A Unified Statistical Approach for Simulation, Modeling, Analysis and Mapping of Environmental Data

Alessandro Fassò
Michela Cameletti
University of Bergamo
Viale Marconi n. 5
24044 Dalmine (BG)
alessandro.fasso@unibg.it

In this paper, hierarchical models are proposed as a general approach for spatio-temporal problems, including dynamical mapping, and the analysis of the outputs from complex environmental modeling chains. In this frame, it is easy to define various model components concerning both model outputs and empirical data and to cover with both spatial and temporal correlation. Moreover, special sensitivity analysis techniques are developed for understanding both model components and mapping capability. The motivating application is the dynamical mapping of airborne particulate matters for risk monitoring using data from both a monitoring network and a computer model chain, which includes an emission, a meteorological and a chemical-transport module. Model estimation is determined by the Expectation-Maximization (EM) algorithm associated with simulation-based spatio-temporal parametric bootstrap. Applying sensitivity analysis techniques to the same hierarchical model provides interesting insights into the computer model chain.

Keywords: Hierarchical modeling, spatio-temporal process, EM algorithm, sensitivity analysis, particulate matters

1. Introduction

Thanks to the increase in development and use of simulation models for environmental studies, computational and statistical models have become increasingly coupled together.

1.1 General Remarks

On the one hand, statistical methods can be successfully used for the analysis of environmental computer models [1], in particular for planning computer simulations by means of Monte Carlo or more general computer experiments [2] and for modeling and analysis of the uncertainty of model outputs and sensitivity analysis [3].

Moreover, statistical modeling has been proved useful for constructing model emulators [4]. These are simplified versions of more complex environmental computer models and can be used for model interpretation [5, 6] and approximated code runs. The latter are especially useful for expensive and time-consuming code runs, which include meteorology, transport, etc. Statistical modeling is also being increasingly used to integrate simulated and observed data in the so-called 'data assimilation' problem which may be tackled by statisticians using the order-reduced Kalman filtering approach [7]. In some other cases, the Bayesian approach is useful to model the uncertainty of mechanistic models [8].

On the other hand, simulation is becoming an important part of statistical estimation of environmental models. When these simulations involve a large number of replications of complex spatio-temporal model runs, a distributed computing environment is called for.

1.2 Case Study

We show how the general approach of hierarchical spatio-temporal models may be used in mapping and understanding airborne particulate matter concentrations, when data are available on a daily frequency from different sources.

SIMULATION, Vol. 86, Issue 3, March 2010 139–154
© 2010 The Society for Modeling and Simulation International
DOI: 10.1177/0037549709102150
Figure 1 appears in color online: http://sim.sagepub.com

These sources include land description, a monitoring network and a computer model chain (EMCT) which includes modules for emissions, meteorology, chemistry and transport. Moreover, we show how to assess the simulation model outputs and obtain a surrogate model which may be useful for further simulations as the full model chain is computationally expensive. Hence, following the approach of [9] and [10], we demonstrate how the abovementioned problems can be unified to some extent and the resulting uncertainty assessed.

1.3 Hierarchical Models

As described in [11–13], the hierarchical approach is useful for complex environmental processes. These vary in time and space and depend on several variables that interact on a wide variety of scales. As a matter of fact, this approach makes it possible to take a conditional viewpoint for which the joint probability distribution of the spatio-temporal process can be expressed as the product of some simpler conditional distributions defined at each hierarchical stage. Hierarchical models can be tackled using a Bayesian or a classical point of view, the latter being the main focus of this paper. In particular, both cases rely heavily on simulations and this may result in huge computational challenges.

This modeling approach has already been used in various environmental applications. For example, spatio-temporal modeling for calibration of radar rainfall data by means of a ground-truth monitoring network have been considered [14]. Similarly, the calibration of particulate matters measurements from heterogeneous networks has been investigated [15]; daily sulphur dioxide data have been analyzed [16] and near-surface wind modeling has been developed [7]. Moreover, in hydrology, the concurrent estimation of model parameters and missing data in river runoff series has been considered [17]. [18] demonstrate that, for hourly air quality data, non-linear models seem more appropriate; however, on the daily scale it is common practice to use linear models on transformed data.

Using the Bayesian paradigm, Markov Chain Monte Carlo (MCMC) algorithms are required for sampling from the posterior distributions of the parameters while, in the classical framework, optimization and resampling methods are used for estimation and uncertainty assessment. For example, the EM algorithm [19, 20] is a sequential estimation algorithm, while bootstrap resampling is an embarrassingly parallel problem with coarse grains.

1.4 Sensitivity Analysis

The statistical approach to sensitivity analysis (SA) is essentially based on some appropriate variance decompositions. According to this, the sensitivity of a model output to certain model inputs is assessed in terms of output uncertainty which can be apportioned to each input.

The classical approach to statistical SA is based on model repeated simulation [2], which may be based on Monte Carlo or other sampling plans. Application of SA ranges from meteorology to econometrics, see e.g. [2]. In waste water treatment, SA has been considered [5] for sizing the treatment plant and understanding filter life span. Moreover, SA in recreational water quality monitoring has been considered [1]. In this paper, a model-based SA for correlated inputs is proposed. This extends the classical SA for independent inputs and also extends the results of [10], as a new concept of conditional SA is introduced.

1.5 Structure of Article

This article is organized as follows: the next four sections are on methods and one large section is on the case study. In particular, Section 2 discusses a rather general spatio-temporal model, which encompasses various model components and covers both spatial and temporal correlation. The model under consideration has three levels of hierarchy and takes into account a measurement error as well as a spatio-temporal dynamical field with both stochastic and systematic components. Temporal and spatial random effects are introduced in the second level and, in the last stage, a Markovian process is used for modeling the temporal dynamics of the latent process.

Section 2.3 defines the details of a new version of the EM algorithm for model coefficient estimation, which extends [21] to deterministic trends with covariates.

Section 3 considers the spatial interpolation problem for mapping and shows that this can be solved by the Kriging method implemented using the plug-in approach. The methodological part ends with Section 4, where the spatio-temporal parametric bootstrap is used for obtaining the parameter standard errors and for evaluating map uncertainty.

After briefly reviewing the basic concepts of SA, Section 5 introduces the concept of conditional SA, which is especially useful for hierarchical models in order to adjust for the latent components.

Section 6 discusses the application of daily data to airborne particulate matters for the Piemonte Region, Italy, where data from a monitoring network and an EMCT model chain are integrated. After model identification, the mapping capabilities are discussed and the model is interpreted in terms of various uncertainty decompositions. A Conclusion closes the paper.

2. Statistical Modeling

In this section, after introducing the structure of the hierarchical spatio-temporal model, the iterative estimation algorithm based on the Gaussian maximum likelihood is defined in detail.

2.1 Model Setup

In this work a three-stage hierarchical model takes into account measurement errors as well as deterministic and stochastic spatio-temporal dynamical fields. In particular, after using the first level for measurement error, temporal and spatial random effects are introduced at the second level. At the third level, a Markovian process models the temporal dynamics.

Suppose that a certain phenomenon, e.g. particulate matters concentration, is observed at location $s \in D$ and day $t = 1, 2, \ldots, T$ by the following measurement equation:

$$z(s,t) = u(s,t) + \varepsilon(s,t) \tag{1}$$

where $u(s,t)$ is the underlying 'true' local pollution level with the structure:

$$u(s,t) = X(s,t)\beta + K(s)\,y_t + \omega(s,t). \tag{2}$$

In Equation (2), $X(s,t)$ is a d-dimensional spatio-temporal field of known covariates observed at time t at location s including, for example, land features (which are purely spatial) and spatio-temporal fields that can be observed or simulated. The p-dimensional vector y_t, which is constant in space, is related to the 'global true' pollution level; the matrix $K(s)$ defines a $p-$dimensional field of known coefficients able to 'localize' the global level; for example, it may be based on the observed data through an EOF decomposition [7] or, in other cases, it may be constant over the geographical space D [22].

The process $\varepsilon(s,t)$ is a typical Gaussian instrumental error which is white noise in space and time with variance σ_ε^2. The Gaussian process $\omega(s,t)$ is the spatial small-scale component and is a white noise in time, but is correlated over space with a covariance function depending on the parameter θ, namely

$$E\left[\omega(s,t) \times \omega\left(s',t\right)\right] = \sigma_\omega^2 C_\theta(h)$$

where $h = \left\|s - s'\right\|$ is the Euclidean distance between sites s and s'. As the covariance function depends only on h, the spatial process $\omega(s,t)$ is second-order stationary and isotropic. Various examples of spatial covariance functions are discussed in [23, chapter 1] and a typical case is given by the following exponential function

$$C_\theta(h) = \exp(-\theta h). \tag{3}$$

Moreover, y_t has stable Markovian temporal dynamics given by

$$y_t = Gy_{t-1} + \eta_t \tag{4}$$

where η_t is a p-dimensional Gaussian white noise process with variance-covariance matrix Σ_η. The process starts from y_0 which is given by a p-dimensional Gaussian vector with mean μ_0 and variance-covariance matrix Σ_0.

