Journal of the Royal Statistical Society
SERIES C Applied Statistics

Appl. Statist. (2013)
62, *Part* 2, *pp.* 287–308

A model-based framework for air quality indices and population risk evaluation, with an application to the analysis of Scottish air quality data

Francesco Finazzi,

University of Bergamo, Italy

E. Marian Scott

University of Glasgow, UK

and Alessandro Fassò

University of Bergamo, Italy

[Received July 2011. Final revision September 2012]

Summary. The paper is devoted to the development of a statistical framework for air quality assessment at the country level and for the evaluation of the ambient population exposure and risk with respect to airborne pollutants. The framework is based on a multivariate space–time model and on aggregated indices defined at different levels of aggregation in space and time. The indices are evaluated, uncertainty included, by considering both the model outputs and the information on the population spatial distribution. The framework is applied to the analysis of air quality data for Scotland for 2009 referring to European and Scottish air quality legislation.

Keywords: Air quality indices; Ambient exposure and risk; Multivariate space–time models; Unbalanced networks

1. Introduction

European legislation on air quality (Directive 2008/50/EC) identifies the needs for improved monitoring and assessment of air quality, including 'to provide information to the public'.

The aim of this paper is to provide a model-based statistical framework for air quality assessment and the evaluation of population exposure and risk in a national context and applied to Scotland. The air quality strategy for Scotland is based on the European Commission directives, so a further aim of this paper is to develop and use a national Scottish air quality index as a scientific tool and as a source of public information.

The framework is applied to observed air quality data for Scotland for 2009 to provide a retrospective analysis of air quality and its expected effect on population at the country level. Indeed, the ultimate role of air quality assessment should be, on the one hand, to evaluate whether any actions that are undertaken to improve air quality have been successful or not (see Scott (2007)) and, on the other, to provide information about population risk and exposure. In particular, the

Address for correspondence: Francesco Finazzi, Department of Information Technology and Mathematical Methods, University of Bergamo, viale Marconi 5, 24044 Dalmine BG, Italy.
E-mail: francesco.finazzi@unibg.it

 0035–9254/13/62287

main focus is on the definition of air quality, exposure and risk indices and on their evaluation at different levels of spatial and temporal aggregation. Each index represents a different aspect of pollution. The role of the air quality index is to provide concise information about the air quality level *per se* without any reference to the population exposure. However, exposure and risk indices consider also anthropological information and, hence, identify dangerous situations for the population's health.

The problem of assessing air quality over large regions and that of deriving indices is complex in several respects. The main complication arises from the way that airborne pollutants are measured in the field. The high economic costs of installation and maintenance of monitoring networks usually prevent pollutants from being measured with a spatial resolution that is adequate to assess exposure and risk with homogeneous accuracy all over the region.

The problem becomes more prominent when unbalanced monitoring networks are considered, i.e. when not all the pollutants are measured at each monitoring station. In such a case, it is not always clear how to define aggregated indices for the whole region and how to evaluate their uncertainty. Moreover, it is not straightforward to compare across years when, in each year, the structure of the monitoring network (sites included) and the quantity of missing data differ.

The above-mentioned problems are addressed by adopting, as the basis of the statistical framework, the dynamic coregionalization model (DCM) that has been introduced by Fassò and Finazzi (2011) and which can handle both unbalanced monitoring networks and missing data automatically. The DCM is used to evaluate the space–time correlation of the pollutants and their cross-correlation whereas the model output is used to define the indices and their uncertainty. In particular, the model output, in terms of the estimated pollutant concentrations, and the information about population distribution are combined to derive the exposure and risk indices.

In this work, the focus is on so-called ambient exposure rather than on personal exposure. A review of ambient exposure estimation methods can be found in Jerret *et al.* (2005), though limited to the intraurban case. A good example of personal exposure estimation, in contrast, can be found in Zidek *et al.* (2007), though it would be impractical to extend the personal exposure approach from city-size to country-size regions.

On the contrary, we aim to provide high resolution ambient exposure maps at the country level. Aware that this may introduce an ecological bias (according to Freedman (2001), 'The ecological fallacy consists in thinking that relationships observed for groups necessarily hold for individuals'), we point out that our approach is an improvement with respect to the current air quality legislation which is based only on temporal averages of the measured pollutant concentration at the monitoring stations.

The rest of the paper is organized as follows. Section 2 is dedicated to the DCM. Parameter estimation and space–time pollutant concentration mapping are discussed for multivariate data observed in a heterotopic configuration. In Section 3, the problem of defining aggregated air quality indices for state-size regions is introduced and a model derived from the DCM is considered. Exposure and risk assessment indices based on coupling population spatial distribution and model outputs are defined in Section 4. Air quality data for 2009 that were collected over Scotland are considered in Section 5 and analysed within the statistical framework that is developed in this work.

2. The dynamic coregionalization model

Hierarchical models represent a useful statistical approach for the analysis of environmental data

and they have been applied profitably in both frequentist and Bayesian contexts. Examples can be found in Banerjee *et al.* (2004) and Cressie and Wikle (2011) for the univariate and multivariate case. A comparison of different space–time hierarchical models in terms of prediction error is provided by Cameletti *et al.* (2011) though the comparison is limited to a particular data set. Since no general optimality results are available, we opt for the DCM which is flexible and easy to interpret. Moreover, as detailed in Finazzi and Fassò (2012), a software implementation of the DCM is readily available to download from `http://code.google.com/p/d-stem/`.

The DCM is a hierarchical multivariate space–time model based on latent variables which can handle both heterotopic data (non-colocated data) and missing data in a natural way. As a consequence, model estimation is based on the original data set as it has been acquired in the field without the need for any preliminary interpolation or missing data imputation.

Let $\mathbf{y}(\mathbf{s},t)=(y_1(\mathbf{s},t),\ldots,y_q(\mathbf{s},t))'$ be the q-dimensional data response vector at the spatial location $\mathbf{s}\in\mathcal{D}\subset\mathbb{R}^2$ and at time $t\in\mathbb{N}^+$. The general form of the model is

$$\mathbf{y}(\mathbf{s},t)=X(\mathbf{s},t)\boldsymbol{\beta}+K\,\mathbf{z}(t)+\boldsymbol{\gamma}\odot\mathbf{u}(\mathbf{s},t)+\boldsymbol{\delta}\odot\mathbf{w}(\mathbf{s},t)+\boldsymbol{\varepsilon}(\mathbf{s},t) \tag{1}$$

where $X(\mathbf{s},t)$ is a matrix of known covariates, $\boldsymbol{\beta}=(\boldsymbol{\beta}_1',\ldots,\boldsymbol{\beta}_q')'$ is a vector of global coefficients and '$\odot$' is the Hadamard product. The p-dimensional latent temporal state $\mathbf{z}(t)=(z_1(t),\ldots,z_p(t))'$ has the Markovian dynamics

$$\mathbf{z}(t)=G\,\mathbf{z}(t-1)+\boldsymbol{\eta}(t) \tag{2}$$

with G a stable transition matrix and Gaussian innovation $\boldsymbol{\eta}\sim N(0,\Sigma_\eta)$. The $q\times p$ matrix K is the loading matrix of known coefficients. The latent spatial component is modelled by both $\mathbf{u}(\mathbf{s},t)=(u_1(\mathbf{s},t),\ldots,u_q(\mathbf{s},t))'$ and $\mathbf{w}(\mathbf{s},t)=(w_1(\mathbf{s},t),\ldots,w_q(\mathbf{s},t))'$ which are independent and identically distributed over time. For each fixed t, $u_i(\mathbf{s},t)$, $1\leqslant i\leqslant q$, are independent latent zero-mean and unit variance Gaussian processes with spatial correlation function $\Gamma_i=\mathrm{cov}\{u_i(\mathbf{s},t),u_i(\mathbf{s}',t)\}=\rho_i(h,\boldsymbol{\theta}_i)$, where ρ_i is a valid correlation function parameterized by $\boldsymbol{\theta}_i$ and $h=\|\mathbf{s}-\mathbf{s}'\|$ is the Euclidean distance between s and s'. In contrast, $\mathbf{w}(\mathbf{s},t)$ is described by a q-dimensional linear coregionalization model (LCM) of c components

$$\mathbf{w}(\mathbf{s},t)=\sum_{j=1}^{c}\mathbf{w}^j(\mathbf{s},t) \tag{3}$$

where $\mathbf{w}^j(\mathbf{s},t)$, $1\leqslant j\leqslant c$, are independent latent zero-mean and unit variance Gaussian processes with correlation matrix function $\Gamma_j^{\mathrm{C}}=\mathrm{cov}\{w_i^j(\mathbf{s},t),w_{i'}^j(\mathbf{s}',t)\}=V_j\,\rho_j^{\mathrm{C}}(h,\boldsymbol{\theta}_j^{\mathrm{C}})$, $1\leqslant i,i'\leqslant q$, $1\leqslant j\leqslant c$. Each V_j is a correlation matrix and each ρ_j^{C} is, again, a valid correlation function (see Wackernagel (2003) for an introduction to the LCM).

The vectors $\boldsymbol{\gamma}=(\gamma_1,\ldots,\gamma_q)'$ and $\boldsymbol{\delta}=(\delta_1,\ldots,\delta_q)'$ contain the scale parameters. Finally, $\boldsymbol{\varepsilon}(\mathbf{s},t)=(\varepsilon_1(\mathbf{s},t),\ldots,\varepsilon_q(\mathbf{s},t))'$ is the measurement error which is assumed to be white noise in space and time. In particular, $\varepsilon_i(\mathbf{s},t)\sim N(0,\sigma_{\varepsilon,i}^2)$, $1\leqslant i\leqslant q$. The parameter set to be estimated is

$$\Psi=\{\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\delta},\boldsymbol{\sigma}_{\boldsymbol{\varepsilon}}^2;G,\Sigma_{\boldsymbol{\eta}};\boldsymbol{\theta};\boldsymbol{\theta}^{\mathrm{C}},\mathbf{V}\} \tag{4}$$

where $\boldsymbol{\sigma}_{\boldsymbol{\varepsilon}}^2=(\sigma_{\varepsilon,1}^2,\ldots,\sigma_{\varepsilon,q}^2)'$, $\boldsymbol{\theta}=(\boldsymbol{\theta}_1',\ldots,\boldsymbol{\theta}_q')'$, $\boldsymbol{\theta}^{\mathrm{C}}=((\boldsymbol{\theta}_1^{\mathrm{C}})',\ldots,(\boldsymbol{\theta}_c^{\mathrm{C}})')'$ and $\mathbf{V}=\{V_1,\ldots,V_c\}$.

