Space-Time Integration of Heterogeneous Networks in Air Quality Monitoring [1]

Integrazione spazio temporale di reti etereogenee per il monitoraggio della qualità dell'aria

Alessandro Fassò, Orietta Nicolis [2]
Dipartimento di Ingegneria Gestionale e dell'Informazione
Università di Bergamo
alessandro.fasso@unibg.it, orietta.nicolis@unibg.it

Riassunto: Lo scopo di questo lavoro è di proporre un modello per l'integrazione di dati provenienti da reti eterogenee di monitoraggio al fine di valutare la qualità dell'aria. Per esempio, la rete di monitoraggio del PM_{10} nel Nord Italia è composta per la gran parte da centraline che si basano su due sistemi di rilevamento, TEOM and LVG. Mentre i dati rilevati con il metodo TEOM sottostimano il "vero" livello di $PM_{10,}$ le centraline LVG sono più precise e per questo sono state scelte dalla Comunità Europea come "strumenti di riferimento". L'idea su cui si basa il lavoro è di utilizzare le concentrazioni giornaliere dei PM_{10}, misurate con gli strumenti più precisi, per correggere le misure meno esatte, rilevate da centraline "non equivalenti" a quelle gravimetriche e che, non necessariamente si trovano nella stessa zona in cui sono situati i primi. A tal fine, introduciamo un modello di calibrazione multivariato spazio temporale che abbiamo denominato *Geostatistical Dynamical Calibration model* (GDC). La principale ipotesi su cui si basa il modello è che entrambi gli strumenti siano contaminati da errori di misura e che le rilevazioni TEOM siano distorte, rispetto alle "vere" concentrazioni, per un fattore addittivo ed uno moltiplicativo. Si assume, inoltre, che "vero" livello di PM_{10} sia un processo spazio temporale latente, rappresentato dall'equazione di stato nella formulazione *state space*. Le stime dei valori calibrati si ottengono dall'applicazione del filtro di Kalman. Questo approccio può essere considerato un'estensione geostatistica del modello DDC (*Dynamical Displaced Calibration*) di Fassò and Nicolis (2004).

Keywords: Calibration; spatio-temporal modelling; Kalman filter; Geostatistical Dynamical Calibration model (GDC).

1. Introduction

In recent years, atmospheric pollution has been of great concern for many countries of the world, as a result of studies that have verified the negative effects on human health. This has carried some public institutions to invest in instruments of measure in order to find the levels of concentrations, widening the monitoring net, and setting limit values for assessing air quality standards (see, for example, C.D. (1996)).

The pollutants that have recently aroused greater worries are the fine particulate matters. These have often exceeded the limits of attention and alarm for human health established by European legislation. Studies have shown (Gerrity *et al.* (1979)) that such

[1] Work partially supported by Italian MIUR grant, PRIN-Cofin 2004.
[2] Indirizzo per corrispondenza: Viale Marconi, 5, 24044 Dalmine (BG), Italy.

particles can easily be inhaled into the upper human respiratory tract where they can remain for weeks or even months before being excreted. The smallest of these particles, those with a diameter of 10 mm and 2.5 mm or less (i.e. PM2.5) are of even greater concern since they penetrate deep into the lungs and leave any of the substances of which they may be comprised.

Short term space-time statistical forecasting models for PM_{10} have been recently considered by Shaddick and Wakefield (2002) and by Sun *et al.* (2000) from the hierarchical Bayesian point of view and, for PM2.5, by Kolenikov and Smith (2002) using a nonparametric approach.

In this work, we consider the airborne particles with a diameter of 10 mm or less (i.e. PM_{10} in mg/m^3) in the North of Italy. The first instruments for monitoring PM_{10} were installed about ten years ago, but a network with a sufficient number of monitors for doing geostatistical analysis has only recently been set up.

The method of reference for the sampling and the measurement of PM_{10} given by the Italian D.M. (2002) is described in the European norm EN 12341 "*Air quality - Determination of the PM_{10} fraction of suspended particulate matter. Reference method and field test procedure to demonstrate reference equivalence of measurement methods*". The measurement principle is based on the collection of the PM_{10} on a filter and the determination of its mass for gravimeter way. We denote by LVG (Low Volume Gravimeter) the reference instrument.

In the North of Italy the monitoring networks are composed of different monitor types, not always based on a gravimeter principle and it is often necessary to apply suitable transformations to make the data equivalent to those gathered by the reference system cited in EN 12341.

The aim of this work is to propose statistical calibration models to integrate heterogeneous networks in air quality monitoring. As a matter of fact, we have relatively dense networks which are based on the well known automatic monitors based on a tapered element oscillating microbalance (TEOM). These monitors are known to underestimate the "true" PM_{10} level given by the reference method. However, the TEOM monitoring system has many advantages consequent to automatic operations and can give hourly data. A correction factor of 1.3 has been proposed, for example, by the APEG (1999) but some preliminary results obtained on the data of the region Piemonte (Fassò and Nicolis (2004)) show that this method gives an overestimation "for low values" and an underestimation "for high values".

Since, in the same areas, we have some gravimeter monitors (LVG), often located in different sites from the TEOM monitors, the idea of this paper is to use the PM_{10} concentrations from LVG monitors to perform a dynamical calibration of the spatially displaced TEOM data. In particular, we introduce a Geostatistical Dynamical Calibration model (GDC), based on a multivariate state - space approach.