Note that the three error components, namely $\varepsilon(s,t)$, $\omega(s,t)$ and η_t, are zero-mean and independent over time as well as mutually independent. Hence, the parameter set which identifies Equations (1–4) and is estimated using observed data is given by

$$\Psi = \left(\beta, \sigma_\varepsilon^2, \theta, \sigma_\omega^2, G, \Sigma_\eta, \mu_0\right). \tag{5}$$

2.2 Matrix Representation

In this section, the matrix notation to be used for the estimation and mapping procedures described in Sections 2.3 and 3, respectively, is introduced. Suppose there is a network of n stations and observations for T consecutive days. Denoting the network information at time t by the n-dimensional column vector $Z_t = (z(s_1,t), \ldots, z(s_n,t))'$ and the full data set by $Z = (Z_1, \ldots, Z_T)$; similarly y is used for the full latent information. Moreover, let X_t denote the corresponding $n \times d$ matrix of known regressors at time t and K be the $n \times p$ loading matrix.

Equations (1), (2) and (4) can be rewritten compactly using the two-stage hierarchical model:

$$Z_t = X_t\beta + Ky_t + e_t \tag{6}$$

$$y_t = Gy_{t-1} + \eta_t \tag{7}$$

which can be considered as a classical *state-space model* [24], where Equation (6) is the measurement equation and Equation (7) is the state equation.

If all the parameters are known, the unobserved temporal process y_t is estimated for each time point t using the *Kalman filter* and *Kalman smoother* techniques with initial conditions given by y_0. In the following, the Kalman smoother outputs are denoted by y_t^T, P_t^T and $P_{t,t-1}^T$ which are the mean, variance and lag-one covariance of the y_t conditional on the complete observation matrix Z, respectively, as defined in detail in [25, appendix A].

In Equation (6), the error $e_t = \omega_t + \varepsilon_t$ has a zero-mean Gaussian distribution with variance-covariance matrix $\Sigma_e = \sigma_\omega^2 \Gamma\left(\left\|s_i - s_j\right\|\right)_{i,j=1,\ldots,n}$, where Γ is the scaled spatial covariance function:

$$\Gamma(h) = \begin{cases} 1 + \frac{\sigma_\varepsilon^2}{\sigma_\omega^2} & h = 0 \\ C_\theta(h) & h > 0. \end{cases} \tag{8}$$

It is interesting to note that the measurement error variance σ_ε^2 can be interpreted in geostatistical terms as the so-called 'nugget effect' of the spatial process $e(s,t)$ for fixed t.

2.3 Estimation using the EM Algorithm

The maximum likelihood (ML) estimation of the unknown parameter set Ψ defined by Equation (5) is performed by optimizing the log-likelihood function which, as shown in [26], is given by

$$\log L\left(\Psi; Z\right) = -\frac{nT}{2}\log\left(2\pi\right) - \frac{1}{2}\sum_{t=1}^{T}\Big[\log\left|\Omega_t\right| + \left(Z_t - \mu_t\right)'\,\Omega_t^{-1}\left(Z_t - \mu_t\right)\Big] \tag{9}$$

where

$$\mu_t = \left(X_t\beta + Ky_t^{t-1}\right),$$

$$\Omega_t = \left(KP_t^{t-1}K' + \Sigma_e\right),$$

$y_1^0 = \mu_0$, $P_1^0 = \Sigma_0$ and the symbol $|.|$ is used for matrix determinant. Since direct maximization of log-likelihood Equation (9) is complex, the Expectation-Maximization (EM) algorithm is used [27, 28]. This method, which is based on the complete log-likelihood Equation (10), is particularly suitable for missing data problems, including the models defined by Equations (6) and (7), where the missing data component is given by the latent process y_t.

Moreover, the EM algorithm is useful for spatio-temporal separable models because the maximization step does not require numerical optimization for the model parameters, except those related to the spatial covariance. Hence, it avoids large Hessian matrix inversions and the related instability and non-positive definiteness which often arise in performing numerical maximization of the likelihood. Missing data are also handled in a natural way.

Apart from an additive constant, the complete log-likelihood is given by

$$\begin{aligned}\log L_c\left(\Psi; \bar{Z}\right) &\propto -\frac{T}{2}\log\left|\Sigma_e\right| \\ &- \frac{1}{2}\sum_{t=1}^{T}\left(Z_t - X_t\beta - Ky_t\right)'\,\Sigma_e^{-1}\left(Z_t - X_t\beta - Ky_t\right) \\ &- \frac{1}{2}\log\left|\Sigma_0\right| - \frac{1}{2}\left(y_0 - \mu_0\right)'\,\Sigma_0^{-1}\left(y_0 - \mu_0\right) \\ &- \frac{T}{2}\log\left|\Sigma_\eta\right| - \frac{1}{2}\sum_{t=1}^{T}\left(y_t - Gy_{t-1}\right)'\,\Sigma_\eta^{-1} \\ &\times\ \left(y_t - Gy_{t-1}\right)\end{aligned} \tag{10}$$

where $\bar{Z} = (y_0, \ldots, y_T, Z_1, \ldots, Z_T)$ is the complete dataset. At each iteration $k = 1, 2, \ldots$ the EM algorithm consists of an expectation step (E) and a maximization step (M) which are described extensively in the following sections. Given the current values of the parameters $\Psi^{(k)}$, the E-step computes the expected value of the complete log-likelihood function $\log L_c\left(\Psi; \bar{Z}\right)$ conditional on the observation matrix Z and $\Psi^{(k)}$, that is

$$Q\left(\Psi; \Psi^{(k)}\right) = E_{\Psi^{(k)}}\left[\log L_c\left(\Psi; \bar{Z}\right) \mid Z\right].$$

At the M-step, a value $\Psi^{(k+1)}$ is chosen so that $Q\left(\Psi^{(k+1)}; \Psi^{(k)}\right) \geq Q\left(\Psi^{(k)}; \Psi^{(k)}\right)$.

2.3.1 E-step

With reference to the complete log-likelihood Equation (10), it is easy to implement the E-step and to compute the function $Q\left(\Psi; \Psi^{(k)}\right)$ which is reported in the equation as follows:

$$\begin{aligned}-\,2Q\left(\Psi; \Psi^{(k)}\right) &= -\,2E_{\Psi^{(k)}}\left[\log L_c\left(\Psi; \bar{Z}\right) \mid Z\right] \\ &= \tilde{Q} + \log\left|\Sigma_0\right| + T\log\left|\Sigma_\eta\right| \\ &+ tr\left\{\Sigma_0^{-1}\left[\left(y_0^T - \mu_0\right)\left(y_0^T - \mu_0\right)' + P_0^T\right]\right\} \\ &+ tr\left\{\Sigma_\eta^{-1}\left[S_{11} - S_{10}G' - GS_{10}' + GS_{00}G'\right]\right\}\end{aligned} \tag{11}$$

where

$$\tilde{Q} = \tilde{Q}\left(\Psi; \Psi^{(k)}\right) = T\log\left|\Sigma_e\right| + tr\left[\Sigma_e^{-1}W\right] \tag{12}$$

and

$$W = \sum_{t=1}^{T}\left[\left(Z_t - X_t\beta - Ky_t^T\right)\left(Z_t - X_t\beta - Ky_t^T\right)'\right] + \sum_{t=1}^{T}KP_t^TK'. \tag{13}$$

Note also that

$$S_{00} = S_{00}^{(k)} = \frac{\sum_{t=1}^{T}\left(y_{t-1}^T y_{t-1}^{T\prime} + P_{t-1}^T\right)}{T},$$

$$S_{10} = S_{10}^{(k)} = \frac{\sum_{t=1}^{T}\left(\hat{y}_t\hat{y}_{t-1}' + P_{t,t-1}^T\right)}{T} \quad \text{and}$$

$$S_{11} = S_{11}^{(k)} = \frac{\sum_{t=1}^{T}\left(y_t^T y_t^{T\prime} + P_t^T\right)}{T},$$

with the Kalman smoother outputs y_t^T, P_t^T and $P_{t,t-1}^T$ computed using $\Psi^{(k)}$ as the 'true' value.