The DCM defined in equation (1) is flexible in the sense that the covariates $X(\mathbf{s},t)$ and the latent variables $\mathbf{z}(t)$, $\mathbf{u}(\mathbf{s},t)$ and $\mathbf{w}(\mathbf{s},t)$ describe different aspects of the phenomenon under study and they can be included and excluded from the model depending on what is the focus of the data analysis. The covariates (if available) and the latent spatial variables $\mathbf{u}(\mathbf{s},t)$ and $\mathbf{w}(\mathbf{s},t)$ are to be included when the model is used for mapping whereas the latent temporal variable $\mathbf{z}(t)$ should be included when the focus is on the temporal dynamics.

The main difference between $\mathbf{u}(\mathbf{s},t)$ and $\mathbf{w}(\mathbf{s},t)$ is that each $u_i(\mathbf{s},t)$, $1\leqslant i\leqslant q$, is characterized by

its own correlation function $\rho_i(h, \boldsymbol{\theta}_i)$ whereas the LCM imposes, for each of the c components, a unique correlation function across variables. It can be said that $\mathbf{u}(\mathbf{s}, t)$ is the direct component whereas $\mathbf{w}(\mathbf{s}, t)$ is the interaction component. Although the model can be estimated with both the components included, preliminary studies suggest that, when real data are considered, it is important to choose one component or the other depending on the spatial correlation structure of the data. Indeed, the LCM should be included solely when data are known to be spatially cross-correlated whereas the direct component should be considered solely when the correlation functions describing the q-variables are expected to be different. In this latter case, it is still worthwhile to consider the model in its multivariate form since the q-variables may be temporally cross-correlated. Finally, if the flexibility of model (1) must be increased, a version of the DCM with spatiotemporal varying coefficients can be considered as detailed in Finazzi and Fassò (2011).

The matrix $Y = Y(\mathcal{S}, \mathcal{T}) = (Y(\mathcal{S}_1, \mathcal{T})', \ldots, Y(\mathcal{S}_q, \mathcal{T})')'$ is the $N \times T$ matrix of all the observations collected at locations $\mathcal{S} = \{\mathcal{S}_1, \ldots, \mathcal{S}_q\}$ and time $\mathcal{T} = \{1, \ldots, T\}$. Here, $\mathcal{S}_i$ is the collection of n_i locations where the variable y_i is observed and $N = \sum_{i=1}^{q} n_i$. The n_i are not constrained to be equal; nor are the $\mathcal{S}_i$. Thus, the fully heterotopic case is admitted, i.e. the variables can be observed over disjoint sets $\mathcal{S}_i$ of spatial locations. This includes the above-mentioned case of unbalanced monitoring networks. Also note that the matrix Y may include missing data.

The maximum likelihood estimate of Ψ is obtained by the expectation–maximization (EM) algorithm as described in Fassò and Finazzi (2011). The whole estimation procedure has been proven to be stable even when large data sets or large parameter sets are considered. This is largely due to the quasi-closed form property of the EM steps for this model. Note that the standard deviations of the estimated model parameters are directly obtained from an approximated Fisher information matrix $\hat{\mathcal{J}}$ computed as in Fassò *et al.* (2009).

Let $\hat{\Psi}$ be the maximum likelihood estimate of Ψ; then the concentration of the ith pollutant at a new set of sites $\mathcal{S}_0 \notin \mathcal{S}$ and time $t \in \mathcal{T}$ is evaluated by means of a plug-in approach as

$$\hat{\mathbf{y}}_i(\mathcal{S}_0, t) = X_i(\mathcal{S}_0, t)\hat{\boldsymbol{\beta}}_i + K_i\,\mathbf{z}^{\mathrm{T}}(t) + \hat{\gamma}_i\,\mathbf{u}_i^{\mathrm{T}}(\mathcal{S}_0, t) + \hat{\delta}_i\,\mathbf{w}_i^{\mathrm{T}}(\mathcal{S}_0, t) \tag{5}$$

where $\{\hat{\boldsymbol{\beta}}_i, \hat{\gamma}_i, \hat{\delta}_i\} \subset \hat{\Psi}$, $\mathbf{z}^{\mathrm{T}}(t) = E_{\hat{\Psi}}\{\mathbf{z}(t)|Y\}$ is the Kalman smoother output, $\mathbf{u}_i^{\mathrm{T}}(\mathcal{S}_0, t) = E_{\hat{\Psi}}\{\mathbf{u}_i(\mathcal{S}_0, t)|Y\}$ and $\mathbf{w}_i^{\mathrm{T}}(\mathcal{S}_0, t) = E_{\hat{\Psi}}\{\mathbf{w}_i(\mathcal{S}_0, t)|Y\}$ are the estimated latent spatial variables, $X_i(\mathcal{S}_0, t)$ is the matrix of covariates and K_i is again the loading matrix. Note that the Kalman smoother provides a fast algorithm for the evaluation of $E_{\hat{\Psi}}\{\mathbf{z}(t)|Y\}$ using the state space representation of model (1) (see for instance Shumway and Stoffer (2006)). In contrast, the conditional expectations of the latent spatial variables with respect to the observed data are evaluated through the usual formulae of the multivariate normal distribution adapted for the missing data case as detailed in Fassò *et al.* (2009).

The spatial prediction variance–covariance matrix of $\hat{\mathbf{y}}_i(\mathcal{S}_0, t)$ is given by

$$\Sigma_{\hat{\mathbf{y}}_i}(\mathcal{S}_0, t) = \mathrm{var}\{K_i\,\mathbf{z}(t) + \hat{\gamma}_i\,\mathbf{u}_i(\mathcal{S}_0, t) + \hat{\delta}_i\,\mathbf{w}_i(\mathcal{S}_0, t)|Y\}. \tag{6}$$

If the sites in $\mathcal{S}_0$ cover the whole region $\mathcal{D}$ as a fine regular grid, we call $\hat{\mathbf{y}}_i(\mathcal{S}_0, t)$ a map and the ordered collection

$$\hat{\mathbf{Y}}_i(\mathcal{S}_0) = \{\hat{\mathbf{y}}_i(\mathcal{S}_0, 1), \ldots, \hat{\mathbf{y}}_i(\mathcal{S}_0, T)\} \tag{7}$$

a dynamic map for the ith pollutant. If, instead of the set of sites $\mathcal{S}_0$, a tessellation $\mathcal{B}$ of the region $\mathcal{D}$ is considered, the change-of-support problem (see Gotway and Young (2002)) must be addressed and

$$\hat{y}_i(B, t) = E_{\hat{\Psi}}\left\{\frac{1}{|B|}\int_{\mathbf{s}\in B} y_i(\mathbf{s}, t)\,\mathrm{d}s|Y\right\} \tag{8}$$

must be evaluated for each block $B \in \mathcal{B}$. However, if the blocks in $\mathcal{B}$ are not too large with respect to the spatial variability scale of $y_i(\mathbf{s}, t)$, then $\hat{y}_i(B, t)$ can be replaced by $\hat{y}_i(\mathbf{s}^*, t)$, with $\mathbf{s}^*$ the centre of the pixel B. In what follows, the dependence on the estimated parameter set $\hat{\Psi}$ will be dropped to simplify the notation.

It should be noted that the dynamic map carries all the information about the temporal and the spatial dynamics of the ground level pollutant concentration. However, the amount of information is huge and it is rarely useful to decision makers. The following sections describe how aggregate information (uncertainty included) can be derived by taking advantage of the flexibility of the DCM and by considering both the estimated model latent variables and the estimated pollutant concentrations.

3. Global air quality indices

When environmental space–time data are considered, aggregation is often useful over either space or time. Obtaining aggregated results over time is usually straightforward as the data are usually collected at regular time steps. Spatial aggregation is more complex as the spatial sampling locations are irregularly sparse over the region. For these reasons, the focus of this section is on aggregation over space; in particular, the problem of defining global air quality indices with measures of uncertainty is addressed.

With a global air quality index, we refer here to a single number that can describe, for a region $\mathcal{D}$ at time t, the air quality in terms of the monitored pollutants. Hence, a global air quality index represents, on the one hand, a simple and concise reporting measure for the public and, on the other, a measure to compare different temporal periods with respect to air quality easily.

From a statistical point of view, the global air quality index is considered here as a latent variable which manifests itself through the pollutants' concentration measurements collected at the sampling sites. Although in a different context, the same idea has been developed by Chiu *et al.* (2011) in the definition of health factor indices.

To define the latent global air quality index, the following multivariate model is proposed:

$$\mathbf{y}(\mathbf{s}, t) = K(\mathbf{s})\,\mathbf{z}(t) + \varepsilon(\mathbf{s}, t), \tag{9}$$

which is a reduced version of model (1). This kind of model has been used in various environmental applications. For example, for large data sets, using the so-called fixed rank smoothing approach of Cressie and Johannesson (2008), the matrix $K(\mathbf{s})$ is defined by a set of fixed spatial basis functions. In ecological trend analysis, Zuur *et al.* (2007) used the so-called dynamic factor model to estimate the common trend of a non-large number of time series. To do this, the matrix $K(\mathbf{s})$ is estimated by using suitable constraints.

In our case, the global air quality index should represent a common trend at the monitoring stations but the number of times series can be high and the matrix $K(\mathbf{s})$ is poorly identifiable. Hence, we avoid identifiability problems by fixing $K(\mathbf{s})$ to agree with the station averages. The dimensionality of $\mathbf{z}(t)$ is equal to 1 or to the total number of pollutants q. The first case may be considered when the pollutants are highly positively correlated. The $\mathbf{z}(t)$ is hence unidimensional and the global air quality index can be defined as

$$I_1(t) = z^{\mathrm{T}}(t) \tag{10}$$

where $z^{\mathrm{T}}(t)$ is the estimated latent state output of the Kalman smoother. In the second case, when the pollutants are not positively correlated, it is better to rely on a q-variate $\mathbf{z}(t)$ and each pollutant retains its own temporal trend. Following the aggregation approaches of Bruno and Cocchi (2002) and Lee *et al.* (2011), two possible global air quality indices are

$$I_2(t) = \frac{1}{q} \sum_{i=1}^{q} z_i^{\mathrm{T}}(t), \tag{11}$$

$$I_3(t) = \max_{i=1,\ldots,q} \{z_i^{\mathrm{T}}(t)\} \tag{12}$$

where $z_i^{\mathrm{T}}(t)$ is the ith component of the estimated latent state $\mathbf{z}^{\mathrm{T}}(t)$ output of the Kalman smoother.

