MODELLISTICA E CALCOLO SCIENTIFICO

MODELING AND SCIENTIFIC COMPUTING

MOX–Report No. 18/2008

Geostatistical estimate of PM10 concentrations in Northern Italy: validation of kriging reconstructions with classical and flexible variogram models

Luca Bonaventura, Stefano Castruccio,
Paola Crippa, Giovanni Lonati

MOX, Dipartimento di Matematica "F. Brioschi"
Politecnico di Milano, Via Bonardi 29 - 20133 Milano (Italy)

mox@mate.polimi.it http://mox.polimi.it

Geostatistical estimate of PM10 concentrations in Northern Italy: validation of kriging reconstructions with classical and flexible variogram models

Luca Bonaventura ♯, Stefano Castruccio ♯,
Paola Crippa♭, Giovanni Lonati♭

18th July 2008

♯ MOX – Modelling and Scientific Computing,
Dipartimento di Matematica "F. Brioschi", Politecnico di Milano
Via Bonardi 9, 20133 Milano, Italy
luca.bonaventura@polimi.it

♭ Dipartimento di Ingegneria Idraulica Ambientale e del Rilevamento
Politecnico di Milano
Milano, Italy
giovanni.lonati@polimi.it

Keywords: Geostatistical interpolation, statistics of random fields, fine particulate matter, cross validation, flexible variogram models.

AMS Subject Classification: 62-07, 62H99, 62P12, 62P30, 62F03.

Abstract

The applicability of classical geostatistical tools to the reconstruction of PM10 concentration fields over the entire Po Valley has been assessed, based on a large dataset of daily PM10 data spanning the period 2003-2006. The impact of data detrending by the median polish procedure and of the variogram model chosen for the geostatistical estimates have been investigated, by comparison of the results obtained with several isotropic variogram models as well as with anisotropic flexible variogram models. The relative merits of the different approaches were evaluated by cross-validating the resulting reconstructions and performing normality tests on the corresponding residuals. Although exponential and linear variograms yield reliable reconstructions in most of the cases, the analysis has highlighted significant seasonal and interannual variations in the basic features of the estimated concentration fields and residual correlation structure. As a consequence, none of the classical models is able to cope with all the different situations encountered, while the anisotropic flexible variogram models appear to provide a more robust tool for automatic reconstruction of the PM10 concentration fields without expert user intervention.

1 Introduction

The present study is devoted to the assessment of geostatistical techniques for estimation of concentration values of particulate matter with diameter less than 10 μm over Northern Italy and in paerticular over the Po Valley. A precise estimate of the PM10 concentration values is of paramount importance for air quality management, since epidemiological studies have shown adverse human health effects produced by the exposure to high PM10 concentrations (see e.g. [12]). Worldwide guidelines and regulations fix limits of ambient PM10 mass concentrations (see e.g. [9]), which are well known to be exceeded over large periods of time in many areas of Northern Italy. Respiratory and cardiovascular disorders are associated both to the particles toxicity and to their small dimensions, thanks to which particles can penetrate in the respiratory and even in the circulatory system. Geostatistical interpolation techniques are an attractive option to monitor this phenomenon in a systematic and automatic way, especially considering how the need for reliable concentration estimates contrasts with the sparse and inhomogeneous nature of the available measurement network. It would be desirable to develop a tool that can provide such estimates in real time with minimum expert user intervention, in order to support environmental management decisions that have often various economic and social implications.

In this paper, the applicability of kriging reconstruction techniques (see e.g. [5], [13], [14]) to PM10 concentration fields has been assessed, using a large dataset of PM10 pollution data covering the entire Po Valley over a time span of 4 years. In the present study, only more standard geostatistical procedures were considered, in order to establish a first reference for further analyses to be carried out with more advanced techniques using a Bayesian approach along the lines proposed in [2]. The data were analyzed in order to exclude measurement stations not covering a sufficiently large time span in each year and a smaller dataset consisting of approximately 90% of the original data was selected, in order to eliminate possible boundary effects and to exclude isolated stations that were identified as outliers. Seasonal means were then computed, from which the geostatistical estimates were derived.

In order to assess the relative merits of different possible approaches to geostatistical interpolation of these data, ordinary kriging reconstructions were carried out with various variogram models. From a preliminary screening, exponential and linear variogram models were identified as the most appropriate to fit the empirical variograms obtained. Along with classical isotropic variograms, the anisotropic flexible variogram model introduced in [1] was also employed. In spite of not having received much attention since it was proposed, this variogram seems to yield a very attractive possibility for variogram estimation, since it only requires minimal *a priori* assumptions on the variogram functional form and it can very naturally cope with anisotropic data. Furthermore, since the PM10 concentration fields can be assumed to have a significant deterministic mean component depending on meteorological factors and on the nature of the primary emissions, the impact of data detrending was also assessed, whereby detrending was performed by the median polish procedure (see e.g. [5]).

The assessment of the different kriging reconstructions was carried out by performing a cross validation for each seasonally averaged dataset, by repeatedly using one of the data items to compute normalized residual with respect to the kriging prediction obtained on the basis of the remaining items. A Jarque-Bera test (see e.g. [8]) was used to check whether these residuals are normally distributed. In general, this is indeed the case, so that the confidence intervals predicted by the kriging procedure can be assumed to provide a reliable statistical estimate of the concentration values. The effective coverage of these intervals was also assessed *a posteriori* and found to be, in most case, very close to the theoretical one.

However, strong seasonal and interannual variations in the basic features of estimated concentration fields and of the residual correlation structure have been identified, which are even more apparent if time averages over shorter time periods are considered. As a result, there does not seem to be a unique isotropic variogram model capable to yield correct results in all the different situations encountered within the large dataset available. On the other hand, the anisotropic flexible variogram model yield reconstructions that are much less sensitive to these changes, so that they seem to constitute a good option for the development of a fully automated and objective reconstruction procedure that can be used to provide environmental regulators and health authorities with reliable real time estimates of PM10 concentrations.

In section 2, the PM10 dataset we considered is described in detail, along with the procedure used to exclude outliers from the dataset actually used in the geostatistical analysis. In section 3, the stochastic model for the data is introduced and the data detrending procedure is briefly outlined. The variograms reconstructed based on isotropic exponential and linear variogram models are presented and discussed in detail. In section 4, flexible variogram models are briefly reviewed along with the the technique used for their estimation based on the available data. In section 5 the cross validation of the chosen variogram models and detrending technique is discussed. In section 6, the actual concentration values reconstructed by kriging procedure are presented, along with the estimated variance of the underlying stochastic fields. The reconstructions show clearly that, especially in the winter season, concentration values are often well above those allowed by the current legislation, even in areas with relatively low primary production. In section 7, the main results obtained are summarized and some conclusions are drawn on the optimal choice of geostatistical model for this type of pollution data.

2 The PM10 concentration dataset for Northern Italy

PM10 pollution is an environmental issue of great concern in Northern Italy. In most part of the main urban areas of the Po Valley, extended over four regions (Emilia Romagna, Lombardia, Piemonte and Veneto) both long-term an short-term air quality limits are not attained ([10],[11]). Ambient fine particulate matter is both of primary and secondary origin: primary PM 10 is directly emitted by anthropogenic or natural sources, such as combustion processes, mechanical production, and traffic, while secondary PM is produced by chemical-physical trasformations and reactions of gaseous precursors. The morphology of this area induces the formation of a peculiar microclimate, since the Po Valley is surrounded in the North and in the West by the Alpes and in the South by the Appennines. These mountain chains retain air masses in the Po Valley, reducing pollution transport across the mountainous region and retaining pollutants in the lowest layers of the atmosphere. Intense urban pollution events are also favoured by the continental climate of this area, characterized by hot summers, wet winters with minimum temperatures often below zero, persistent fog, very low wind speed and frequent thermal inversions, that reduce the height of the boundary layer and limit the pollutants diffusion in the atmosphere. The meteorological, climatic and morphological uniformity of the area of the Po Valley appears to justify the application of geostatistic interpolation techniques, that rely on stationarity and, often, isotropy hypotheses, to the phenomenon of PM10 air pollution. Furtermore, the application of geostatistical techniques is also justified by the rather uniform PM10 concentration levels over the entire Po valley, as a consequence of the particularly relevant contribution of secondary PM to the measured PM10 mass.

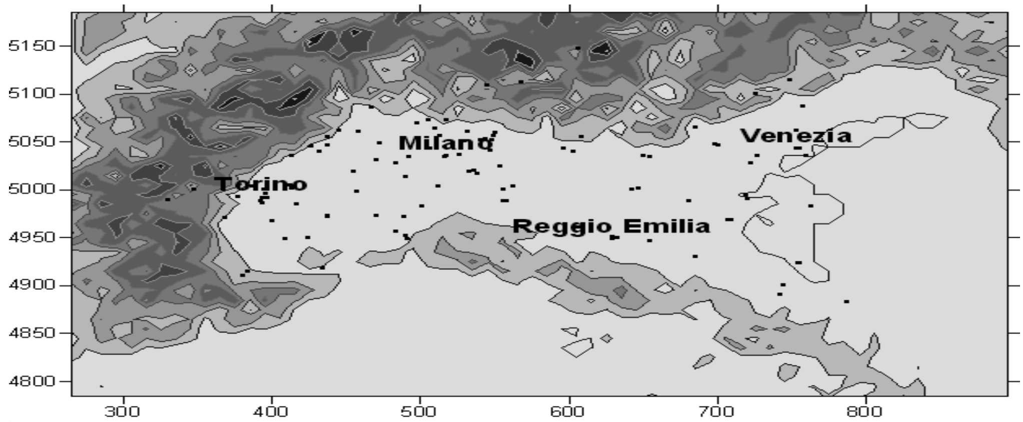


Figure 1: Locations of the measurement stations in the Po Valley, Northern Italy.

