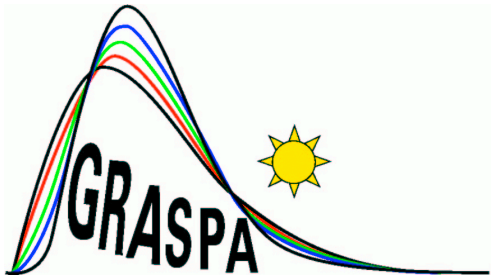
**ISSN: 2037-7738**
GRASPA Working Papers
[online]

Web Working Papers
by
The Italian Group of Environmental Statistics




Gruppo di Ricerca per le Applicazioni della Statistica
ai Problemi Ambientali

www.graspa.org

# Multivariate hidden dynamic geostatistical model for analysing and mapping air quality data in Apulia, Italy

Crescenza Calculli, Alessandro Fassò,
Francesco Finazzi, Alessio Pollice, Annarita Turnone

GRASPA Working paper n. 48, December 2014

# Multivariate hidden dynamic geostatistical model for analysing and mapping air quality data in Apulia, Italy

**Crescenza Calculli**[*1], **Alessandro Fassò**[2], **Francesco Finazzi**[3], **Alessio Pollice**[4], and **Annarita Turnone**[5]

[1] Italian Institute for Nuclear Physics, INFN - Bari, Via E. Orabona n. 4, 70125 Bari, Italy
[2] Dept. of Engineering, University of Bergamo, viale Marconi, 5 - 24044 Dalmine (BG), Italy
[3] Dept. of Management, Economics and Quantitative Methods, University of Bergamo, via dei Caniana 2 - 24127 Bergamo (BG), Italy
[4] Dept. of Economics and Mathematical Methods, University of Bari, Largo Abbazia S. Scolastica, 53 - 70124 Bari, Italy
[5] Agenzia regionale per la prevenzione e la protezione dell'ambiente, ARPA Puglia, Corso Trieste 27 - 70126 Bari, Italy

*Summary*

In this work we propose the multivariate extension of a spatio-temporal model known in the literature and we deduce its maximum likelihood estimator based on the EM algorithm. An illustrating example concerns the joint modeling of air quality and meteorology in Apulia region, Italy. In particular a 8-variate model is fitted for daily particulate matters ($PM_{10}$) and nitrogen dioxides ($NO_2$) concentrations and six non co-located meteorological variables, without the need of preliminary data interpolation. Some preliminary evidence of the model capability to detect a Saharan dust event is also given.

We propose an extension of a univariate hierarchical model with a Markovian latent geostatistical component, which was introduced in the literature as the dynamic spatio-temporal model. Since it is given by a Markovian sequence of spatial random fields plus a measurement error, in this paper we refer to as the hidden dynamic geostatistical model, or HDG model. In the frame of large datasets, a semi-parametric variant of the HDG model was introduced as the fixed rank smoothing spatio-temporal random effect model (FRS-STRE). The univariate HDG has been addressed by both maximum likelihood and Bayesian techniques and some comparison studies showed its good performance in both cases.

For estimation and prediction we develop an HDG extension of the `D-STEM` software which is able to handle multiple variables with heterogeneous spatial supports, covariates, heterotopic monitoring networks and missing data. Uncommon in the EM literature, also the standard deviation of parameter estimates are available. In the air quality example, concentration maps and the related uncertainty maps are estimated in order to highlight the potentiality and usefulness of the proposed approach in the context of official environmental communication.

*Key words:* multivariate hierarchical models, spatio-temporal random effect models, D-STEM, particulate matters, nitrogen dioxides

## 1 Introduction

Environmental and geophysical phenomena are usually characterized by spatial and temporal variability, whose dynamics gives an opportunity to predict events and assess policies over specified areas. By means of hierarchical multivariate spatio-temporal modeling, this paper provides a framework for the analysis of air pollution phenomena which are often an emergency and one of the greatest factors of environmental risk for the human health. In particular, the European Community has defined a complete and congruent set of rules acknowledged by legislation concerning the environment of all member countries. These rules seek to standardize effective control methodologies which include monitoring, towards the quantification of the

* Corresponding author e-mail: crescenza.calculli@ba.infn.it

spatial distribution of the concentration of each pollutant and the evaluation of air quality (i.e. methodologies to measure, calculate, predict or estimate the levels of pollution in the environment). Recently, following the adoption of the new Directive 2008/50/EC of the European Parliament and of the Council of the European Union, the Italian Legislative Decree n. 155/2010 has defined new criteria for the use of evaluation methods different from measurements in fixed sites, with particular reference to modeling techniques. In this Decree potentialities of numerical models are recognized as a valuable approach to the evaluation of the air quality integrated with the analysis of measured data. Modeling techniques can also contribute to the evaluation of plans and programs for pollution reduction, as discussed in Fassò (2013). Moreover Italian rules give local authorities the possibility to neglect air pollution excesses due to the transboundary transport of pollutants. In recent years, a growing number of works has shown the importance of the transboundary contribution to particulate air pollution, especially in Southern Italian regions where dust events due to intrusions of air masses coming from Saharan regions often occur (Amodio *et al.*, 2011; Querol *et al.*, 2004a,b). Therefore the capability to detect dust events is to be considered a relevant feature of numerical models addressing the evaluation of air quality in this area.

In the last decade univariate models for air pollution data have been extensively developed within the hierarchical modelling framework (Sahu *et al.*, 2006; Calder, 2008; McMillan *et al.*, 2010). As far as spatio-temporal data are concerned, univariate hierarchical models are used in several case studies. Smith *et al.* (2003) deal with high percentages of missing particulate matter ($PM_{10}$) concentration data by a simplified Expectation-Maximization (EM) algorithm. In the Bayesian framework, univariate spatio-temporal hierarchical models are proposed to account for rural/background and urban/suburban random effects on $PM_{10}$ concentrations (Sahu *et al.*, 2006), to combine monitoring data and the output from a local-scale air pollution model for health risk assessment (Pirani *et al.*, 2013) and to measure association of nitrogen dioxide ($NO_2$) data with human activity in European urban areas (Shaddick *et al.*, 2013). Multivariate spatio-temporal data are increasingly being used in the analysis of air quality. Pollice and Jona Lasinio (2010) propose the use of a multivariate model for normalized daily concentrations of three pollutants in the Taranto area, Italy, within a Bayesian hierarchical setting including time varying weather covariates and a semi-parametric spatial covariance structure. Finazzi *et al.* (2013) use a three-variate model for mapping daily risk of $PM_{10}$, $NO_2$ and Ozone for an heavily unbalanced monitoring network in Scotland. De Iaco *et al.* (2013) implement spatio-temporal cokriging based on the linear coregionalization model (ST-LCM) to obtain prediction maps of $PM_{10}$ using a three-variate model which includes temperature and wind speed in the Apulia region, Italy.

Suitable software that can handle new univariate and multivariate models with complex spatio-temporal data structures is being developed according to computational improvements. The `spacetime` package (Pebesma, 2012) enables the definition of data structures in space and time and allows the `gstat` functions (Pebesma, 2004) to run spatio-temporal analysis (Pebesma, 2012) in `R`. In the Bayesian framework, the `R-INLA` package (Martino and Rue, 2010) is designed for the Integrated Nested Laplace Approximation (INLA) approach that represents an alternative to inference via MCMC in latent Gaussian models. The approximation of the marginal posteriors of the elements of the latent field, as well as of the model hyperparameters, is based on an efficient combination of Laplace approximations of the full conditionals and numerical integration routines. The `R-INLA` package provides tools for estimation and prediction within a very wide and flexible class of models ranging from (generalized) linear mixed to spatial and spatio-temporal models and is successfully used in a large amount of applications. Recently `D-STEM` has been proposed by Finazzi and Fassò (2014) as a statistical package for multivariate environmental spatio-temporal data analysis and prediction. Like `R-INLA` this software is based on hierarchical models with latent variables but it uses the EM algorithm for model estimation. D-STEM has been successfully tested with various data structures and in several case studies: at the urban scale (Fassò, 2013) for assessing traffic policies in Milan, Italy; at the country scale (Finazzi *et al.*, 2013) for evaluating multi-variable air quality indexes in Scotland; and at the continental scale (Fassò and Finazzi, 2013), considering a large data set of both ground level and remote sensing data over Europe.

In this work we propose to use the EM algorithm for the maximum likelihood estimation of a multivariate extension of the univariate dynamic space–time model introduced by Huang and Cressie (1996). Considering snow precipitation, they focus on spatial optimal prediction and propose a simplified method-of-moment estimate which assumes no measurement errors in the response. In this paper we refer to this model as the hidden dynamic geostatistical (HDG) model, since it is given by a Markovian sequence of spatial random fields plus a measurement error. In a large comparison study Huang *et al.* (2007) conclude that *"In terms of computation and the ability to handle large space–time data sets, a separable and flexible model, such as our Model I* (HDG)*, would be the choice"*. Here parameter estimation is performed using numerical optimization of the log-likelihood, assuming the measurement error variance to be known. Along these lines, considering large datasets such as global $CO_2$ on a regular grid, Katzfuss and Cressie (2011) face the course of dimensionality by fixed rank smoothing (FRS) and a version of HDG called spatio-temporal random effects (STRE) model. The authors propose an EM estimation algorithm which is quite similar to the one in Fassò and Finazzi (2011), but they do not use a spatial covariance model for the latent component. In a fully Bayesian frame, the same authors (Katzfuss and Cressie, 2012) provide the previous STRE-FRS model with a set of appropriate priors and compare their performances in a simulation study. Cameletti *et al.* (2011) consider the STRE-FRS model for air quality assessment in the Piemonte Region, Italy, and compare its performance with five alternative models, including non separable covariance models. They conclude that *"Model C* (HDG) *is the only three-star predictor: this suggests that, in our case study, a model with a complex hierarchical structure is globally preferable to one with a complex spatio-temporal covariance function"*. References to the previous use of models similar to the one we propose include Cameletti *et al.* (2013) who consider INLA-SPDE estimation for the Piemonte dataset and Gelfand *et al.* (2005) who utilize a multivariate dynamic spatial models to analyze precipitations and temperatures in Colorado (USA). The latter is similar to the multivariate HDG model, but it is based a fixed random walk dynamics rather than on the more flexible Markovian one.