There are several approaches to the statistical calibration (inverse linear regression, nonparametric regression, non liner models, etc.) with applications in different fields (see Osborne (1991), for a review). Recently, the state - space approach there has been considerable interest in the calibration for the spatio-temporal measures (see, for example, McBride and Clyde (2003) from the Bayesian point of view for the PM2.5 calibration and Brown *et al.* (2001) for the calibration of the rainfall radar data).

The model that we propose in this paper is based on the hypothesis that both instruments, LVG and TEOM, are affected by measurements errors. In particular, the TEOM measures may be biased relative to the "true" concentrations by an additive and multi-

plicative factor. The "true" measurement is an unobservable spatio temporal process and represents the state equation of the model. The estimates of "calibrated" values are obtained from the Kalman filtering procedure. Since we consider the spatial correlation between data gathered by different monitors, this approach is a geostatistical extension of the Dynamical Dispaced Calibration model of Fassò and Nicolis (2004). In order to reduce the dimensionality of the model, we decompose the "true" process in the K principal fields (Mardia *et al.* (1998)) using the Empirical Orthogonal Function (EOF) decomposition (Wikle and Cressie (1999), Wikle (2002)). The EOF analysis is used by environmental and meteorological statisticians for several reasons, which can be summarized as follows: it permits the extraction of information from the huge spatio-temporal datasets, reducing dimensionality; it is able to account for multiscale dynamical variability across different dynamical variables in space and time, account for various sources of errors; it gives an optimal and separable orthogonal decomposition of a spatio-temporal process (Wikle (2002)).

The paper is organized as follows. In Section 2, we describe the Geostatistical Dynamical Calibration Model both in the scalar and matrix form. In Sections 3-6 we discuss the estimation procedure of the model: the preliminary analysis of data with the Empirical Orthogonal functions, the estimation of the parameters via the maximum likelihood function and the estimation of the "calibrated values". In Section 7 we show some preliminary results. Section 8 concludes the paper with some comments and open problems.

2. Geostatistical Dynamical Calibration Model (GDC)

Let $y_G(t_1, s_1), ..., y_G(t_N, s_p)$ denote the concentrations of PM_{10} measured by the LVG monitor in $\mu g/m^3$ at time points $\{t_1, \dots, t_N\}$ at the stations $\{s_1, ..., s_p\}$, where $s_i \in \mathcal{D}$, $(i = 1, ...)$ with D some spatial domain in d-dimensional Euclidean space, while $y_T(t_1, s_{p+1}), ..., y_T(t_N, s_{p+q})$ denote the PM_{10} measurements of the TEOM monitors at the same time point, but in displaced stations $\{s_{p+1}, ..., s_{p+q}\}$. Assuming that both data sets have a component error, the model can be expressed by the following measurement equations

$$y_G(t, s_G) = \mu(t, s_G) + \varepsilon_G(t, s_G), \quad (1)$$

$$y_T(t, s_T) = \alpha(t) + \beta\mu(t, s_T) + \varepsilon_T(t, s_T), \quad (2)$$

where $\{s_G : s_1, ..., s_p\}$ and $\{s_T : s_{p+1}, ..., s_{p+q}\}$ are spatial points on the irregular grids, M_G and M_T, respectively, so that $card(M_G) = p$, $card(M_T) = q$ and $n = p + q$ is the total number of stations considered. Additionally, we assume that the errors components $\varepsilon_G(t, s_G)$ and $\varepsilon_T(t, s_T)$ are indipendently and Normally distributed with mean zero and standard deviations σ_{ε_G} (for $s_1, ..., s_p$) and σ_{ε_T} (for $s_{p+1}, ..., s_{p+q}$), respectively, with $E[\varepsilon_G(t, s_G), \varepsilon_T(t, s_T)] = 0$. The component $\mu(t, s)$ is an unobservable process that represents the "true" PM_{10} concentration on day t at station s, $\{s : s_1, ..., s_n\}$ while $\alpha(t)$ is a dynamical temporal additive bias and β is a parameter that represents the multiplicative bias. In particular, we assume that $\alpha(t)$ is generated by a first order latent Markovian process

$$\alpha(t) = \tau\alpha(t-1) + \varsigma(t) \quad (3)$$

with $0 \leq \tau \leq 1$ and $\varsigma(t)$ is a iid process with mean zero and standard deviation σ_ς.

We assume that the unobservable process $\mu(t,s)$ can be written

$$\mu(t,s) = \mu_0(t,s) + \mu_k(t,s) \tag{4}$$

where $\mu_0(t,s)$ is a component representing the small-scale spatial variation that does not have a temporally dynamic structure, that is $E[\mu_0(t,s_i),\mu_0(t,s_j)] = C_0(\|s_i - s_j\|)$, for $s_i \neq s_j$. Following the approach of Mardia *et al.* (1998) and Wikle and Cressie (1999) we assume the component $\mu_k(t,s)$ is a temporally dynamic component,

$$\mu_k(t,s) = \int_D w(s)\mu_k(t-1,u)du + \eta(t,s)$$

where $\eta(t,s)$ is a spatially coloured error process and $w(s)$ is a function representing the interaction between the state process at location u and time $t-1$ and μ_k at location s and time t (Wikle and Cressie (1999)). We assume that $\eta(t,s) \sim N(0,\sigma_\eta^2)$ and $E[\mu_k(t-1,u),\eta(t,s)] = 0$ for all u,s,t.