The data set used in this study is formed by PM10 daily averaged concentration values, measured from January 2003 to December 2006 by the air quality monitoring network of the above mentioned regions. The measurement networks evolved over these years, both in terms of number and location of the monitoring sites and in terms of the measurement instruments themselves. For each single year, the present analysis has only used PM10 daily concentrations provided by monitoring stations that had been collecting at least 75% of the annual data. Depending on data availability, the overall number of annual time series considered is about 80. Monitoring stations are mainly located in urban areas (approximately 56% of the stations comprised in the network) with with a roughly equal number at traffic exposed sites and at urban background sites (i.e. sites in urban areas where levels are representative of the exposure of the general urban population). Suburban areas with heavy traffic emissions account for 27% of the stations, while 21% is located in suburban areas with background emissions. Only 5% of the station is located in industrial areas and approximately the same number at rural background sites.

PM10 concentrations have been measured using different instruments, based on different measurement principles such as the gravimetric method, Beta attenuation monitors (BAM) and TEOM (Tapered Element Oscillating Microbalance). The gravimetric method is the reference method for PM10 measurement at European level (EN12341 norm). The same regulation establishes a standard procedure for assessing whether other measurement techniques yield data that can be considered equivalent to those obtained by the reference method. The dataset used in this study consists of values that have been collected at measurement sites validated according to the EN12341 norm. The gravimetric method measures the net mass on a filter, determined by weighing the filter before and after sampling air containing particulate matter, in a temperature and relative humidity controlled environment. Filters are equilibrated for 24 hours at constant relative humidity between 20% and 40% and at constant temperature between 15C and 30C, in order to minimize the liquid water associated with soluble compounds and to minimize the loss of volatile species. The gravimetric method works continuously in time and returns daily average measures. The method based on Beta attenuation determines PM10 concentration by filtering a polluted air volume. The mass concentration is calculated from the level of the Beta radiation absorbed by the clean filter and by the filter with the mass deposition. TEOM is formed by an oscillating monitor with a filter that accumulates particulate matter. The mass deposition induces changes in the instrument oscillation frequency, that are converted in mass concentration measures. The inlet air is heated to $30\mathrm{C} - 50\mathrm{C}$ to keep moisture in the vapour phase, or dried with a diffusion dryer. As a consequence, semivolatile compounds like ammonium nitrate and volatile organics can volatilize and for this reason the use of TEOM can lead to underestimation of the true PM10 concentrations, especially during the colder seasons, if compared to the values obtained by the gravimetric reference method (see e.g. [3]). Potential measurement artifacts reported for TEOM, resulting in PM10 mass underestimation due to the loss of semivolatiles, have been accounted for by correcting measured concentrations by means of proper experimental correction factors.

In order to reduce possible boundary effects, within the complete domain encompassing all the available monitoring sites (approximately, a box of 450×300km size) a smaller domain of approximately 390×240km has been considered, still comprising more than 90% of the available stations. This avoided, for example, estimating correlations with isolated stations in subalpine valleys, typically characterized by exposure to local emissions and peculiar meteorological conditions. Furthermore, a small number of individual stations were classified as outliers and excluded from the analysis. This classification was done on the basis of an empirical comparison, whenever the measured values appeared to be very different from those at measurement stations in the same province. This selection was justified *a posteriori* considering that most of these stations had either been removed from the official measurement network after a relatively short period (in general, a few months) or that they employed TEOM, which in some cases gave significantly lower PM10 concentrations with respect to measurements carried out with instruments based on Beta or Gravimetric techniques.

For this edited dataset, time averages for winter (October-March) and summer (April-September) were computed and all geostatistical analyses discussed in the following sections 3-6 were carried out for these seasonal mean values. Furthermore, averages over a period of three months were also computed, to study in deeper detail the temporal evolution of some statistical properties of the concentration fields.

3 Data model and isotropic variogram estimation

The seasonal concentration averages $y_i, i = 1, \dots, N$ were assumed to be realizations of a random field $Y(\mathbf{x}, \omega) = \mu(\mathbf{x}) + Z(\mathbf{x}, \omega)$, where $\mu(\mathbf{x})$ is a deterministic trend and $Z(\mathbf{x}, \omega)$ is a constant mean, instrinsically stationary random field (see e.g. [4] for a complete review of various stationarity concepts). For the more classical geostatistical reconstructions, the field Z was assumed to be isotropic, thus leading to semivariogram functions $\gamma(h)$, that were only functions of a single scalar variable, while in the case of the flexible variogram model of [1], the field was assumed to be anisotropic and a semivariogram function $\gamma(\mathbf{h}), \mathbf{h} \in \mathbf{R}^2$ was considered.

Since various factors (climatology, terrain morphology, location of primary emission sources) can be assumed to have a steady and direct impact on the measured values, it is reasonable to assume that a non constant deterministic trend is present in the data. In this work, only detrending based on the median polish technique (see e.g. [5]) is applied, leaving investigation of more sophisticated detrending procedures for future analyses. More precisely, a coarse grid of 5×4 points was superimposed to the computational domain, with uniform spacings of approximately 80km and 60km in the x and y directions, respectively. At these points, concentration values were computed by the iterative process described in [5], which entails repeated computation of the median values along rows and columns of the coarse grid. As an initial guess, the concentration values at the measurement stations nearest to each coarse grid point was assumed. A piecewise linear form $\mu(\mathbf{x}) = a + bx + cy$ was then assumed for the deterministic trend over each rectangle having four neighbouring points on the coarse grid as vertices and the coefficients were determined for each rectangle starting from the median polish values at the vertices. The resulting piecewise linear function was used to detrend the data and to compute the values of $z_i = y_i - \mu(\mathbf{x}_i), i = 1, \dots, N$ used for the estimation of the random field Z. The results obtained by performing ordinary kriging on the detrended data have been compared to those of ordinary kriging with unknown constant mean on the data without detrending.

h	2003	2004	2005	2006
10 km	120	150	151	208
20 km	92	140	128	166
30 km	128	172	172	254
45 km	208	306	270	484
60 km	329	390	362	618
80 km	416	536	472	868
100 km	518	608	556	1010
150 km	1124	1278	1340	2148
200 km	842	1022	1022	1494
250 km	630	900	888	586
300 km	390	832	586	584
390 km	390	716	610	352

Table 1: Numerosity of distance classes used for isotropic variogram model estimates: number of pairs in each distance bin for each dataset year.

The first step in the geostatistical estimation procedure is the choice of an appropriate variogram model to describe the spatial correlation structure of the data. For the estimates based on isotropic variogram models, an empirical semivariogram was computed by either the Matheron estimator

$$\gamma(h_k) = \frac{1}{2|\mathcal{N}(h_k)|} \sum_{(i,j)\in\mathcal{N}(h_k)} |Z(\mathbf{x}_i) - Z(\mathbf{x}_j)|^2 \qquad (1)$$

or the more robust estimator

$$\gamma(h_k) = \frac{1}{2\left(0.457 + \frac{0.494}{|\mathcal{N}(h_k)|}\right)} \left(\frac{1}{|\mathcal{N}(h_k)|} \sum_{(i,j)\in\mathcal{N}(h_k)} |Z(\mathbf{x}_i) - Z(\mathbf{x}_j)|^{\frac{1}{2}} \right)^4 . \qquad (2)$$

proposed by Cressie and Hawkins ([6],[7]). Here, $h_k, k = 1, \dots, K$ denotes a finite set of distance ranges for which the variogram is estimated, $\mathcal{N}(h_k)$ denotes the class of all pairs of measuremement points whose distance is comprised in the interval $[h_{k-\frac{1}{2}}, h_{k+\frac{1}{2}})$ (where $h_{k+\frac{1}{2}}$ denotes the arithmetic mean of the neighbouring values), and $|\mathcal{N}(h_k)|$ is the number of pairs in the class h_k. For the present study, distance classes $\mathcal{N}(h)$ were computed for the distance ranges reported in Table 1, along with the number of data pairs belonging to each class. It can be observed how the measurement has evolved over the years, thus increasing the amount of available data. Furthermore, all the classes have sufficiently large size for the subsequent estimates to be statistically significant.

These empirical variograms were used to estimate a valid variogram model by a weighted least squares method. It is to be remarked that the result of this least squares fit was found to be quite sensitive to the initial guess used in the minimization algorithm. For this reasons, 10^4 different initial guesses have been considered and the variogram parameters were chosen that gave the smallest value for the generalized least square cost functional at the end of the minimization process.

Among isotropic variograms, we have considered exponential, linear, Gaussian, spherical and Bessel variogram models. Only exponential and linear variogram models appeared to yield results sufficiently close to the empirical variogram over the whole range of spatial lags. As an example, we show in Figures 2-5 the fitted linear and exponential variograms obtained from the median polish detrended data starting from the Cressie-Hawkins empirical variogram. A significant seasonal and interannual variability of the variogram structure is clearly displayed by these estimates and summarized in Tables 2-3, in which the values of the estimated variogram parameters are reported.

For the data without detrending, higher sill and range values are characteristic of the winter periods, while much lower values are generally obtained for the summer months. This appears to be consistent with the different nature of the emissions and meteorological forcing in the two seasons. During the winter, significant domestic heating emissions are present and the local meteorology is characterized by smaller boundary layer thickness and large scale synoptic systems, which imply larger scale spatial correlations. During the summer, small scale convection dominates, thus inducing smaller scale spatial correlations, while much higher boundary layer values are also responsible for the lower concentration levels (see section 6). On the contrary, detrended data tend to show smaller scale correlation in the winter season than in summer, which seems to point at a stronger dependence of the winter concentrations on deterministic factors such as emissions intensity. The estimated nugget parameter is in most cases quite small, although for some datasets and variogram models it can get as large as the sill value.