2.3.2 M-step

Using the so-called conditional maximization steps [28, chapter 5], the solution of $\frac{\partial Q}{\partial \Psi} = 0$ is approximated by partitioning $\Psi = \left\{\check{\Psi}, \tilde{\Psi}\right\}$. The first result is a closed form solution for the first component:

$$\check{\Psi} = \left(\beta, \sigma_\omega^2, G, \Sigma_\eta, \mu_0\right)$$

holding the second component fixed at its current value $\tilde{\Psi} = \left\{\theta, \sigma_\varepsilon^2\right\}$ and Σ_0 constant. In particular, the closed forms are given by

$$\beta^{(k+1)} = \left[\sum_{t=1}^{T}\left(X_t'\Sigma_e^{-1}X_t\right)\right]^{-1} \times \left\{\sum_{t=1}^{T}\left[X_t'\Sigma_e^{-1}\left(Z_t - Ky_t^T\right)\right]\right\} \tag{14}$$

$$\sigma_\omega^{2(k+1)} = \frac{\sigma_\omega^{2\,(k)}}{Tn} tr\left[\Sigma_e^{-1}W\right] \tag{15}$$

$$G^{(k+1)} = S_{10}S_{00}^{-1} \tag{16}$$

$$\Sigma_\eta^{(k+1)} = S_{11} - S_{10}S_{00}^{-1}S_{10}' \tag{17}$$

$$\mu_0^{(k+1)} = y_0^T \tag{18}$$

where $\Sigma_e = \Sigma_e^{(k)}$ and W is given by Equation (13) with $\beta = \beta^{(k+1)}$. Since there are no closed forms for the remaining parameters $\tilde{\Psi} = \left\{\theta, \sigma_\varepsilon^2\right\}$, the Newton Raphson (NR) algorithm is used for minimizing the quantity $\tilde{Q}$ given by Equation (12). The latter is considered as a function of $\tilde{\Psi}$ only, that is

$$\tilde{Q}\left(\tilde{\Psi}\right) = \tilde{Q}\left(\left\{\breve{\Psi}^{(k+1)}, \tilde{\Psi}\right\}; \Psi^{(k)}\right).$$

At the generic kth iteration of the EM algorithm, the updating formula for the ith iteration of the inner NR algorithm is given by

$$\tilde{\Psi}_{(i+1)} = \tilde{\Psi}_{(i)} - H^{-1}_{\tilde{\Psi}=\tilde{\Psi}_{(i)}} \times \Delta_{\tilde{\Psi}=\tilde{\Psi}_{(i)}} \tag{19}$$

where H and Δ are the Hessian matrix and the gradient vector of $\tilde{Q}\left(\tilde{\Psi}\right)$, respectively, evaluated in $\tilde{\Psi} = \tilde{\Psi}_{(i)}$. In [25, appendix B], the complete calculations required for H and Δ are reported together with the details for the exponential covariance function. Equation (19) is repeated until the NR algorithm converges. Hence the obtained roots, say $\tilde{\Psi}^{(k+1)}$, are used for the next outer EM iteration based on $\Psi^{(k+1)} = \left\{\breve{\Psi}^{(k+1)}, \tilde{\Psi}^{(k+1)}\right\}$.

The EM algorithm converges when the following two convergence criteria are jointly met:

$$\frac{\left\|\Psi^{(k+1)} - \Psi^{(k)}\right\|}{\left\|\Psi^{(k)}\right\|} < \pi$$

and

$$\frac{\left\|\log L\left(\Psi^{(k+1)}; Z\right) - \log L\left(\Psi^{(k)}; Z\right)\right\|}{\left\|\log L\left(\Psi^{(k)}; Z\right)\right\|} < \pi,$$

where π is a small positive *a priori* fixed quantity. The use of these relative criteria instead of some other absolute criteria makes it possible to correct for the different parameter scales.

3. Mapping

In this section, it is shown how to map a process which is continuous in space but is measured only in a limited number of spatial sites. In particular, given the hierarchical model of Section 2, the aim is to predict $u\left(s_0, t\right)$ given the observation vector Z_t, where $s_0 \notin \{s_1, \ldots, s_n\}$ is a new spatial location. Considering, for example, the problem of air pollution, this means that a continuous air quality map is obtained for each time point, given the observations coming from the monitoring network.

Supposing that all the model parameters are known, the spatial predictor is obtained by the joint $(n+1)$-dimensional Gaussian conditional distribution:

$$\left(\begin{array}{c} Z \\ u\left(s_0, t\right) \end{array} \mid y_t\right) \sim N_{n+1}\left[\left(\begin{array}{c}\mu_1 \\ \mu_2\end{array}\right), \left(\begin{array}{cc}\Sigma_e & \Omega \\ \Omega' & \sigma_\omega^2\end{array}\right)\right]$$

where $\mu_1 = X_t\beta + Ky_t$, $\mu_2 = X\left(s_0, t\right)\beta + K\left(s_0\right)y_t$ and $X\left(s_0, t\right)$ is the covariate vector observed at time t at site s_0. The quantity $K\left(s_0\right)$ is a p-dimensional loading vector which can be computed using the *loess* method as in [15], or can be fixed to one as in [22]. The covariance vector Ω is constant in time and contains elements for $i = 1, \ldots, n$ given by

$$Cov\left[z\left(s_i, t\right), u\left(s_0, t\right)\right] = \Gamma\left(\left|s_i - s_0\right|\right)$$

where Γ is the spatial covariance function of Equation (8).

From the standard theory of the multivariate Gaussian distribution, the conditional random variable $\left(u\left(s_0, t\right) \mid Z_t, y_t\right)$ has an univariate Gaussian distribution with mean $\hat{u}\left(s_0, t\right)$ and variance $\hat{\sigma}_K^2\left(s_0\right)$ given by

$$\hat{u}\left(s_0, t\right) = \mu_2 + \Omega'\Sigma_e^{-1}\left(Z_t - \mu_1\right) \tag{20}$$

$$\hat{\sigma}_K^2\left(s_0\right) = \sigma_\omega^2 - \Omega'\Sigma_e^{-1}\Omega. \tag{21}$$

It is interesting to note that conditional mean (20) and variance (21) correspond to the simple Kriging predictor and its prediction error variance, respectively [29].

Since the parameter set Ψ is not known and y_t is a latent process, the plug-in approach is used. This means that Ψ and y_t are substituted by the ML estimate $\hat{\Psi}$ and the Kalman smoother output y_t^T, respectively. However, this solution requires taking into account the estimate and the latent process uncertainty. This could be achieved substituting the Kriging variance of Equation (21) with a more general measure that considers all the uncertainty sources and that can be computed using the spatio-temporal bootstrap, as described in the next section.

4. Bootstrapping Space-time Data

The spatio-temporal bootstrap is used here for parameter uncertainty assessment, including confidence intervals computed without normality assumptions. In addition, with reference to a Kriging spatial interpolator, it is applied for computing map uncertainty and data roughness assessment.

In this section, a sampling scheme for bootstrapping data which are dependent in space and time is proposed. In the literature, only purely spatial or temporal bootstrap techniques have been discussed [e.g. 30, 31].

The resampling spatio-temporal strategy introduced here is very simple and is based on the estimated parametric model of Section 2. In particular, samples are drawn directly from the Gaussian distributions involved and are used in Equations (6) and (7), with Ψ replaced by its ML estimate $\hat{\Psi}$ for obtaining the bootstrap samples $Z_b^\star$, $b = 1, \ldots, B$. The procedure starts from a p-dimensional vector $y_0^\star$ simulated from $N\left[\hat{\mu}_0, \Sigma_0\right]$.

In this way, B bootstrap samples are simulated and, for each of them, the ML estimate $\hat{\Psi}_b^\star$ and the spatial prediction $\hat{u}_b^\star(s_0, t)$ are computed using the EM algorithm and the spatial prediction technique described in Sections 2.3 and 3, respectively. The bootstrap replications $\hat{\Psi}_1^\star, \ldots, \hat{\Psi}_B^\star$ and $\hat{u}_1^\star(s_0, t), \ldots, \hat{u}_B^\star(s_0, t)$ are then used for computing the standard error of each parameter and spatial prediction. Moreover, percentile confidence intervals and full empirical distributions can be easily calculated.

5. Conditional Sensitivity Analysis

The modern approach to statistical SA is based on an appropriate variance decomposition [e.g. 2, 5, 6]. If repeated code runs are possible, sensitivity analysis design makes it possible to define an appropriate input sampling plane, e.g. Latin hypercubes. This in turn implies orthogonal inputs and the variance decomposition is easier. If the code runs are expensive and/or the model requires observational data, repeated inputs are not allowed and one has to adapt the variance decomposition to the input structure at hand.

In our case, the hierarchical model of Section 2 can be used for simulation but the ECMT outputs are difficult to repeat since ECMT is not a cheap code.

Generally speaking, using the notation in [10], a linear model with three correlated input sets is considered, namely:

$$z = u + \varepsilon = \beta_1' x_1 + \beta_2' x_2 + \beta_3' x_3 + \varepsilon. \tag{22}$$

This model is similar to Equation (1) with known latent variables y and ω.

If the three input sets are independent, then

$$Var(z) = \sum_{j=1}^{3} \beta_j' V\left(x_j\right) \beta_j + V(\varepsilon)$$

and sensitivity of z to the input sets x_j, $j = 1, 2, 3$, is simply given by

$$S_j = \frac{\beta_j' V\left(x_j\right) \beta_j}{Var(z)}$$

where $Var(z)$ is the variance of scalar random variable z and $V(x)$ is the variance-covariance matrix of the stochastic vector x. From a more general point of view, this problem could also be attached by regression elements [32] or by additive elements of the likelihood function [33].

When x_3 is not present in Equation (22), giving $u = \beta_1' x_1 + \beta_2' x_2$, and the input sets x_1 and x_2 are not independent, [10] suggests starting from the general variance decomposition

$$Var(u) = Var(E(u|x_1)) + E(Var(u|x_1))$$

which also holds for correlated x's. This gives two sensitivity indexes for assessing the effect of, say, x_1. The first is the total effect of x_1:

$$S_1 = \frac{Var(E(u|x_1))}{Var(z)}$$

which also incorporates the effect of x_2 due to correlation between x_1 and x_2. The second index is the net effect of x_1 adjusted for x_2, namely

$$S_{1|2} = \frac{E(Var(u|x_2))}{Var(z)}.$$

If the input sets x_1 and x_2 are linearly related, i.e. $E(x_2|x_1) = b_{21}' x_1$, the above sensitivity indexes are simply given by

$$S_1 = \frac{b_1' V(x_1) b_1}{Var(z)}$$

and

$$S_{1|2} = \frac{\beta_1' V(x_1|x_2) \beta_1}{Var(z)}.$$

The coefficients involved in these definitions are the standard conditional Gaussian quantities or least square quantities, namely

$$\begin{aligned} b_{21} &= V(x_1)^{-1} Cov\left(x_1, x_2'\right), \\ b_1 &= \beta_1 + b_{21}\beta_2 \end{aligned}$$

and

$$V(x_1|x_2) = V(x_1) - Cov\left(x_1, x_2'\right) V(x_2)^{-1} Cov\left(x_2, x_1'\right),$$

the latter being the standard residual variance-covariance matrix.

