ORIGINAL ARTICLE

A dynamic structural equation approach to estimate the short-term effects of air pollution on human health

Dani Gamerman[1] | **Luigi Ippoliti**[2] | **Pasquale Valentini**[2]

[1]Departamento de Métodos Estatísticos, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil

[2]Department of Economics, University G. d'Annunzio, Chieti-Pescara, Pescara, Italy

Correspondence
Luigi Ippoliti, Department of Economics University G. d'Annunzio, Chieti-Pescara, Pescara, Italy.
Email: luigi.ippoliti@unich.it

Abstract

Detailed knowledge on the effects of air pollutants on human health is a prerequisite for the development of effective policies to reduce the adverse impact of ambient air pollution. However, measuring the effect of exposure on health outcomes is an extremely difficult task as the health impact of air pollution is known to vary over space and over different exposure periods. In general, standard approaches aggregate the information over space or time to simplify the study but this strategy fails to recognize important regional differences and runs into the well-known risk of confounding the effects. However, modelling directly with the original, disaggregated data requires a highly dimensional model with the curse of dimensionality making inferences unstable; in these cases, the models tend to retain many irrelevant components and most relevant effects tend to be attenuated. The situation clearly calls for an intermediate solution that does not blindly aggregate data while preserving important regional features. We propose a dimension-reduction approach based on latent factors driven by the data. These factors naturally absorb the relevant features provided by the data and establish the

link between pollutants and health outcomes, instead of forcing a necessarily high-dimensional link at the observational level. The dynamic structural equation approach is particularly suited for this task. The latent factor approach also provides a simple solution to the spatial misalignment caused by using variables with different spatial resolutions and the state-space representation of the model favours the application of impulse response analysis. Our approach is discussed through the analysis of the short-term effects of air pollution on hospitalization data from Lombardia and Piemonte regions (Italy).

KEYWORDS

Bayesian dynamic factor models, cardiovascular and respiratory diseases, environmental epidemiology, impulse response analysis, spatio-temporal models, structural equation models

1 | INTRODUCTION

It is well known that exposure to high levels of air pollution can cause a variety of adverse effects on human health. Although air quality has been generally improved over the last decades, many recent epidemiological studies have consistently shown positive associations between low-level exposure to air pollution and health outcomes. Hence, adverse health effects of air pollution, even at relatively low levels, remain a public concern.

The statistical literature on health care research is very rich and includes a plethora of models referring to different types of study designs. Most of these studies are usually based on time series studies, developed both in single and multisites frameworks (e.g. Dominici et al., 2002; Blangiardo et al., 2019; Peng et al., 2006), cohort studies (e.g. Dockery et al., 1993; Lepeule et al., 2012) and areal unit studies (e.g. Bruno et al., 2016; Lee & Sahu, 2016; Lee & Shaddick, 2010; Lee et al., 2014). However, because air pollution concentrations vary at fine spatio-temporal scales, quantifying the impact of air pollution appears more as an inherently spatio-temporal problem (see, for example, Greven et al., 2011; Knorr-Held, 2000; Lawson et al., 2012; Shaddick & Zidek, 2015).

The statistical literature on spatio-temporal designs, however, is sparse and, to the best of our knowledge, only a few papers consider the short-term impact of single pollutants on single diseases (see, for example, Choi et al., 2009; Liu et al., 2020). Only a limited number of papers consider the joint long-term effects of multiple air pollutants on single diseases (see, for example, Huang et al., 2018; Rushworth et al., 2014). As a result, this paper intends to contribute to the existing literature on spatio-temporal models by assessing the short-term effects of multiple air pollutants on multiple health outcomes. We mainly do this by exploiting the advantages of factor models, which cope with many sites and variables via dimension reduction, and impulse response analysis, which describes the temporal health effects of a given impulse on pollution.

Short-term effects are usually estimated by regressing daily counts of disease cases against air pollution concentrations using ecological regression models. However, directly modelling

multivariate, spatially and temporally referenced data, poses challenges to statistical accuracy, model interpretability and computational complexity. Aggregating the information over space, or time, simplifies the study but it fails to recognize important regional or temporal interval differences and runs into the well-known risk of confounding the effects. Here, as an intermediate solution, we introduce a Bayesian *Generalized Spatio-Temporal Structural Equation* model (GST-SEM) which we feel is appealing on several grounds.