With regard to the uncertainty that is related to the above indices, this can be evaluated by considering the variance–covariance matrix $P^{\mathrm{T}}(t)$ related to $\mathbf{z}(t)$. In particular, the variance of $I_2(t)$ can be evaluated as

$$\mathrm{var}\{I_2(t)\} = \frac{1}{q^2} \sum_{i,j=1}^{q} p_{ij}(t)$$

where $p_{ij}(t)$ is the (i, j)th element of the matrix $P^{\mathrm{T}}(t)$. From $\mathbf{z}(t)|Y \sim N_q\{\mathbf{z}^{\mathrm{T}}(t), P^{\mathrm{T}}(t)\}$, a 95% confidence interval for $I_2(t)$ can be evaluated as $I_2(t) \pm 1.96\sqrt{\mathrm{var}\{I_2(t)\}}$. Confidence intervals for $I_3(t)$ do not have, in general, a simple closed form but they can be easily evaluated by considering the quantiles of $N_q\{\mathbf{z}^{\mathrm{T}}(t), P^{\mathrm{T}}(t)\}$.

When plotted against time, the indices I_1, I_2 and I_3 provide an immediate view of the air quality trend over the region considered. This allows comparison across days and years and to test whether air quality is either improving or worsening over time. From an epidemiological point of view, however, this kind of information is not sufficiently rich to derive conclusions about the potential effect of pollution on population health, which is the object of the next section.

4. Population exposure and risk assessment

Mapping pollutant concentration over space and time is important to identify critical areas with respect to air quality. To evaluate the potential effect of airborne pollution on population health, however, the spatial distribution of the population must also be considered. Hence, population exposure and population risk are evaluated by analysing the interaction between the spatial distributions of the pollutants and population.

4.1. Exposure index

Exposure and risk are related concepts and the respective indices may carry similar information. However, in particular contexts, exposure and risk might differ substantially. In a way, the risk index should be related to the chance of extreme events whereas the exposure index should be related to the number of people who are exposed to a given pollution level.

The exposure index for the ith pollutant, the block $B \in \mathcal{B}$ and the temporal frame $\tilde{\mathcal{T}} = \{t_1, \ldots, t_2\} \subset \mathcal{T}$ are defined here by

$$\varkappa_i(B, \tilde{\mathcal{T}}) = \bar{y}_i(B)\, d(B) \tag{13}$$

where

$$\bar{y}_i(B) = \frac{1}{t_2 - t_1 + 1} \sum_{t \in \tilde{\mathcal{T}}} \hat{y}_i(B, t)$$

is the estimated temporal average concentration of the ith pollutant and $d(B)$ is the (time invariant) population count of block B. In this case, we prefer to evaluate a temporally averaged index since, as said before, the exposure index should reflect the long-term effect. For instance, the set

$\tilde{\mathcal{T}}$ can represent a month, a whole season or a year. If needed, the exposure index $\varkappa_i(B, \tilde{\mathcal{T}})$ can be aggregated over space to define the following average exposure index for region $\mathcal{D}$:

$$\varkappa_i(\tilde{\mathcal{T}}) = \frac{1}{D} \sum_{B \in \mathcal{B}} \varkappa_i(B, \tilde{\mathcal{T}}) \tag{14}$$

where D is the total population count of region $\mathcal{D}$.

To evaluate how the spatial distributions of population and pollutant concentration interact, an interesting picture is provided by the cumulative exposure distribution, which is given by

$$\hat{H}_i(y) = \frac{1}{D} \sum_{B \in \mathcal{B}: \hat{y}_i(B,t) \leqslant y} d(B). \tag{15}$$

4.2. Risk index

The risk index is defined here by considering a concentration threshold L for a pollutant above which the effect on human health is known to be significant. The threshold L can be, for example, the concentration level which causes respiratory hospital admissions to increase with respect to a baseline rate. Given this threshold, two central quantities are the probability that the pollutant concentration y_i exceeds the L-threshold in block B and time t, namely

$$\pi_i(B,t) = P\{y_i(B,t) > L\}, \tag{16}$$

and the probability that the number of days for which L is exceeded in block B exceeds M, i.e.

$$\varpi_i(B) = P\{|\check{\mathcal{T}}_i(B)| > M\} \tag{17}$$

with $\check{\mathcal{T}}_i(B) \subset \tilde{\mathcal{T}}$ the set of days for which the exceedance occurs.

The following risk indices are considered:

$$r_i(B,t) = \pi_i(B,t)\, d(B), \tag{18}$$

$$\tilde{r}_i(B) = \varpi_i(B)\, d(B). \tag{19}$$

In particular, the risk index that is defined in equation (19) reflects the current air quality norm, which usually prescribes a maximum number of days that the pollutant concentration can exceed a threshold L. The risk index of equation (19) can be evaluated by noting that

$$|\check{\mathcal{T}}_i(B)| \sim \sum_{t \in \tilde{\mathcal{T}}} \mathrm{Be}\{\pi_i(B,t)\} \tag{20}$$

with $\mathrm{Be}(p)$ the Bernoulli random variable with parameter p. Since the sum of independent Bernoulli random variables with varying parameter p has no simple closed formula, expression (20) can be evaluated numerically. For instance, the distribution of the number of days $|\check{\mathcal{T}}_i(B)|$ can be evaluated by Monte Carlo simulation. In practice, the probabilities in equations (18) and (19) are computed by using the estimated parameter set $\hat{\Psi}$.

As a final remark, it is worth noting that the exposure and risk indices that were defined above are conditioned on the observed covariates and the estimated latent variables $\mathbf{u}$, $\mathbf{w}$ and $\mathbf{z}$. In other words, both indices must be applied only in retrospective analysis and they cannot be considered as characteristics of a particular spatial site $\mathbf{s} \in \mathcal{D}$ independent of time.

4.3. Exceedance probability evaluation

A key aspect in assessing risk is the evaluation of the exceedance probability $\pi_i(B,t) = P\{y_i(B,t) > L\}$. Although, for fixed $\hat{\Psi}$, $P\{\hat{y}_i(B,t) > L\}$ can be easily evaluated, it gives a conservative

estimate of $\pi_i(B,t)$ as it does not take into account any model misspecification error and $\hat{y}_i(B,t)$ is smoother than the real pollutant concentration $y_i(B,t)$. Moreover, we are interested in evaluating confidence intervals for $\pi_i(B,t)$, which are not provided by the dynamic kriging. For these reasons, the following procedure is considered.

(a) The leave one site out cross-validation technique is applied and the cross-validation residuals $e(\mathbf{s},t), \mathbf{s} \in S_i$, are considered. In particular, $e(\mathbf{s},t)$ is computed by using data $\mathbf{Y}_{-\mathbf{s}}$ which is the data matrix omitting all the data from site $\mathbf{s}$.
(b) Residuals are Studentized with respect to the dynamic kriging variance $\hat{\sigma}^2$, namely

$$\tilde{e}(\mathbf{s},t) = \frac{e(\mathbf{s},t)}{\hat{\sigma}(\mathbf{s},t)}, \qquad \mathbf{s} \in \mathcal{S}_i, \quad t \in \mathcal{T}. \tag{21}$$

(c) Considering all the Studentized residuals $\tilde{\mathcal{E}} = \{\tilde{e}(\mathbf{s},t) : \mathbf{s} \in \mathcal{S}_i, t \in \mathcal{T}\}$, their cumulative distribution function $F_{\tilde{\mathcal{E}}}$ is obtained by kernel smoothing.
(d) For each block B and time t, the exceedance probability is evaluated as

$$\pi_i(B,t) \cong 1 - F_{\tilde{\varepsilon}} \left\{ \frac{L - \hat{y}_i(\mathbf{s}^*,t)}{\hat{\sigma}(\mathbf{s}^*,t)} \right\} \tag{22}$$

with $\hat{y}_i(\mathbf{s}^*,t)$ the kriged pollutant concentration under the estimated model with parameter set $\hat{\Psi}$.

The cross-validation residuals are presumed to take into account the model misspecification error and they are characterized by a higher variance with respect to classical residuals. The transformation at step (b) of this procedure is not a real Studentization (Cook, 1982) since $\sigma^2(\mathbf{s},t)$ is not an estimate of the residual standard deviation $\sigma_e^2(\mathbf{s},t)$. However, $\hat{\sigma}^2(\mathbf{s},t) \propto \sigma_e^2(\mathbf{s},t)$ and the Studentization procedure is applied here to homogenize the model residuals which, on their own, are not homoscedastic with respect to space. Indeed, the cumulative distribution function $F_{\tilde{\mathcal{E}}}$ can be evaluated by considering all the Studentized residuals provided that they are white noise in both space and time. Finally, in step (d) the approximations $\hat{y}_i(B,t) \simeq \hat{y}_i(\mathbf{s}^*,t)$ and $\hat{\sigma}^2(B,t) \simeq \hat{\sigma}^2(\mathbf{s}^*,t)$ are considered negligible since it is assumed that $\mathcal{B} \ni B$ is a fine tessellation of the region $\mathcal{D}$.

To evaluate whether the probabilities given by expression (22) are reliable, we provide confidence intervals as follows. We sample from the asymptotic distribution of the estimated parameter set, $\hat{\Psi} \sim N(\hat{\Psi}, \hat{\mathcal{J}}^{-1})$, where $\hat{\mathcal{J}}$ is the approximated Fisher information matrix of Section 2. As a consequence, we obtain a collection of R parameter sets $\boldsymbol{\Psi} = (\Psi^{(1)}, \ldots, \Psi^{(R)})$. For each $\Psi^{(j)}$ we evaluate bootstrap replications of $\hat{y}_i(\mathbf{s}^*,t)$ and $\hat{\sigma}(\mathbf{s}^*,t)$ and we compute $\pi^{(j)}(B,t)$ by using equation (22). The confidence intervals are based on the sample quantiles of $\pi^{(1)}(B,t), \ldots, \pi^{(R)}(B,t)$.

5. Analysis of the Scottish air quality data for year 2009

The methodology that was discussed in the previous sections is applied here to Scottish air quality data for the year 2009. The aims are to obtain better insight into the spatiotemporal dynamics of the principal airborne pollutants, to evaluate population exposure and risk and to understand whether the air quality monitoring network is appropriate to answer the above questions or whether it should be strengthened in terms of number of monitoring stations and spatial distribution. The air pollution standards and the air quality objectives that are considered in the analysis are based on the 'Air quality standards (Scotland) regulations 2007 for the purpose of local air quality management'. A summary of the current standards and objectives can be found in Department for Environment, Food and Rural Affairs (2009).