The comparison between the two input sets therefore may be determined by comparing $\left(S_1, S_{1|2}\right)$ to $\left(S_2, S_{2|1}\right)$.

In the case of Equation (22) with three input sets, the sensitivity analysis of the first two sets (as above) is influenced by x_3. In other words, if x_3 is correlated with the other inputs, then spurious sensitivity conclusions may occur.

To avoid this, the conditional SA is proposed by applying the above SA indexes for correlated inputs to adjusted variables $x_{1|3} = x_1 - E(x_1|x_3)$, $x_{2|3} = x_2 - E(x_2|x_3)$ and $z_3 = z - E(z|x_3)$, which are easily estimated under linear assumptions using least square residuals and are uncorrelated with x_3.

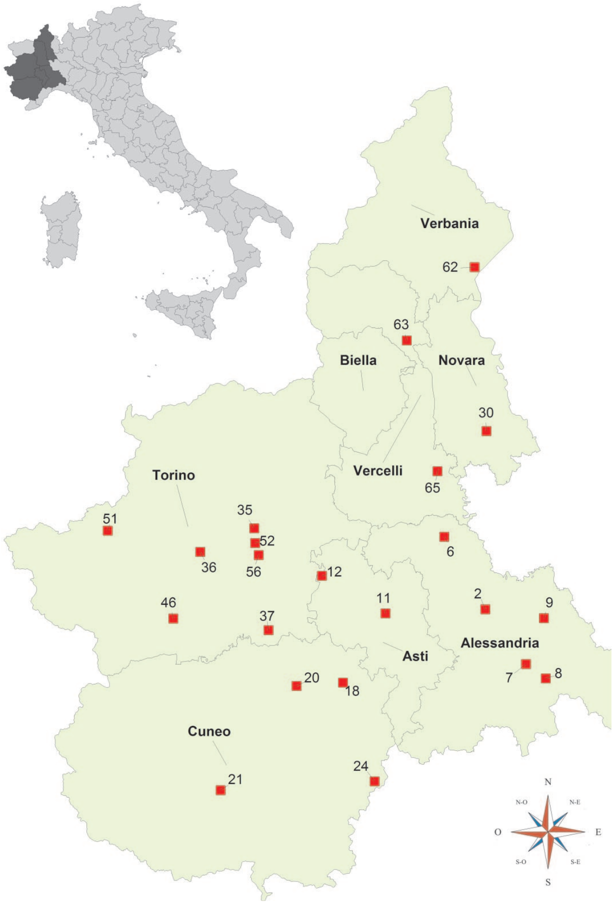

Figure 1. Piemonte region localization and monitoring network (stations are identified by the ID code reported in Table 1)

Table 1. PM_{10} mean and standard error (SE) by station and season (year 2004)

		Winter		Summer	
Station	Station ID	Mean	SE	Mean	SE
Alba	18	50.27	25.30	25.67	10.85
Alessandria	2	62.94	32.30	35.78	13.68
Asti	11	63.51	30.86	26.73	12.20
Borgaro	35	59.04	30.34	31.11	14.65
Borgosesia	63	43.14	24.22	26.43	17.15
Bra	20	68.95	32.46	32.06	14.43
Buttigliera Alta	36	53.66	28.45	31.45	15.68
Buttigliera d'Asti	12	53.17	26.77	28.15	13.31
Carmagnola	37	57.86	33.26	34.10	13.46
Casale Monferrato	6	50.68	26.30	26.76	12.00
Cuneo	21	37.65	25.62	27.98	13.91
Novara	30	66.73	31.73	37.39	16.79
Novi Ligure	7	57.38	29.89	32.51	12.21
Pinerolo	46	46.88	27.97	28.55	15.47
Saliceto	24	40.00	20.59	19.54	8.20
Serravalle Scrivia	8	39.25	23.31	24.46	9.65
Susa	51	34.20	21.96	25.86	12.99
Torino Grassi	52	89.73	39.35	51.85	22.31
Tortona	9	62.91	30.96	34.42	12.34
Torino Consolata	56	78.13	34.18	38.12	16.48
Vercelli	65	70.85	32.32	36.29	15.34
Verbania	62	36.10	22.01	21.51	11.91

6. Particulate Matters Case Study

The general approach of the previous sections is now used for mapping risky particulate matters concentrations in Piemonte, Italy and understanding relationships with data from the EMCT model chain.

6.1 Data Description

The Piemonte region is located in northwest Italy (see Figure 1) which covers an area of 25.399 km^2 of which more than 40% is highlands. Piemonte is surrounded by mountains (the Alps to the north and west and the Apennines to the south) while to the east there is the river Po Valley. This area, in particular, is densely populated and is characterized by metropolitan and industrialized zones as well as roads with heavy traffic.

Regarding air quality, two aspects need to be considered: emissions and atmospheric conditions. Firstly, as expected, industries and road transport (the main contributing sectors to primary particulate matters emissions) are mainly located on the plain. Moreover, the mountain chains that surround the region shelter the area from mass flow circulation. This leads to stable atmospheric conditions (especially in winter) which reduce pollutant dispersion. All the plain zone is therefore characterized by critical particulate concentration levels which are more severe in the urban centers. This effect is reinforced by increased emissions due to building heating.

The rest of this section discusses the various data entering our mapping model. These are different in nature; monitoring data are daily data collected on an irregular grid. Land information is constant over time. Data from the EMCT simulation model concerning emissions, meteorology and particulate matters concentrations are on a regular grid.

6.1.1 Monitoring Network

The regional environmental agency (ARPA Piemonte) is responsible for the entire air quality system and manages the regional monitoring network and the data supply. For

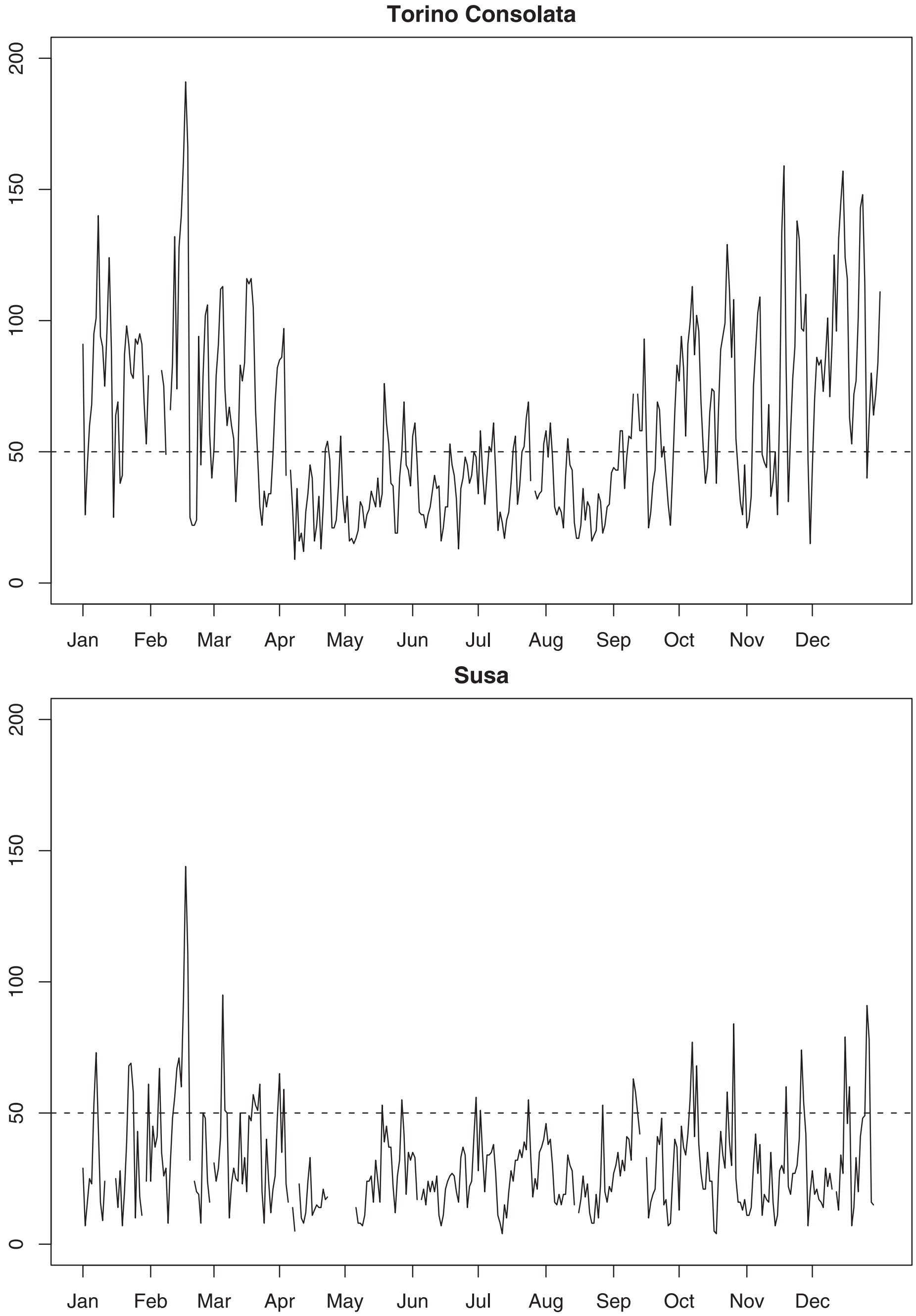

Figure 2. PM_{10} time series for Torino Consolata and Susa stations (year 2004)

the year 2004, the daily particulate matters concentrations (in $\mu g\,m^{-3}$) with an aerodynamic diameter of less than 10 μm (PM_{10}) are examined, measured by 22 Low Volume Gravimetric (LVG) stations. As can be seen in Figure 1, even if the stations are mainly located in the most populated towns the network spatial coverage is good and stations can also be found in rural plain areas and urbanized alpine valleys.