To reduce the dimensionality of the state equations (Mardia *et al.* (1998)), the component $\mu_k(t,s)$ can be decomposed into K dominant components, known as principal fields,

$$\mu_k(t,s) = \sum_{k=1}^{K} \phi_k(s)a_k(t) \tag{5}$$

where $\{a_k(t) : k = 1,...,K\}$ are zero-mean time series, and $\{\phi_k(s) : k = 1,...,K\}$ are deterministic basis functions that are complete and orthonormal, that is

$$\int_{\mathcal{D}} \phi_k(s)\phi_l(s) = \begin{cases} 1 & \text{for } k = l \\ 0 & \text{otherwise.} \end{cases}$$

Denoting by $\phi(s) = (\phi_1(s),...,\phi_K(s))'$ and $\mathbf{a}(t) = (a_1(t),...,a_K(t))$,the temporal dynamical component can be written as

$$\mu_k(t,s) = \phi(s)'\mathbf{a}(t) \tag{6}$$

where $\mathbf{a}(t)$ is assumed to evolve according to the state equation,

$$\phi(s)\mathbf{a}(t) = b(s)\mathbf{a}(t-1) + \eta(t,s) \tag{7}$$

where $b(s) = (b_1(s),...,b_K(s))'$ are unknown but nonstochastic parameters. Substituting (6) and (4) in equations (1) and (2) the measurements equations can be written as

$$\begin{aligned} y_G(t,s_G) &= \mu_0(t,s_G) + \phi(s_G)'\mathbf{a}(t) + \varepsilon_G(t,s_G), & (8)\\ y_T(t,s_T) &= \alpha(t) + \beta\left[\mu_0(t,s_T) + \phi(s_T)'\mathbf{a}(t)\right] + \varepsilon_T(t,s_T). & (9) \end{aligned}$$

Assuming that we have a matrix $m = n \times N$ of observations at locations $\{s_1,\ldots,s_n\}$ with $n = p+q$ and at time points $\{1,\ldots,N\}$, we can rewrite the measurement equations in the matrix form

$$\begin{aligned} \mathbf{y}_G(t) &= \mu_{0G}(t) + \Phi_G\mathbf{a}(t) + \varepsilon_G(t) & (10)\\ \mathbf{y}_T(t) &= \alpha(t) + \beta\mu_{0T}(t) + \beta\Phi_T\mathbf{a}(t) + \varepsilon_T(t) & (11) \end{aligned}$$

while the state equations are

$$\alpha(t) = \tau\alpha(t-1) + \varsigma(t) \quad (12)$$
$$\Phi\mathbf{a}(t) = \mathbf{B}\mathbf{a}(t-1) + \eta(t) \quad (13)$$

where

$$\mathbf{y}_G(t) = \begin{pmatrix} y_G(t,s_1) \\ \vdots \\ y_G(t,s_p) \end{pmatrix};\ \mathbf{y}_T(t) = \begin{pmatrix} y_T(t,s_{p+1}) \\ \vdots \\ y_T(t,s_{p+q}) \end{pmatrix}$$

are the observations for LVG and TEOM and

$$\mu_{0G}(t) = \begin{pmatrix} \mu_0(t,s_1) \\ \vdots \\ \mu_0(t,s_p) \end{pmatrix};\ \mu_{0G}(t) = \begin{pmatrix} \mu_0(t,s_{p+1}) \\ \vdots \\ \mu_0(t,s_{p+q}) \end{pmatrix}$$

are the small-scale components; we note that the matrices

$$\Phi_G = \{\phi(s_1),...,\phi(s_p)\}' \text{ and } \Phi_T = \{\phi(s_{p+1}),...,\phi(s_{p+q})\}'$$

are $(p \times K)$ and $(q \times K)$, respectively. The elements of Φ_G and Φ_T are the basis functions $\phi_1,...,\phi_K$ evaluated at each TEOM and LVG location. Combining Φ_G and Φ_T we obtain the $n \times K$ matrix $\Phi = \{\phi(s_1),...,\phi(s_n)\}'$. The coefficient of the state equations are expressed by parameter τ and by the transition matrix $\mathbf{B} = \{b(s_1),...,b(s_n)\}'$. In this formulation the errors components have a Multinormal distribution: $\varepsilon(t) \sim N_n(0,\Sigma_\varepsilon)$, with $\Sigma_\varepsilon = diag(\sigma^2_{\varepsilon_G},...,\sigma^2_{\varepsilon_T})$; $\mu_0(t) \sim N_n(0,\Gamma)$, is the small-scale component with a spatial covariance and $\Gamma_{ij} = C_{\mu_0}(\|s_i - s_j\|)$ for $\{s_i : s_1,...,s_n\}$ is the generic element of the covariance matrix Γ; finally, $\eta(t) \sim N_K(0,\Sigma_\eta)$, with $\Sigma_\varepsilon = diag(\sigma^2_{\eta_1},...,\sigma^2_{\eta_K})$.

The state equation (13) can be written as

$$\mathbf{a}(t) = \mathbf{H}\mathbf{a}(t-1) + \mathbf{J}\eta(t) \quad (14)$$

where $\mathbf{J} = (\Phi'\Phi)^{-1}\Phi'$ and $\mathbf{H} = \mathbf{JB}$.