Here we do not enter in the unfruitful discussion about full Bayesian versus maximum likelihood estimation. Rather we consider HDG a popular model and the EM algorithm a valuable estimation computational tool, which is still being enriched with new features, as for example constrained estimation (Holmes, 2013) and missingness in the covariates (Naranjo *et al.*, 2013). Hence, in this work we develop the extension of the `D-STEM` software (Finazzi and Fassò, 2014) for estimation and prediction of the multivariate HDG model with missing data and non co-located monitoring networks. The illustrating example concerns the concentrations of $PM_{10}$ and $NO_2$ and the measurements of five meteorological variables in the Apulia region, Italy. As it is often the case, pollution and meteorological information come from two monitoring networks, differing in the number of stations and their spatial locations. In particular, we consider both pollutant concentrations and meteorological variables as response variables in the model formulation, so that no data imputation or interpolation is required. The Kriging approach is then applied with the estimated model for daily mapping pollutant concentrations over the study region. Although not reported in this paper, the model can also be used to obtain daily maps of the meteorological variables.

The remaining part of the paper is structured as follows: the whole Section 2 is dedicated to the definition of the multivariate HDG model and its maximum likelihood estimation by means of the EM algorithm. The use of this model is illustrated in Section 3 through a case study concerning air pollution in the Apulia region, Italy. The results of estimation and mapping procedures are discussed in Section 3.1 with particular reference to a Saharan dust event. Finally, Section 4 provides the main concluding considerations and some remarks on the proposed approach.

## 2 The multivariate hidden dynamic geostatistical model

Let $\mathbf{y}(\mathbf{s},t) = (y_1(\mathbf{s},t),\ldots,y_q(\mathbf{s},t))'$ be the $q$-variate response variable at site $\mathbf{s} \in \mathcal{D} \subset \mathbb{R}^2$ and discrete time $t = 1,\ldots,T$, where $\mathcal{D}$ is a region of space. We say that a spatio-temporal stochastic process is a

hidden dynamic geostatistical (HDG) model if it is defined by

$$\begin{array}{rcl} \mathbf{y}(\mathbf{s},t) & = & \mathbf{X}_{\boldsymbol{\beta}}(\mathbf{s},t)\,\boldsymbol{\beta} + \mathbf{X}_{\mathbf{z}}(\mathbf{s},t)\,\mathbf{A}\mathbf{z}(\mathbf{s},t) + \boldsymbol{\varepsilon}(\mathbf{s},t) \\ \mathbf{z}(\mathbf{s},t) & = & \mathbf{G}\mathbf{z}(\mathbf{s},t-1) + \boldsymbol{\eta}(\mathbf{s},t) \end{array} \tag{1}$$

where $\mathbf{z}(\mathbf{s},t) = (z_1(\mathbf{s},t),...,z_p(\mathbf{s},t))'$ is a latent random variable with Markovian dynamics, $\boldsymbol{\eta}(\mathbf{s},t) = (\eta_1(\mathbf{s},t),\ldots,\eta_p(\mathbf{s},t))'$ is the spatially correlated temporal innovation while $\boldsymbol{\varepsilon}(\mathbf{s},t) = (\varepsilon_1(\mathbf{s},t),...,\varepsilon_q(\mathbf{s},t))'$ is the random measurement error.

In general, $p$ and $q$ can be different, but the case $p = q$ is considered here, as discussed in subsection 2.1. In Equation (1) $\boldsymbol{\eta}(\mathbf{s},t)$ is a Gaussian process, namely $\boldsymbol{\eta}(\mathbf{s},t) \sim \text{GP}_q(\mathbf{0},\boldsymbol{\Gamma})$, with matrix correlation function $\boldsymbol{\Gamma}$ given by

$$\boldsymbol{\Gamma} = \mathbf{V}\rho(\|\mathbf{s}-\mathbf{s}'\|;\theta,\nu)$$

where $\mathbf{V}$ is a valid $q-$dimensional correlation matrix with elements $v_{i,j}$, while

$$\rho(\|\mathbf{s}-\mathbf{s}'\|;\theta,\nu) = \frac{1}{\tilde{\Gamma}(\nu)2^{\nu-1}}\left(\sqrt{2\nu}\frac{\|\mathbf{s}-\mathbf{s}'\|}{\theta}\right)^{\nu} K_{\nu}\left(\sqrt{2\nu}\frac{\|\mathbf{s}-\mathbf{s}'\|}{\theta}\right) \tag{2}$$

is the Matérn spatial covariance function. In particular, $\|\mathbf{s}-\mathbf{s}'\|$ is the distance between two generic spatial locations, $\theta > 0$ and $\nu > 0$ are parameters, $\tilde{\Gamma}$ is the gamma function and $K_\nu$ is the modified Bessel function of the second kind. The range parameter $\theta$ is here assumed unknown while the smoothing parameter $\nu$ is fixed. The measurement errors $\varepsilon_i(\mathbf{s},t)$ are independent and white noise in space and time but each variable retains its own variance, namely $\text{Var}(\boldsymbol{\varepsilon}(\mathbf{s},t)) = \text{diag}(\boldsymbol{\sigma}^2) = \text{diag}(\sigma_1^2,...,\sigma_q^2)$.

The fixed effect design matrix $\mathbf{X}_{\boldsymbol{\beta}}(\mathbf{s},t)$ is such that each observed variable $y_i(\mathbf{s},t)$ can be related to $b_i$ covariates. In particular

$$\mathbf{X}_{\boldsymbol{\beta}}(\mathbf{s},t) = \text{blockdiag}\left(\mathbf{x}_{\boldsymbol{\beta},1}(\mathbf{s},t),...,\mathbf{x}_{\boldsymbol{\beta},q}(\mathbf{s},t)\right)$$

where blockdiag is the block diagonal operator and $\mathbf{x}_{\boldsymbol{\beta},i}(\mathbf{s},t)$, $i = 1,...,q$ are $1\times b_i$ vectors, $b_1+...+b_q = b$. Instead, the random effect design matrix

$$\mathbf{X}_{\mathbf{z}}(\mathbf{s},t) = \text{diag}\left(x_{\mathbf{z},1}(\mathbf{s},t),...,x_{\mathbf{z},q}(\mathbf{s},t)\right)$$

is a $q \times q$ diagonal matrix, set to the identity matrix in the present simple case.

Finally $\mathbf{A} = \text{diag}(\alpha_1,...,\alpha_q)$, $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1',...,\boldsymbol{\beta}_q'\right)'$, where $\boldsymbol{\beta}_i = (\beta_{i,1},...,\beta_{i,b_i})'$ are vectors of unknown coefficients, while $\mathbf{G}$ =diag $(g_1,...g_q)$ is the diagonal transition matrix.

The parameter set to be estimated is then given by

$$\Psi = \left\{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\sigma}^2,\mathbf{G},\mathbf{V},\theta\right\}$$

where, with abuse of notation, we use $\mathbf{V}$ for its lower triangular submatrix and $\mathbf{G} = (g_1,...g_q)$, which define the unique parameters to be estimated for these two matrices.

### 2.1 Comments on some features of the multivariate HDG model

The assumption $p = q$ is made here to ease model interpretability, as it allows to univocally relate each observed variable $y_i(\mathbf{s},t)$ to one and only one latent component $z_i(\mathbf{s},t)$. In general this assumption is not necessary, another important special case being the common factor latent model with $p = 1$: sharing the same $z$ component among different $y_i$'s is useful for example in data fusion, see e.g. Fassò and Finazzi (2013) and Berrocal *et al.* (2012).

Under the assumption that the transition matrix $\mathbf{G}$ is diagonal, the $q-$dimensional VAR model for $\mathbf{z}(\mathbf{s},t)$ in (1) may be splitted into $q$ separated equations, nonetheless the components $z_i(\mathbf{s},t)$ are correlated, due to

the correlation among the components of $\boldsymbol{\eta}_t$ given by matrix $\mathbf{V}$. Yule-Walker recursive formulas relating the auto-covariance matrices of $\mathbf{z}(\mathbf{s},t)$ and the variance covariance matrix of the innovation $\boldsymbol{\eta}_t$ are easily obtained, see e.g. Huang and Cressie (1996).

The HDG model is shown to have a multiplicative separable covariance function by Cameletti *et al.* (2011). If one prefers additive separability, this is obtained by a variant of HDG model in which the product $\boldsymbol{a}(\mathbf{s},t) = \mathbf{X}_{\boldsymbol{z}}(\mathbf{s},t)\,\mathbf{A}\mathbf{z}(\mathbf{s},t)$ is substituted by the sum

$$\boldsymbol{a}(\mathbf{s},t) = \mathbf{X}_{\boldsymbol{\zeta}}(\mathbf{s},t)\,\boldsymbol{\zeta}(t) + \mathbf{X}_{\boldsymbol{\eta}}(\mathbf{s},t)\,\boldsymbol{\eta}(\mathbf{s},t) \tag{3}$$

where $\boldsymbol{\zeta}(t)$ is a purely temporal Markovian process common to all locations, $\boldsymbol{\eta}(\mathbf{s},t)$ is an iid sequence of random fields defined as in model (1), $\mathbf{X}_{\boldsymbol{\eta}}(\mathbf{s},t)$ and $\mathbf{X}_{\boldsymbol{\zeta}}(\mathbf{s},t)$ are suitable known quantities. This is essentially the model considered in the original version of D-STEM software.