In order to provide a brief description of the PM_{10} data, Figure 2 shows two time series plots from an urban station (Torino Consolata) located in the plain and a suburban station (Susa) from the Alpin area. The former shows high PM_{10} concentration levels which exceed the limit value of 50 $\mu g\,m^{-3}$ for almost all the winter season (following the European directive n. 1999/30, this standard should not be exceeded more than 35 days a year). On the contrary, the

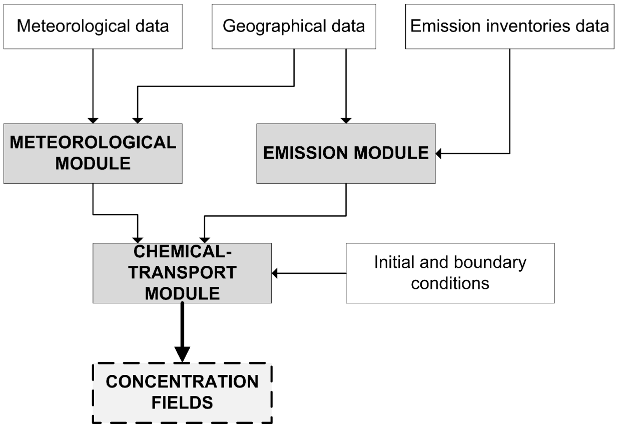

Figure 3. Flowchart of EMCT model chain

second station shows a less severe situation, even if the limits are occasionally exceeded.

Table 1 highlights the well-known seasonality of PM pollution, showing higher levels in winter in accordance with different atmospheric and emission conditions.

6.1.2 Simulation Model Data

ARPA Piemonte implements a nested system of deterministic computer-based models for air quality assessment. The final output is given by hourly concentration fields of some primary and secondary pollutants defined on a regular 4 km by 4 km grid. In particular, the EMCT modeling system, which is schematically shown in Figure 3 and extensively described in [34, 35], is composed of three main modules:

1. *Meteorological module* based on *Minerve* and *Surfpro* models (developed by Aria Technologies and Arianet, respectively): both models use meteorological data and geographical information, such as landuse (given by the *Corine Land Cover* project).

2. *Emission module* based on *Emission Manager* model (developed by Arianet) which uses data coming from the regional and national Emission Inventories. As these data are defined yearly on a city scale, the emission model disaggregates them spatially and temporally in order to obtain emission rates for different pollutants and for the whole grid.

3. *Chemical-transport module* based on the chemical-transport model *FARM* (Flexible Air Quality Regional Model) by Arianet. FARM is a three-dimensional Eulerian transport model that takes into account transport, chemical transformation and the deposition and dispersion of atmospheric pollutants.

Table 2. Seasonal variance decomposition for log-scale data

	PPM	SimPM	PM_{10}
Winter			
Total variance	0.53	0.37	0.42
Spatial variability (%)	90.59	59.16	18.00
Temporal variability (%)	9.41	40.84	82.00
Summer			
Total variance	0.78	0.36	0.29
Spatial variability (%)	86.15	38.83	17.36
Temporal variability (%)	13.85	61.17	82.64

Although the quality and the reliability of such computer-based data is an important point discussed by [35], this issue is not taken into consideration here. The covariates to be used in the trend $X_t\beta$ of Equation (6) are chosen from a set of gridded variables that are the intermediate or final output of the EMCT deterministic model chain. In particular, the set of daily variables under consideration includes meteorological fields, particulate primary emissions (PPM in $\mathrm{g\,s^{-1}\,km^{-2}}$) and concentrations (SimPM in $\mu\mathrm{g\,m^{-3}}$) for the year 2004.

6.2 Preliminary Data Analysis

A preliminary description of the relationships among network data and EMCT outputs is considered here, taking into account the seasonal component (which may be the source of spurious correlations) and the spatial component required for interpolation.

6.2.1 Spatial and Seasonal Variability

As a preliminary analysis, using the logarithmic transformation discussed in Section 6.3, the raw data variability in winter and summer is compared. Each seasonal variance is decomposed into a spatial and a temporal component. In each column of Table 2, the famous decomposition of the total variance to within-station variance and between-stations variance is used, namely

$$Var\,(z) = E\,(Var\,(z|s)) + Var\,(E\,(z|s))$$

where z is the generic seasonal response, i.e. emissions (PPM), simulated concentrations (SimPM) or observed concentrations (PM_{10}) and $s = s_1, \ldots, s_n$ spans the monitoring network. In particular, the between-station variability is connected with spatial heterogeneity, while the within-station variability is related to the non-seasonal temporal dynamics on a daily scale.

Although the variability of emissions is higher in summer, the opposite is true for observed concentrations. For simulated concentrations, however, there is no relevant seasonal difference. Consistent with the steady-state pattern of the emission inventory, the spatial variability of

Table 3. Local correlations between EMCT outputs and observed concentrations (PM_{10}) for log-scale data

	Mixing Height-PM_{10}			$PPM-PM_{10}$			$SimPM-PM_{10}$		
Station	Total	Winter	Summer	Total	Winter	Summer	Total	Winter	Summer
Alba	–0.39	0.04	0.04	0.52	0.22	0.12	0.54	0.29	0.54
Alessandria	–0.39	–0.08	–0.06	0.43	0.20	0.27	0.57	0.60	0.42
Asti	–0.55	–0.14	–0.06	0.60	0.18	0.32	0.57	0.48	0.46
Borgaro	–0.40	–0.20	–0.02	0.46	0.30	0.10	0.56	0.53	0.34
Borgosesia	–0.29	–0.16	0.09	0.40	0.02	0.26	0.49	0.36	0.37
Bra	–0.42	0.02	0.13	0.48	0.17	0.27	0.56	0.37	0.49
Buttigliera Alta	–0.36	–0.23	0.05	0.39	0.10	0.25	0.55	0.44	0.47
Buttigliera d’Asti	–0.37	0.02	0.04	0.44	0.02	0.02	0.57	0.42	0.47
Carmagnola	–0.21	0.12	0.10	0.35	0.09	0.27	0.22	0.07	0.33
Casale Monferrato	–0.37	–0.11	0.09	0.42	0.16	–0.07	0.42	0.40	0.21
Cuneo	–0.04	0.20	0.09	0.06	–0.21	–0.06	0.31	0.06	0.51
Novara	–0.43	–0.18	–0.05	0.49	0.09	0.26	0.57	0.38	0.54
Novi Ligure	–0.41	–0.17	–0.02	0.46	0.27	0.11	0.51	0.49	0.42
Pinerolo	–0.25	0.03	–0.02	0.27	0.06	–0.05	0.37	0.08	0.41
Saliceto	–0.48	–0.33	0.01	0.49	0.15	–0.05	0.29	0.20	0.51
Serravalle Scrivia	–0.27	–0.02	0.12	0.24	–0.09	–0.19	0.44	0.36	0.34
Susa	–0.22	–0.24	–0.04	0.12	–0.05	–0.11	0.36	0.24	0.42
Torino Grassi	–0.41	–0.27	0.0001	0.57	0.30	0.42	0.66	0.59	0.47
Tortona	–0.48	–0.15	–0.15	0.45	0.28	0.28	0.55	0.57	0.43
Torino Consolata	–0.53	–0.23	–0.07	0.55	0.19	0.02	0.68	0.46	0.44
Vercelli	–0.51	–0.18	–0.11	0.59	0.22	0.37	0.59	0.42	0.53
Verbania	–0.46	–0.33	–0.21	0.42	0.26	0.14	0.56	0.32	0.65
Mean	–0.37	–0.12	0.002	0.42	0.13	0.13	0.50	0.37	0.44
R^2	0.14	0.01	0.0003	0.18	0.02	0.02	0.25	0.14	0.20

emissions is much higher than the temporal variability in both seasons. The opposite holds for the observed concentrations, as stations are primarily located in human risk areas and meteorological effects increase the temporal variability. These two aspects are more balanced for simulated concentrations, as both the meteorological effects and remote low polluted areas are present in the gridded data. This analysis highlights the need for a spatio-temporal model which can cope with such complex heterogeneity.