In general, the exponential model seems appropriate to recover the spatial correlation structure in most of the cases, but the linear variogram model appears to fit better the summer data in at least two of the four considered years. In particular, for the 2004 summer season the detrended data yield an empirical variogram so close to the linear model that For both models, clear discrepancies from the empirical variograms can be seen in at least one of the seasonally averaged data. This motivates the attempt, carried out in the next section, to apply flexible variogram models, in order to obtain accurate estimates with a tecnique that can better adapt to the large variability displayed by the data.

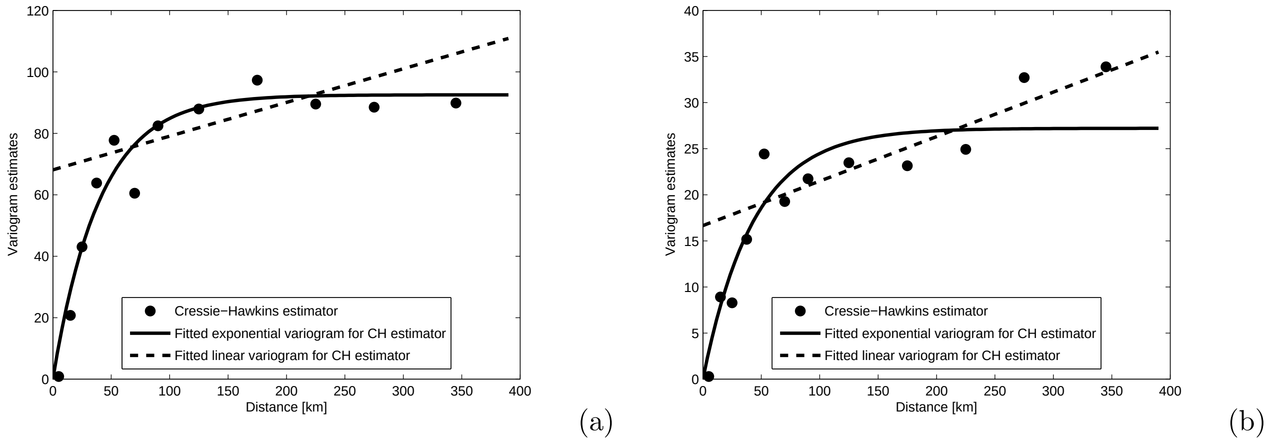


Figure 2: Exponential and linear variogram functions fitted to Median Polish detrended data for winter (a) and summer (b) of 2003

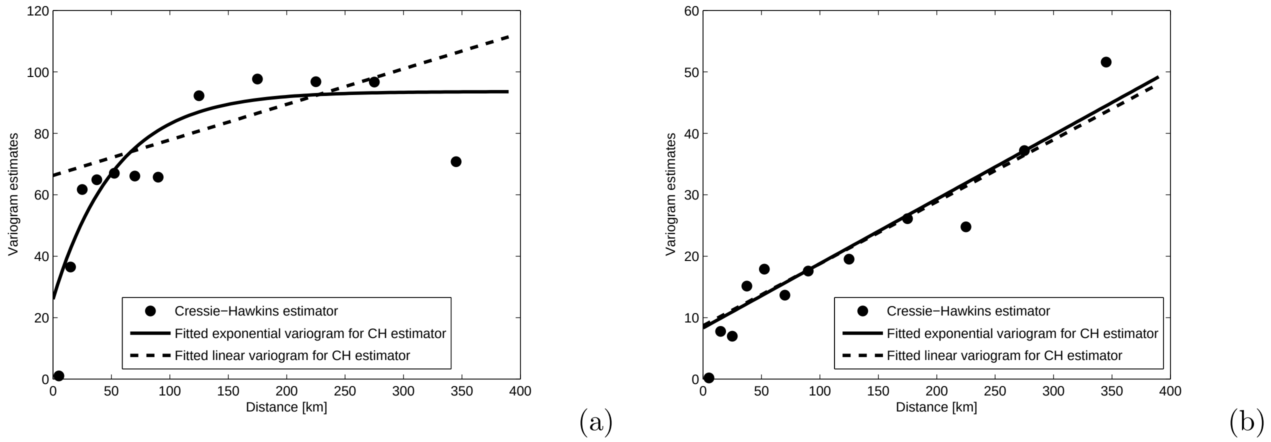
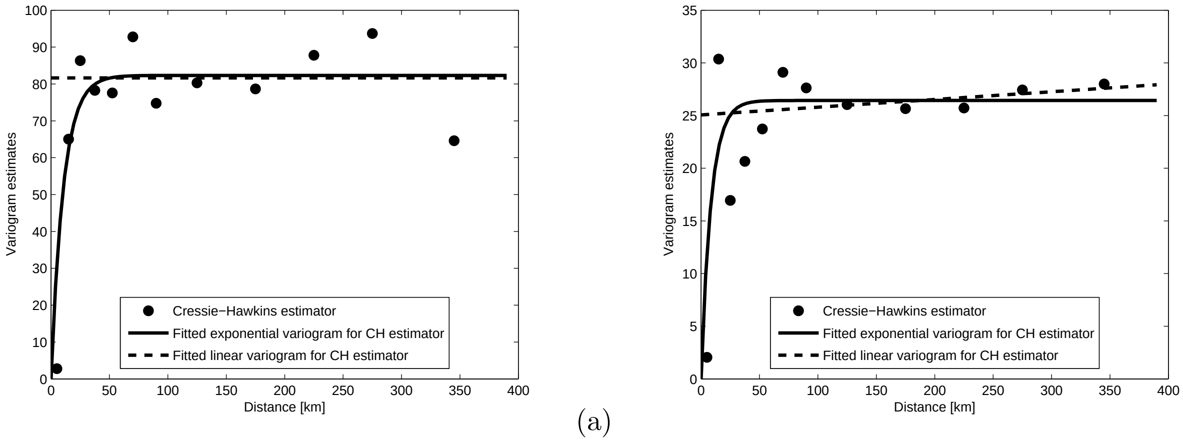


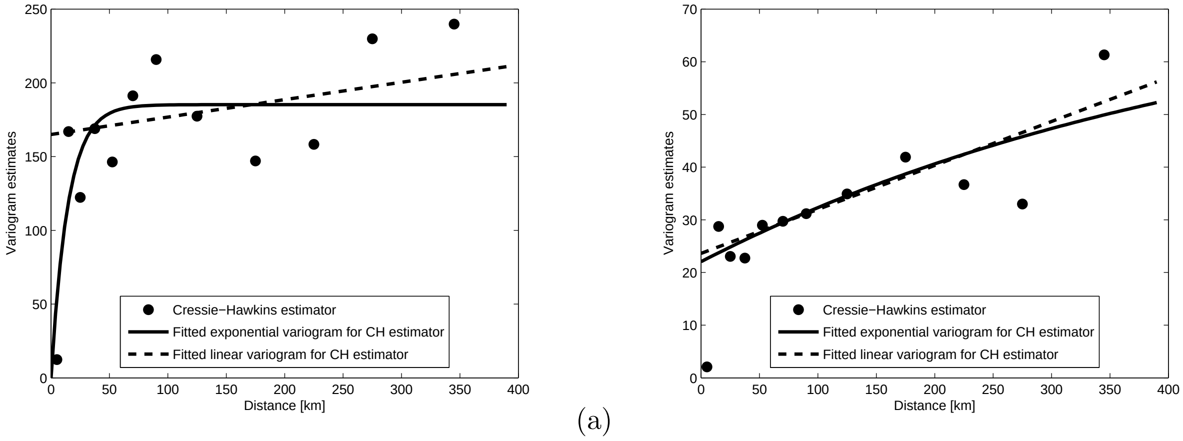
Figure 3: Exponential and linear variogram functions fitted to median polish detrended data for winter (a) and summer (b) of 2004



(a)

(b)

Figure 4: Exponential variogram functions fitted to median polish detrended data for winter (a) and summer (b) of 2005



(a)

(b)

Figure 5: Exponential and linear variogram functions fitted to Median Polish detrended data for winter (a) and summer (b) of 2006.

4 Flexible variogram models and their estimation

Geostatistical interpolation is usually formulated assuming the data to consist in a realization of a random field $Z : D \times \Omega \to \mathbb{R}, D \subset \mathbb{R}^d$ with a valid semivariogram function $\gamma(\mathbf{h})) = \mathbb{E}[(Z(\mathbf{x}+\mathbf{h})-Z(\mathbf{x}))^2]/2$. The classical characterizazion of valid variograms is given in terms of conditionally negative definite functions. In general, a piecewise linear function (in more than one dimension, a piecewise multilinear one) is not conditionally negative defi-

Dataset	Nugget	Sill	Range [km]
Winter 2003	0	642	368.45
Summer 2003	0	30	53.14
Winter 2004	20	88	55.19
Summer 2004	2	27	116.63
Winter 2005	0	129	21.41
Summer 2005	0	44	24.11
Winter 2006	0	244	24.11
Summer 2006	11	43	40.05

Table 2: Temporal evolution of variogram parameters for exponential variogram model fitted to Cressie Hawkins estimator based on raw data.

Dataset	Nugget	Sill	Range [km]
Winter 2003	0	93	40.21
Summer 2003	16	66	834.87
Winter 2004	26	94	53.82
Summer 2004	0	10^9	10^{13}
Winter 2005	0	82	10.6
Summer 2005	21	27	35.24
Winter 2006	0	185	14.49
Summer 2006	22	76	470.71

Table 3: Temporal evolution of variogram parameters for exponential variogram model fitted to Cressie Hawkins estimator based on median polish detrended data.

nite, so that simple interpolation of the values of an empirical variogram estimator does not yield a valid variogram function.