In the following Apulian case study the identity matrix is a natural choice for $\mathbf{X}_{\boldsymbol{z}}(\mathbf{s},t)$ with $p = q$. Nevertheless the role of a more general random effect design matrix $\mathbf{X}_{\boldsymbol{z}}(\mathbf{s},t)$ can be quite relevant. When $x_{\mathbf{z},i}(\mathbf{s},t)$, $i = 1,...,q$ are *strictu sensu* covariates, then the respective $z_i(\mathbf{s},t)$ can be seen as the corresponding spatio-temporal varying coefficients. Moreover, following a semiparametric approach, if $\boldsymbol{\eta}(\mathbf{s},t)$ in (1) is white noise in space and time and $x_{\mathbf{z},i}(\mathbf{s},t)$ are some appropriate basis functions, e.g. splines, the fixed rank smoothing spatio-temporal random effects (FRS-STRE) model of Katzfuss and Cressie (2011) is obtained.

### 2.2 Complete-data likelihood function

Suppose that the $q$ variables are observed at the sets of spatial locations $\mathcal{S}_i = \{\mathbf{s}_{i,1},...\mathbf{s}_{i,n_i}\}$, $i = 1,...,q$ and let $\mathbf{y}_i(\mathcal{S}_i,t) = \text{stack}\,(y_i(\mathbf{s}_{i,1},t),...,y_i(\mathbf{s}_{i,n_i},t))$, where stack is the stacking operator. Then, for each time $t$, the following $n \times 1$ observation vector is obtained

$$\mathbf{y}_t = \left(\mathbf{y}_1(\mathcal{S}_1,t)',...,\mathbf{y}_q(\mathcal{S}_q,t)'\right)' \tag{4}$$

with $n = n_1 + ... + n_q$. In general, $\mathcal{S}_i \neq \mathcal{S}_j$, that is, each variable can be observed at a different set of spatial locations. Moreover, the vector $\mathbf{y}_t$ may include missing values. Model (1) can be given the following representation

$$\begin{aligned} \mathbf{y}_t &= \boldsymbol{\mu}_t + \boldsymbol{\varepsilon}_t \\ \boldsymbol{\mu}_t &= \mathbf{X}_t^{\boldsymbol{\beta}}\boldsymbol{\beta} + \boldsymbol{X}_t^{\boldsymbol{z}}\tilde{\mathbf{A}}\mathbf{z}_t \\ \mathbf{z}_t &= \tilde{\mathbf{G}}\mathbf{z}_{t-1} + \boldsymbol{\eta}_t \end{aligned} \tag{5}$$

with $\mathbf{X}_t^{\boldsymbol{\beta}} = \text{blockdiag}\,(\mathbf{X}_{\boldsymbol{\beta},1}(\mathcal{S}_1,t),...,\mathbf{X}_{\boldsymbol{\beta},q}(\mathcal{S}_q,t))$, where $\mathbf{X}_{\boldsymbol{\beta},i}(\mathcal{S}_i,t) = \text{stack}\,(\mathbf{x}_{\boldsymbol{\beta},i}(\mathbf{s}_{i,1},t),...,\mathbf{x}_{\boldsymbol{\beta},i}(\mathbf{s}_{i,n_i},t))$, and $\mathbf{X}_t^{\mathbf{z}} = \text{blockdiag}\,(\mathbf{X}_{\mathbf{z},1}(\mathcal{S}_1,t),...,\mathbf{X}_{\mathbf{z},q}(\mathcal{S}_q,t))$, where $\mathbf{X}_{\mathbf{z},i}(\mathcal{S}_i,t) = \text{diag}\,(x_{\mathbf{z},i}(\mathbf{s}_{1,1},t),...,x_{\mathbf{z},i}(\mathbf{s}_{1,n_i},t))$, for $i = 1,...,q$. Vectors $\boldsymbol{\mu}_t$, $\mathbf{z}_t$, $\boldsymbol{\eta}_t$ and $\boldsymbol{\varepsilon}_t$ are defined as $\mathbf{y}_t$ in (4). Finally $\tilde{\mathbf{A}} = \text{blockdiag}\left(\alpha_1\mathbf{I}_{n_1},...,\alpha_q\mathbf{I}_{n_q}\right)$ and $\tilde{\mathbf{G}} = \text{blockdiag}\left(g_1\mathbf{I}_{n_1},...,g_q\mathbf{I}_{n_q}\right)$, where $\mathbf{I}_{n_i}$ is the identity matrix of dimension $n_i$. In order to define the complete-data log-likelihood function, notice that the following Gaussian distributions are implied by the model structure:

$$\begin{aligned} \mathbf{y}_t \mid \mathbf{z}_t &\sim N_n(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_\varepsilon) \\ \mathbf{z}_t \mid \mathbf{z}_{t-1} &\sim N_n\left(\tilde{\mathbf{G}}\mathbf{z}_{t-1}, \boldsymbol{\Sigma}_{\boldsymbol{\eta}}\right) \\ \mathbf{z}_0 &\sim N_n(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \end{aligned}$$

with variance-covariance matrices given by $\boldsymbol{\Sigma}_\varepsilon = \text{blockdiag}\left(\sigma_1^2\mathbf{I}_{n_1},...,\sigma_q^2\mathbf{I}_{n_q}\right)$ and $\boldsymbol{\Sigma}_{\boldsymbol{\eta}} = \left(v_{i,j}\rho(\mathbf{H}_{ij})\right)_{i,j=1,...,q}$, where $\mathbf{H}_{ij} = d(\mathcal{S}_i,\mathcal{S}_j)$ is the distance matrix between the spatial locations in $\mathcal{S}_i$ and in $\mathcal{S}_j$, and the variance-covariance matrix $\boldsymbol{\Sigma}_0$ is known.

With this notation, the complete-data log-likelihood function for observations $\mathbf{Y} = \{\mathbf{y}_1, ..., \mathbf{y}_T\}$ and $\mathbf{Z} = \{\mathbf{z}_0, \mathbf{z}_1, ..., \mathbf{z}_T\}$ is given by:

$$\begin{aligned} -2l\left(\Psi; \mathbf{Y}, \mathbf{Z}\right) &= T \log |\boldsymbol{\Sigma}_\varepsilon| + \sum_{t=1}^{T} \mathbf{e}_t' \boldsymbol{\Sigma}_\varepsilon^{-1} \mathbf{e}_t \\ &\quad + \log |\boldsymbol{\Sigma}_0| + (\mathbf{z}_0 - \boldsymbol{\mu}_0)' \boldsymbol{\Sigma}_0^{-1} (\mathbf{z}_0 - \boldsymbol{\mu}_0) \\ &\quad + T \log |\boldsymbol{\Sigma}_{\boldsymbol{\eta}}| + \sum_{t=1}^{T} \left(\mathbf{z}_t - \tilde{\mathbf{G}} \mathbf{z}_{t-1}\right)' \boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{-1} \left(\mathbf{z}_t - \tilde{\mathbf{G}} \mathbf{z}_{t-1}\right) \end{aligned}$$

where $\mathbf{e}_t = \mathbf{y}_t - \boldsymbol{\mu}_t$. Notice that, due to the hierarchical structure of HDG model, $l\left(\Psi; \mathbf{Y}, \mathbf{Z}\right)$ may be written as $l\left(\Psi\right) = l\left(\Psi_1\right) + l\left(\mu_0, \Sigma_0\right) + l\left(\Psi_2\right)$ where $\Psi_1 = \left\{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \mathbf{G}\right\}$ and $\Psi_2 = (\mathbf{V}, \theta)$, reducing the optimization dimensionality in the maximization step of the EM algorithm.

### 2.3 Missing data handling

In order to deal with missing data, the observation vector at time $t$, $\mathbf{y}_t$, is partitioned as

$$\tilde{\mathbf{y}}_t = \left(\left(\mathbf{y}_t^{(1)}\right)', \left(\mathbf{y}_t^{(2)}\right)'\right)',$$

where $\mathbf{y}_t^{(1)} = \mathbf{L}_t \mathbf{y}_t$ is the sub-vector of non-missing data at time $t$ and $\mathbf{L}_t$ is the elimination matrix. Vector $\tilde{\mathbf{y}}_t$ is thus a permutation of $\mathbf{y}_t$ and $\mathbf{y}_t = \mathbf{D}_t \tilde{\mathbf{y}}_t$, with $\mathbf{D}_t$ the commutation matrix. In the sequel, given $\mathbf{b}_t$ a generic $n \times 1$ vector and $\mathbf{B}_t$ a generic $n \times n$ matrix at time $t$, $\mathbf{b}_t^{(1)}$ and $\mathbf{B}_t^{(1)}$ will stand for $\mathbf{L}_t \mathbf{b}_t$ and $\mathbf{L}_t \mathbf{B}_t \mathbf{L}_t'$, respectively. On the other hand, if $\mathbf{B}_t$ is a $n \times m$ matrix, then $\mathbf{B}_t^{(1)} = \mathbf{L}_t \mathbf{B}_t$. Finally, $\mathbf{0}_n$ and $\mathbf{0}_{n \times m}$ will be used to define the $n \times 1$ vector and the $n \times m$ matrix of all zeros, respectively.

The first equation in (5) becomes $\mathbf{y}_t^{(l)} = \boldsymbol{\mu}_t^{(l)} + \boldsymbol{\varepsilon}_t^{(l)}$, $l = 1, 2$ and the variance-covariance matrix of the permuted errors is conformably partitioned, namely:

$$\text{Var} \begin{pmatrix} \boldsymbol{\varepsilon}_t^{(1)} \\ \boldsymbol{\varepsilon}_t^{(2)} \end{pmatrix} = \begin{pmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{R}_{12}' & \mathbf{R}_{22} \end{pmatrix}$$

Since $\boldsymbol{\Sigma}_\varepsilon$ is diagonal, it follows that $\mathbf{R}_{11}$ and $\mathbf{R}_{22}$ are diagonal matrices while $\mathbf{R}_{12} = \mathbf{0}_{(n-u_t) \times u_t}$, with $u_t$ the number of non-missing data at time $t$.