6.2.2 Auxiliary Variables

The basic idea of regression-based mapping is to use some covariates as spatial predictors in the interpolation process. Table 3 shows that although the EMCT outputs are poorly locally correlated with the observed concentrations in some cases, in other cases the local correlation is relevant. In this paper, a local property is intended to hold for a single station, hence a local correlation is a Pearson correlation coefficient computed with data from the same station. A similar interpretation holds for local average, etc.) Supposing that the local concentration averages are known and, using standard regression reasoning, their interpolating capability in terms of explained variance is rather small as shown by the last line of Table 3. The challenge for this project is to obtain better results.

6.3 Model Specification

The measured concentrations described in Section 6.1.1 are now fitted to the model described in Section 2. In order to reduce heteroskedasticity and data long tails, the logarithmic transformation for the three particulate variables (emissions, simulated and observed concentrations) is used.

The covariates are chosen by a preliminary regression analysis, using Akaike’s information criterion (AIC) and parameter significance. The results show that the variables to be included in the model within the considered set of EMCT outputs of Section 6.1.2 are:

1. daily *particulate primary emissions* (PPM) in $g\,s^{-1}\,km^{-2}$, an intermediate output of EMCT;
2. daily *simulated concentrations* (SimPM) in $\mu g\,m^{-3}$, the final output of EMCT;
3. daily *mixing height* (the height to which the lower atmosphere undergoes mechanical or turbulent mix-

Table 4. Seasonal parameter estimates, standard errors (SE) and 95% bootstrap confidence interval bounds

	Estimate	SE	95% CI	bounds
Winter				
Intercept	3.237	0.046	3.147	3.325
PPM	0.040	0.012	0.017	0.062
SimPM	0.239	0.019	0.203	0.275
Mixing height	–0.133	0.108	–0.364	0.072
Altitude	–0.822	0.060	–0.948	–0.701
Summer				
Intercept	2.417	0.071	2.185	2.649
PPM	0.093	0.008	0.080	0.109
SimPM	0.233	0.016	0.204	0.268
Mixing height	0.191	0.076	0.047	0.335
Altitude	–0.252	0.052	–0.348	–0.146

Table 5. Non-seasonal parameter estimates, standard errors (SE) and 95% bootstrap confidence interval bounds

	Estimate	SE	95% CI	bounds
σ_ω^2	0.078	0.001	0.075	0.080
θ	0.023	0.002	0.019	0.026
σ_ε^2	0.078	0.002	0.074	0.082
G	0.747	0.038	0.651	0.806
Σ_η	0.054	0.004	0.045	0.062
μ_0	–0.434	1.074	–2.544	1.551

ing producing a nearly homogeneous air mass, related to the height where relatively vigorous mixing and pollutant dispersion occurs) in km;

4. *altitude* in km, which is not a simulated variable.

To cope with the strong seasonality of air quality data, a seasonal model with different β coefficients for winter and summer is used. Moreover, according to an unreported performance analysis which is similar to the cross-validation described in Section 6.6, one dimensional underlying process y_t with $p = 1$ was chosen. Finally, the spatial covariance function is the exponential term given by Equation (3).

6.4 Model Estimation and Description

Tables 4 and 5 report the estimates computed using the EM algorithm described in Section 2.3. The estimates, which are also used as a basis of $B = 500$ bootstrap replications, are given together with the corresponding bootstrap standard errors and the bounds of the 95% confidence intervals.

Examining the size of confidence intervals, apart from the initial value μ_0 which is a nuisance parameter for the model with no substantial interest, it can be observed that all but one of the parameters are characterized by a high level of accuracy.

The coefficients for altitude are both negative, consistently with less anthropized highlands and *ceteris paribus*, this effect is stronger in winter.

Generally speaking, mixing height is expected to be negatively correlated with pollutant concentrations. For example, using the Piemonte data, the correlation for the whole year between mixing height and $\log(PM_{10})$ is –0.32. Things change after deseasonalizing as the winter correlation is reduced to –0.16 while the summer correlation is +0.06 which is non-significantly positive. A similar result holds for our model, where the conditioning on mixing height induced by the other variables is stronger than simply splitting winter and summer as above. The result is therefore further modified so that in winter the coefficient is non-significantly negative and in summer is moderately positive.

The values of the intercept give information about the average regional pollution level. In particular, returning to the original scale, the hypothetical difference between winter and summer (with all the other variables at zero) would be 14 $\mu\,\mathrm{g}\,\mathrm{m}^{-3}$.

As expected, the coefficient for emissions (PPM) is significantly positive and doubles in summer. The concentrations are more sensitive to variations of emissions in summer than in winter, when pollution is more persistent.

Similarly to the emissions, the coefficient for simulated concentrations (SimPM) is significantly positive but, contrary to the emissions, it is more stable over the year with almost the same value in winter and summer. This is consistent with the point that both SimPM and the network PM_{10} measure the same quantity but the spatial resolution is different. A deeper comparison of the roles of emissions and simulated concentrations is given in Section 6.6.

Considering the non-seasonal part of the estimated model, note that the variances σ_ε^2 and σ_ω^2 are quite close. Moreover, from the spatial correlation parameter θ we note that at 50 km the spatial correlation is about 0.3 and at 90 km is about 0.1. Finally, the temporal coefficient G, being less than one, is in the stationarity range and its positive value confirms the well-known temporal persistence of particulate matters even after adjusting for all the covariates. In this sense, the spatial and temporal persistence coefficients, θ and G, complement and clarify the preliminary analysis of Section 6.2.1.

6.5 Mapping

In this section, the mapping of the PM_{10} field measured by the monitoring network is taken into consideration. The problem of network design is not examined at this point. However, it is worth mentioning that the stations are often localized for assessing risk and, in this sense, risky concentrations are being mapped here.

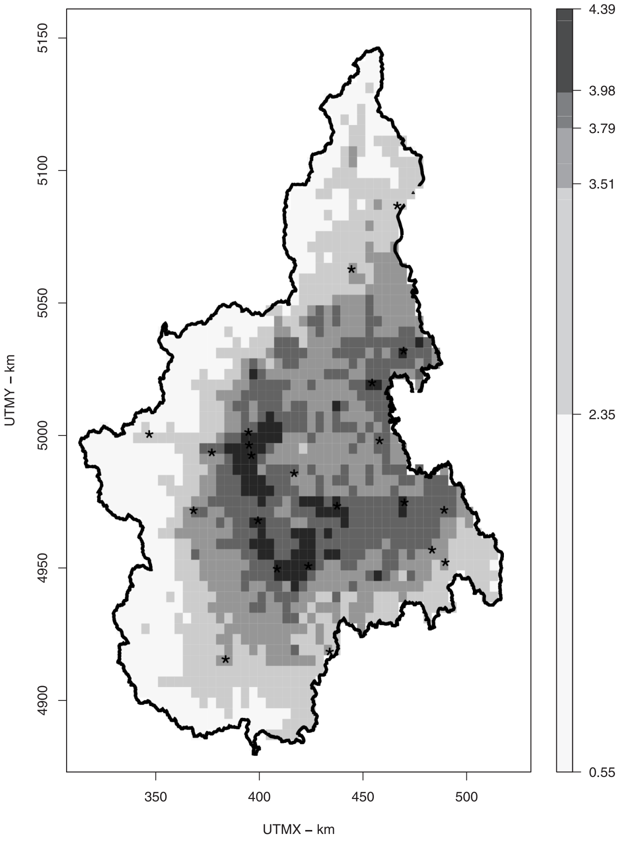

Figure 4. Concentration map for 30 January 2004 for log-scale data

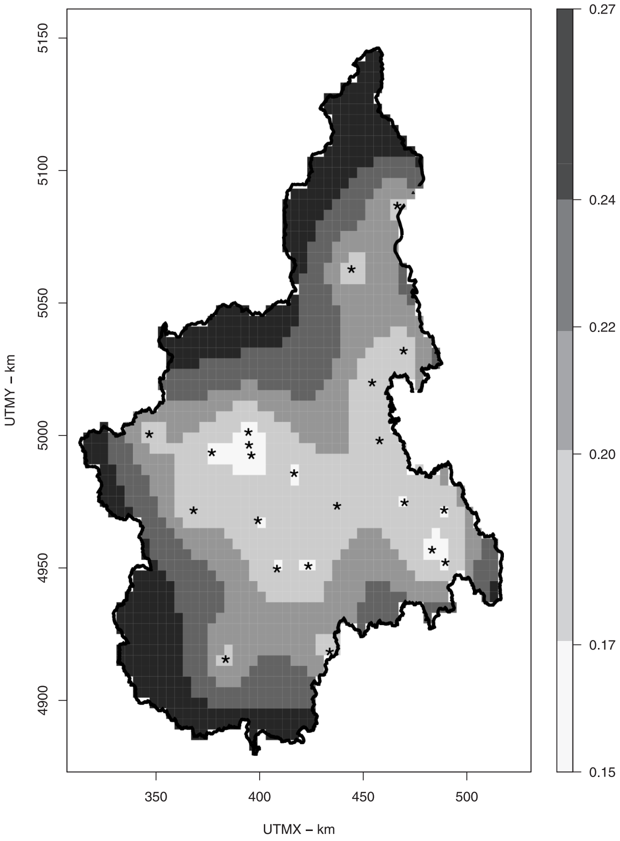

Figure 5. Kriging standard error map for 30 January 2004 for log-scale data

Figure 4 is the regional log-concentration map for 30 January 2004 obtained using Equation (20) and the gridded data $X(s_0, t)$ of Section 6.1.2. The concentration map shows that the more heavily polluted areas are located in the plain around Torino and near the urban centers of the southern part of the region, while the lowest concentrations are in the boundary mountain areas. In the central plain part of the region, a homogeneous concentration level can be seen. This is in accordance with the value of the spatial parameter θ discussed previously.