Assuming that all the parameters of the model are known, determining the "calibrated" values is the same as finding the optimal predictor in a state space model by the Kalman filter, given observations up to and including time t.

3. Estimation

In order to estimate the parameters of the GDC model, it is convenient to express the model in the following standard state-space form,

$$\mathbf{Y}(t) = \mathbf{M}(t) + \Theta \cdot \mathbf{A}(t) + \varepsilon(t) \quad (15)$$
$$\mathbf{A}(t) = \Psi \cdot \mathbf{A}(t-1) + \mathbf{J}^{*\prime} \cdot \eta^*(t) \quad (16)$$

where

$$\begin{aligned}\mathbf{Y}(t) &= \begin{pmatrix} \mathbf{y}_G(t) \\ \mathbf{y}_T(t) \end{pmatrix}; \mathbf{M}(t) = \begin{pmatrix} \mu_0(t,s_1) \\ \vdots \\ \beta\mu_0(t,s_n) \end{pmatrix}; \varepsilon(t) = \begin{pmatrix} \varepsilon_G(t) \\ \varepsilon_T(t) \end{pmatrix}; \\ \mathbf{A}(t) &= \begin{pmatrix} \alpha(t) \\ \mathbf{a}(t) \end{pmatrix}; \eta^*(t) = \begin{pmatrix} \varsigma(t) \\ \eta(t) \end{pmatrix}\end{aligned}$$

and $\mathbf{J}^* = (1, \mathbf{J}^*)$. The matrices of the measurement and state equations can be written as

$$\Theta = \begin{pmatrix} \mathbf{0}_p & \Phi_G(s) \\ 1_q & \beta\Phi_T(s) \end{pmatrix} \text{ and } \Psi = \begin{pmatrix} \tau & 0 \\ 0 & \mathbf{H} \end{pmatrix}.$$

The error components are distribuited as follows: $\varepsilon(t) \sim N_n(0, \Sigma_\varepsilon)$, and $\eta^*(t) \sim N_{K+1}(0, \Sigma_{\eta*})$, where

$$\Sigma_{\eta*} = \begin{pmatrix} \sigma_\varsigma^2 & 0 \\ 0 & \Sigma_\eta \end{pmatrix} \text{ and } \Sigma_\varepsilon = \begin{pmatrix} \Sigma_{\varepsilon_G} & 0 \\ 0 & \Sigma_{\varepsilon_T} \end{pmatrix}$$

are the variances for the errors of measurement and state equations, respectively.

The estimation procedure for the parameters of the model (15) can be divided into three main steps:

1. first we compute the matrices Φ_G and Φ_T by EOF analysis, given LVB observations, $\mathbf{y}_G(t)$ (see Section 4);
2. by the optimization of the of maximum likelihood obtained by the Kalman filter recursion, we estimate the parameters of the GDC model, given the matrices Φ_G and Φ_T resulting from step 1 (see Section 5);
3. finally, the Kalman filter algorithm provides the estimates of the state equation $\mathbf{A}(t)$ and the calibrated values $\widehat{\mu}(t, s_i)$, for $i = 1, ..n$ and $t = 1, ..N$, conditionally on the matrices Φ_G and Φ_T and the parameter estimates (see Section 6).

4. EOF analysis

We denote by Φ the $n \times K$ matrix including the basis function for the TEOM and LVG data,

$$\Phi = \begin{pmatrix} \Phi_G \\ \Phi_T \end{pmatrix}.$$

The evaluation of Φ is given by the Empirical Orthogonal Function (EOF) analysis of the LVG data, $\mathbf{y}_G$. EOF analysis can be seen as the geophysicist's manifestation of the classic eigenvalue/eigenvector decomposition of a correlation (or covariance) matrix. In its discrete formulation, EOF analysis is simply the Principal Component Analysis (PCA) decomposition, while in the continuous framework, it is simply a Karhunen-Loéve (K-L) expansion (see Wikle (2002)). Although the continuous K-L representation of EOFs is the most realistic from a physical point-of-view, it is only rarely considered in applications

due to the discrete nature of data observations and the added difficulty of solving the K-L integral equation.

In general, in a discrete EOF analysis, if we know $\mu(s,t)$ at each location $\{s : s_1, ..., s_n\}$ and time point $T = 1, ..., N$, $\mu(s,t) = (\mu(t,s_1), ..., \mu(t,s_n))\prime$, we can define the k-th EOF ($k = 1, ..., p$) to be $\phi_k = (\phi_k(s_1), ..., \phi_k(s_n))'$, where ϕ_k is the vector in the linear combination $a_k(t) = \phi_k'\mu(t,s)$. Furthermore, ϕ_1 is the vector that allows $var[a_1(t)]$ to be maximized subject to the constraint $\phi_1'\phi_1 = 1$. Then ϕ_2 is the vector that maximizes $var[a_2(t)]$ subject to the constraint $\phi_2'\phi_2 = 1$ and $cov\,[a_1(t), a_2(t)] = 0$. Thus, ϕ_k is the vector that maximizes $var[a_k(t)]$ subject to the orthogonality constraint $\phi_k'\phi_k = 1$ and $cov\,[a_k(t), a_j(t)] = 0$ for all $k \neq j$. This is equivalent to solving the eigensystem