### 2.4 EM algorithm

The EM algorithm is considered here to obtain the maximum likelihood estimate of the model parameter set $\Psi$. The algorithm is known to be based on the iteration of an expectation step and a maximization step. The expectation step amounts at computing $Q\left(\Psi, \Psi^{\langle m \rangle}\right) = E_{\Psi^{\langle m \rangle}}\left[-2l\left(\Psi; \mathbf{Y}, \mathbf{Z}\right) \mid \mathbf{Y}^{(1)}\right]$ where $\mathbf{Y}^{(1)} = \left\{\mathbf{y}_1^{(1)}, ..., \mathbf{y}_T^{(1)}\right\}$, and the maximization step, namely $\Psi^{\langle m+1 \rangle} = \arg\max_{\Psi} Q\left(\Psi, \Psi^{\langle m \rangle}\right)$, gives the updating formulas for the model parameters. A full derivation of the updating formulas is given in the appendix. Interestingly, for the non geostatistical part of $\Psi$, namely $\Psi_1 = \left\{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \mathbf{G}\right\}$, the updating

formulas are available in closed form, as follows:

$$\begin{aligned}
\alpha_i^{\langle m+1\rangle} &= \frac{\sum_{t=1}^{T} \text{tr}\left[\tilde{\mathbf{I}}_i \left(\mathbf{y}_t^{(1)} - \mathbf{X}_t^{\boldsymbol{\beta},(1)}\boldsymbol{\beta}\right)\left(\mathbf{X}_t^{\mathbf{z},(1)}\mathbf{z}_t^T\right)'\right]}{\sum_{t=1}^{T} \text{tr}\left[\tilde{\mathbf{I}}_i \mathbf{X}_t^{\mathbf{z},(1)}\left(\mathbf{z}_t^T\left(\mathbf{z}_t^T\right)' + \mathbf{P}_t^T\right)\left(\mathbf{X}_t^{\mathbf{z},(1)}\right)'\right]}, \; i = 1, ..., q \\
\boldsymbol{\beta}^{\langle m+1\rangle} &= \left[\sum_{t=1}^{T}\left(\mathbf{X}_t^{\boldsymbol{\beta},(1)}\right)'\mathbf{X}_t^{\boldsymbol{\beta},(1)}\right]^{-1}\left(\sum_{t=1}^{T}\left(\mathbf{X}_t^{\boldsymbol{\beta},(1)}\right)'\left[\mathbf{y}_t^{(1)} - \mathbf{X}_t^{\mathbf{z},(1)}\tilde{\mathbf{A}}\mathbf{z}_t^T\right]\right) \\
\left(\sigma_i^2\right)^{\langle m+1\rangle} &= \frac{1}{n_i T}\text{tr}\left[\tilde{\mathbf{I}}_i \sum_{t=1}^{T}\begin{pmatrix} \boldsymbol{\Omega}_t^{(1)} & \mathbf{0}_{u_t\times(n-u_t)} \\ \mathbf{0}_{u_t\times(n-u_t)} & \mathbf{R}_{22}\end{pmatrix}\right], \; i = 1, ..., q \\
g_i^{\langle m+1\rangle} &= \frac{\text{tr}\left[\tilde{\mathbf{I}}_i \mathbf{S}_{10}\right]}{\text{tr}\left[\tilde{\mathbf{I}}_i \mathbf{S}_{00}\right]}, \; i = 1, ..., q \\
\boldsymbol{\mu}_0^{\langle m\rangle} &= \mathbf{z}_0^T.
\end{aligned}$$

where $\mathbf{z}_t^T = E\left(\mathbf{z}_t \mid \mathbf{Y}^{(1)}\right)$, $\mathbf{P}_t^T = \text{Var}\left(\mathbf{z}_t \mid \mathbf{Y}^{(1)}\right)$ and $\mathbf{P}_{t,t-1}^T = \text{Cov}\left(\mathbf{z}_t, \mathbf{z}_{t-1} \mid \mathbf{Y}^{(1)}\right)$ denote the output of the Kalman smoother, $\mathbf{S}_{10} = \sum_{t=1}^{T}\mathbf{z}_t^T\left(\mathbf{z}_{t-1}^T\right)' + \mathbf{P}_{t,t-1}^T$ and $\mathbf{S}_{00} = \sum_{t=1}^{T}\mathbf{z}_{t-1}^T\left(\mathbf{z}_{t-1}^T\right)' + \mathbf{P}_{t-1}^T$ are the well known EM second moments (see e.g. Shumway and Stoffer (2010)), matrix $\boldsymbol{\Omega}_t^{(1)}$ is given in the Appendix and $\tilde{\mathbf{I}}_i$ is a selector matrix defined by

$$\tilde{\mathbf{I}}_i = \text{diag}\left(\mathbf{0}'_{n_1}, ..., \overbrace{\mathbf{1}'_{n_i}}^{i-th\ position}, ..., \mathbf{0}'_{n_q}\right)$$

and used to set to zero all the diagonal elements of the matrix pre-multiplied by $\tilde{\mathbf{I}}_i$ and not related to the $i$-th variable. Also notice that $\boldsymbol{\mu}_0$ is not considered as a model parameter but it is needed to start the Kalman recursion.

Finally, the geostatistical model parameters $\Psi_2 = (\mathbf{V}, \theta)$ are updated through numerical optimization. In particular

$$\left\{\mathbf{V}^{\langle m+1\rangle}, \theta^{\langle m+1\rangle}\right\} = \arg\max_{\mathbf{V},\theta} T\log|\boldsymbol{\Sigma}_{\boldsymbol{\eta}}| + \text{tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{-1}\left(\mathbf{S}_{11} - \mathbf{S}_{10}\tilde{\boldsymbol{G}}' - \tilde{\mathbf{G}}\mathbf{S}_{10}' + \tilde{\mathbf{G}}\mathbf{S}_{00}\tilde{\mathbf{G}}'\right)\right] \quad (6)$$

where $\mathbf{S}_{11} = \sum_{t=1}^{T}\mathbf{z}_t^T\left(\mathbf{z}_t^T\right)' + \mathbf{P}_t^T$.

Starting from an initial value $\Psi^{\langle 0\rangle}$ for the model parameter set, the EM algorithm is iterated until convergence, that is, until the observed data log-likelihood or the elements of the model parameters stop changing significantly.

### 2.5 Mapping with the HDG model

Given the maximum likelihood estimates of the parameter set $\Psi$, predictions of the $i$-th response at new sites $\mathcal{S}_0$ and time $t = 1, ..., T$ are obtained:

$$\hat{\mathbf{y}}_i(\mathcal{S}_0, t) = \mathbf{X}_{\beta,i}(\mathcal{S}_0, t)\hat{\boldsymbol{\beta}}_i + \mathbf{X}_{\mathbf{z},i}(\mathcal{S}_0, t)\hat{\mathbf{A}}_i(\mathcal{S}_0)\mathbf{z}_t^{T,i}(\mathcal{S}_0) \quad (7)$$

while the variance-covariance matrix of $\hat{\mathbf{y}}_i(\mathcal{S}_0, t)$ is given by

$$\boldsymbol{\Sigma}_{\hat{\mathbf{y}}_i}(\mathcal{S}_0, t) = \hat{\mathbf{A}}_i(\mathcal{S}_0)^2\mathbf{X}_{\mathbf{z},i}(\mathcal{S}_0, t)\mathbf{P}_{t,t-1}^{T,i}(\mathcal{S}_0)\mathbf{X}_{\mathbf{z},i}(\mathcal{S}_0, t)' \quad (8)$$

where $\mathbf{z}_t^{T,i}(\mathcal{S}_0)$ and $\mathbf{P}_{t,t-1}^{T,i}(\mathcal{S}_0)$ are the Kalman smoother outputs extended over sites $\mathcal{S}_0$ and for the $i$-th variable while $\hat{\mathbf{A}}_i(\mathcal{S}_0) = \hat{\alpha}_i \mathbf{I}_{|\mathcal{S}_0|}$. If $\mathcal{S}_0$ overlays the entire region $\mathcal{D}$ as a fine regular grid, $\hat{\mathbf{y}}_i(\mathcal{S}_0, t)$ allows to draw a map and the ordered collection

$$\hat{\mathbf{Y}}_i(\mathcal{S}_0) = \{\hat{\mathbf{y}}_i(\mathcal{S}_0, 1), \ldots, \hat{\mathbf{y}}_i(\mathcal{S}_0, T)\} \tag{9}$$

represents a dynamic map for the $i$-th variable.

## 3 The Apulia case study

The methodology discussed in the previous sections is applied to air quality data collected in Apulia for jointly analyzing and mapping concentrations of two troublesome pollutants over the study area: $PM_{10}$ and $NO_2$. Apulia (South-eastern Italy) is a wide, mostly coastal, territory with a high demographic density. The region is characterized by some of the highest environmental risk areas in Italy, deeply investigated for the presence of several industrial activities (as petrochemical plants, oil refineries, navy arsenals, electric power plants) near urban centers (Martuzzi *et al.*, 2002). It is well known that high concentrations of both $PM_{10}$ and $NO_2$ are measured in those urban areas characterized by high population density, heavy traffic and industries. Whereas particulate matter and the exceedance of the related limiting values have attracted considerable public attention during the last years, the $NO_2$ problem is a relatively new one becoming mature through the introduction of new European limiting values in 2010 (EU Directive 2008/50/EC). We analyze the daily average concentrations (in $\mu g/m^3$) of both $PM_{10}$ and $NO_2$ and the measurements of some meteorological variables that may drive the pollutant diffusion and advection. Pollutant concentrations and the meteorological data are provided by the Apulia Region Environmental Protection Agency (ARPA Puglia) and by the Agrometeorological Service of the Apulia region (ASSOCODIPUGLIA), respectively. We consider a three months study period ($T = 92$ days) corresponding to the summer season 2012 (from $1^{st}$ July to $30^{th}$ September). This period was characterized by high temperatures and an event of intrusion of long range transported air mass from the African continent, with a large number of monitoring sites exceeding the daily $PM_{10}$ limit value of 50 $\mu g/m^3$ (according to the Legislative Decree 155/2010). The focus on this time period is particularly relevant as the Italian law gives local authorities the possibility to neglect air pollution excesses due to the transboundary transport that was recently shown to provide a substantial contribution to particulate air pollution in Southern Italian regions, where intrusion of air masses coming from Saharan dust events often occurs (Amodio *et al.*, 2011; Querol *et al.*, 2004a,b).