The concept of flexible variogram model introduced in [1] relies instead on a different characterization of valid variograms. It was proven by these authors that, for $d = 1$, under the assumption that the semivariogram is constant for $h > c$, with $c > 0$ given, the function 2γ can be represented as

$$2\gamma(h) = \int_{\mathbb{R}} [f(x \mid a_1, \dots, a_k, c, k) - f(x - h \mid a_1, \dots, a_k, c, k)]^2 dx, \tag{3}$$

where f is a measurable function. The main point of the flexible variogram model consists in choosing a piecewise constant function f, to yield as a consequence piecewise linear valid variograms. More specifically, for $k > 0$ and $a_1, \dots, a_k > 0$, define the function f with support $[0, c]$ by

$$f(x \mid a_1, \dots, a_k, c, k) = \sum_{j=1}^{k} a_j I\left(\frac{(j-1)c}{k} < x \leq \frac{jc}{k}\right), \tag{4}$$

where $I(a, b)$ denotes the indicator function of the interval (a, b). Hence, f is a piecewise constant function. Using (4) in the representation theorem (3), after some algebra we have an explicit expression of the semivariogram. For convenience, values at a finite set of points are computed first, and the remaining values are then recovered by linear interpolation, which is justified since the resulting function is indeed piecewise linear. The resulting definition of the semivariogram can then be described as follows:

- if exists an integer m such that $h = mc/k$,

$$2\gamma(h) = \frac{c}{k}\sum_{i=1}^{k} a_i^2 - \frac{2c}{k}\sum_{i=m+1}^{k} a_i a_{i-m}, \tag{5}$$

- if $h < c$, but is not an integer multiple of c/k,

$$2\gamma(h) = (1 - V)2\gamma\left(\frac{m_l c}{k}\right) + V2\gamma\left(\frac{m_u c}{k}\right), \tag{6}$$

  where $m_l = \lfloor hc/k \rfloor$ and $m_u = \lceil hc/k \rceil$ and $V = (h - m_l c/k)/(c/k)$, that is, the value of the semivariogram is given by linear interpolation of the two values at the nearest multiple integers of c/k enclosing h;

- if $h > c$

$$2\gamma(h) = \frac{2c}{k}\sum_{i=1}^{k} a_i^2. \tag{7}$$

In [1], specific variograms were then obtained by fixing *a priori* k and c and estimating the a_i from the data, starting from the same empirical estimators $\hat{\gamma}$ introduced in the previous section and using a weighted least square algorithm. The integer k represents the number of equal size intervals in which $[0, c]$ is divided and over which the variogram is represented by a different linear function. In general, k will have to be smaller than the number of different lags used in an empirical variogram estimator.

The representation theorem introduced above also holds in the multidimensional case, so that for $d > 1$ one has

$$2\gamma(\mathbf{h}) = \int_{\mathbb{R}^d} (f(\mathbf{x} \mid a_1, \ldots, a_k, c, k) - f(\mathbf{x} - \mathbf{h} \mid a_1, \ldots, a_k, c, k))^2 d\mathbf{x}. \tag{8}$$

Along the lines of [1], in the two dimensional case that is relevant for the present application we can define piecewise constant functions on the rectangular domain $[0, c] \times [0, d]$ as

$$\begin{aligned} & f(x, y \mid a_{i,j}, c, d, m, n) = \\ & \sum_{i=1}^{m} \sum_{j=1}^{n} a_{i,j} I \left[\left(\frac{c(i-1)}{m} \leq x < \frac{ci}{m} \right) \left(\frac{d(j-1)}{n} \leq y < \frac{dj}{n} \right) \right], \end{aligned} \tag{9}$$

for integer m, n and $a_{i,j} > 0$. Substituting (9) into (8) yields then a valid variogram function. As for the one dimensional case, values at special points and all the others are exactly recovered by bilinear interpolation. With this respect, it should be noticed that the original formula given in [1] for determination of $\gamma(\mathbf{h}) = \gamma(h_1, h_2)$ at points whose coordinates are integer multiples of c/m and d/n is only valid in the special case in which both these numbers are positive. This is sufficient to cover the one dimensional case, for which $\gamma(h) = \gamma(-h)$ for all $h \in \mathbb{R}$. In two dimensions, however, although $2\gamma(h_1, h_2) = 2\gamma(-h_1, -h_2)$, in general $2\gamma(h_1, h_2) \neq 2\gamma(-h_1, h_2)$, so using the original formula in [1] for a generic field could yield a non valid variogram. In the more general case, if h_1, h_2 are integer multiples of c/m and d/n, the appropriate formula is

$$2\gamma(h_1, h_2) = \sum_{i=1}^{m} \sum_{j=1}^{n} a_{i,j} a_{i-\operatorname{sgn}(h_1)\left\lfloor \frac{|h_1| c}{m} \right\rfloor, j-\operatorname{sgn}(h_2)\left\lfloor \frac{|h_2| d}{n} \right\rfloor}, \tag{10}$$

if $0 \leq |h_1| < c$ and $0 \leq |h_2| < d$, where $\operatorname{sgn}(\cdot)$ denotes the signum function. For arbitrary lag values, the semivariogram is computed by bilinear interpolation between the values of the variogram on the corners of the rectangle containing (h_1, h_2) whose vertices are the nearest integer multiples of c/m and d/n.

In the anisotropic case, the two dimensional analog of (1) was computed. As an example, the distance lags in the x and y directions are reported in Table 4, along with the number of data pairs belonging to each class for the year 2006 only. It can be observed that, in order to obtain sufficiently populated distance bins, a smaller number of distance classes has to be employed.

Subsequently, also the anisotropic, flexible variogram models introduced in [1] have been fitted to the anisotropic empirical variogram using a procedure entirely analogous to the one described above for the isotropic case. The flexible variogram model was assumed to be given by a piecewise bilinear function defined on by its values at 5×4 regularly spaced points over the domain spanned by the data. As an example, the contour levels of the anisotropic variogram fitted for the 2003 winter data are shown in figure 6, highlighting the different length scales in the two coordinate directions,

h_x/h_y	15 km	30 km	60 km	150 km
15 km	104	68	103	117
30 km	80	75	112	106
90 km	231	216	367	454
150 km	166	158	298	412
250 km	182	143	284	385
400 km	75	58	114	152

Table 4: Numerosity of distance classes used for anisotropic variogram model estimates: number of pairs in each distance bin for dataset year 2006.

that are naturally recovered by the estimation process with no need for expert user intervention. Due to the difficulty of displaying the and empirical variogram and fitted function in the anisotropic case, no direct comparison with the empirical variogram is shown here. However, results of the cross validation reported in section 5 will demonstrate the good performance of the flexible anisotropic model.

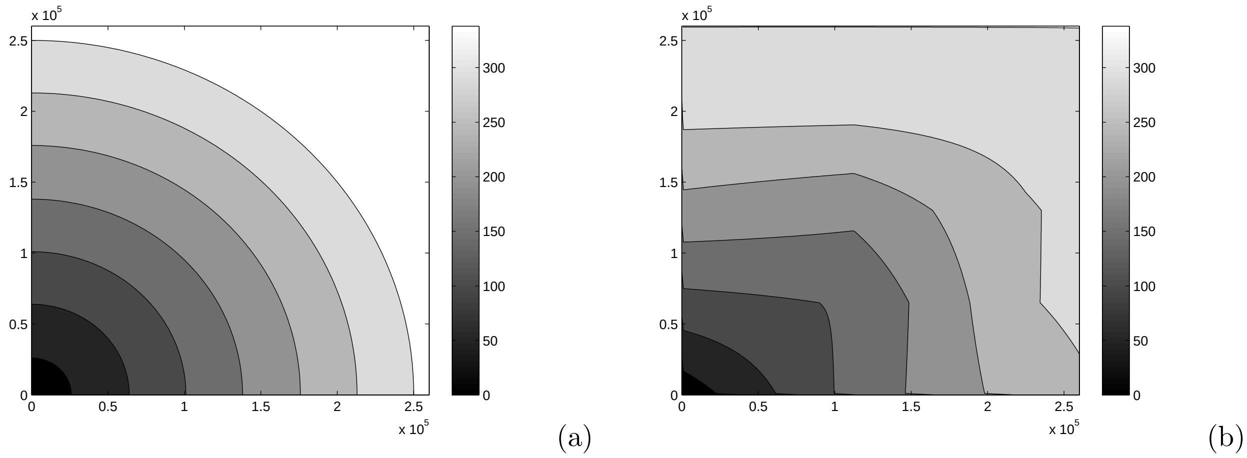


Figure 6: Exponential isotropic (a) and flexible anisotropic (b) variogram functions fitted for winter 2003 data.

5 Cross validation and normality tests

In order to assess the accuracy of the spatial predictions obtained by geostatistical reconstruction using the previously estimated variograms, a cross validation procedure has been carried out. More precisely, for each measurement site $\mathbf{x}_i, i = 1, \ldots, N$ the variogram estimation and the kriging reconstruction were carried out on the basis of the remaining $N - 1$ sites and estimates $\hat{Z}_i, \hat{\sigma}_i$ were obtained for the field and standard deviation values, respectively. The normalized residuals $\zeta_i = (z_i - \hat{Z}_i)/\hat{\sigma}_i$ were then computed and analyzed. Although these quantities do not need to follow a Gaussian distribution for the kriging reconstruction to be consistent, if this is the case, assuming the reconstructed field to be Gaussian is an hypothesis compatible with the available data. For Gaussian fields, the variogram describes completely the field stochastic structure, so that the confidence intervals built on the basis of the kriging variance are a reliable estimate of the reconstructed field uncertainty. On the other hand, if the residuals do follow a Gaussian distributions, this can be seen as an indicator that the kriging predictions not provide in that case a complete information on the field uncertainty.