$$\mathbf{C}_\mu \Phi = \Phi \Lambda$$

where $\mathbf{C}_\mu = E\,[\mu(t), \mu(t)]$, $\Phi = (\phi_1, ..., \phi_n)'$ with $\phi_k = (\phi_k(s_1), ..., \phi_k(s_p))'$,($k = 1, ..., n$), $\Lambda = diag(\lambda_1, ..., \lambda_p)$ and $var\,[a_i(t)] = \lambda_i$, $i = 1, ..., n$. The solution is obtained by a symmetric decomposition

$$\mathbf{C}_\mu = \Phi \Lambda \Phi'.$$

Considering the K eigenvalues we obtain the truncated expansion of equation (6). Since the EOF analysis depends on the decomposition of a covariance matrix, it is necessary to estimate this matrix in practice. The traditional approach is based on the method of moments estimation procedure (Wikle (2002)).

Since, in our model, the process $\mu(s,t)$ is unobservable, we estimate the $C_\mu(s_i, s_j)$ using the observations deriving from the more accurate instrument, that is, the LVG data $\mathbf{y}_G$. In practice, we use a simple space-time prediction scheme to obtain smooth predictions of $\mu(s,t)$, which we denote by $\tilde{\mu}(s,t)$, on a grid including the n locations (LVG and TEOM), given the $\mathbf{y}_G$ data. So the estimate of $C_\mu(s_i, s_j)$ is obtained evaluating the empirical variance of $\tilde{\mu}(s,t)$. We then apply the singular value decomposition to the estimated covariance matrix $\widehat{\mathbf{C}}_\mu$ to obtain the estimation of Φ and consequently the estimation of Φ_G and Φ_T. Choosing the K eigenvalues different from zero, the covariance function becomes

$$\widehat{C}_{\mu_K}(s_i, s_j) = \sum\nolimits_{k=1}^{K} \lambda_k \phi_k(s_i) \phi_k(s_j).$$

We assume that the covariance model for $\tilde{\mu}(s,t)$ is composed of a traditional geostatistical model (stationary, isotropic) and a truncated EOF expansion (Nychka and Saltzman (1998)),

$$C_{\tilde{\mu}}(s_i, s_j) = C_{\mu_0}(s_i, s_j) + C_{\mu_K}(s_i, s_j) \tag{17}$$

where $C_{\mu_0}(s_i, s_j)$ is an isotropic covariance function (such as exponential, Mathèrn, etc.), characterized by the spatial parameter set θ. In practice, the parameter set θ can be estimated preliminarily by the following difference

$$\widehat{C}_{\mu_0}(s_i, s_j) = \widehat{C}_{\tilde{\mu}}(s_i, s_j) - \widehat{C}_{\mu_k}(s_i, s_j)$$

or can be included in the parameter set of the Kalman filter maximum likelihood estimator.

5. Estimation of the parameters

The GDC model described by equations (15)-(16) is estimated by the maximum likelihood procedure. The parameters to be estimated in this model are the following: the temporal parameters β and τ; the spatial parameter set θ; the variances of the measurement and the state error components $\sigma^2_{\varepsilon_G}$, $\sigma^2_{\varepsilon_T}$, σ^2_{ς}, the elements of the matrices $\Sigma_\eta = diag(\sigma^2_{\eta_1}, ..., \sigma^2_{\eta_K})$ and $\mathbf{H}$. In Wikle and Cressie (1999), the matrix $\mathbf{H}$ is estimated by a preliminary procedure based on assuming that the variances of error components are known. Since, in our case, these components are estimated inside the model and $a_1(t), ..., a_K(t)$ are uncorrelated, we assume that $\mathbf{H}$ has a diagonal structure, given by $\mathbf{H} = diag(h_1, ..., h_K)$.

We denote by Ω the set of the unknown parameters of the model to be estimated, $\Omega = (\beta, \tau, h_1, ...h_K, \theta, \sigma^2_{\varsigma}, \sigma^2_{\eta_1}, ..., \sigma^2_{\eta_K}, \sigma^2_{\varepsilon_G}, \sigma^2_{\varepsilon_T})$. The maximum likelihood estimator uses the zero mean prediction error and its covariance obtained by the Kalman filter recursion. Let $\widehat{\mathbf{A}}(t|t-1)$ and $\widehat{\mathbf{A}}(t|t)$ be estimates of $\mathbf{A}(t-1)$ and $\mathbf{A}(t)$ up to the time $t-1$ and t, respectively, with covariances given by $\mathbf{P}(t|t-1)$ and $\mathbf{P}(t|t)$.The prediction equations for the state equations of the Kalman filter are

$$\begin{aligned} \widehat{\mathbf{A}}(t|t-1) &= \Psi \widehat{\mathbf{A}}(t|t-1) \\ \mathbf{P}(t|t-1) &= \Psi \mathbf{P}(t-1|t-1)\Psi' + \mathbf{J}^{*'}\Sigma_{\eta^*}\mathbf{J}^*. \end{aligned}$$

The one step prediction error is

$$\widetilde{\varepsilon}(t|t-1) = \mathbf{Y}(t) - \Theta \cdot \mathbf{A}(t|t-1) \tag{18}$$

and its covariance is

$$\Sigma_{\widetilde{\varepsilon}}(t|t-1) = \Theta \mathbf{P}(t|t-1)\Theta' + \Sigma_\varepsilon + \Sigma_{\mu_0}. \tag{19}$$