$PM_{10}$ and $NO_2$ concentrations come respectively from 73 and 68 monitoring sensors of the Apulia multipollutant air quality network. Figures 1(a) - 1(b) show the spatial distribution of the two pollutants measuring sites. The missing data rates for each pollutant are reported in Table 1. Meteorological data come from 58 stations (Figure 1(c)) and no missing data occur in the period. We consider the daily average temperature (TEMP, in °C), relative humidity (RH, in $\%$), atmospheric pressure (AP, in hPa), wind speed ($WIND_s$, in $m/s$), east-west ($WIND_u$, in $m/s$) and north-south wind component ($WIND_v$, in $m/s$). Precipitations are not considered, due to the lack of rainy days except for a short rainfall in the last week of July. Pollutant concentrations and meteorological variables are integrated by a set of time invariant covariates as population counts (*pop*), monitoring station coordinates (*lat* and *lon*, UTM coordinates in km) and land elevation (*elev*, in m). Population data are considered as a proxy of pollution emissions and are drawn from the LandScan™ ambient population count database, updated to the year 2008 (Bhaduri *et al.*, 2007), which provides global population data with approximately 1 km resolution (30"$\times$ 30"). Land elevation data come from the GTOPO30 global digital elevation model (DEM) with a horizontal grid spacing of 30 arc-seconds (approximately 1 kilometer).

Pollutant concentrations and meteorological data are obtained by non co-located monitoring networks (see Figure 1), giving an example of heterotopic spatial data, where different variables are observed at different locations (Wackernagel, 2003). In order to exploit the meteorological information avoiding data

**Table 1** Summary statistics of the pollutant concentrations from $1^{st}$ Jul to $30^{th}$ Sept, 2012.

| Pollutant | Number of stations | Mean ($\mu g/\mathrm{m}^3$) | Standard deviation ($\mu g/\mathrm{m}^3$) | Missing (%) |
|---|---|---|---|---|
| $PM_{10}$ | 68 | 27.18 | 3.68 | 24.20 |
| $NO_2$ | 73 | 43.90 | 8.01 | 21.05 |

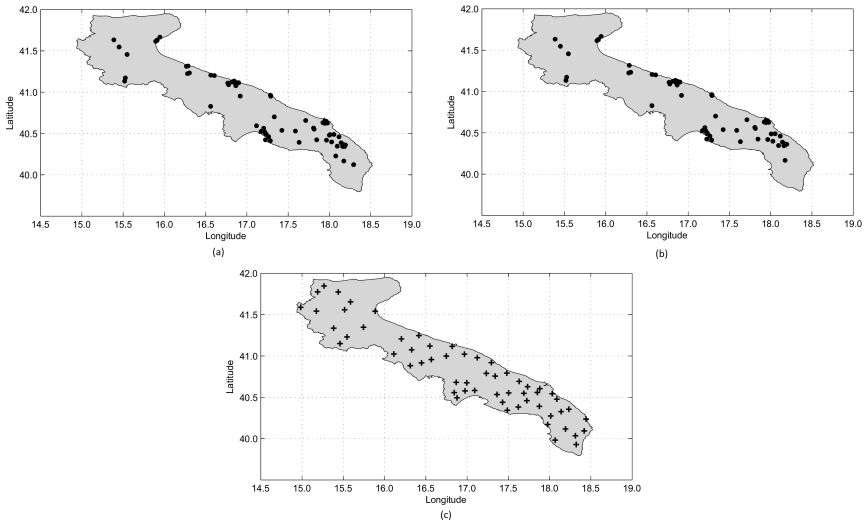



**Fig. 1** Spatial distribution of the Apulia region air quality network: (a) $NO_2$ (73 sites); (b) $PM_{10}$ (68 sites). Locations of the meteorological network sites: (c) meteorological stations (58 sites).

interpolation, we choose to consider both air quality and meteorological variables as a 8-dimensional response vector. Traffic-type stations are excluded from the analysis to avoid possible overestimation of the concentration of the two pollutants over the entire region, as traffic and road data are not currently available.

Maximum likelihood parameter estimation and spatial predictions for the multivariate HDG model are implemented by an extension of the `D-STEM` software previously introduced by Finazzi and Fassò (2014), based on the EM algorithm which reaches convergence even when the number of parameters to be estimated is high as in the case study. Model estimations and mapping are performed on an Intel(R) Core(TM) i3- 2370M CPU laptop, with 2.40 GHz and 8GB RAM. The estimation time is $\sim$ 27 minutes and the EM algorithm takes 42 iterations to converge.

### 3.1 Estimation and mapping results

In this Section we report the results of the EM maximum likelihood estimation of the multivariate HDG model for the case study. Preliminary data analysis (not reported, but available from the authors upon request) led to log-transform pollutant concentrations to reduce heteroskedasticity and long tails in the data distributions. All response variables and covariates were also standardized in order to improve numerical stability and an simplify the comparison between the parameter values across pollutants and meteorological variables.

Table 2 reports the maximum likelihood estimates of the parameters in $\boldsymbol{\Psi}$. The estimated $\hat{\boldsymbol{\beta}}$-coefficients suitably describe the physics of air pollution. As expected land elevation (*elev*) has a significant effect in reducing both $PM_{10}$ and $NO_2$ concentrations. A positive relationship can be seen between the covariate

**Table 2** Estimated parameters for the multivariate HDG model. Standard deviations in parentheses.

| | $\hat{\beta}_{elev}$ | $\hat{\beta}_{pop}$ | $\hat{\beta}_{lat}$ | $\hat{\beta}_{lon}$ | $\hat{g}_i$ | $\hat{\alpha}_i$ | $\hat{\sigma}_i^2$ |
|---|---|---|---|---|---|---|---|
| $NO_2$ | $-0.163_{(0.028)}$ | $0.022_{(0.021)}$ | $0.017_{(0.129)}$ | $-0.207_{(0.142)}$ | $0.621_{(0.025)}$ | $1.192_{(0.021)}$ | $0.449_{(0.013)}$ |
| $PM_{10}$ | $-0.196_{(0.036)}$ | $0.169_{(0.024)}$ | $-0.025_{(0.153)}$ | $-0.046_{(0.171)}$ | $0.833_{(0.012)}$ | $0.708_{(0.029)}$ | $0.289_{(0.008)}$ |
| TEMP | $-0.145_{(0.022)}$ | | $-0.174_{(0.110)}$ | $-0.188_{(0.131)}$ | $0.884_{(0.005)}$ | $0.381_{(0.013)}$ | $0.005_{(0.000)}$ |
| RH | $-0.046_{(0.019)}$ | | $0.116_{(0.098)}$ | $0.350_{(0.118)}$ | $0.809_{(0.006)}$ | $0.478_{(0.013)}$ | $0.013_{(0.001)}$ |
| AP | $-0.928_{(0.025)}$ | | $-0.015_{(0.096)}$ | $-0.042_{(0.108)}$ | $0.948_{(0.003)}$ | $0.234_{(0.013)}$ | $0.000_{(0.000)}$ |
| $WIND_u$ | $-0.094_{(0.021)}$ | | $0.190_{(0.113)}$ | $0.258_{(0.137)}$ | $0.605_{(0.016)}$ | $0.901_{(0.016)}$ | $0.126_{(0.004)}$ |
| $WIND_v$ | $-0.070_{(0.024)}$ | | $0.142_{(0.127)}$ | $0.248_{(0.155)}$ | $0.553_{(0.017)}$ | $1.095_{(0.017)}$ | $0.225_{(0.006)}$ |
| $WIND_s$ | $0.191_{(0.025)}$ | | $0.057_{(0.138)}$ | $-0.104_{(0.167)}$ | $0.572_{(0.013)}$ | $1.180_{(0.014)}$ | $0.081_{(0.003)}$ |

**Table 3** Estimated $\hat{\mathbf{V}}$ correlation matrix of the multivariate HDG model. Standard deviations in parentheses.

| | $NO_2$ | $PM_{10}$ | TEMP | RH | AP | $WIND_u$ | $WIND_v$ | $WIND_s$ |
|---|---|---|---|---|---|---|---|---|
| $NO_2$ | 1.00 | $0.637_{(0.033)}$ | $-0.021_{(0.073)}$ | $-0.123_{(0.057)}$ | $-0.578_{(0.094)}$ | $0.210_{(0.031)}$ | $0.181_{(0.026)}$ | $-0.299_{(0.021)}$ |
| $PM_{10}$ | | 1.00 | $0.223_{(0.141)}$ | $0.089_{(0.117)}$ | $0.001_{(0.231)}$ | $0.270_{(0.065)}$ | $-0.002_{(0.057)}$ | $-0.515_{(0.041)}$ |
| TEMP | | | 1.00 | $-0.617_{(0.079)}$ | $-0.000_{(0.192)}$ | $-0.063_{(0.068)}$ | $-0.094_{(0.056)}$ | $0.057_{(0.044)}$ |
| RH | | | | 1.00 | $0.004_{(0.163)}$ | $0.042_{(0.056)}$ | $0.003_{(0.046)}$ | $-0.016_{(0.036)}$ |
| AP | | | | | 1.00 | $0.155_{(0.102)}$ | $-0.047_{(0.087)}$ | $-0.191_{(0.064)}$ |
| $WIND_u$ | | | | | | 1.00 | $0.132_{(0.026)}$ | $0.082_{(0.022)}$ |
| $WIND_v$ | | | | | | | 1.00 | $-0.524_{(0.014)}$ |
| $WIND_s$ | | | | | | | | 1.00 |