This section is organized as follows. Section 5.1 describes the data considered in terms of pollutants, population distribution and covariates. The global air quality index for Scotland is evaluated in Section 5.2 whereas population exposure and risk indices are developed in Section 5.3.

5.1. Description of the data

The sources of data that are considered in this work are essentially three: the ground level concentration of airborne pollutants, measured by the Scottish automatic urban network, the population spatial distribution downloaded from the Oak Ridge National Laboratory and the meteorological covariates downloaded from the Nasa Global Modeling and Assimilation Office. Each source of data is characterized by a different spatial and temporal resolution, as described hereafter.

5.1.1. Pollutant concentrations

The Scottish automatic urban network provides hourly mean data on six main airborne pollutants, namely nitrogen dioxide (NO_2), ozone (O_3), carbon monoxide (CO), sulphur dioxide (SO_2) and particulate matters PM_{10} and $PM_{2.5}$. In this work, only the NO_2, O_3 and PM_{10} concentration data are considered since they are measured at sufficient monitoring stations to justify a space–time analysis. Moreover, the hourly mean data are averaged to work with daily data.

For the year 2009, the number of monitoring stations is 81. Each station measures only a subset of the three pollutants considered and missing data are possible, due to temporary breakdowns of either the station or the single measuring instrument. Days with less than 75% hourly data (18 h) are considered as days with missing data. The exact number of monitoring stations for each pollutant is reported in Table 1, showing that the network is unbalanced in the sense of Bodnar *et al.* (2008). The respective spatial distributions are reported in Figs 1(a)–1(c), from which it is clear that the monitoring stations are not evenly distributed over Scotland as they are mainly in the most populated areas.

5.1.2. Population distribution

The population distribution has a twofold role here. It is considered as a time invariant covariate and it is used to evaluate the exposure and risk indices that were discussed in Section 4. The Oak Ridge National Laboratory manages the LandScan^{TM} ambient population count database, currently updated to the year 2008 (see Bhaduri *et al.* (2007)). The database provides estimates of 24-h average population counts over the entire world with $30'' \times 30''$ resolution (approximately 1 km $\times$ 1 km). The population spatial distribution for Scotland is reported in

Table 1. Summary statistics of the pollutant concentration data for 2009

Pollutant	*Number of stations*	*Mean* ($\mu g\, m^{-3}$)	*Standard deviation* ($\mu g\, m^{-3}$)	*Missing* (%)
NO_2	66	32.19	23.35	12.7
O_3	10	55.80	18.99	12.1
PM_{10}	60	16.60	8.58	16.1

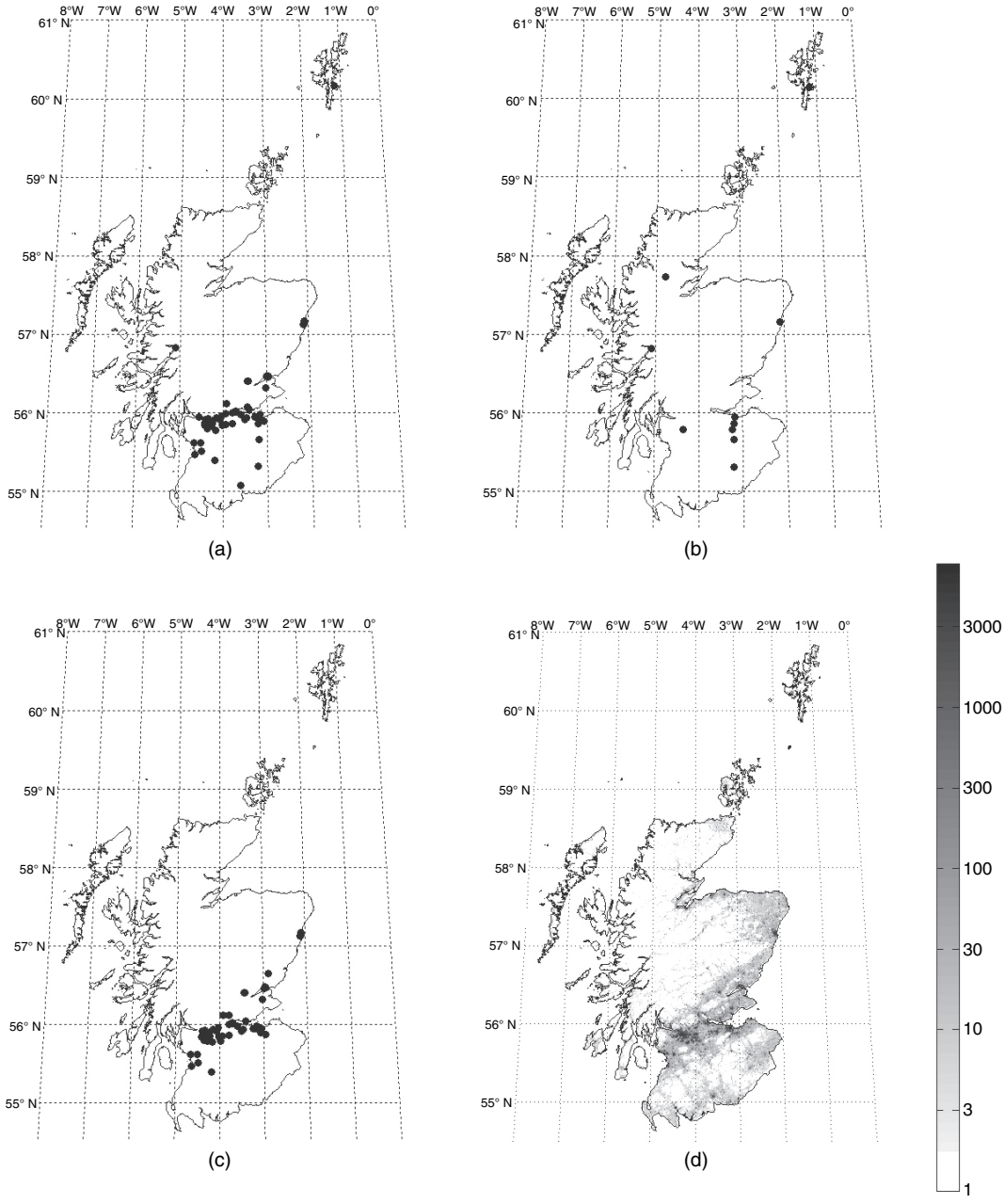

Fig. 1. Spatial distributions of the Scottish automatic urban network monitoring sites: (a) NO_2 (66 sites); (b) O_3 (10 sites); (c) PM_{10} (60 sites); (d) population spatial distribution

Fig. 1(d), from which it is clear that most of the population is in the central belt along the Glasgow–Edinburgh parallel.

5.1.3. Morphological and meteorological covariates

Pollutant concentrations are known to be related to some anthropological and meteorological covariates owing to the physical processes that drive the pollutant diffusion and advection. In this work, we consider population count pop, sea level pressure slp, temperature t, specific humidity sh, wind speed ws and boundary layer height blh. In particular, the boundary layer height is the height of the lowest part of the troposphere that is directly influenced by the ground and it determines the volume that is available for pollutants to disperse. Note that the population count is a good proxy of both the pollutant emissions and the site type (urban, suburban and rural). The meteorological covariates are downloaded from the Nasa Global Modeling and Assimilation Office. In particular, the 'Modern era retrospective analysis for research and applications' product (see Rienecker *et al.* (2011)) is considered, which is characterized by a temporal resolution of 1 h and a spatial resolution of $\frac{2}{3}^{\circ}$ longitude by $\frac{1}{2}^{\circ}$ latitude. Since the pollutant concentrations are daily averages, the meteorological covariates are also averaged over 24 h and they are interpolated at $30'' \times 30''$ resolution for mapping purposes.

As a final remark we point out that, though the concentration data might be preferentially sampled (see Diggle *et al.* (2010)), the problem is largely mitigated by the covariates and in particular by the population distribution which acts as a proxy for the pollutant emissions. Net of the covariates, the 'residual' data $Y(\mathcal{S}_i, t) - X(\mathcal{S}_i, t)\hat{\boldsymbol{\beta}}_i$ can be assumed to be not preferentially sampled with respect to the residual random field $y_i(\mathbf{s}, t) - \mathbf{x}(\mathbf{s}, t)\hat{\boldsymbol{\beta}}_i$, $\forall i, t$. Moreover, since the monitoring stations are not very high in number, we believe that the population distribution, which is available at high spatial resolution, is more informative than the network itself on the preferential sampling.

5.2. Global air quality index estimation

The methodology that was discussed in Section 3 is here considered to evaluate a global air quality index for Scotland for 2009. Since NO_2, O_3 and PM_{10} are known to have different temporal dynamics over the year, the global air quality index (12) is considered. With regard to the estimation of the latent temporal state $\mathbf{z}(t)$, model (9) is considered with

$$K(\mathcal{S}) = \begin{pmatrix} \dfrac{\bar{\mathbf{y}}_{NO_2}(\mathcal{S}_{NO_2})}{\bar{y}_{NO_2}} & 0 & 0 \\ 0 & \dfrac{\bar{\mathbf{y}}_{O_3}(\mathcal{S}_{O_3})}{\bar{y}_{O_3}} & 0 \\ 0 & 0 & \dfrac{\bar{\mathbf{y}}_{PM_{10}}(\mathcal{S}_{PM_{10}})}{\bar{y}_{PM_{10}}} \end{pmatrix} \tag{23}$$

where

$$\bar{\mathbf{y}}_i(\mathcal{S}_i) = T^{-1} \sum_{t=1}^{T} \mathbf{y}_i(\mathcal{S}_i, t)/F_i$$

is the temporal average scaled pollutant concentration at the sampling sites $\mathcal{S}_i$ for the ith pollutant, $i \in \{NO_2, O_3, PM_{10}\}$ whereas $\bar{y}_i = |\mathcal{S}_i|^{-1} \Sigma_{\mathcal{S}_i} \bar{\mathbf{y}}_i(\mathcal{S}_i)$ is the network average scaled pollutant concentration. The scaling factors F_i have a different role from that of the concentration thresholds L. The latter are usually provided by the air quality legislation as thresholds that should