By the updating equations of Kalman filter, we get the smoothed estimates of the states

$$\begin{aligned} \widehat{\mathbf{A}}(t|t) &= \widehat{\mathbf{A}}(t|t-1) + \mathbf{G}(\mathbf{t})\widetilde{\varepsilon}(t|t-1) \\ &= \widehat{\mathbf{A}}(t|t-1) + \mathbf{G}(\mathbf{t})\left\{\mathbf{Y}(t) - \Theta\widehat{\mathbf{A}}(t|t-1)\right\} \end{aligned} \tag{20}$$

and its covariance matrix,

$$\begin{aligned} \mathbf{P}(t|t) &= \mathbf{P}(t|t-1) - \mathbf{G}(t)\Sigma_{\widetilde{\varepsilon}}(t|t-1)G(t) \\ &= \mathbf{P}(t|t-1) - \mathbf{G}(t)\Theta\mathbf{P}(t|t-1) \end{aligned} \tag{21}$$

where $\mathbf{G}(t)$ is the Kalman Gain and it is given by

$$\begin{aligned} G(t) &= P(t|t-1)\Theta\Sigma_{\widetilde{\varepsilon}}(t|t-1)^{-1} \\ &= P(t|t-1)\Theta\left\{\Sigma_\varepsilon + \Sigma_{\mathbf{M}} + \Theta\mathbf{P}(t|t-1)\Theta'\right\}^{-1} \end{aligned}$$

with $\Sigma_{\mathbf{M}} = var(M)$. The log-likelihood for Θ is given by

$$\ln L_Y(\Omega) = -\frac{1}{2}\sum_{t=1}^{N}\log|\Sigma_{\widetilde{\varepsilon}}(t|t-1)| - \frac{1}{2}\sum_{t=1}^{N}\widetilde{\varepsilon}(t|t-1)'\Sigma_{\widetilde{\varepsilon}}(t|t-1)^{-1}\widetilde{\varepsilon}(t|t-1). \tag{22}$$

Since (22) is non linear in the parameters, we need numerical algorithms to optimize it.

Figure 1: *Map of the PM_{10} monitoring network in the Piemonte region: low volume (LV or LVB), TEOM, high volume (HV) and BETA monitors.*

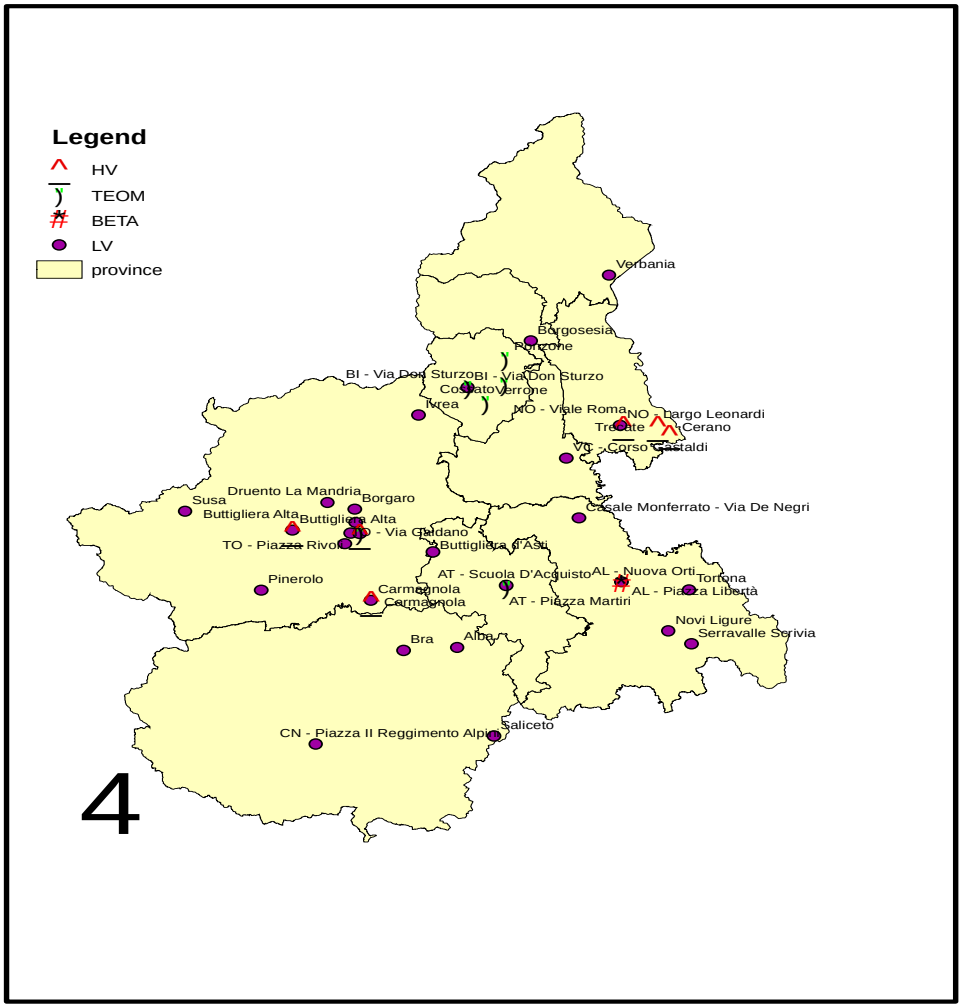

Table 1: *Principal Component Analysis: Standard Deviation, Proportion of Variance Explained and Cumulative Proportion of Variance Explained.*