The normality of the residuals can be investigated by inspecting the quantile-quantile plot of the residuals against the unit normal distributions. An example of such plot is shown in Figure 7, highlighting that for most of the measurement sites no significant deviations from Gaussianity arise. Similar plots arise for all the considered datasets.

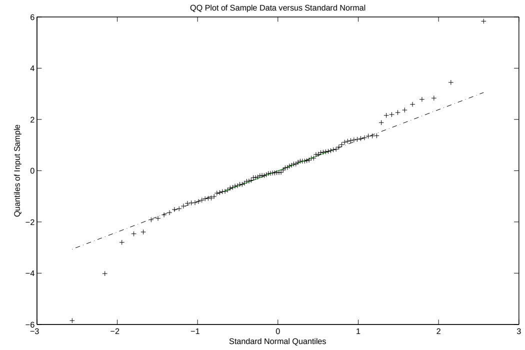


Figure 7: Quantile plot of the residual distribution based on winter 2006 Median Polish detrended data and exponential variogram model, against the standard normal distribution.

Furthermore, a Jarque-Bera normality test has been performed (see e.g. [8]). The Jarque-Bera test is a two-sided goodness-of-fit test that is suitable when a fully specified null distribution is unknown and its parameters must be estimated. The statistical test is given by $JB = N\left\{S^2 + (K-3)^2/4\right\}/6$, where N is the sample size, S is the sample skewness, and K is the sample kurtosis. For large sample sizes, the statistical test has a χ^2 distribution with two degrees of freedom. The Jarque-Bera test uses a table of critical values computed by Monte-Carlo simulation for sample sizes less than 2000 and significance levels between 0.001 and 0.50. Critical values for the test are computed by interpolating from the table values and using the analytic χ^2 approximation only when extrapolating for larger sample sizes. It is to be remarked that this test is applicable also to dependent random variables, which is indeed the case for the residuals of an estimated random field with non trivial correlation structure.

The MATLAB `jbtest` function has been used, that returns the $p-$value computed using inverse interpolation from the table of critical values. Small

Dataset	Confidence interval width	Width variance	Coverage%	$p-$value %
Winter 2003	39.16	569.61	81.69	>50
Summer 2003	10.76	16.76	70.42	12.08
Winter 2004	38.96	0.22	95.18	19.64
Summer 2004	8.95	4.61	71.08	7.14
Winter 2005	44.05	0.20	96.25	>50
Summer 2005	19.37	33.69	87.50	7.11
Winter 2006	60.12	0.27	93.68	>50
Summer 2006	17.00	24.15	83.16	40.83

Table 5: Validation of kriging with exponential variogram model based on raw data.

Dataset	Confidence interval width	Width variance	Coverage %	$p-$value %
Winter 2003	22.95	30.70	81.69	13.32
Summer 2003	16.88	0.28	91.55	31.84
Winter 2004	33.87	0.85	92.77	42.03
Summer 2004	10.64	0.78	80.72	10.42
Winter 2005	42.05	0.34	98.75	10.61
Summer 2005	24.97	0.12	100.00	32.36
Winter 2006	55.65	0.55	100.00	40.50
Summer 2006	22.55	0.10	94.74	>50

Table 6: Validation of kriging with linear variogram model based on raw data.

values of p cast doubt on the validity of the null hypothesis. Using the cross validation, the Jarque-Bera test provides a p-value representative of the threshold below which the null hypothesis of the normality of residuals is not refused. Thus, a p-value higher than 5% is a good index of the normality of residuals. Along with the normality test, the effective *a posteriori* coverage of the 95% significance level confidence intervals $[\hat{Z}_i - 3\hat{\sigma}_i, \hat{Z}_i + \hat{\sigma}_i]$ predicted by kriging has been estimated for each set predicted values. Jarque-Bera p-values and coverage percentages are shown in tables 5-5 for the exponential and linear model applied to either detrended or raw PM10 seasonal data. Also the average width of the confidence intervals and their sample variance are displayed.

It can be observed that, in most cases, the Gaussian residual hypothesis is not rejected. Furthermore, the effective *a posteriori* coverage of these intervals is in general close to the theoretical one. Detrending is clearly important in reducing the prediction uncertainty, especially in conjunction with the exponential model, while it seems to affect less the flexible variogram predictions. In general, based on this seasonally averaged data the exponential model applied to detrended data would appear as the best candidate to perform reliable reconstructions.

However, the time variability of the characteristics of the sampled field

Dataset	Confidence interval width	Width variance	Coverage	$p-$value %
Winter 2003	36.31	0.17	94.37	>50
Summer 2003	19.59	0.09	92.96	19.64
Winter 2004	36.35	0.12	97.59	>50
Summer 2004	21.65	0.10	96.39	30.06
Winter 2005	35.43	0.11	97.50	16.04
Summer 2005	20.10	0.05	92.50	>50
Winter 2006	52.89	0.25	93.68	24.04
Summer 2006	23.63	0.03	95.79	>50

Table 7: Validation of kriging with exponential variogram model based on Median Polish detrended data.

Dataset	Confidence interval width	Width variance	Coverage %	$p-$value %
Winter 2003	32.70	0.99	95.77	>50
Summer 2003	16.40	0.20	91.55	39.35
Winter 2004	31.35	7.56	89.16	>50
Summer 2004	12.59	0.59	84.34	30.06
Winter 2005	-	-	-	-
Summer 2005	19.68	0.06	98.75	35.87
Winter 2006	19.49	0.19	90.53	>50
Summer 2006	48.20	24.44	95.79	>50

Table 8: Validation of kriging with linear variogram model based on Median Polish detrended data.

Dataset	Confidence interval width	Width variance	Coverage %	$p-$value %
Winter 2003	29.41	11.42	94.37	2.80
Summer 2003	13.09	2.06	88.73	>50
Winter 2004	26.99	5.65	90.36	49.93
Summer 2004	10.49	3.38	80.72	40.80
Winter 2005	35.44	4.67	100.00	12.43
Summer 2005	22.23	0.87	97.50	3.91
Winter 2006	33.74	15.76	85.26	42.01
Summer 2006	19.43	0.88	93.68	35.62

Table 9: Validation of kriging with anisotropic flexible variogram model based on raw data.

Dataset	Confidence interval width	Width variance	Coverage %	$p-$value %
Winter 2003	25.05	5.66	91.55	2.51
Summer 2003	12.78	1.90	87.32	45.07
Winter 2004	28.13	2.61	91.57	>50
Summer 2004	11.36	1.50	90.36	31.78
Winter 2005	31.64	2.13	96.25	11.78
Summer 2005	19.28	0.22	96.25	7.06
Winter 2006	37.41	12.84	87.37	>50
Summer 2006	19.24	0.49	94.74	45.07

Table 10: Validation of kriging with anisotropic flexible variogram model based on Median Polish detrended data.

leads to significant variations in the effective coverage of the predicted confidence intervals, as well as of their width. In some cases, as for example winter 2005 for the linear model applied to detrended data, the fitted variogram profile consists of an almost horizontal line, thus leading to entirely unrealistic prediction values. These variations become even more dramatic if the same validation is carried out on datasets consisting of concentration values averaged over each year quarter. In this case, the exponential model is unable to provide useful estimates for 5 of the 16 datasets, yielding either very large confidence intervals or almost singular kriging matrices. The linear model performs better for those cases in which exponential fails, but it also displays analogous behaviour on a relevant portion of the data. As a result, there does not seem to be a unique isotropic variogram model capable to yield correct results in all the different situations encountered within the large dataset available.

On the other hand, as shown in Table 5, the anisotropic flexible variogram model yield reconstructions that appear to be much less sensitive to these changes. In view of the development of a a fully automated and objective reconstruction procedure that can be used to provide environmental regulators and health authorities with reliable real time estimates of PM10 concentrations, this anisotropic model seems to constitute an appropriate choice.

6 Reconstruction of concentration and standard deviation fields

The PM10 concentrations fields have been reconstructed using ordinary kriging, starting from either the detrended or raw data, with the exponential and flexible variogram models discussed in the previous sections. As an example, the following figures represent the PM10 reconstructed fields in winter and summer seasons of the year 2005. The reconstructed fields are qualitatively and quantitatively similar, although the standard deviation values obtained with the flexible variogram model appear to be uniformly smaller over larger areas. Concentration peaks are located over urban and