With reference to the uncertainty maps for the same day, the Kriging and the bootstrap standard error are plotted in Figures 5 and 6, respectively. The former, which is computed using Equation (21), is the pure prediction error which increases when the distance from the monitoring stations increases. This is in accordance with the Kriging theory stating that where there are less data there is more uncertainty [29]. The latter, which is based on the bootstrap replications $\hat{u}(s_0, t)$, is higher near the network sites where the bootstrap variability is higher. It decreases at greater distances from the network stations. Moreover, it can be seen that, as expected, the bootstrap standard error is higher than the Kriging error because it takes into account the parameter and latent variables uncertainty.

6.6 Model Discussion and Sensitivity Analysis

In this section, the estimated model is discussed with particular reference to the role of simulated emissions and concentrations. Firstly in Section 6.6.1, conditionally on the observed covariate field X, this is done in terms of spatial interpolation capability of the risk field measured by the network and by assessing the usefulness of our model with respect to universal Kriging. Then in Section 6.6.2, the EMCT outputs are compared in terms of sensitivity analysis conditional to the model latent components. The model itself is discussed in terms of the uncertainty sources given by space, time and estimation.

6.6.1 Cross Validation

To evaluate the spatial capability of models with different covariate fields, a leave-one-out cross-validation analysis

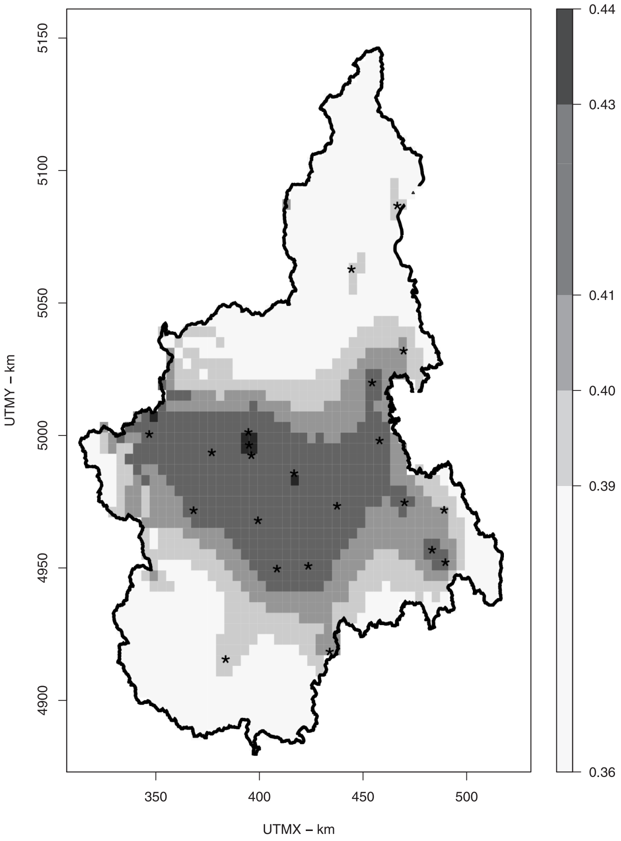

Figure 6. Bootstrap standard error map for 30 January 2004 for log-scale data

is used following the approach proposed in [15]. The procedure, which removes one station at a time and predicts PM_{10} concentrations in the removed site, is structured as follows: (a) removing the jth station ($j = 1, \ldots, 22$) from the observed data matrix Z, the reduced data matrix Z_{-j} is obtained; and (b) using Z_{-j} the ML estimate $\hat{\Psi}_{-j}$ and the spatial prediction $\hat{u}_{-j}\left(s_j, t\right)$ in site s_j for all the time points $t = 1, \ldots, T$ are calculated.

The performance analysis is then based on the network average mean absolute error (MAE), mean square error (MSE) and root mean square error (RMSE) of the daily bias which is defined as $e\left(s_j, t\right) = z(s_j, t) - \hat{u}_{-j}\left(s_j, t\right)$.

Model comparison starts from the naive interpolation based on a day-by-day universal Kriging [29], which can be related to a model given by $Z_t = X_t\beta_t + e_t$ and corresponds to performing T independent spatial Krigings. Although this modeling approach has vague theoretical properties, it is the easiest way for spatial interpolation of time dependent data using all the covariates of Section 6.3.

The cross-validation results summarized in Table 6 show that the hierarchical model is preferable to the naive universal Kriging model, as the model interpretation is clearer and MSE is reduced by 20%.

Table 6. Performance results based on cross-validation and the AIC index for log-scale data

	MAE	MSE	RMSE	AIC
Universal kriging	0.333	0.189	0.435	–
Covariate field with:				
PPM	0.290	0.152	0.390	7636.7
SimPM	0.289	0.151	0.389	7292.8
PPM and SimPM	0.287	0.150	0.388	7162.8

6.6.2 Input SA

The assessment of the role of simulated emissions and concentrations is more complex. On the one hand, as assessed by the AIC column of Table 6, simulated concentrations are better than emissions alone. However, from the last two rows of Table 6, it can be seen that emissions give additional information with respect to simulated concentrations (at least from the likelihood point of view).

On the other hand, with reference to the mapping capability for the monitoring network data, the EMCT outputs under consideration seem to have essentially the same role because the differences in all the three cross-validation statistics of Table 6 are very small. This can be partially explained by the fact that the measurement error of our model given by σ_ε^2 is not small.

In order to further deepen the analysis, the conditional SA argument of Section 5 is used. Here, the conditioning set x_3 is given by the mixing height, the unidimensional latent temporal component y_t^T and the estimated spatial small-scale component, namely $\hat{\omega}(s, t)$.

In particular, a local and a global approach are used for computing a network average of station-by-station SA and a global all-stations SA, respectively, with the resulting seasonal sensitivity indexes given in Table 7. Considering the network mean SA, it can be observed that simulated concentrations are more important as they explain a larger quota of variability. On the other side, EMCT emissions have a limited role and, especially in winter, could be neglected. This is consistent with the estimated β of Table 4, where the coefficient for PPM is smaller in winter than in summer. Considering the global SA, which also takes into account the between-station spatial variability, the role of emissions remains the same while the net effect of simulated concentrations is reduced in both seasons. This can be related to the results reported in Table 2 and to the limited ability of concentrations to explain the spatial variability.

These conclusions are not affected by the approximations arising from the estimation procedure as a large number of degrees of freedom is available. The role of

Table 7. Seasonal conditional SA results: subscript 1 represents emissions PPM while subscript 2 represents simulated concentrations SimPM

	S_1	$S_{1\|2}$	S_2	$S_{2\|1}$
Winter				
Network mean SA	0.028	0.003	0.306	0.281
Global SA	0.065	0.003	0.121	0.059
Summer				
Network mean SA	0.102	0.035	0.390	0.323
Global SA	0.125	0.035	0.178	0.088

model estimation uncertainty is further considered in the following section.

6.6.3 SA of Model Components

With reference to the mapping procedure of Section 6.5, the aim here is to assess the smoothing uncertainty related to y_t, the spatial uncertainty and the estimation uncertainty related to $\hat{\beta}$. To do this, an approximate total variance decomposition for $\tau(s,t)^2 = E((z(s,t) - \hat{u}(s,t))^2|X)$ is considered, which is conditional on the covariate field X and is given by

$$\tau^2(s,t) \cong X(s,t)' V\left(\hat{\beta}\right) X(s,t) + V\left(y_t^T\right) + \hat{\sigma}_K^2(s) + \hat{\sigma}_\varepsilon^2 \qquad (23)$$

where, using $V()$ for the variance-covariance matrix operator, the first summand is the model estimation uncertainty, the second is the time smoothing uncertainty (which does not depend on mapping pixel $s \in D$) and the third is the pure Kriging effect given by Equation (21).

Note that this model-based variance decomposition is conditional on Z and X and is mainly aimed at data roughness understanding and internal model validation. Moreover, note that a source of approximation in Equation (23) arises from neglecting the estimation uncertainty on the second and third terms of the right-hand side of Equation (20). The neglected estimation uncertainty is usually dominated by $V(\hat{\beta})$ which is small in our case study. A second source of approximation in Equation (23) arises from considering the three terms of the right-hand side of Equation (20) as orthogonal ones. Since the three components in Equation (2) are uncorrelated, here the orthogonality assumption is weakened again only by estimation. From the results reported in Table 8, it can be observed that the residual roughness and the spatial prediction error are the main components.