	$C1$	$C2$	$C3$	$C4$	$C5$	$C6$	$C7$	$C8$	$C9$
Std. Dev.	98.35	27.17	23.22	18.73	15.51	14.03	12.77	12.30	11.75
Prop. of Var.	0.75	0.06	0.04	0.03	0.02	0.02	0.01	0.01	0.01
Cum. Prop.	0.75	0.81	0.85	0.88	0.89	0.91	0.92	0.93	0.94
	$C10$	$C11$	$C12$	$C13$	$C14$	$C15$	$C16$	$C17$	$C18$
Std. Dev.	11.25	10.51	10.05	9.53	8.74	8.20	7.83	7.17	6.23
Prop. of Var.	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.00	0.00
Cum. Prop.	0.95	0.96	0.97	0.98	0.98	0.99	0.99	1.00	1.00

Table 2: *Principal Component Analysis: four principal components.*

Centraline	C1	C2	C3	C4
AL-Nuova Orti	0.23	-0.44	0.04	-0.14
Alba	0.26	-0.02	-0.12	0.25
Borgaro	0.22	0.14	-0.08	-0.10
Borgosesia	0.20	0.10	0.08	-0.01
Bra	0.27	0.07	0.04	-0.02
Buttigliera Alta	0.22	0.19	0.02	0.16
Buttigliera d'Asti	0.22	0.06	0.31	-0.02
Carmagnola	0.29	0.09	0.06	-0.04
Casale Monferrato	0.25	-0.23	0.07	-0.14
CN-P.II Regg.Alp.	0.14	0.34	0.35	0.44
Novi Ligure	0.24	-0.55	0.03	0.36
Pinerolo	0.18	0.23	0.11	0.23
TO-Piazza Rivoli	0.26	0.15	-0.43	0.00
TO-Via Consolata	0.32	0.17	-0.22	-0.17
TO-Via Gaidano	0.28	0.04	-0.50	-0.09
Tortona	0.21	-0.36	0.09	0.20
VC-Corso.Gastaldi	0.24	0.01	0.41	-0.63
Ponzone	0.12	0.12	0.25	0.07

6. The calibrated values

The estimation of the process $\mu(t, s_i)$ for $i = (s_1, ..., s_n)$ is based on the Kalman filter predictor $\widehat{\mathbf{A}}(t|t)$ of equation (20). Since

$$\widehat{\mathbf{A}}(t|t) = \begin{pmatrix} \widehat{\alpha}(t|t) \\ \widehat{\mathbf{a}}(t|t) \end{pmatrix}$$

with $\widehat{\mathbf{a}}(t|t) = (\widehat{a}_1(t|t), ..., \widehat{a}_K(t|t))'$, the Kalman filter prediction of $\widehat{\mu}(t, s_i)$

$$\widehat{\mu}(t, s_i) = \phi(s_i)'\widehat{\mathbf{a}}(t|t) + C_{\mu_0}(s_i)' \left(C_{y_G}\right)^{-1} y_G(t)$$

where $\phi(s_i) = (\phi_1(s_i), ..., \phi_K(s_i))'$, $C_{y_G}(s_i, s_j) = cov\{y_G, y_G\}$, and

$$C_{\mu_0}(s_i) = E\{\mu_0(t, s_i), \mu_0(t)\} = (C_{\mu_0}(s_i, s_1), ..., C_{\mu_0}(s_i, s_p))'.$$

We note that $C_{\mu_0}(s_i, s_j)$ is a spatial covariance structure (see Section 4) and the second component of $\widehat{\mu}(t, s)$ is a type of simple kriging applied to the LVG spatial error term $\mu_0(t, s_G)$ (Wikle and Cressie (1999)).

7. The case of the Piemonte region: some preliminary results

The PM_{10} monitoring network of the Piemonte region (in the North of Italy) is composed by different measurement monitors (see Figure 1): low volume (LV or LVB), TEOM, high volume (HV) and BETA monitors[3] (Section 1). The data of this analysis consist

[3]The data considered in this work have been gathered by the Piemonte AriaWeb informative system that is a branch of the *Sistema Regionale di Rilevamento della Qualità dell'Aria (SRQA)*.

Table 3: *Estimated paramters of a GDC model.*

Log-likelihood:	25859.2		
	Value	Std.Error	t-value
α	41.1400	0.52870	77.82
h_1	0.8624	0.02711	31.81
h_2	0.7692	0.04077	18.87
β	0.5775	0.02200	26.26
$\ln(\sigma^2_{\eta_1})$	7.8230	0.08312	94.11
$\ln(\sigma^2_{\eta_2})$	5.4340	0.13970	38.91
$\ln(\sigma^2_{\epsilon})$	5.2110	0.01898	274.50

of daily concentrations of PM_{10}, measured by LVG monitors and TEOM monitors for the period 1th January 2003 to 31th December 2003 ($T = 365$). Since some monitors stations were composed by a large number of missing values, we only selected the stations with more than 90% of the validated data. In particular, we consider a number of 16 LVG monitors situated in the following sites: *AL-Nuova Orti, Alba, Borgaro, Borgosesia, Bra, Buttigliera Alta, Buttigliera d'Asti, Carmagnola, Casale Monferrato, CN-Piazza.II.Reggimento Alpini, Novi Ligure, Pinerolo, TO-Piazza Rivoli, TO-Via Gaidano, Tortona, VC-Corso Gastaldi* and 2 TEOM monitors situated in *TO-Via Consolata* and *Ponzone*.