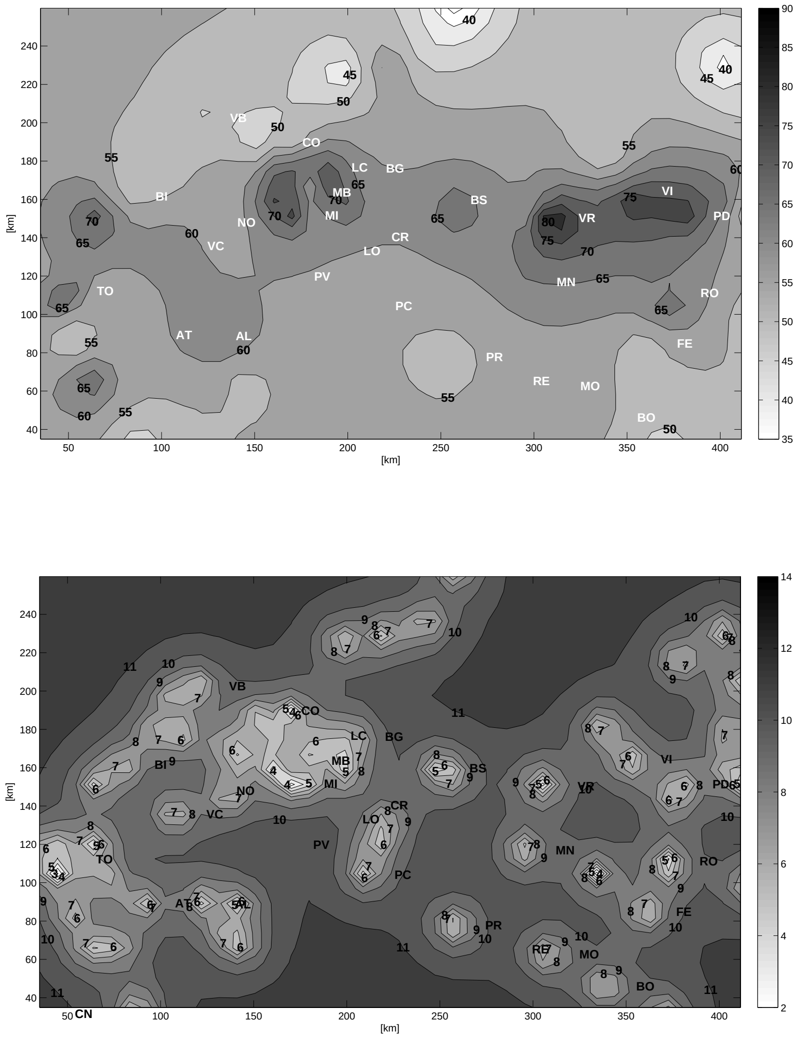
Dataset	Confidence interval width	Width variance	Coverage	p-value %
First quarter 2003	22.39	7.63	84.51	<0.01
Second quarter 2003	9.30	4.69	64.79	48.25
Third quarter 2003	12.10	2.40	77.46	>50
Fourth quarter 2003	22.71	6.55	91.55	>50
First quarter 2004	25.30	9.82	84.34	>50
Second quarter 2004	9.40	4.18	66.27	31.90
Third quarter 2004	9.94	3.48	73.49	>50
Fourth quarter 2004	25.83	6.06	85.54	22.00
First quarter 2005	37.87	5.93	89.87	26.14
Second quarter 2005	20.53	0.23	94.94	2.62
Third quarter 2005	19.11	0.36	93.67	0.24
Fourth quarter 2005	33.61	0.22	88.61	22.31
First quarter 2006	35.10	29.30	79.38	>50
Second quarter 2006	22.46	0.44	94.85	0.4867
Third quarter 2006	19.10	0.54	95.88	>50
Fourth quarter 2006	34.88	19.92	85.57	>50

Table 11: Validation of kriging with anisotropic flexible variogram model based on Median Polish detrended data: quarterly averaged data.

industrialized areas, but occasionally also in less densely populated, mostly rural areas, such as the Lodi (LO) and Cremona (CR) provinces. The peak values and their locations vary remarkably from year to year, with the Turin (TO), Milan (MI) and Verona (VR) provinces displaying in turn the highest concentration values. Especially in the winter season (see e.g. the reconstruction obtained using the flexible variogram model for winter 2006), the estimated values, that can be considered reliable with high probability given the analyses presented in the previous sections, are remarkably higher than the maximum concentration values allowed by the current legislation.

7 Conclusions

The applicability of classical geostatistical tools to the reconstruction of PM10 concentration fields over Northern Italy has been assessed, based on a large dataset of daily PM10 data spanning the years 2003-2006. The impact of the variogram model chosen for the geostatistical estimates has been investigated, by comparing the results obtained with several isotropic variogram models as well as with anisotropic flexible variogram models. Although exponential and linear variograms yield reliable reconstructions in most of the cases, the analysis has highlighted significant seasonal and interannual variations in the basic features of the estimated concentration fields and residual correlation structure. None of the classical models appears able to cope with all the different situations encountered, while the



(a)

(b)

Figure 8: PM10 concentration values (a) and standard deviations (b) reconstructed using ordinary kriging with exponential variogram in winter 2005.

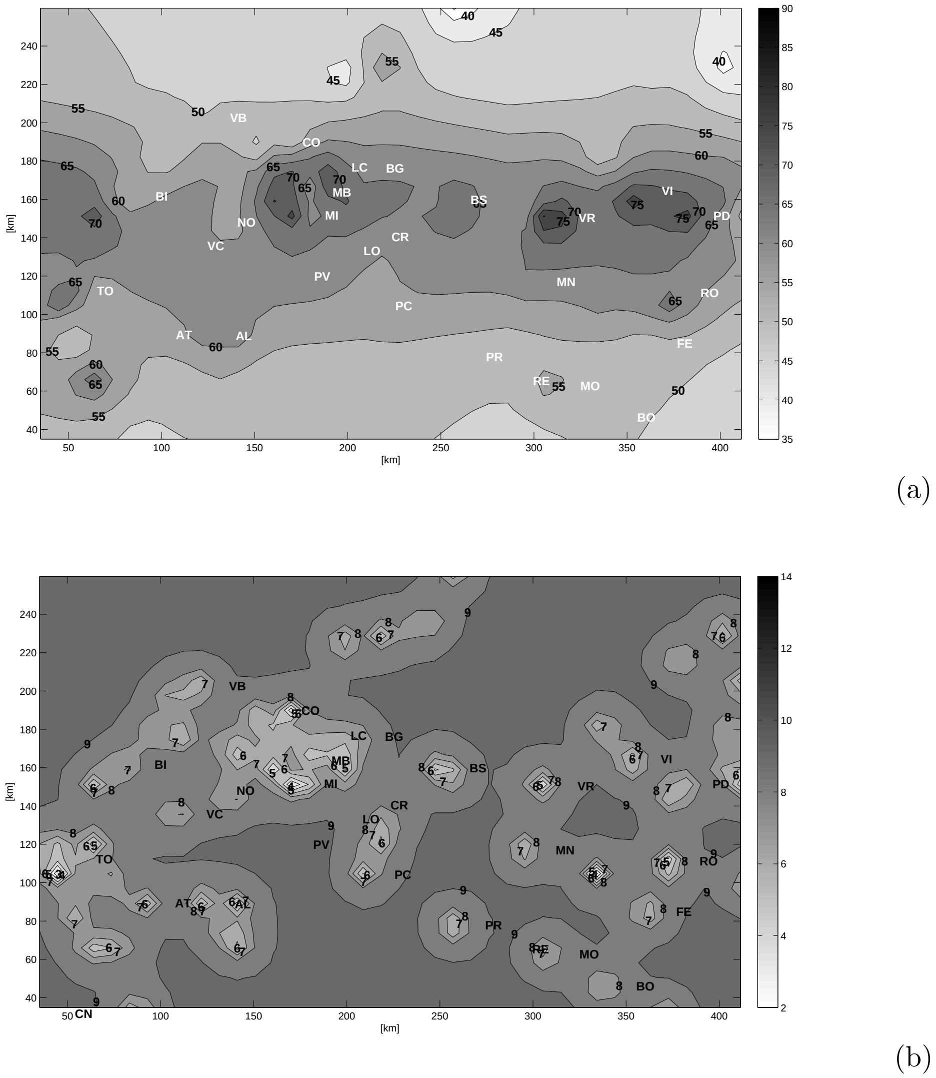


Figure 9: PM10 concentration values (a) and standard deviations (b) reconstructed using median polish kriging with exponential variogram in winter 2005.

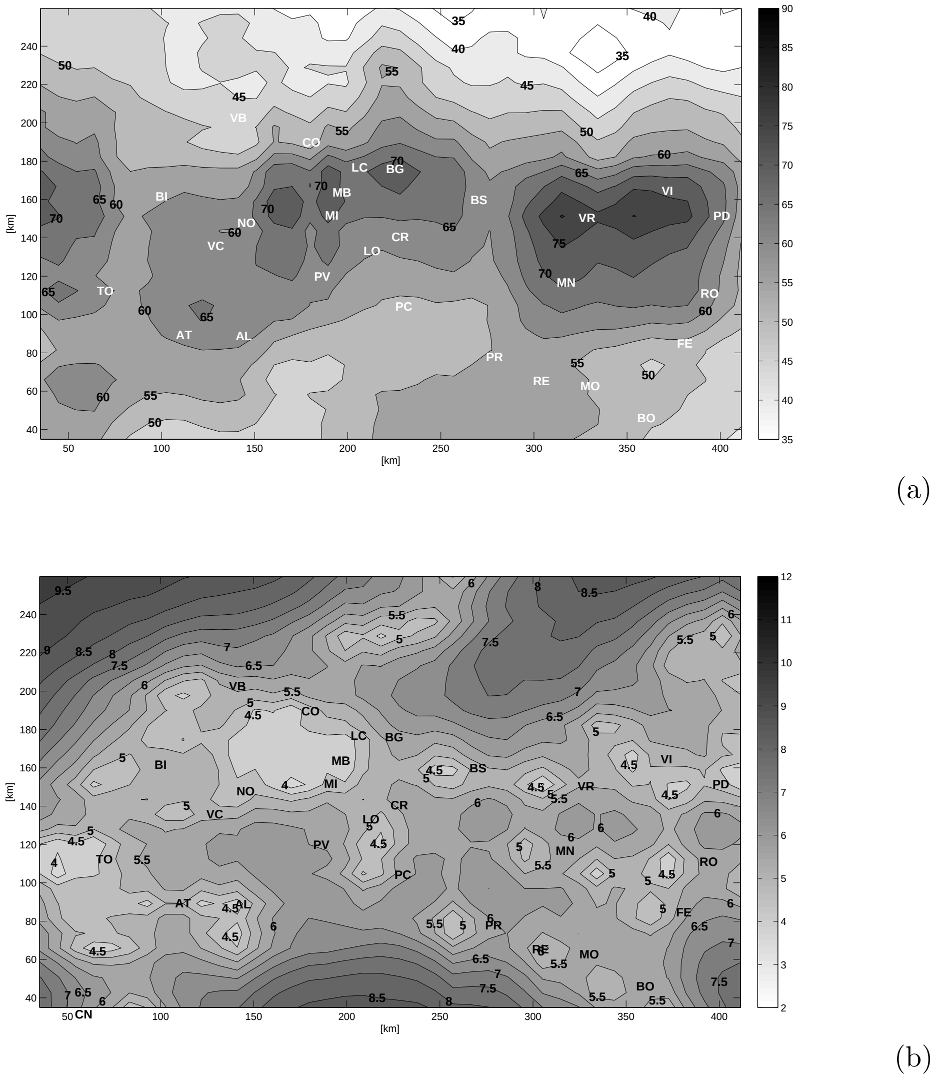


Figure 10: PM10 concentration values (a) and standard deviations (b) reconstructed using median polish kriging with anisotropic flexible variogram in winter 2005.