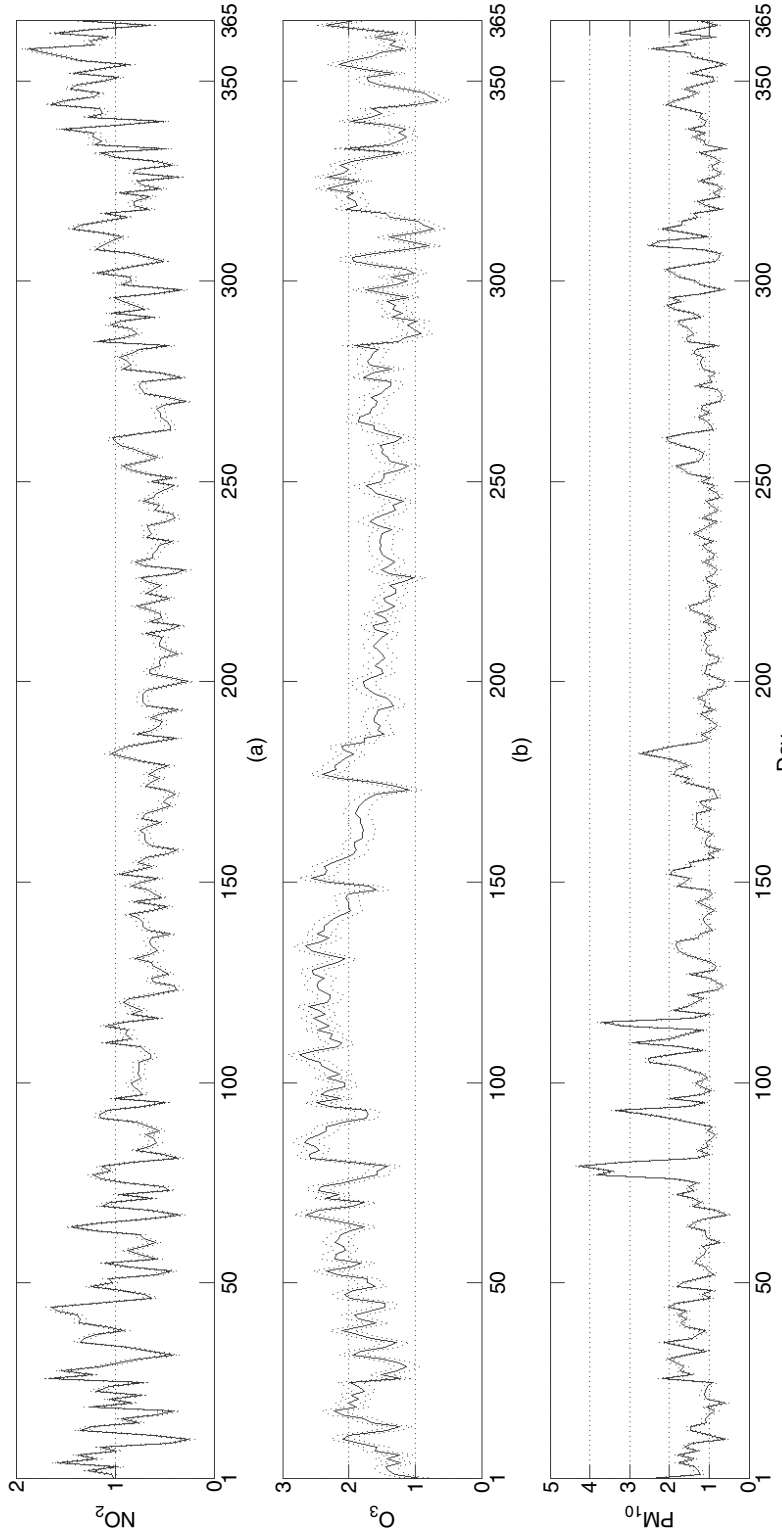

Fig. 2. Kalman smoother output $\mathbf{z}^T(t)$ of model (9) ($\cdots\cdots$, error bounds $z_i^T(t) \pm 2\sqrt{p_i(t)}$): (a) NO_2 component; (b) O_3 component; (c) PM_{10} component

not be exceeded whereas the former are used to homogenize the pollutants in the case that they are provided by using different statistics (average, running average, maximum etc.).

The UK air quality index and banding system (IBS) approved by the UK Committee on Medical Effects of Air Pollution Episodes is characterized by a 1–10-index divided into four bands, namely low, moderate, high and very high. Each index value corresponds to a range of concentration where the pollutant concentration can fall when measured over a period of time ΔT. The limits of each range depend on the particular pollutant as well as ΔT. For PM_{10}, an index value equal to 10 corresponds to a running 24-h mean concentration ${}_{24h}\bar{y}(\mathbf{s})$ equal to or higher than 128 μg m^{-3}. Although the data that are considered in this work are daily average concentrations rather than running 24-h means, it makes sense to consider $F_{PM_{10}} = 128/10 = 12.8$. The division by 10 is introduced to keep the index I_3 comparable with respect to the index of the IBS. The scaling factors for NO_2 and O_3 are not immediately available since the IBS prescribes ranges for the running 8-h mean ${}_{8h}\bar{y}(\mathbf{s})$ for O_3 and the hourly mean ${}_{1h}\bar{y}(\mathbf{s})$ for NO_2. Preliminary analysis not reported here suggests that

$$ {}_{24h}\bar{y}_{NO_2}(\mathbf{s}) \simeq \frac{1}{1.91}\max_{day}\{{}_{1h}\bar{y}_{NO_2}(\mathbf{s})\} $$

and

$$ {}_{24h}\bar{y}_{O_3}(\mathbf{s}) \simeq \frac{1}{1.15}\max_{day}\{{}_{8h}\bar{y}_{O_3}(\mathbf{s})\}. $$

The scaling factors are then chosen to be $F_{NO_2} = 764/19.1 = 40.0$ and $F_{O_3} = 360/11.5 = 31.3$, where 764 μg m^{-3} and 360 μg m^{-3} are the concentrations corresponding to an index value equal to 10 in the IBS for NO_2 and O_3 respectively.

Using the resulting loading matrix $K(\mathcal{S})$, model (9) is estimated by the EM algorithm and the related temporal component $\mathbf{z}^{\mathrm{T}}(t)$ is depicted in Fig. 2, where error bounds are defined as $z_i^{\mathrm{T}}(t) \pm 2\sqrt{p_i(t)}$, with $p_i(t)$ the ith diagonal element of $\mathrm{var}\{\mathbf{z}(t)|Y\}$. Note that each pollutant is characterized by a different temporal dynamic as expected. O_3 peaks in March or April whereas NO_2 peaks in winter. The particulate matter PM_{10} does not show a clear trend and the peaks are a consequence of unfavourable meteorological conditions. Fig. 3 shows the evaluated air quality index $I_3(t)$ which is representative of Scotland as a whole. By analysing the temporal series of $I_3(t)$, it can be concluded that, during the year 2009, air pollution over Scotland remained low with the exception of three events spread over 7 days during March and April. All the events can be associated with moderate concentration levels of PM_{10} due to adverse meteorological conditions. Note, moreover, that the decisive pollutants are O_3 and PM_{10} whereas NO_2 is identified in Fig. 3 only three times.

5.3. Population exposure and risk evaluation

As discussed in Section 4, the evaluation of population exposure and risk is based on coupling population spatial distribution and estimated pollutant concentrations obtained as output of the DCM. The DCM defined in equation (1) allows joint modelling of the space–time correlation of all the pollutants considered. However, to define the parametric structure of the multivariate model better, it is useful first to estimate one univariate model for each pollutant. In fact, the dimension p of the latent temporal state $\mathbf{z}(t)$, the inclusion of either $\mathbf{u}(\mathbf{s}, t)$ or $\mathbf{w}(\mathbf{s}, t)$ (or both) and the number c of coregionalization components must be decided before estimating the model. Although multivariate models can be compared by means of cross-validation techniques, the number of possible model parameterizations may be large and it is useful to

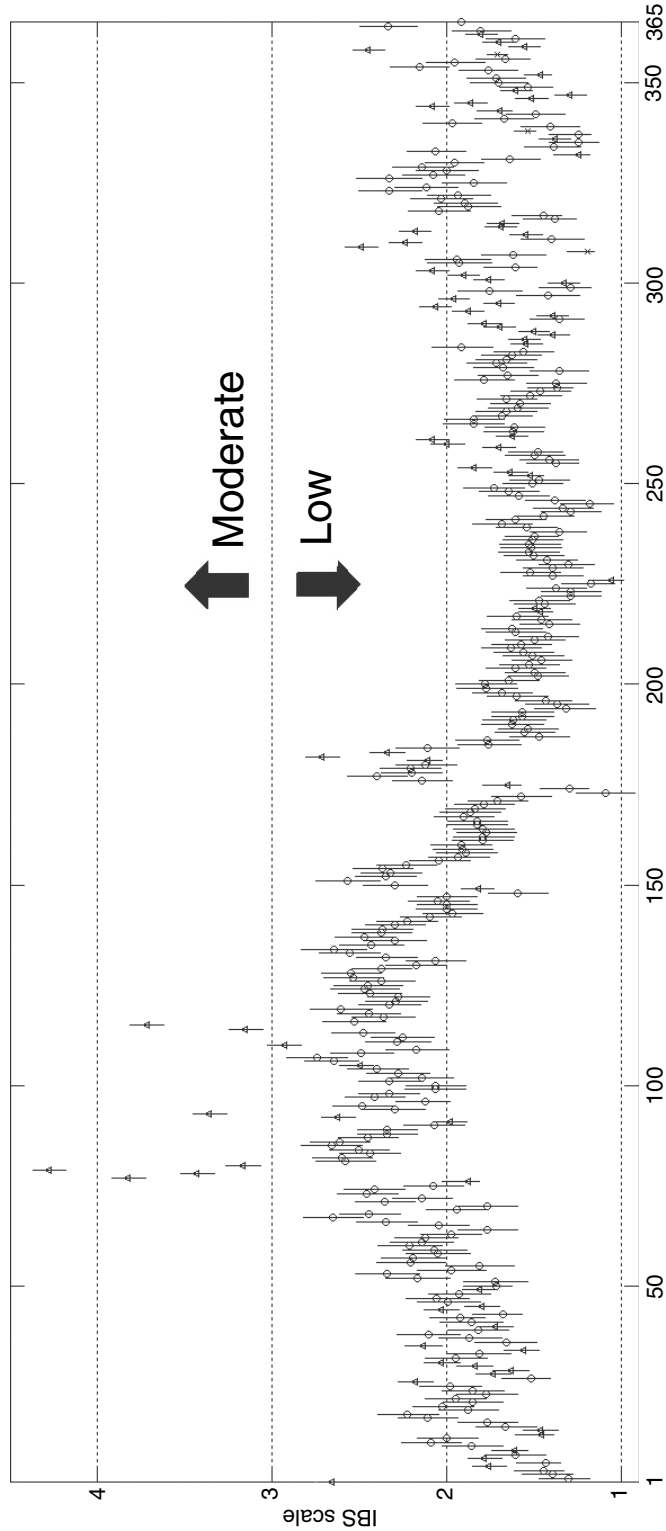

Fig. 3. Scotland daily air quality assessment through the air quality index I_3 for the year 2009 (the pollutant that gives rise to the maximum indicated by the marker): |, 95% confidence interval; x, NO_2; O, O_3; Δ, PM_{10}

consider the univariate models as a guide to define the parameterization of the multivariate model. As far as the spatial correlation structures is concerned, the exponential correlation function has been considered, namely $\rho(h, \boldsymbol{\theta}_i) = \exp(-h/\theta_i)$, $h \in \mathbb{R}^+$, $\theta_i \in \mathbb{R}^+$.