In *TO-Via Consolata* we also had the LVG readings, but we utilized them only for the validation of the GDC model. The application of the GDC model to the Piemonte PM_{10} concentrations can be summarized into the followings steps: (i) the evaluation of the principal components, that is, the Φ matrix; (ii) the estimation of the parameters and (iii) the determination of the calibrated values. Since our aim was to obtain some preliminary results, we considered a simplified model, in which we assume that the α parameter of Equation (11) is constant, the small scale-variation is zero and the variance of LVG errors ($\sigma^2_{\epsilon_G}$) is equal to the variance of TEOM errors ($\sigma^2_{\epsilon_T}$).

From the application of the principal component analysis to the LVG data we have obtained the results shown in Table 1 and Table 2. Since we didn't know the gravimeter measure of the TEOM sites, we utilized, as a preliminary estimation, the TEOM readings multiplied by 1.3 (see Section 1). From the results of Table 1, we can see that the first two components sum up to more than 80% of the variance. The estimates of the GDC model with two principal components are represented in Table 3. We denoted by h_1 and h_2 the elements of the matrix Ψ. To avoid the problem of negative variances in the estimation procedure, we considered the logarithm transformations. To estimate the model we consider the LVG data, after removing a constant trend. It is interesting to note the small standard deviations of each parameter estimate. Finally, we compared the calibrated values for TO - Via Consolata with the LVG measures in the same site, obtaining a R-square value of about 0.90.

8. Conclusions and further developments

In this work we have introduced a geostatistical state - space approach to the calibration problem of PM_{10} data in the Piemonte region. The GDC model is able to correct the TEOM data in each site by using the observations gathered in any other site of the LVG

monitors and considering both temporal dynamics and spatial correlations. Thanks to the EOF analysis, it is possible to reduce the dimensionality of the model. From preliminary results, we can see that the GDC model produces significant parameter estimates. We intend to apply the model to a large database including all monitors of the North of Italy and considering the small-scale spatial component. However, we think that the model could further be expanded to consider exogenous variables (for example, temperature and wind direction) and different basis functions (wavelets, non parametric functions, etc.) in the principal field decomposition.

References

APEG (1999) *Source Apportionment of Airborne Particulate Matter in the United Kingdom*, Report of the Airborne Particles Expert Group.

Brown P.E., Diggle P.J., Lord M.E. and Young P. (2001) Space-time calibration of radar-rainfall data, *Journal of the Royal Statistical Society, Series C*, 50, 221–241.

C.D. (1996) Council directive 96/62/ec of 27 september 1996 on ambient air quality assessment and management, *Official Journal*, 296, 55–63.

D.M. (2002) Decreto ministeriale 2 aprile 2002, n. 60. recepimento della direttiva 1999/30/ce del consiglio del 22 aprile 1999 concernente i valori limite di qualità dell'aria ambiente per il biossido di zolfo, il biossido di azoto, gli ossidi di azoto, le particelle e il piombo e della direttiva 2000/69/ce relativa ai valori limite di qualità dell'aria ambiente per il benzene ed il monossido di carbonio, *Suppl. n. 77 alla G.U. n. 87 del 13 aprile 2002*, 296, 55–63.

Fassò A. and Nicolis O. (2004) Modelling dynamics and uncertainty in assessment of quality standards for fine particulate matters, *Working Paper n.21, GRASPA*.

Gerrity T.R., Lee P.S., Hass E.J., Marinelli A., Werner P. and Lourenco R.V. (1979) Calculated deposition of inhaled particles in the airway generations of normal subjects, *Journal of Applied Physiology*, 49, 867–873.

Kolenikov S. and Smith R. (2002) Spatio-temporal modeling of the longitudinal pm2.5 data with missing values, *ASA Proceedings of Joint Statistical Meetings*.

Mardia K., Goodall C., Redfern E. and Alonso F. (1998) The kriged kalman filter, *Sociedad de Estadistica e Investigacion Operativa Test*, 7, 217–285.

McBride S. and Clyde M. (2003) Design of air quality monitoring networks, *Working Paper 03-23, Duke University*.

Nychka D. and Saltzman N. (1998) Design of air quality monitoring networks, *Lecture Notes in Statistics: Case Studies in Environmental Statistics*.

Osborne C. (1991) Statistical calibration: A review, *International Statistical Review*, 59, 309–366.

Shaddick G. and Wakefield J. (2002) Modelling daily multivariate pollutant data at multiple sites, *Journal of the Royal Statistical Society, Series C*, 51, 351–372.

Sun L., Zidek J.V., Le N.D. and Özkaynak H. (2000) Interpolating vancouver's daily ambient pm10 field, *EnvironMetrics*, 11, 651–663.

Wikle C.K. (2002) Spatio-temporal models in climatology, *to appear in Encyclopedia of Life Support Systems*.

Wikle C.K. and Cressie N. (1999) A dimension-reduced approach to space-time kalman filltering, *Biometrika*, 86, 815–824.