*pop* and $PM_{10}$, showing higher concentrations of this pollutant in most populated areas. The signs of estimated coefficients related to meteorological variables are also in line with their physical behavior, e.g. the estimated $\hat{\beta}_{elev}$ is negative for the TEMP, RH and AP responses, representing the tendency of these variables to decrease with growing altitude. Notice the systematic higher estimated sd's of the effects of the *lon* coordinate with respect to *lat*, possibly highlighting the presence of anisotropy due to the stretched shape characterizing the observation area. The analysis of the $\hat{g}_i$-values suggests that response variables have specific temporal dynamics. In particular, the $PM_{10}$ has a slower temporal dynamics compared to $NO_2$, while directional wind components are characterized by lower persistence. Estimates of $\alpha_i$'s account for the importance of the spatio-temporal interaction term $\mathbf{z}(\mathbf{s},t)$ and show an inverse proportional compensation behavior with $\hat{g}_i$'s. The measurement errors $\hat{\sigma}_i^2$ are higher for $NO_2$ than $PM_{10}$ and generally lower for all meteorological variables, except for the North-South wind component, corresponding to peaks of intensity of winds blowing from North-East over the study area. The estimated parameter $\hat{\theta}$ represents the range (in km) of the exponential spatial correlation ($\nu = 1/2$) common to all response variables and expressed by the latent random vector $\boldsymbol{\eta}(\mathbf{s},t)$. It amounts to 182.08 km (sd = 4.313), implying a smooth spatial behavior. The estimated matrix $\hat{\mathbf{V}}$ is reported in Table 3, representing the cross-correlations of the elements of $\boldsymbol{\eta}(\mathbf{s},t)$. A positive correlation between the two pollutants is noticeable, legitimizing the choice of the multivariate model considering the two pollutants altogether.

In order to evaluate the predictive capability of the multivariate HDG model, a leave one site out cross-validation (CV) is used, following the approach proposed in Fassò *et al.* (2007). The procedure consists in recursively estimating the multivariate HDG model removing one site at a time, then predicting concentrations of both pollutants at the removed site for the 92 time points. Cross-validation mean squared errors (CMSE) for the two pollutants, obtained averaging the squared residuals for each day and site, are respectively $CMSE_{NO_2}$ = 0.663 and $CMSE_{PM_{10}}$ = 0.565, confirming the different predictive capability of the multivariate HDG model shown by the estimated measurement errors $\hat{\sigma}_i^2$ in Table 2.

The descriptive capability of the multivariate HDG model is exemplified focusing on a specific atmospheric event by the dynamic maps $\hat{\mathbf{Y}}_{NO_2}(\mathcal{S}_0)$ and $\hat{\mathbf{Y}}_{PM_{10}}(\mathcal{S}_0)$ (see Equation 9, Section 2.5) evaluated over a fine regular grid $\mathcal{S}_0$ with high spatial resolution (1km $\times$ km approximatively) within the Apulia boundaries. Maps in Figure 2 show the estimated daily average concentrations of $PM_{10}$ and $NO_2$ and the associated standard deviations referred to 29 September 2012. Back trajectory modeling based on Modis

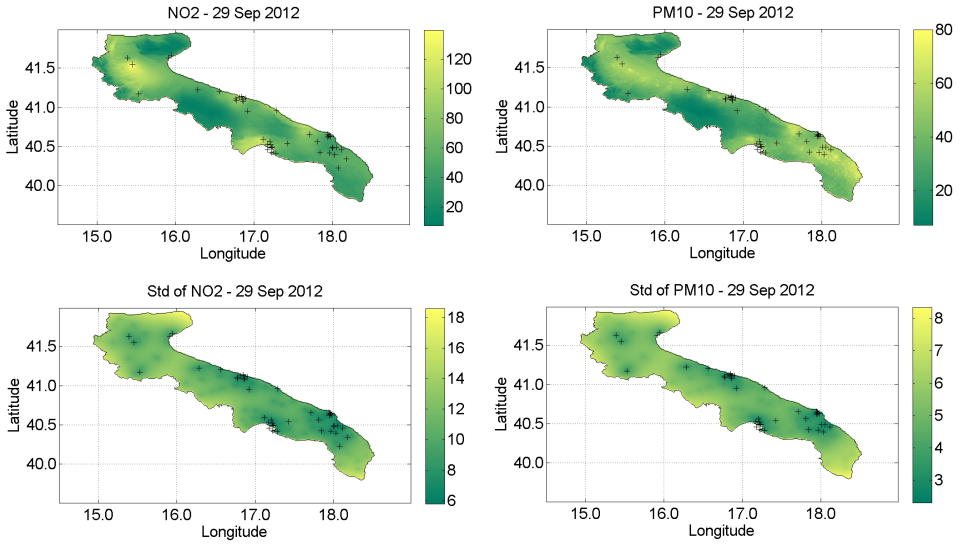



**Fig. 2** Estimated $NO_2$ concentration and standard deviation (top and bottom left) and estimated $PM_{10}$ concentration and standard deviation (top and bottom right) for the dust day (September 29, 2012) over the Apulia region. The '+' symbol represents the air quality network stations.

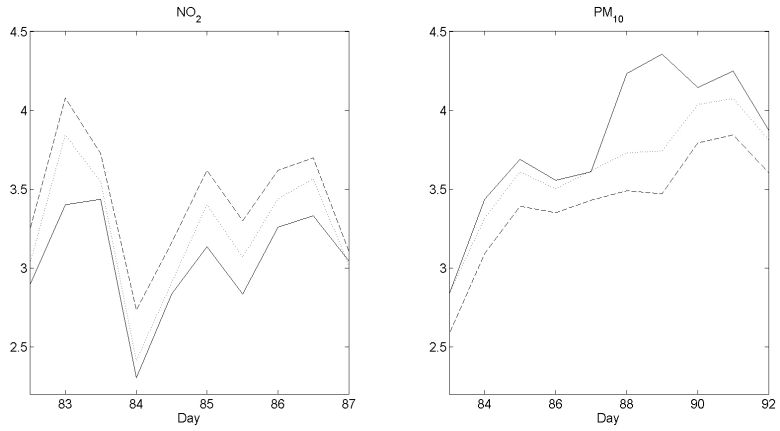



**Fig. 3** Log-transformed standardized concentrations of the two pollutants at the Campi Salentina monitoring station on 21-30 September 2012: observed (solid lines), model predictions (dotted) and CV predictions (dashed).

satellite images (Draxler and Rolph, 2011) and synoptic analysis indicate a probable advection of African dust particles over the the Apulia region reaching its maximum on 29 September. Such events do not have a constant evolution in time and can involve only part of the area under consideration, due to its geographical configuration. Maps obtained by the procedure defined in Section 2.5 focused on the dust event show that on 29 September the estimated $PM_{10}$ concentrations exceed the threshold of 50 $\mu g/m^3$ over the Salento peninsula (south of Apulia) and in the metropolitan areas of Taranto, Bari and Foggia. $NO_2$ estimated concentrations are not as high, proving that this pollutant is not as affected by the advection of dust particles. Uncertainties related to both pollutant maps are quite small, excluding the unmonitored areas and those along the borders of the region. A further example of the multivariate HDG model capabilities is obtained comparing observed concentrations of the two pollutants with model predictions obtained as in

(9) and CV predictions at the Campi Salentina monitoring station, located in the Southern part of the study area, mostly exposed to the dust event. As expected, due to the outlying character of particulate matter in these days, the model underestimates $PM_{10}$. Instead, due to positive cross correlation among pollutants, it overestimates $NO_2$ as shown in Figure 3. This behavior can be considered as the basis to define a statistical detector of dust events.

## 4 Conclusion and discussion

In this work we propose the maximum likelihood estimation of a multivariate spatio-temporal model which can be used for the analysis and mapping of air quality data. We consider a model specification based on a first order auto-regressive spatial component to jointly model spatial and temporal dependencies. The use of the multivariate HDG model is illustrated for a case study concerning air quality over the Apulia region, Italy. Concentrations of $NO_2$ and $PM_{10}$ and a set of meteorological variables are considered in a multivariate setting. The proposed approach allows to handle missing data and non co-located monitoring networks in a natural way. For the purpose we extend the `D-STEM` software based on the EM algorithm giving model parameter estimates, dynamic maps and cross-validation results.

It is worth pointing out that the HDG model provides a flexible structure for the analysis of air quality in several data frameworks. For instance, this modeling approach could be used for the temporal forecasting of concentration levels possibly coupled with a meteorological forecaster. A line of further investigation is related to almost real-time mapping of air quality data, obtained through daily recursive model estimation. This is accomplished considering a sliding data window of size *n* days in Calculli *et al.* (2014). The case study suggests that joint HDG modeling of particulate and gaseous air pollutants may be used to detect dust events on a statistical basis and further research on this subject is needed to devise statistical methods for transboundary adjustment of particulate matters. The availability of traffic and road data would allow the integration of the case study data base with traffic-type sensor pollutant concentration data, avoiding possible overestimation of the anthropogenic contribution to $NO_2$ and $PM_{10}$ pollution over the entire region.

*Acknowledgements.* The authors thank ASSOCODIPUGLIA - Associazione Regionale dei Consorzi di Difesa della Puglia for providing the meteorological data. The analysis utilized the LandScan™ high resolution global population data set copyrighted by UT-Battelle, LLC, operator of Oak Ridge National Laboratory under contract DE-AC05-00OR22725 with the US Department of Energy. The US Government has certain rights on this data set. Neither UT-Battelle, LLC, nor the US Department of Energy, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of the data set. Finazzi was funded for this work through the FIRB2012 project "Statistical modelling of environmental phenomena: pollution, meteorology, health and their interactions" (RBFR12URQJ).