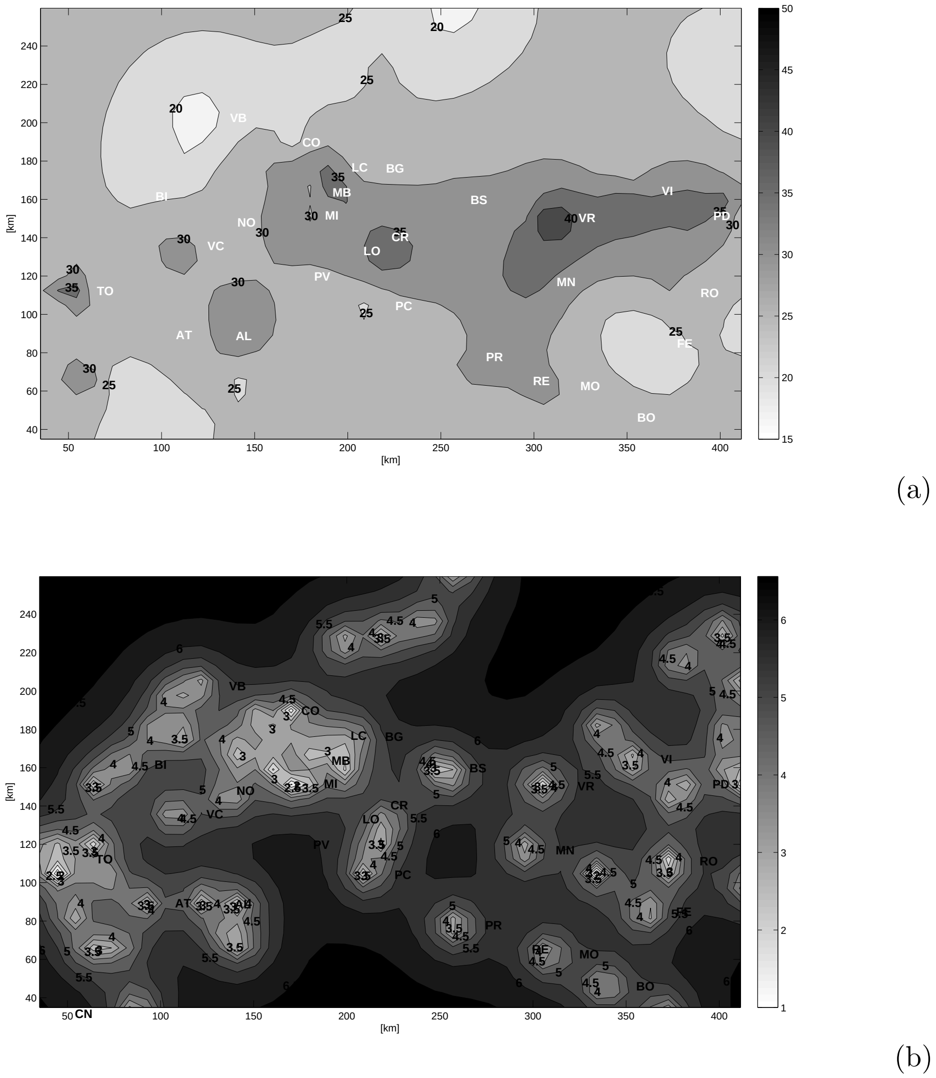


Figure 11: PM10 concentration values (a) and standard deviations (b) reconstructed using ordinary kriging with exponential variogram in summer 2005.

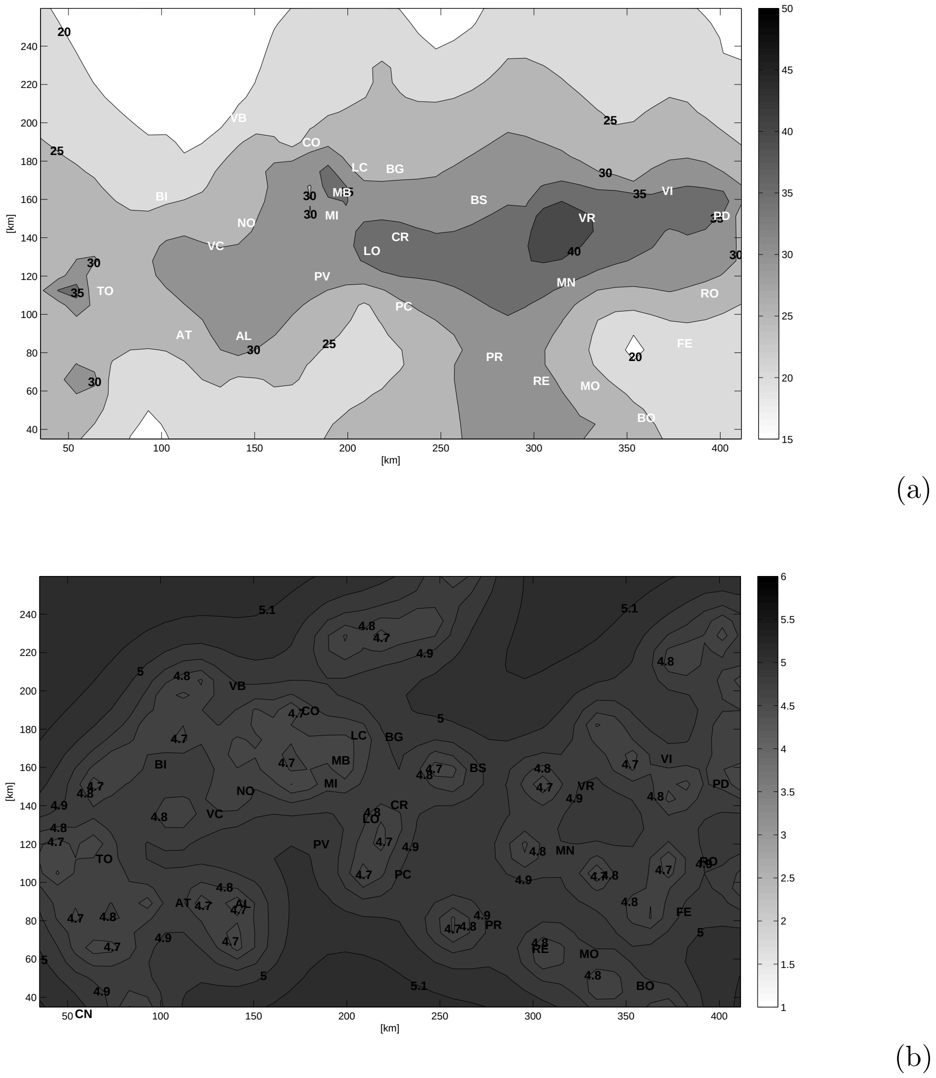


Figure 12: PM10 concentration values (a) and standard deviations (b) reconstructed using median polish kriging with exponential variogram in summer 2005.

