-temporal misalignment in the large datasets”. *Atmospheric Environment*, 72, pp. 60-69.

LIM, Y.-R.; BAE, H.-J.; LIM, Y.-H.; YU, S.; KIM, G.-B. and CHO, Y.-S. (2014). “Spatial analysis of PM10 and cardiovascular mortality in the Seoul metropolitan area”. *Environmental Health and Toxicology*, 29, pp. 9-16.

LU, H.C. and FANG, G.C. (2003). “Predicting the exceedances of a critical PM10 concentration, a case study in Taiwan”. *Atmospheric Environment,* 37, pp. 3491-3499.

MALINA, C. and FISCHER, F. (2012). “The impact of low emission zones on PM10 levels in urban areas in Germany”. *CAWM Discussion Paper No.* 58.

MENAFOGLIO, A.; SECCHI, P. and DALLA ROSA, M. (2013). “A Universal Kriging predictor for spatially dependent functional data of a Hilbert Space”. *Electronic Journal of Statistics*, 7, pp. 2209-2240.

MONTERO, J.M.; MINGUEZ, R. and DURBAN, M. (2012). “SAR models with nonparametric spatial trends. A P-spline approach”. *Estadística Española*, 54(177), pp. 89-111.

MONTERO-LORENZO, J.M.; FERNÁNDEZ-AVILÉS, G.; MONDÉJAR-JIMÉNEZ, J. and VARGAS-VARGAS, M. (2013). “A spatio-temporal geostatistical approach to predicting pollution levels: The case of mono-nitrogen oxides in Madrid”. *Computers, Environment and Urban Systems*, 37, pp 95-106.

MONTERO LORENZO, J.M.; GARCÍA-CENTENO, M.C and FERNÁNDEZ-AVILÉS, G. (2011). “A Threshold Autoregressive Asymmetric Stochastic Volatility Strategy to Alert of Violations of the Air Quality Standards”. *International Journal of Environmental Research*, 11, pp.155-177.

MONTERO, J.M.; FERNÁNDEZ-AVILÉS, G. and MATEU, J. (2015). *Spatial and Spatio-Temporal Geostatistical Modeling and Kriging*. Hardback (USA): Wiley.

MYERS, D.E. (1982). “Matrix formulation of co-kriging”. *Journal of the International Association for Mathematical Geology* 14(3), pp. 249-257.

NGUYEN, Q.T.; SKOV, H.; SØRENSEN, L.L., JENSEN, B.J.; GRUBE, A.G.; MASSLING, A.; GLASIUS, M. and NØJGAARD, J.K. (2013). “Source apportionment of particles at Station Nord, North East Greenland during 2008-2010 using COPREM and PMF analysis”. *Atmospheric Chemistry and Physics*, 13, pp. 35-49.

OMAY, R.E and CANPOLAT, E. (2013). “NonParametric Fixed Effect Panel Data Models: Relationship between Air Pollution and Income for Turkey”. *Anadolu University Journal of Science and Technology -A, Applied Sciences and Engineering,* 14(1), pp. 91-103.

POLLICE, A. and LASINIO, G.J. (2010). “Spatiotemporal analysis of the PM10 concentration over the Taranto area”. *Environmental Monitoring and Assessment*, 162(1-4), pp 177-190.

POPE, C.A. and DOCKERY, D.W. (2013). “Air pollution and life expectancy in China and beyond”. *Proceedings of the National Academy of Sciences, USA*, 110, pp. 12861-12862.

RAHMAN, A.F.M and PORNA, A.K. (2014). “Growth Environment Relationship: Evidence from Data on South Asia”. *Journal of Accounting, Finance and Economics*, 4(1), pp. 86-96.

RAMSAY, J. and SILVERMAN, B. (2005). *Functional data analysis*, second edition. New York: Springer.

ROBESON, S.M. and STEYN, D.G. (1990). “Evaluation and comparison of statistical forecast models for daily maximum ozone concentrations”. *Atmospheric Environment*, Part B, 24(2), pp. 303-312.

ROBERTS, E.M. (1979a). “Review of statistics extreme values with applications to air quality data. Part I. Review”. *Journal of the Air Pollution Control Association*, 29, pp. 632-637.