## References

Amodio M, Andriani E, Angiuli L, Assennato G, De Gennaro G, Gilio AD, Giua R, Intini M, Menegotto M, Nocioni A, Palmisani J, Perrone MR, Placentino CM, Tutino M, 2011. Chemical characterization of PM in the Apulia region: local and long-range transport contributions to particulate matter. *Boreal Environmental Research* **16**: 251–261.

Berrocal V, Gelfand A, Holland D, 2012. Space-time data fusion under error in computer model output: an application to modeling air quality. *Biometrics* **68**: 837–848.

Bhaduri B, Bright E, Coleman P, Urban M, 2007. Landscan USA: a high resolution geospatial and temporal modeling approach for population distribution and dynamics. *GeoJournal* **69**: 103–117.

Calculli C, Turnone A, Finazzi F, Pollice A, Morabito A, 2014. Model based spatio-temporal analysis and mapping of Apulia air quality data. *Proceedings of 1st International Conference on Atmospheric Dust,* Castellaneta Marina - Italy, June 1-6, 2014.

Calder CA, 2008. A dynamic process convolution approach to modeling ambient particulate matter concentrations. *Environmetrics* **19**(1): 39–48.

Cameletti M, Ignaccolo R, Bande S, 2011. Comparing spatio-temporal models for particulate matter in Piemonte. *Environmetrics* **22**(8): 985–996.

Cameletti M, Lindgren F, Simpson D, Rue H, 2013. Spatio-temporal modeling of particulate matter concentration through the SPDE approach. *AStA Advances in Statistical Analysis* **97**(2): 109–131.

De Iaco S, Palma M, Posa D, 2013. Prediction of particle pollution through spatio-temporal multivariate geostatistical analysis. *AStA Advances in Statistical Analysis* **97**(2): 133–150.

Draxler R, Rolph GD, 2011. HYSPLIT (HYbrid Single-Particle Lagrangian Integrated Trajectory). *NOAA Air Resources Laboratory, Silver Spring, MD.* Available at: `http://ready.arl.noaa.gov/HYSPLIT.php`.

Fassò A, 2013. Statistical assessment of air quality interventions. *Stochastic Environmental Research and Risk Assessment* **27**(7): 1651–1660.

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

Fassò A, Finazzi F, 2011. Maximum likelihood estimation of the dynamic coregionalization model with heterotropic data. *Environmetrics* **22**(6): 735–748.

Fassò A, Finazzi F, 2013. A varying coefficients space-time model for ground and satellite air quality data over Europe. *Statistica & Applicazioni, Special online issue* : 45–56.

Finazzi F, Fassò A, 2014. D-STEM: A software for the analysis and mapping of environmental space-time variables. *Journal of Statistical Software.* Accepted.

Finazzi F, Scott EM, Fassò A, 2013. A model-based framework for air quality indices and population risk evaluation, with an application to the analysis of Scottish air quality data. *Journal of The Royal Statistical Society: Series C* **62**(2): 287–308.

Gelfand AE, Banerjee S, Gamerman D, 2005. Spatial process modelling for univariate and multivariate dynamic spatial data. *Environmetrics* **16**(5): 465–479.

Holmes E, 2013. Derivation of an EM algorithm for constrained and unconstrained multivariate autoregressive state-space (MARSS) models. *arXiv:1302.3919v1* .

Huang H, Cressie N, 1996. Spatio-temporal prediction of snow equivalent using the Kalman filter. *Computational Statistics and Data Analysis* **22**: 159–175.

Huang H, Martinez F, Mateu J, Montes F, 2007. Model comparison and selection for stationary space–time models. *Computational Statistics and Data Analysis* **51**: 4577–4596.

Katzfuss M, Cressie N, 2011. Spatio-temporal smoothing and EM estimation for massive remote-sensing data sets. *Journal of Time Series Analysis* **4**(1): 430–446.

Katzfuss M, Cressie N, 2012. Bayesian hierarchical spatio-temporal smoothing for very large datasets. *Environmetrics* **23**(1): 94–107.

Martino S, Rue H, 2010. Implementing approximate Bayesian inference using Integrated Nested Laplace Approximation: a manual for the INLA program. Available at: http://www.math.ntnu.no/hrue/GMRFsim/manual.pdf.

Martuzzi M, Mitis F, Biggeri A, Terracini B, Bertollini R, 2002. Environment and health status of the population in areas with high risk of environmental crisis in italy. *Epidemilogia e Prevenzione* **26**(6): 1–53.

McMillan N, Holland DM, Morara M, Feng J, 2010. Combining numerical model output and particulate data using Bayesian space-time modeling. *Environmetrics* **21**: 48–65.

Naranjo A, Trindade A, Casella G, 2013. Extending the state-space model to accommodate missing values in responses and covariates. *Journal of the American Statistical Association* **108**(501): 202–216.

Pebesma EJ, 2004. Multivariable geostatistics in S: the gstat package. *Computers & Geosciences* **30**: 683–691.

Pebesma EJ, 2012. spacetime: Spatio-temporal data in R. *Journal of Statistical Software* **51**(7): 1–30.

Pirani M, Gulliver J, Fuller GW, Blangiardo M, 2013. Bayesian spatiotemporal modelling for the assessment of short-term exposure to particle pollution in urban areas. *Journal of Exposure Science & Environmental Epidemiology* **27**: 1–9, doi: 10.1038/jes.2013.85.

Pollice A, Jona Lasinio G, 2010. A multivariate approach to the analysis of air quality in a high environmental risk area. *Environmetrics* **21**(7-8): 741–754.

Querol X, Alastuey A, Rodríguez S, Viana MM, Artínano B, Salvador P, Mantilla E, Santos S, Patier R, De la Rosa J, De la Campa AS, Menéndez M, Gil JJ, 2004a. Levels of PM in rural, urban and industrial sites in Spain. *The Science of Total Environment* **334**: 359–376.

Querol X, Alastuey A, Viana MM, Rodríguez S, Artínano B, Salvador P, Do Santos SG, Patier R, Ruiz C, De la Rosa J, De la Campa AS, Menendez M, Gil JJ, 2004b. Speciation and origin of PM10 and PM2.5 in Spain. *Aerosol Science* **35**: 1151–1172.

Sahu SK, Gelfand AE, Holland D, 2006. Spatio-temporal modeling of fine particulate matter. *Journal of Agricultural Biological and Environmental Statistics* **11**(1): 61–86.

Shaddick G, Yan H, Vienneau D, 2013. A Bayesian hierarchical model for assessing the impact of human activity on nitrogen dioxide concentrations in europe. *Environmental and Ecological Statistics,* **20**: 553–570, doi: 10.1007/s10651-012-0234-z.

Shumway R, Stoffer D, 2010. *Time series analysis and its applications, with R examples, 3rd edn*. Springer, New York.

Smith RL, Kolenikov S, Cox LH, 2003. Spatio-temporal modeling of PM2.5 data with missing values. *Journal of Geophysical Research* **108**(D24), doi: 10.1029/2002JD002914.

Wackernagel H, 2003. *Multivariate Geostatistics: An introduction with applications, 2nd edn*. Springer, New York.

## Appendix A

In this appendix a derivation of the closed form formulas of section 2.4 is given. The starting point is the following representation of the conditional expectation of the complete-data log-likelihood, or the E step:

$$\begin{aligned} Q\left(\Psi, \Psi^{\langle m \rangle}\right) &= E_{\Psi^{\langle m \rangle}}\left[-2l\left(\Psi; \mathbf{Y}, \mathbf{Z}\right) \mid \mathbf{Y}^{(1)}\right] \\ &= E_{\Psi^{\langle m \rangle}}\left[E_{\Psi^{\langle m \rangle}}\left[-2l\left(\Psi; \mathbf{Y}, \mathbf{Z}\right) \mid \mathbf{Y}^{(1)}, \mathbf{Z}\right] \mid \mathbf{Y}^{(1)}\right]. \end{aligned}$$

In what follows let $E\left(\cdot \mid \cdot\right) \equiv E_{\Psi^{\langle m \rangle}}\left(\cdot \mid \cdot\right)$ and $\text{Var}\left(\cdot \mid \cdot\right) \equiv \text{Var}_{\Psi^{\langle m \rangle}}\left(\cdot \mid \cdot\right)$. Considering the inner conditional expectation, the following result holds

$$\begin{aligned} &E\left[-2l\left(\Psi; \mathbf{Y}, \mathbf{Z}\right) \mid \mathbf{Y}^{(1)}, \mathbf{Z}\right] = T \log\left|\boldsymbol{\Sigma}_\varepsilon\right| + \text{tr}\left[\boldsymbol{\Sigma}_\varepsilon^{-1} \sum_{t=1}^{T}\left(E\left(\mathbf{e}_t \mid \mathbf{Y}^{(1)}, \mathbf{Z}\right) E\left(\mathbf{e}_t \mid \mathbf{Y}^{(1)}, \mathbf{Z}\right)' + \text{Var}\left(\mathbf{e}_t \mid \mathbf{Y}^{(1)}, \mathbf{Z}\right)\right)\right] \\ &+ \log\left|\boldsymbol{\Sigma}_0\right| + \text{tr}\left[\boldsymbol{\Sigma}_0^{-1}\left(\mathbf{z}_0 - \boldsymbol{\mu}_0\right)\left(\mathbf{z}_0 - \boldsymbol{\mu}_0\right)'\right] \\ &+ T \log\left|\boldsymbol{\Sigma}_{\boldsymbol{\eta}}\right| + \text{tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{-1} \sum_{t=1}^{T}\left(\mathbf{z}_t - \tilde{\mathbf{G}}\mathbf{z}_{t-1}\right)\left(\mathbf{z}_t - \tilde{\mathbf{G}}\mathbf{z}_{t-1}\right)'\right] \end{aligned} \tag{10}$$