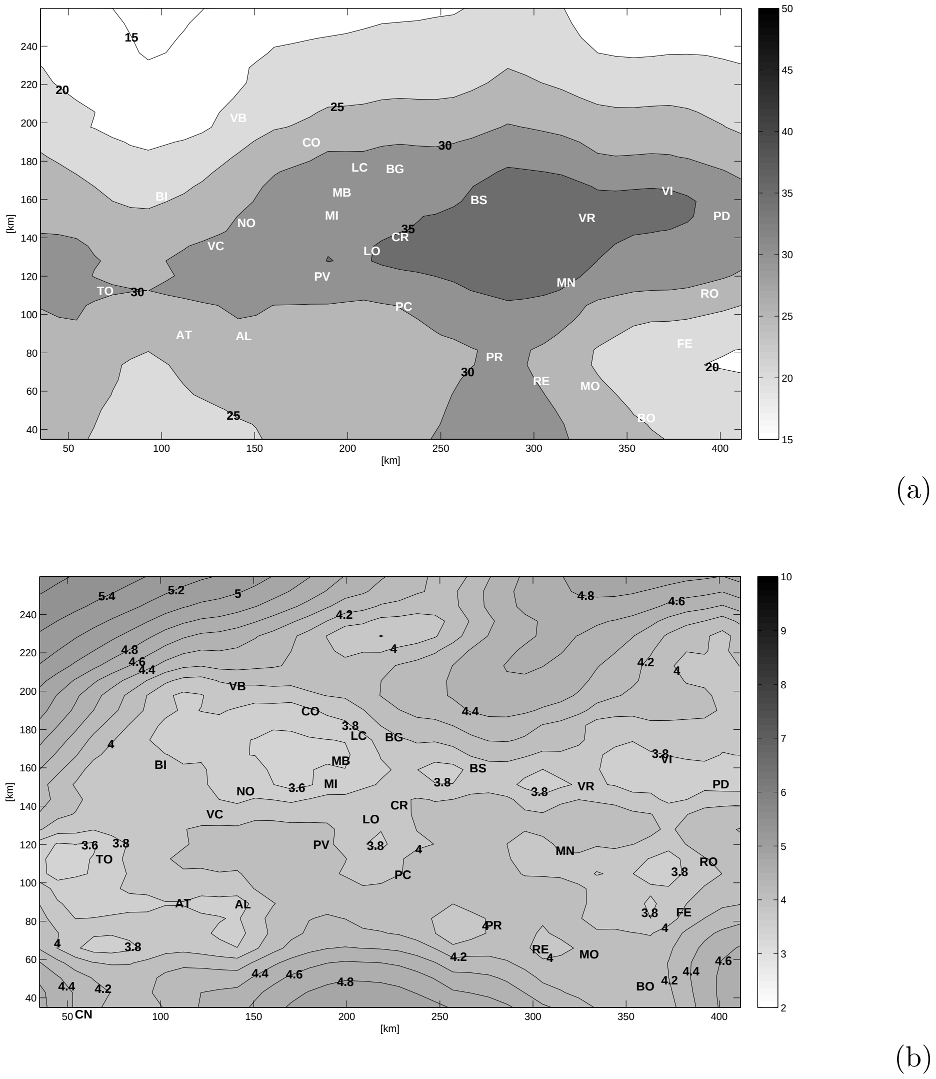


Figure 13: PM10 concentration values (a) and standard deviations (b) reconstructed using median polish kriging with anisotropic flexible variogram in summer 2005.

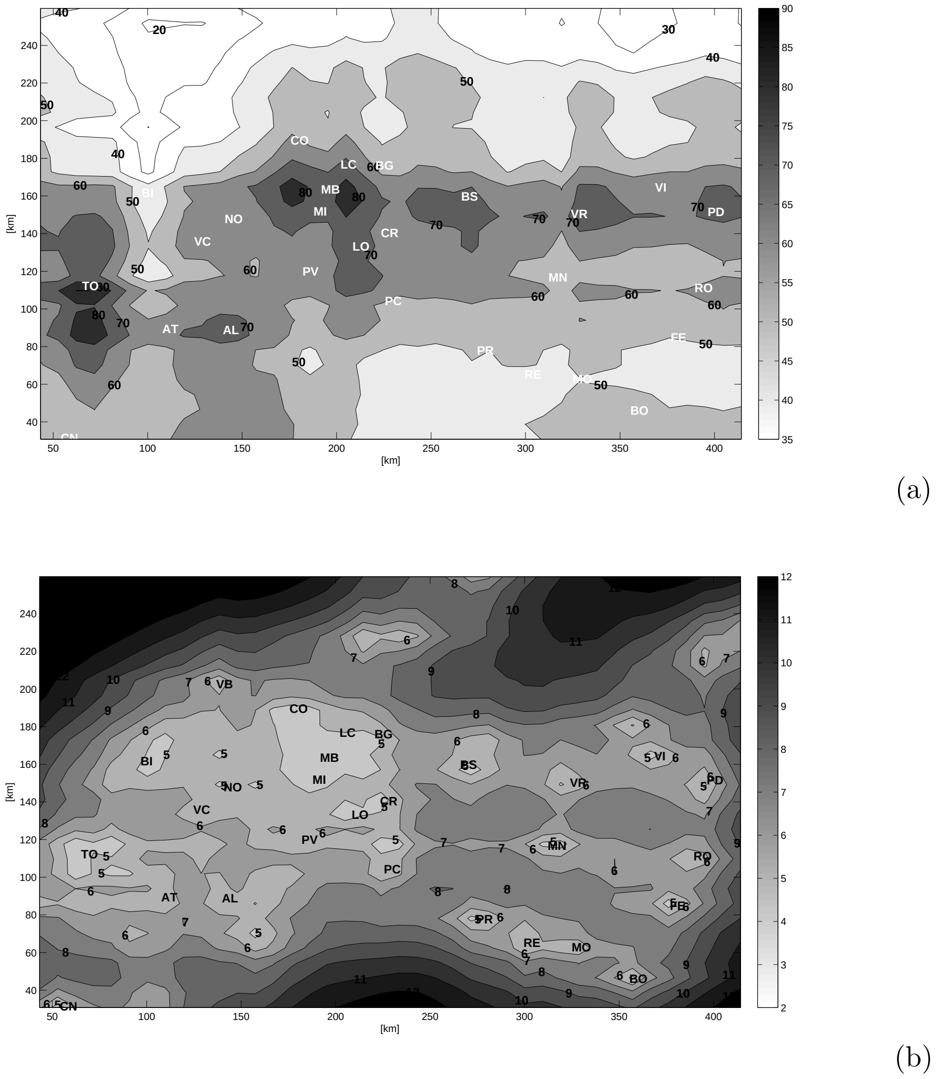


Figure 14: PM10 concentration values (a) and standard deviations (b) reconstructed using median polish kriging with anisotropic flexible variogram in winter 2006.

anisotropic flexible variogram models appear to constitute a more robust tool for automatic reconstruction of the PM10 concentration fields without expert user intervention.

As a prosecution of this work, the development of such a tool is envisaged, in which the use of anisotropic, flexible variogram models will be complemented by a Bayesian approach along the lines of [2], in order to provide environmental regulators and health authorities with reliable real time estimates of PM10 concentrations.

8 Acknowledgements

The PM10 concentration data were provided by ARPA Lombardia. This research was supported by Politecnico di Milano in the context of the Lecco Campus Point initiative and was partly carried out in the framework of the Master Thesis work by Paola Crippa and Stefano Castruccio. The authors are grateful to A. Paganoni and P. Secchi for a careful reading of the manuscript and for several useful suggestions.

References

[1] R.P. Barry and J.M. Ver Hoef. Blackbox kriging: Spatial prediction without specifying variogram models. *Journal of Agricultural, Biological, and Environmental Statistics*, 1:297–322, 1996.

[2] S. Castruccio, L. Bonaventura, P. Secchi, and L. Sangalli. A Bayesian approach to geostatistical interpolation with flexible variogram models. Mox report, to appear, MOX - Politecnico di Milano, 2008.

[3] A. Charron, R.M. Harrison, S. Moorcroft, and J. Booker. Quantitative interpretation of divergence between PM10 and PM2.5 mass measurement by TEOM and gravimetric (Partisol) instruments. *Atmospheric Environment*, 38:415–423, 3 2004.

[4] R. Christensen. *Linear Models for Multivariate, Time Series and Spatial data.* Springer Verlag, 1991.

[5] N. Cressie. *Statistics for spatial data.* Wiley, 1991.

[6] N. Cressie and D.M. Hawkins. Robust estimation of the variogram. *Journal of the International Association for Mathematical Geology*, 12:115–125, 1980.

[7] D.M. Hawkins and N. Cressie. Robust kriging-a proposal. *Journal of the International Association for Mathematical Geology*, 16:3–18, 1984.

[8] C.M. Jarque and A.K. Bera. A test for normality of observations and regression residuals. *International Statistical Review*, 55:1–10, 1987.

[9] World Health Oorganization. *WHO air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulphur dioxide.* World Health Organization, 2006.

[10] European Parliament. Directive 1999/30/EC of the European Parliament on limit values for particulate matter. *Official Journal of the European Union*, L 163, 1999.

[11] European Parliament. Directive 2008/50/EC of the European Parliament on ambient air quality. *Official Journal of the European Union*, L 152, 2008.

[12] A. Pope. Epidemiology of fine particulate and human health: Biological mechanisms and who is at risk? *Environmental Health Perspective*, 108, 2000.

[13] H. Wackernagel. *Multivariate Geostatistics.* Springer Verlag, Berlin, 1995.

[14] A. T. Walden and P.Guttorp. *Statistics in the environmental and earth sciences.* Arnold, 1992.

MOX Technical Reports, last issues

Dipartimento di Matematica "F. Brioschi",
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

18/2008 L. Bonaventura, S. Castruccio, P. Crippa, G. Lonati:
Geostatistical estimate of PM10 concentrations in Northern Italy: validation of kriging reconstructions with classical and flexible variogram models

17/2008 A. Ern, S. Perotto, A. Veneziani:
Hierarchical model reduction for advection-diffusion-reaction problems

16/2008 L. Formaggia, E. Miglio, A. Mola, A. Scotti:
Numerical simulation of the dynamics of boat by a variational inequality approach

15/2008 S. Micheletti, S. Perotto:
An anisotropic mesh adaptation procedure for an optimal control problem of the advection-diffusion-reaction equation

14/2008 C. D'Angelo, P. Zunino:
A finite element method based on weighted interior penalties for heterogeneous incompressible flows

13/2008 L.M. Sangalli, P. Secchi, S. Vantini, V. Vitelli:
K-means alignment for curve clustering

12/2008 T. Passerini, M.R. de Luca, L. Formaggia, A. Quarteroni, A. Veneziani:
A 3D/1D geometrical multiscale model of cerebral vasculature

11/2008 L. Gerardo Giorda, L. Mirabella, F. Nobile, M. Perego, A. Veneziani:
A model preconditioner for the Bidomain problem in electrocardiology

10/2008 N. Grieco, E. Corrada, G. Sesana, G. Fontana, F. Lombardi, F. Ieva, A.M. Paganoni, M. Marzegalli:
Predictors of the reduction of treatment time for ST-segment elevation myocardial infarction in a complex urban reality. The MoMi^2 survey

9/2008 P. Secchi, E. Zio, F. Di Maio:
Quantifying Uncertainties in the Estimation of Safety Parameters by Using Bootstrapped Artificial Neural Networks