Table 2 reports the value of $\hat{\Psi}$ computed by means of the EM algorithm for each of the univariate models. All the variables and the covariates have been log-transformed and standardized. Standardization helps numerical stability and allows a direct comparison of the parameter values across pollutants.

By comparing the values of the estimated $\hat{\boldsymbol{\beta}}$-parameters with respect to their standard deviations, it can be seen that all the covariates are significant except $\hat{\beta}_{\text{slp}}$ for NO_2. The estimated parameters $\hat{\theta}_i$ are expressed in kilometres and they describe the strength of the spatial correlation of the latent variables $u_i(\mathbf{s}, t)$. In particular, in the case of the exponential correlation function, the spatial correlation between any two points in space is around 0.05 when their distance is $3\hat{\theta}_i$. Note that $\hat{\theta}_{O_3}$ has the highest value but also the highest standard deviation. This is because the O_3 monitoring network is very sparse and cannot capture a possible high spatial frequency of $u_{O_3}(\mathbf{s}, t)$. This consideration suggests that an LCM may be appropriate for modelling a spatial latent component that is common to all the pollutants, even if there is not enough evidence to conclude that the $\hat{\theta}_i$ are equal. In contrast, the $\hat{g}_i$-values are significantly different, suggesting that each pollutant should retain its own temporal dynamics; hence $p = 3$ for the multivariate model. The optimal number c of coregionalization components has been assessed through cross-validation, suggesting $c = 1$. The cross-validation mean-squared errors cmse that are obtained by applying the leave one site out technique are reported in Table 2.

Considering now the multivariate model (1), the estimated $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\sigma}}_{\boldsymbol{\epsilon}}^2$ and $\hat{\boldsymbol{\delta}}$ are reported in Table 3 whereas the remaining parameters are

$$\hat{G} = \begin{pmatrix} 0.969_{(0.041)} & -0.021_{(0.017)} & -0.014_{(0.027)} \\ -0.168_{(0.064)} & 0.873_{(0.027)} & 0.110_{(0.035)} \\ 0.275_{(0.118)} & 0.126_{(0.047)} & 0.579_{(0.052)} \end{pmatrix}, \tag{24}$$

Table 2. Estimated parameters for the univariate DCMs and respective cross-validation mean-squared error cmse

Pollutant	$\hat{\beta}_{\text{pop}}$	$\hat{\beta}_{\text{slp}}$	$\hat{\beta}_t$	$\hat{\beta}_{\text{sh}}$	$\hat{\beta}_{\text{ws}}$	$\hat{\beta}_{\text{blh}}$
NO_2	0.464	−0.040	0.321	−0.473	−0.197	−0.220
Standard deviation	0.005	0.021	0.065	0.055	0.015	0.017
O_3	−0.090	−0.229	0.392	−0.297	0.216	0.166
Standard deviation	0.016	0.026	0.082	0.069	0.018	0.021
PM_{10}	0.121	0.224	0.289	−0.294	−0.084	−0.222
Standard deviation	0.006	0.038	0.078	0.068	0.020	0.021
	$\hat{\sigma}_{\varepsilon}^2$	$\hat{g}$	$\hat{\sigma}_{\eta}^2$	$\hat{\gamma}$	$\hat{\theta}\,(km)$	*cmse*
NO_2	0.367	0.901	0.010	0.463	62.30	0.483
Standard deviation	0.003	0.038	0.001	0.010	4.28	
O_3	0.274	0.951	0.027	0.427	136.47	0.561
Standard deviation	0.014	0.019	0.002	0.020	21.96	
PM_{10}	0.286	0.674	0.137	0.516	88.36	0.366
Standard deviation	0.002	0.054	0.019	0.017	8.91	

Table 3. Subset of the estimated parameters for the multivariate DCM

Pollutant	$\hat{\beta}_{\text{pop}}$	$\hat{\beta}_{\text{slp}}$	$\hat{\beta}_t$	$\hat{\beta}_{\text{sh}}$	$\hat{\beta}_{\text{ws}}$	$\hat{\beta}_{\text{blh}}$	$\hat{\sigma}^2_\varepsilon$	$\hat{\delta}$	*cmse*
NO_2	0.447		0.309	−0.415	−0.192	−0.211	0.317	0.567	0.439
Standard deviation	0.005		0.056	0.015	0.018	0.017	0.004	0.010	
O_3	−0.166	−0.196	0.368	−0.251	0.208	0.188	0.243	0.451	0.478
Standard deviation	0.016	0.025	0.079	0.066	0.017	0.020	0.014	0.019	
PM_{10}	0.089	0.266	0.382	−0.285	−0.064	−0.227	0.244	0.571	0.339
Standard deviation	0.005	0.034	0.077	0.068	0.020	0.022	0.003	0.011	

$$\hat{\Sigma}_\eta = \begin{pmatrix} 0.006_{(0.002)} & 0.013_{(0.002)} & 0.011_{(0.004)} \\ & 0.063_{(0.005)} & -0.026_{(0.007)} \\ & & 0.170_{(0.016)} \end{pmatrix}, \tag{25}$$

$$\hat{\theta}_1^{\text{C}} = 40.991_{(2.446)}, \tag{26}$$

$$\hat{V}_1 = \begin{pmatrix} 1 & -0.792_{(0.058)} & 0.711_{(0.049)} \\ & 1 & -0.613_{(0.093)} \\ & & 1 \end{pmatrix} \tag{27}$$

with the standard deviations in parentheses. The $\hat{\beta}$-coefficients related to the covariates are all significantly different from 0 and their sign is in line with the physics of the pollution phenomenon. For instance, as expected, $\hat{\beta}_{\text{pop}}$ is positive for NO_2 and PM_{10} whereas it is negative for O_3, implying, on average, a lower concentration of O_3 in low populated areas. The analysis of the matrix $\hat{\Sigma}_\eta$, and in particular the variances on its diagonal, suggests a low contribution of the latent temporal variable $\mathbf{z}(t)$ in explaining the NO_2 and O_3 variability whereas the contribution is more relevant for PM_{10}. In contrast, the analysis of $\hat{G}$ reveals the slow temporal dynamics of NO_2 and O_3 (diagonal elements close to 1) which can be related to the temporal persistence of the pollutants not accounted for by the covariates. The parameter $\hat{\theta}_1^{\text{C}}$ of the spatial correlation function is common for all the pollutants considered and it is also expressed in kilometres. The matrix $\hat{V}_1$ shows that the component of $\mathbf{w}(\mathbf{s}, t)$ that is related to O_3 is negatively correlated with the components that are related to the other pollutants. Note that $\hat{\delta}_i \simeq 0.5$ for each pollutant, namely $\mathbf{w}(\mathbf{s}, t)$ accounts for about half of the data variability. By comparing the cmse-values reported in Tables 2 and 3, the gain in terms of prediction capability for the multivariate model can be appreciated. The reduction in cmse is particularly evident for O_3 which has the sparsest monitoring network and benefits more from the spatiotemporal correlation with the other pollutants. Note that this is a major benefit of using a multivariate model and in particular the DCM.

The estimated multivariate model is then used to evaluate population exposure and risk with respect to each pollutant. In particular, the dynamic maps $\hat{\mathbf{Y}}_{NO_2}(\mathcal{S}_0)$, $\hat{\mathbf{Y}}_{O_3}(\mathcal{S}_0)$ and $\hat{\mathbf{Y}}_{PM_{10}}(\mathcal{S}_0)$ are evaluated over the regular grid $\mathcal{S}_0$ with spatial resolution $30'' \times 30''$ within the Scottish boundaries.

As a first result, Fig. 4 shows the monthly and the yearly average population exposure evaluated by considering the exposure index (14), which can be related to an average Scottish person. Moreover, Fig. 5 displays, for each pollutant, the yearly average exposure distribution based on the cumulative distribution (15) and evaluated by considering a Gaussian kernel smoother with