where, recalling that $\mathbf{R}_{12} = \mathbf{0}_{(n-u_t)\times u_t}$, we have

$$E\left(\mathbf{e}_t \mid \mathbf{Y}^{(1)}, \mathbf{Z}\right) = \mathbf{D}_t \begin{pmatrix} \mathbf{e}_t^{(1)} \\ \mathbf{R}_{21}\mathbf{R}_{11}^{-1}\mathbf{e}_t^{(1)} \end{pmatrix} = \mathbf{D}_t \begin{pmatrix} \mathbf{e}_t^{(1)} \\ \mathbf{0}_{(n-u_t)\times 1} \end{pmatrix}$$

and

$$\text{Var}\left[\mathbf{e}_t \mid \mathbf{Y}^{(1)}, \mathbf{Z}\right] = \mathbf{D}_t \begin{pmatrix} \mathbf{0}_{u_t\times u_t} & \mathbf{0}_{u_t\times(n-u_t)} \\ \mathbf{0}_{(n-u_t)\times u_t} & \mathbf{R}_{22} - \mathbf{R}_{21}\mathbf{R}_{11}^{-1}\mathbf{R}_{12} \end{pmatrix} \mathbf{D}_t' = \mathbf{D}_t \begin{pmatrix} \mathbf{0}_{u_t\times u_t} & \mathbf{0}_{u_t\times(n-u_t)} \\ \mathbf{0}_{(n-u_t)\times u_t} & \mathbf{R}_{22} \end{pmatrix} \mathbf{D}_t'.$$

Moreover, $\boldsymbol{\mu}_0 \equiv \boldsymbol{\mu}_0^{\langle m \rangle}$, $\tilde{\mathbf{A}} \equiv \tilde{\mathbf{A}}^{\langle m \rangle}$, $\tilde{\mathbf{G}} \equiv \tilde{\mathbf{G}}^{\langle m \rangle}$, $\boldsymbol{\Sigma}_\eta \equiv \boldsymbol{\Sigma}_\eta^{\langle m \rangle}$ and $\boldsymbol{\Sigma}_\varepsilon \equiv \boldsymbol{\Sigma}_\varepsilon^{\langle m \rangle}$, that is, vectors and matrices are evaluated using the value of the model parameters at the $m$-th iteration of the EM algorithm.

Applying the outer conditional expectation to the right hand side of (10) it follows that

$$\begin{aligned} &Q\left(\Psi, \Psi^{\langle m \rangle}\right) = T \log\left|\boldsymbol{\Sigma}_\varepsilon\right| + \text{tr}\left(\boldsymbol{\Sigma}_\varepsilon^{-1} \sum_{t=1}^{T} \boldsymbol{\Omega}_t\right) \\ &+ \log\left|\boldsymbol{\Sigma}_0\right| + \text{tr}\left[\boldsymbol{\Sigma}_0^{-1}\left\{\left[E\left(\mathbf{z}_0 \mid \mathbf{Y}^{(1)}\right) - \boldsymbol{\mu}_0\right]\left[E\left(\mathbf{z}_0 \mid \mathbf{Y}^{(1)}\right) - \boldsymbol{\mu}_0\right]' + \text{Var}\left(\mathbf{z}_0 \mid \mathbf{Y}^{(1)}\right)\right\}\right] \\ &+ T \log\left|\boldsymbol{\Sigma}_{\boldsymbol{\eta}}\right| + \text{tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{-1}\left(\mathbf{S}_{11} - \mathbf{S}_{10}\tilde{\mathbf{G}}' - \tilde{\mathbf{G}}\mathbf{S}_{10}' + \tilde{\mathbf{G}}\mathbf{S}_{00}\tilde{\mathbf{G}}'\right)\right]. \end{aligned} \tag{11}$$

In Equation (11), $\mathbf{S}_{00}$, $\mathbf{S}_{10}$ and $\mathbf{S}_{11}$ are EM second moments (Shumway and Stoffer, 2010) and matrix $\boldsymbol{\Omega}_t$ is derived as follows:

$$\begin{aligned} \boldsymbol{\Omega}_t &= E\left[E\left(\mathbf{e}_t \mid \mathbf{Y}^{(1)}, \mathbf{Z}\right) E\left(\mathbf{e}_t \mid \mathbf{Y}^{(1)}, \mathbf{Z}\right)' + \text{Var}\left(\mathbf{e}_t \mid \mathbf{Y}^{(1)}, \mathbf{Z}\right) \mid \mathbf{Y}^{(1)}\right] \\ &= E\left[E\left(\mathbf{e}_t \mid \mathbf{Y}^{(1)}, \mathbf{Z}\right) E\left(\mathbf{e}_t \mid \mathbf{Y}^{(1)}, \mathbf{Z}\right)' \mid \mathbf{Y}^{(1)}\right] + \text{Var}\left(\mathbf{e}_t \mid \mathbf{Y}^{(1)}, \mathbf{Z}\right) \\ &= \mathbf{D}_t \begin{pmatrix} \boldsymbol{\Omega}_t^{(1)} & \boldsymbol{\Omega}_t^{(1)}\mathbf{R}_{11}^{-1}\mathbf{R}_{21} \\ \mathbf{R}_{12}\mathbf{R}_{11}^{-1}\left(\boldsymbol{\Omega}_t^{(1)}\right)' & \mathbf{R}_{21}\mathbf{R}_{11}^{-1}\boldsymbol{\Omega}_t^{(1)}\mathbf{R}_{11}^{-1}\mathbf{R}_{21} \end{pmatrix} \mathbf{D}_t' + \mathbf{D}_t \begin{pmatrix} \mathbf{0}_{u_t\times u_t} & \mathbf{0}_{u_t\times(n-u_t)} \\ \mathbf{0}_{(n-u_t)\times u_t} & \mathbf{R}_{22} \end{pmatrix} \mathbf{D}_t' \\ &= \mathbf{D}_t \begin{pmatrix} \boldsymbol{\Omega}_t^{(1)} & \mathbf{0}_{u_t\times(n-u_t)} \\ \mathbf{0}_{u_t\times(n-u_t)} & \mathbf{R}_{22} \end{pmatrix} \mathbf{D}_t' \end{aligned}$$

In the definition of $\mathbf{\Omega}_t$ and in the updating formulas in section 2.4, matrix $\mathbf{\Omega}_t^{(1)}$ is given by

$$\mathbf{\Omega}_t^{(1)} = E\left(\mathbf{e}_t^{(1)} \mid \mathbf{Y}^{(1)}\right) E\left(\mathbf{e}_t^{(1)} \mid \mathbf{Y}^{(1)}\right)' + \text{Var}\left(\mathbf{e}_t^{(1)} \mid \mathbf{Y}^{(1)}\right) \tag{12}$$

where

$$E\left(\mathbf{e}_t^{(1)} \mid \mathbf{Y}^{(1)}\right) = E\left(\mathbf{y}_t^{(1)} - \boldsymbol{\mu}_t^{(1)} \mid \mathbf{Y}^{(1)}\right) = \mathbf{y}_t^{(1)} - \mathbf{X}_t^{\boldsymbol{\beta},(1)}\boldsymbol{\beta} - \mathbf{X}_t^{\mathbf{z},(1)}\tilde{\mathbf{A}}\mathbf{z}_t^T \tag{13}$$

and

$$\text{Var}\left[\mathbf{e}_t^{(1)} \mid \mathbf{Y}^{(1)}\right] = \text{Var}\left[\mathbf{X}_t^{\mathbf{z},(1)}\tilde{\mathbf{A}}\mathbf{z}_t \mid \mathbf{Y}^{(1)}\right] = \tilde{\mathbf{A}}^2\mathbf{X}_t^{\mathbf{z},(1)}\mathbf{P}_t^T\left(\mathbf{X}_t^{\mathbf{z},(1)}\right)'. \tag{14}$$

Notice that, due to the hierarchical structure of HDG model, $Q\left(\Psi, \Psi^{\langle m\rangle}\right)$ in (11) is composed by three summands related to different parameters, say

$$Q\left(\Psi, \Psi^{\langle m\rangle}\right) = Q_1\left(\Psi_1, \Psi^{\langle m\rangle}\right) + Q_0\left((\mu_0, \Sigma_0), \Psi^{\langle m\rangle}\right) + Q_2\left(\Psi_2, \Psi^{\langle m\rangle}\right).$$

This simplifies the maximization step as the closed form updating formulas for $\Psi_1 = \left\{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \mathbf{G}\right\}$ can be derived by solving

$$\frac{\partial}{\partial \Psi_1} Q_1\left(\Psi_1, \Psi^{\langle m\rangle}\right) = \frac{\partial}{\partial \Psi_1} T \log|\mathbf{\Sigma}_\varepsilon| + \text{tr}\left(\mathbf{\Sigma}_\varepsilon^{-1}\sum_{t=1}^{T}\mathbf{\Omega}_t\right) = \mathbf{0}.$$

Finally the geostatistical parameters $\Psi_2 = (\mathbf{V}, \theta)$ are updated through numerical optimization. In particular

$$\left\{\mathbf{V}^{\langle m+1\rangle}, \theta^{\langle m+1\rangle}\right\} = \underset{\Psi_2}{\arg\max} Q_2\left(\Psi_2, \Psi^{\langle m\rangle}\right) = \underset{\mathbf{V},\theta}{\arg\max} T \log|\mathbf{\Sigma}_{\boldsymbol{\eta}}| + \text{tr}\left[\mathbf{\Sigma}_{\boldsymbol{\eta}}^{-1}\left(\mathbf{S}_{11} - \mathbf{S}_{10}\tilde{\boldsymbol{G}}' - \tilde{\mathbf{G}}\mathbf{S}_{10}' + \tilde{\mathbf{G}}\mathbf{S}_{00}\tilde{\mathbf{G}}'\right)\right]$$

as in (6).