ROBERTS, E.M. (1979b). “Review of statistics extreme values with applications to air quality data. Part II. Applications”. *Journal of the Air Pollution Control Association*, 29, pp. 733-740.

ROST, J.; HOLST, T.; SAHN, E.; KLINGNER, M.; ANKE, K.; AHRENS, D. and MAYER, H. (2009). “Variability of PM_{10} concentrations dependent on meteorological conditions”. *International Journal of Environment and Pollution,* 36(1-3), pp. 3-18.

SFETSOS, A.; ZORAS, S.; BARTZIS, J.G. and TRIANTAFYLLOU, A.G. (2006). “Extreme value modelling of daily PM_{10} concentrations in an industrial area”. *Fresenius Environmental Bulletin*,15(8), pp. 841-845.

SELDEN, T.M. and SONG, D. (2013). “Environmental Quality and Development: Is There a Kuznets Curve for Air Pollution Emissions?”. *Journal of Environmental Economics and Management*, 27(2), pp. 147-162.

SHARMA, S.; BARAI, S.V. and DIKSHIT, A.K (2003). “Studies of air quality predictors based on neural networks”. *International Journal of Environment and Pollution*, 19(5), pp. 442-453.

SHARMA, P.; KHARE, M. and CHAKRABARTI, S.P. (1999). “Application of extreme value theory for predicting violations of air quality standards for an urban road intersection”. *Transportation Research*, Part D, 4, pp. 201-216.

SHERMAN, M. (2011). *Spatial Statistics and Spatio-Temporal Data*. Chichester (U.K.): Wiley.

SO, M.K.J.; LI, W.K. and LAM, K. (2002). “A threshold stochastic volatility model”. *Journal of Forecasting*, 21, pp. 473-500.

SURMAN, P.G.; BODERO, J. and SIMPSON, R.W. (1987). “The prediction of the numbers of violations of standards and the frequency of air pollution episodes using extreme value theory”. *Atmospheric Environment.*, 21, pp.1843-1848.

TURÓCZI, B.; HOFFER, A.; TÓTH, Á.; KOVÁTS, N.; ÁCS, A.; FERINCZ, Á.; KOVÁCS, A. and GELENCSÉR, A.(2012). “Comparative assessment of ecotoxicity of urban aerosol”. *Atmospheric Chemistry and Physics*, 12, pp. 7365-7370.

van AALST, R. and de LEEUW, F. (1997). “National ozone forecasting systems and international data exchange in northwest Europe”. *EEA Technical Report*, 9, p. 50.

WANG, D. and LU, W.-Z. (2006). “Interval estimation of urban ozone level and selection of influential factors by employing automatic relevance determination model”. *Chemosphere*, 62, pp. 1600-1611.

WELSCH, H. (2002). “Preferences over Prosperity and Pollution: Environmental Valuation Based on Happiness Surveys”. *Kyklos*, 55, pp. 473-494.

WELSCH, H. (2006). “Environment and Happiness: Valuation of Air Pollution Using Life Satisfaction Data”. *Ecological Economics*, 58, pp. 801-813.

WORLD HEALTH ORGANIZATION (2002). *Reducing Risks, Promoting Healthy Life*. Geneva: The World Health Organization.

WORLD HEALTH ORGANIZATION (2011). “Car emissions: A problem in Spain”. *Why Newsletter*, 4, p. 2.

YANOSKY, J.D.; PACIOREK, CH.J. ; SCHWARTZ, J.; LADEN, F.; PUETT, R. and SUH, H.H. (2008). “Spatio-temporal modeling of chronic PM10 exposure for the Nurses' Health Study”. *Atmospheric Environment*, 42(18), pp. 4047-4062.

ZHANG, K.; LARSON, T.V.; GASSETT, A.; SZPIRO, A.A.; DAVIGLUS, M.; BURKE, G.L.; KAUFMAN, J.D.; and ADAR, S.D. (2014). “Characterizing Spatial Patterns of Airborne Coarse Particulate (PM10–2.5) Mass and Chemical Components in Three Cities: The Multi-Ethnic Study of Atherosclerosis”. *Environmental Health Perspectives*, 122(8), pp. 823-831.