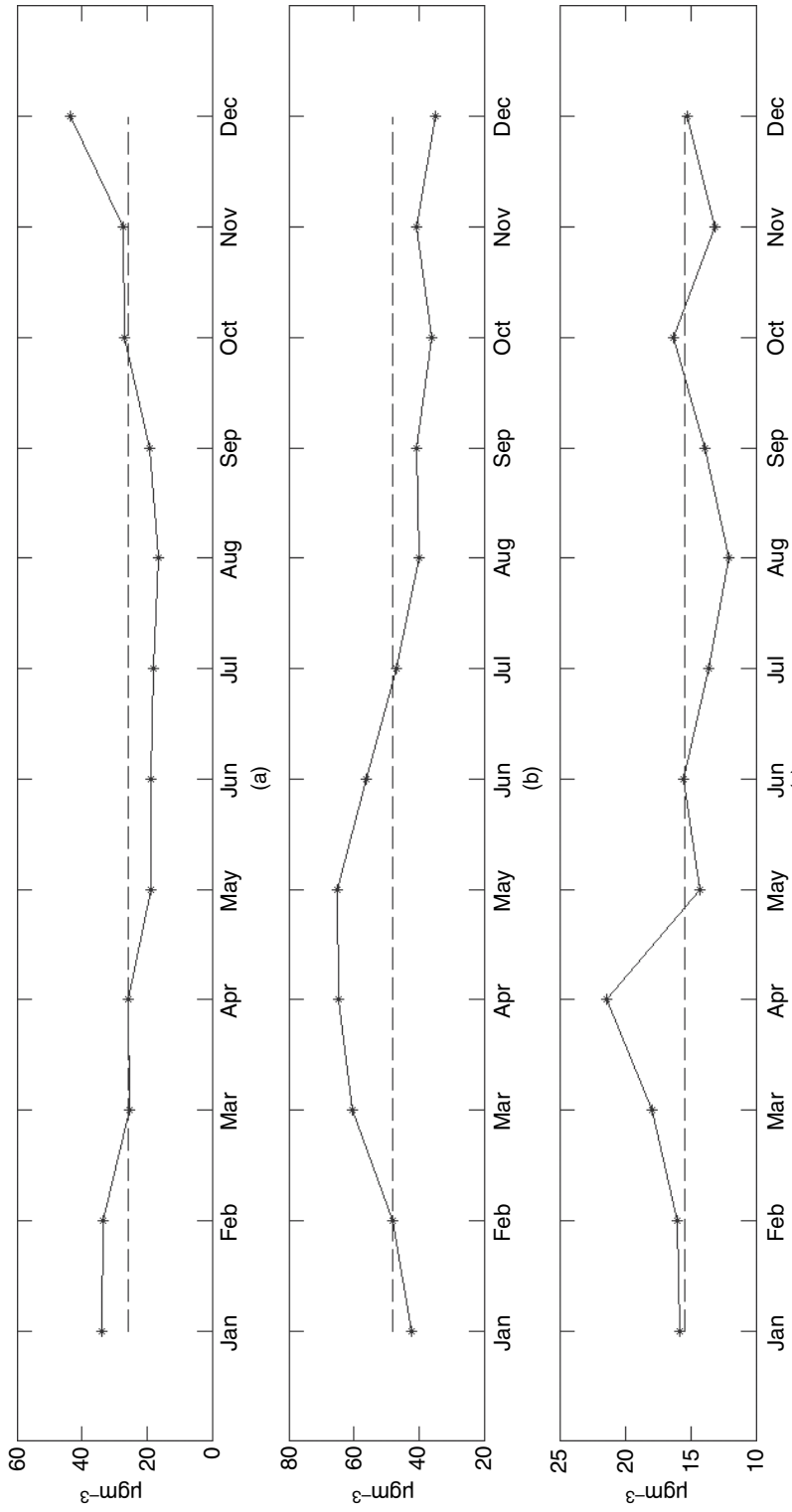

Fig. 4. Monthly and yearly (– – –) average population exposure: (a) NO_2; (b) O_3; (c) PM_{10}

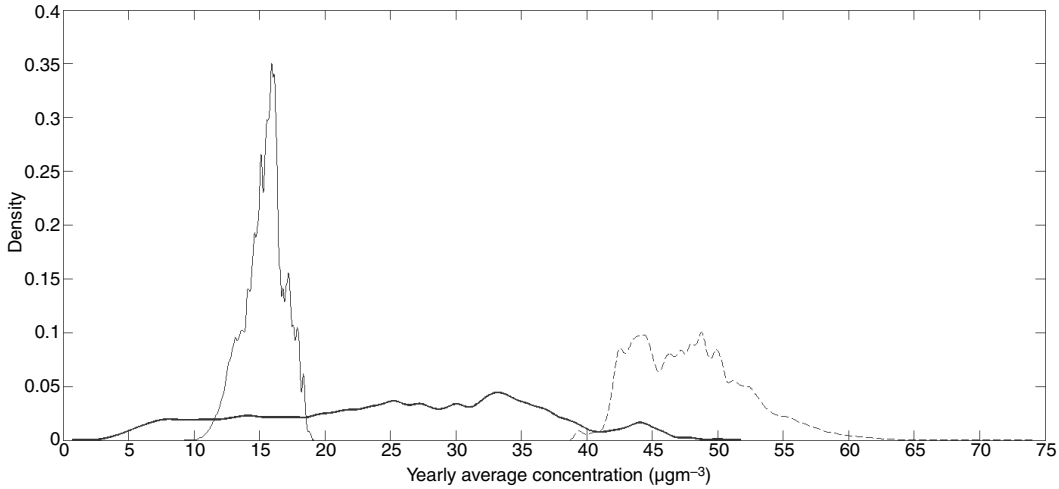

Fig. 5. Yearly average exposure distribution: ——, NO_2; – – –, O_3; ——, PM_{10}

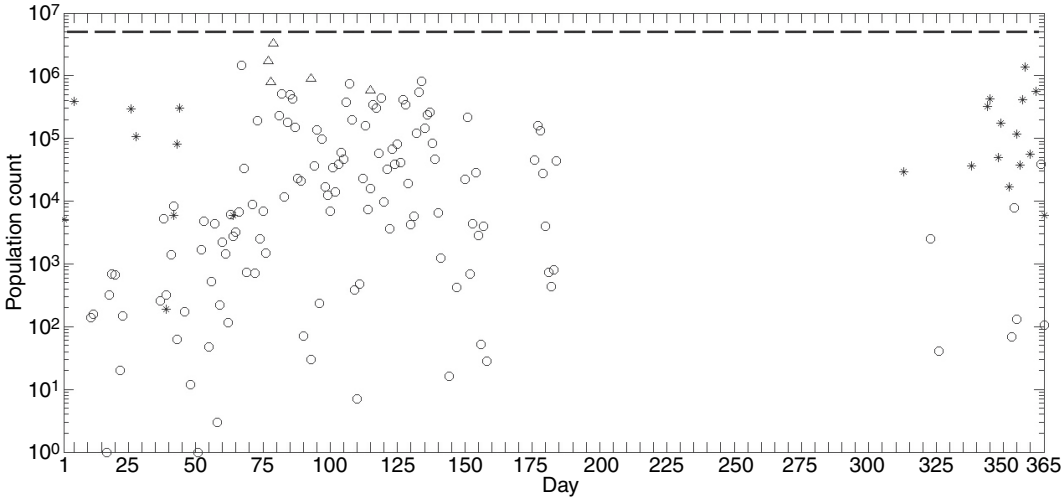

Fig. 6. Daily exposure time series (number of people exposed to a pollutant concentration exceeding the threshold): – – –, total population of Scotland; O, $O_3 > 87\ \mu g\ m^{-3}$; Δ, $PM_{10} > 50\ \mu g\ m^{-3}$; ∗, $NO_2 > 105\ \mu g\ m^{-3}$

bandwidth $0.5\,\mu g\ m^{-3}$. By looking at the results of Fig. 5, it can be noted that the exposure distributions differ greatly across pollutants. In particular, most of the Scottish people are exposed to the same yearly average PM_{10}-concentration $\pm 5\,\mu g\ m^{-3}$, whereas this is not true for NO_2 which is more spread out. Moreover, the exposure distribution of O_3 is characterized by a prominent right tail representing people living in rural areas, where the concentration of O_3 is higher. Although the graphs are reported on the same axis, they are not directly comparable in terms of health effects.

The daily exposure time series given by $D\{1 - \hat{H}_i(L)\}$ are reported in Fig. 6, where D is the total population count and $\hat{H}_i(\cdot)$ is given by equation (15). The thresholds L are 105, 87 and 50 $\mu g\ m^{-3}$ for NO_2, O_3 and PM_{10} respectively and they have been derived from the 'Air quality standards (Scotland) regulations 2007' following arguments similar to those of the previous

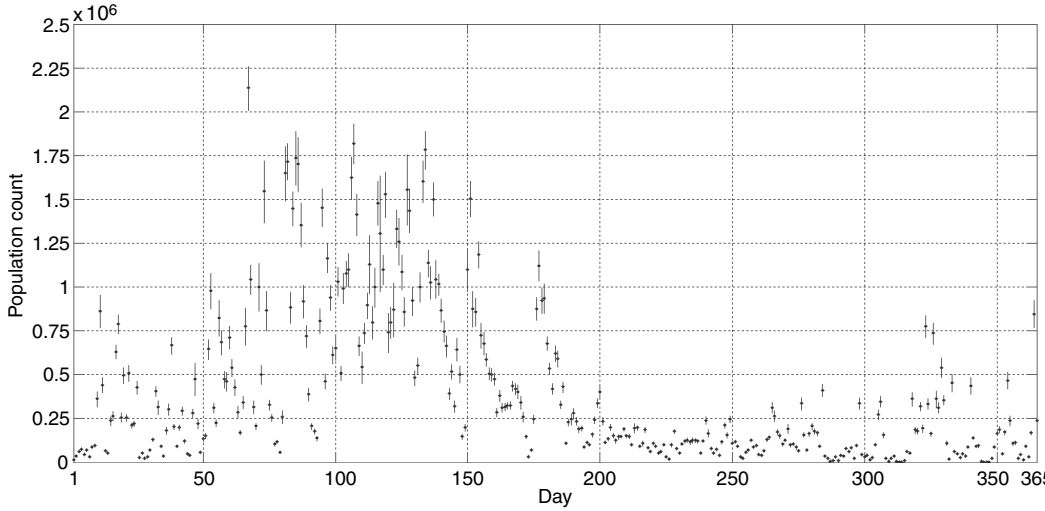

Fig. 7. Aggregated O_3 risk index and 95% confidence intervals (|) (with respect to $L = 87\ \mu g\ m^{-3}$)

paragraph. The time series disappear between day 190 and day 315 since the thresholds are never exceeded from July to October.

To compute the risk indices (18) and (19), the exceedance probabilities and their respective confidence intervals have been evaluated for each day and each pollutant at $30'' \times 30''$ of resolution. In particular, the confidence intervals have been obtained by considering $R = 100$ bootstrap replications as detailed in Section 4.3 and by analysing, for each map pixel and each day, the respective empirical cumulative distribution function.

The daily exceedance probabilities can be used for evaluating daily risk maps or aggregated risk indices. As an example, Fig. 7 shows the daily time series and the respective 95% confidence interval of the risk index for O_3 and $L = 87\ \mu g\ m^{-3}$ aggregated at the country level.

With regard to the number of days of exceedance, Fig. 8(a) reports the map of the average number of days of exceedance for PM_{10} whereas Fig. 8(b) reports the probability map that the threshold L has been exceeded for more than 7 days, which is required to evaluate the risk indicator that is defined in equation (19). Note that the 7-days limit represents one of the objectives of the Scotland national air quality strategy to be achieved by December 31st, 2010. Both the average number of days of exceedance and the probability of exceedance of the 7-days limit have been evaluated by considering the daily exceedance probability maps and by simulating from the distribution that is defined in equation (20). In particular, the results that are reported in Fig. 8(a) are based on 500 Monte Carlo simulation runs (see Section 4.2). The measures of uncertainty that are reported in Figs 8(c) and 8(d) have been obtained by repeating the same simulation approach for the $R = 100$ bootstrap replications of daily exceedance probability maps.