6.7 Implementation and Distributed Computing

Table 8. Variance decomposition for 30 January 2004

Source	Piemonte average
$X(s_0,t)' V\left(\hat{\beta}\right) X(s_0,t)$	0.003
$V(y_t\|Z)$	0.015
$\sigma_K^2(s_0)$	0.049
σ_ε^2	0.078
$\tau^2(s_0,t)$	0.147

All the code for bootstrap simulation, estimation, mapping and performance analysis is written in `R` software [36]. The computer-intensive bootstrap procedure for this model, being embarrassingly parallel with coarse grains, is implemented on a Pentium-based computer cluster with Linux environment. The parallel computing procedure is entirely handled by the `R` packages `RMPI` and `SNOW`. The former is an interface to MPI (Message-Passing Interface) which is a standardized and portable message-passing system which defines the cluster and the coordination of the node work. The latter provides a high-level interface for delivering the job through the cluster.

7. Conclusions

A unified modeling approach for handling various aspects of spatio-temporal environmental data has been presented. The first basic result is on mapping precision, which is related to interpolation properties. The cross-validation analysis shows that using a model which is not only able to use the observed covariates but also to cover for unobserved spatial and temporal components reduces the 'spatialforecasting' errors.

Although mapping is of great importance in practice, our approach returns an estimated model which can be easily interpreted and gives insight into the problem being studied. Moreover, the associated sensitivity analysis is useful for further understanding. In particular, the case study on air quality considers the role of two EMCT model chain outputs for 'predicting' the concentrations measured by the monitoring network, namely primary particulate emissions (an intermediate EMCT output) and particulate concentrations (a final output).

From the preliminary and the cross-validation analysis, correlations between the simulated and observed concentrations are not very high and the mapping precision is almost the same for both outputs. Despite this, it is apparent that (as expected) the information content of simulated emissions not taken into account by the second part of EMCT is very low or negligible. This is especially true in winter, as shown by the conditional sensitivity analysis technique which extends existing SA techniques for correlated inputs. This can be considered as a validating issue for EMCT model chain.

Finally, the model estimation procedure based on EM is effective in obtaining reliable estimates even under highly repeated model simulations by the bootstrap method.

8. Acknowledgements

The work is partially supported by PRIN 2006 project n.2006131039 *Statistical analysis of spatial and temporal dynamics and health impact of particulate matters* and Regione Piemonte project CIPE 2004 *Statistical methods and spatio-temporal models for atmospheric pollution monitoring*.

9. References

[1] Fassò, A. 2007. Statistical sensitivity analysis and water quality. *In* L. J. Wymer, (Ed.) *Statistical Framework for Recreational Water Quality Criteria and Monitoring*, chapter 11, pp. 211–230. New York: Wiley.

[2] Saltelli, A., K. Chan, and M. Scott. 2000. *Sensitivity Analysis*. New York: Wiley.

[3] Fassò, A. and P. Perri. 2002. Sensitivity analysis. *In* A. El-Shaarawi and W. Piegorsch, (Eds.) *Encyclopedia of Environmetrics*, volume 4, pp. 1968–1982. New York: Wiley.

[4] O'Hagan, A. 2006. Bayesian analysis of computer code outputs: A tutorial. *Reliability Engineering and System Safety* 91(10–11), 1290–1300.

[5] Fassò, A., E. Esposito, E. Porcu, A. Reverberi, and F. Vegli. 2003. Statistical sensitivity analysis of packed column reactors for contaminated wastewater. *Environmetrics* 14(8), 743–759.

[6] Saltelli, A., S. Tarantola, F. Campolongo, and M. Ratto. 2004. *Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models*. New York: Wiley.

[7] Wikle, C. K. and N. Cressie. 1999. A dimension-reduced approach to space-time Kalman filtering. *Biometrika* 86(4), 812–829.

[8] Berliner, M. 2003. Physical-statistical modeling in geophysics. *Journal of geophysical research* 108(D24), 3.1–3.10.

[9] Fassò, A. 2006. Sensitivity analysis for environmental models and monitoring networks. *In Voinov, A., Jakeman, A., Rizzoli, A. (eds). Proceedings of the iEMSs Third Biennial Meeting: Summit on Environmental Modelling and Software. International Environmental Modelling and Software Society, Burlington, USA, July 2006*. CDROM. Internet: www.iemss.org/iemss2006/sessions/all.html.

[10] Fassò, A., M. Cameletti, and P. Bertaccini. 2007. Uncertainty decompositions in environmental modelling and mapping. *In Proceedings of The Summer Computer Simulation Conference, San Diego (CA-USA), 15–18 July 2007.*

[11] Clark, J. and A. Gelfand. 2006. A future for models and data in environmental science. *Trends in Ecology and Evolution* 21(7), 375–380.

[12] Wikle, C. K. 2003. Hierarchical models in environmental science. *International Statistical Review* 71(2), 181–199.

[13] Wikle, C. K., L. Berliner, and N. Cressie. 1998. Hierarchical bayesian space-time models. *Journal of Environmental and Ecological Statistics* 5, 117–154.

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

[15] Fassò, A., M. Cameletti, and O. Nicolis. 2007. Air quality monitoring using heterogeneous networks. *Environmetrics* 18(3), 245–264.

[16] Mardia, K., C. Goodall, E. Redfern, and F. Alonso. 1998. The Kriged Kalman filter. *Test* 7, 217–285.

[17] Amisigo, B. A. and N. C. Van De Giesen. 2005. Using a spatio-temporal dynamic state-space model with the EM algorithm to patch gaps in daily riverflow series. *Hydrology and Earth System Sciences* 9, 209–224.

[18] Fassò, A. and I. Negri. 2002. Nonlinear statistical modelling of high frequency ground ozone data. *Environmetrics* 13(3), 225–241.

[19] Shumway, R. and D. Stoffer. 1982. An approach to time series smoothing and forecasting using the EM algorithm. *Journal of Time Series Analysis* 3, 253–264.

[20] Wu, L., J. Pai, and J. Hosking. 1996. An algorithm for estimating parameters of state-space models. *Statistics and Probability Letters* 28(2), 99–106.

[21] Xu, K. and C. K. Wikle. 2007. Estimation of parameterized spatio-temporal dynamic models. *Journal of Statistical Inference and Planning* 137(2), 567–588.

[22] Cameletti, M. 2007. *Modelli spazio-temporali per dati ambientali*. Ph.D. thesis, University of Milano Bicocca.

[23] Banerjee, S., B. Carlin, and A. Gelfand. 2004. *Hierarchical Modeling and Analysis for Spatial Data*. Monographs on Statistics and Applied Probability. New York: Chapman and Hall.

[24] Durbin, J. and S. Koopman. 2001. *Time Series Analysis by State Space Methods*. New York: Oxford University Press.

[25] Fassò, A., Cameletti M. 2009. The EM algorithm in a distributed computing environment for modelling environmental space-time data. *Environmental Modelling & Software*. Accepted for publication. DOI: 10.1016/j.envsoft.2009.02.009.

[26] De Jong, P. 1988. The likelihood for a state space model. *Biometrika* 75, 165–169.

[27] Little, R. and D. Rubin. 2002. *Statistical Analysis with Missing Data*. New York: Wiley.

[28] McLachlan, G. J. and T. Krishnan. 1997. *The EM Algorithm and Extensions*. New York: Wiley.

[29] Cressie, N. 1993. *Statistics for Spatial Data*. New York: Wiley.

[30] Solow, A. 1985. Bootstraping correlated data. *Mathematical Geology* 17(7), 769–775.

[31] Buhlmann, P. 2002. Bootstraps for time-series. *Statistical Science* 17(1), 52–72.

[32] Newton, R. and D. Spurrell. 1967. A development of multiple regression for the analysis of routine data. *Applied Statistics* 16(1), 51–64.

[33] Whittaker, J. 1984. Model interpretation from the additive elements of the likelihood function. *Applied Statistics* 33(1), 52–64.

[34] Muraro, M. and R. De Maria. 2005. Modelling applications and developments for air quality assessment in piemonte. *In Air quality assessment and management in the Piemonte Region according to European Legislation*. Torino, 28 October 2005.

[35] Finardi, S., R. De Maria, A. D'Allura, C. Cascone, G. Calori, and F. Lollobrigida. 2008. A deterministic air quality forecasting system for Torino urban area, Italy. *Environmental Modelling and Software* 23(3), 344–355.

[36] R Development Core Team. 2006. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL www.R-project.org.

Alessandro Fassò *is professor of Statistics and Head of the Department of Information Technology and Mathematical Methods, University of Bergamo, secretary of The International Environmetrics Society (TIES) and Principal Investigator of the Italian PRIN2006 project* Statistical modeling, impact and risk analysis of environmental phenomena with spatial and temporal components. *He is a member of the editorial board of the* Journal of the German Statistical Society (AStA) *and of* Statistica & Applicazioni. *His recent research interests include environmetrics, sensitivity analysis of computer models, environmental time-series, spatio-temporal data, stochastic monitoring, industrial statistics and quality control.*

Michela Cameletti *is a postdoctoral researcher in Environmental Statistics at the Department of Information Technology and Mathematical Methods, University of Bergamo. Her PHD (2007) was on* Spatio-temporal models for environmental data *at this department. She is a member of the Italian research group on environmental statistics named GRASPA (www.graspa.org). Her research interests include statistical modeling of particulate matters in space and time.*