The probability that the threshold of $50\ \mu g\ m^{-3}$ has been exceeded for more than 7 days is higher in the Grampian region (north-east) and along the southern border of Scotland rather than in cities such as Glasgow or Edinburgh. However, this is a consequence of the fact that those regions are poorly covered by monitoring stations and the uncertainty in the estimated concentration of pollutant is high. The north-west regions are not covered as well but they are characterized by a very low PM_{10}-concentration and the exceedance probability is not so high despite the uncertainty. These last considerations suggest that the Scottish air quality network

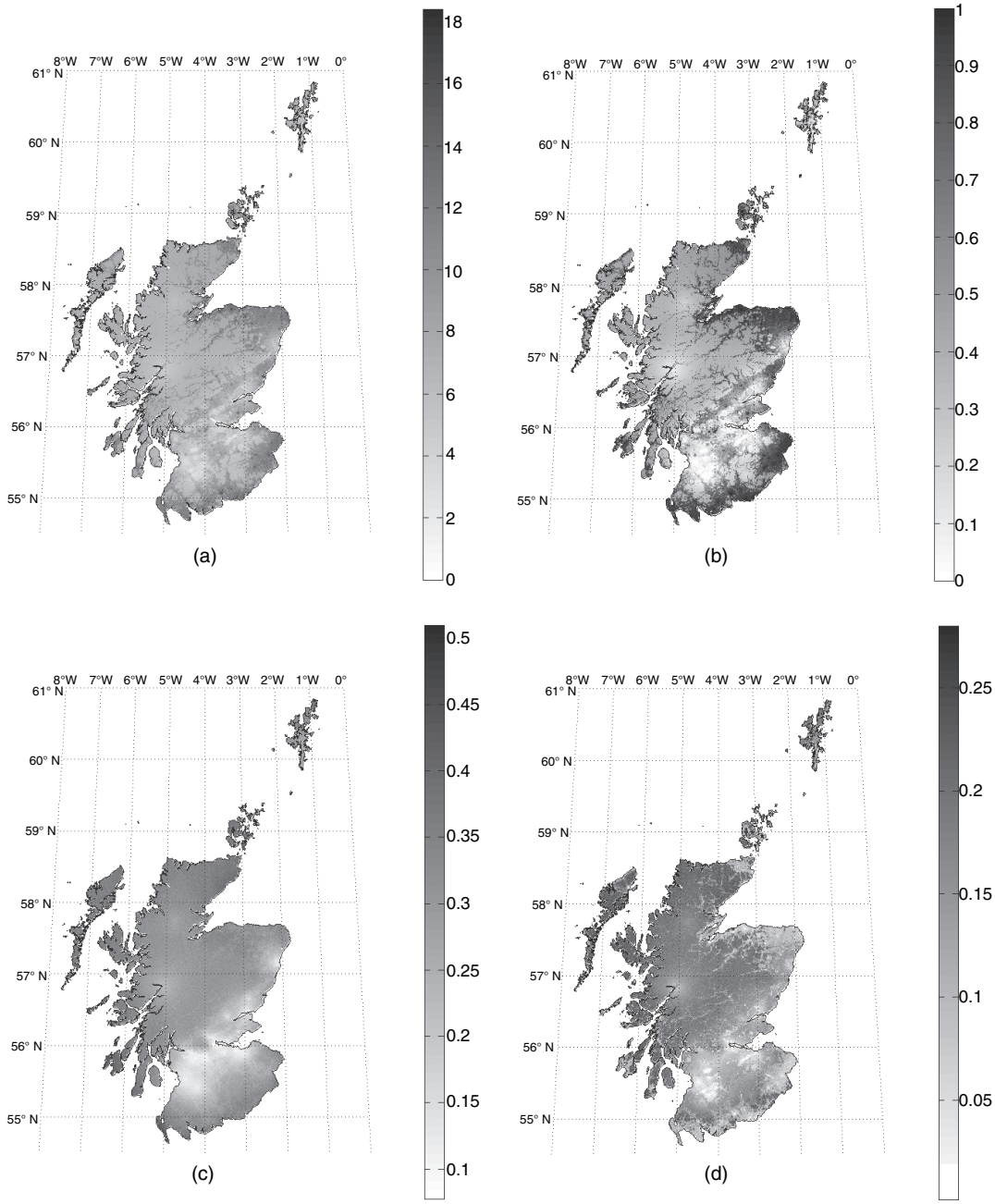

Fig. 8. (a) Map of the estimated average number of days of exceedances for PM_{10}, (b) map of the probability of exceedance of the 7-days limit, (c) standard deviation of map (a) and (d) range of the 95% confidence interval on map (b)

should be improved by installing additional monitoring stations in those areas which are uncovered and are characterized by a high risk and/or a high uncertainty. This is particularly true for O_3 that is currently monitored at only 10 locations.

6. Conclusions

Assessing air quality at the country level and evaluating the related population exposure are challenging tasks. In this paper, both problems have been addressed by developing a statistical framework based on a flexible multivariate space–time model, the DCM, and on a set of aggregated indices built by considering both the model output and the information of the population spatial distribution.

The DCM is sufficiently flexible to accommodate the data complexity related to ground level unbalanced networks and it naturally copes with the inevitable missing data problem. The indices are accompanied by measures of uncertainty and they can be provided at different levels of temporal and spatial aggregation to study different aspects of the pollution phenomenon. The global air quality indices allow us to compare different time periods with respect to the air quality at the country level easily and readily. In contrast, the exposure and risk indices provide an effective way to identify critical areas with respect to air quality and useful information to improve the ground level monitoring network.

As a whole, the statistical framework has been proven able to assimilate the current air quality legislation, both at the European and at the national levels, and to provide easily interpretable results for decision makers. The statistical framework has been successfully applied to the analysis of the Scottish air quality data for the year 2009, shedding light on aspects related to population exposure and risk that have never been investigated before.

Acknowledgements

The analysis was done utilizing the LandScan 2008^{TM} 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 in 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.

This research is part of project EN17 'Methods for the integration of different renewable energy sources and impact monitoring with satellite data', funded by Lombardia Region.

References

Banerjee, S., Carlin, B. and Gelfand, A. (2004) *Hierarchical Modeling and Analysis for Spatial Data*. Boca Raton: Chapman and Hall–CRC.

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

Bodnar, O., Cameletti, M., Fassò, A. and Schmid, W. (2008) Comparing air quality in Italy, Germany and Poland using bc indexes. *Atmosph. Environ.*, **42**, 8412–8421.

Bruno, F. and Cocchi, D. (2002) A unified strategy for building simple air quality indices. *Environmetrics*, **13**, 243–261.

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

Chiu, G. S., Guttorp, P., Westveld, A. H., Khan, S. A. and Liang, J. (2011) Latent health factor index: a statistical modeling approach for ecological health assessment. *Environmetrics*, **22**, 243–255.

Cook, R. (1982) *Residuals and Influence in Regression*. New York: Chapman and Hall.

Cressie, N. and Johannesson, G. (2008) Fixed rank kriging for very large spatial data sets. *J. R. Statist. Soc.* B, **70**, 209–226.

Cressie, N. and Wikle, C. (2011) *Statistics for Spatio-temporal Data*. Hoboken: Wiley.

Department for the Environment, Food and Rural Affairs (2009) Scottish Government local air quality management policy guidance 2009. Department for the Environment, Food and Rural Affairs. (Available from `http://www.scotland.gov.uk/topics/environment/waste-and-pollution/pollution-1/16215/pg09`.)

Diggle, P. J., Menezes, R. and Su, T. (2010) Geostatistical inference under preferential sampling (with discussion). *Appl. Statist.*, **59**, 191–232.

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

Fassò, A., Finazzi, F. and D'Ariano, C. (2009) Integrating satellite and ground level data for air quality monitoring and dynamical mapping. *Technical Report Graspa WP 34*. Gruppo di Ricerca per le Applicazioni della Statistica ai Problemi Ambientali, Bergamo. (Available from `www.graspa.org`.)

Finazzi, F. and Fassò, A. (2011) EM estimation of the dynamic coregionalization model with varying coeffcients. In *Proc. Spatial 2: Spatial Data Methods for Environmental and Ecological Processes, Foggia, Sept. 1st–2nd*, 2nd edn (ed. B. Cafarelli). Bergamo: Gruppo di Ricerca per le Applicazioni della Statistica ai Problemi Ambientali. (Available from `www.graspa.org`.)

Finazzi, F. and Fassò, A. (2012) D-STEM—a statistical software for multivariate space-time environmental data modeling. In *METMA 6: Proc. Int. Wrkshp Spatio-temporal Modelling* (eds A. M. Goncalves, I. Sousa, L. Machado, P. Pereira, R. Menezes and S. Faria).

Freedman, D. (2001) Ecological inference and the ecological fallacy. In *International Encyclopedia of the Social and Behavioral Sciences*, pp. 4027–4030. Amsterdam: Elsevier.

Gotway, C. and Young, L. (2002) Combining incompatible spatial data. *J. Am. Statist. Ass.*, **97**, 632–648.

Jerret, M., Arain, A., Kanaroglou, P., Beckerman, B., Potoglou, D., Sahsuvaroglu, T., Morrison, J. and Giovis, C. (2005) A review and evaluation of intraurban air pollution exposure models. *J. Expos. Anal. Environ. Epidem.*, **15**, 185–204.

Lee, D., Ferguson, C. and Scott, E. M. (2011) Constructing representative air quality indicators with measures of uncertainty. *J. R. Statist. Soc.* A, **174**, 109–126.

Rienecker, M. M., Suarez, M. J., Gelaro, R., Todling, R., Bacmeister, J., Liu, E., Bosilovich, M. G., Schubert, S. D., Takacs, L., Kim, G.-K., Bloom, S., Chen, J., Collins, D., Conaty, A., da Silva, A., Gu, W., Joiner, J., Koster, R. D., Lucchesi, R., Molod, A., Owens, T., Pawson, S., Pegion, P., Redder, C. R., Reichle, R., Robertson, F. R., Ruddick, A. G., Sienkiewicz, M. and Woollen, J. (2011) Merra—NASA's modern-era retrospective analysis for research and applications. *J. Clim.*, **24**, 3624–3648.

Scott, E. M. (2007) Setting and evaluating the effectiveness of environmental policy. *Environmetrics*, **18**, 333–343.

Shumway, R. and Stoffer, D. (2006) *Time Series Analysis and Its Applications, with R Examples*. New York: Springer.

Wackernagel, H. (2003) *Multivariate Geostatistics: an Introduction with Applications*, 2nd edn. New York: Springer.

Zidek, J., Shaddick, G., Meloche, J., Chatfield, C. and White, R. (2007) A framework for predicting personal exposures to environmental hazards. *Environ. Ecol. Statist.*, **14**, 411–431.

Zuur, A., Ieno, E. N. and Smith, G. M. (2007) *Analysing Ecological Data*. New York: Springer.