For example, it extends to the space-time framework the models in Liu et al. (2005) and Blangiardo et al. (2019), which have been developed in a pure spatial or temporal setting. Also, by using a fully Bayesian framework, where the uncertainty is propagated through all model components, it overcomes the shortcomings of the two-step approach of Huang et al. (2018).

Furthermore, we avoid the direct use of the health regression model for estimating the impact of multiple correlated pollutants simultaneously. In practice, our model enables the modelling of the temporal relationship between dependent (health data) and regressor (air pollution) variables in a latent space. The observed processes can thus be described by a potentially small set of common dynamic latent factors with the advantage of overcoming the difficulties related to collinearity and low signal-to-noise ratio issues usually found with multiple pollutants. As discussed in Dominici et al. (2010), Wilson et al. (2018), Huang et al. (2018) and Blangiardo et al. (2019), this represents a challenge for scientific research and air quality management.

By modelling the spatial variation via spatially structured factor loadings, we entertain the possibility of identifying clusters of spatial sites that share common temporal trends of risk. Also, from time series studies, it is known that an important scientific objective is to understand how the risk of hospitalization propagates over multiple days in the future for an increase of air pollution (Peng et al., 2009). In the following, we show that our model allows for the application of impulse response analysis at different levels of the hierarchy and that this facilitates the study of the distribution of the risk of hospitalization both at global (i.e. whole area of interest) and local (i.e. specific areal unit) scales for a given impulse over some or all the common latent factors summarizing the pollutants.

Lastly, our model offers a simple approach suitable for handling variables and indicators measured at different scales and coming from independent sources. Health outcomes are typically collected over a discrete, areal spatial domain whereas pollution levels are collected over a continuous, pointwise spatial domain. Linking the two different types of spatial resolution poses the well-known spatial misalignment problem (Gotway & Young, 2002). A naíve solution for this problem involves pre-specification of sets of point-level measurements to be related to areal measurements (e.g. point level pollution measurements from a single areal unit are linked only to health measurements from the same areal unit). Other approaches, based for example on block average calculations, are discussed in Banerjee and Gelfand (2002) and Gelfand et al. (2001). Ren and Banerjee (2013) also use factor analysis with spatial misalignment but their factors vary with space and their misalignment is restricted to a single spatial resolution, designed mainly to handle varying-location settings.

Our approach is simple as the relations between health data and pollutants are established at a latent level, where the effect of space is marginalized. Therefore no pre-specified choices of locations or block average calculations are required.

The remainder of the paper is organized as follows. In Section 2, we provide a brief description of the motivating example and the data used in this study. In Section 3, we introduce the statistical model while Section 4.1 provides details on prior specification with particular reference to the estimation of the factor loading matrices and common factors. In Sections 4.2 and 4.3, we

consider further inferential issues and related computational problems and in Section 4.4 we discuss possible uses of the model. In Section 5, we provide results from a simulation study and in Section 6 we fit the model to real data. Finally, Section 7 concludes the paper with a discussion.

2 | MOTIVATING EXAMPLE

This paper is motivated by the study of the association between urban pollution and hospital admissions observed in Lombardia and Piemonte regions (Italy) in 2011. The hospital admissions are aggregated by Health Authorities at regional level and, for a given health outcome, daily time series are constructed by cause and by summing the number of emergency admissions in each district. These data provide the basis for challenging modelling issues and, to our knowledge, this is the first time that a statistical analysis has been conducted on them.

Pollution data refer to daily-average concentration levels of CO, NO_2, PM_{10} and O_3. Throughout the paper, these variables are denoted as X_l, and the entire set of information for these data, observed at a specific monitoring site, $\mathbf{u}_j$, and time point t, is denoted by $\left\{ X_l(\mathbf{u}_j, t), \mathbf{u}_j \in \mathcal{D}_x \subseteq \mathcal{R}^2 \right\}$, $l = 1, \dots, n_x; j = 1, \dots, N_{x_l}; t = 1, \dots, T$, where n_x is the number of available pollutants, N_{x_l} is the number of sites available for each variable X_l, and T is the length of the time series.

It is common practice in spatial statistics to treat these variables as geostatistical data and, as such, the X_l are collected at a number of monitoring sites[1] within a region of interest $\mathcal{D}_x$. A relevant issue here is that, since the variables only share some of the sample locations in which they have been observed, we work under a *partial heterotopy* framework (Wackernagel, 2003). The number of the monitoring sites (N_{x_l}) available for each pollutant is 94 for CO, 168 for NO_2, 96 for PM_{10} and 94 for O_3. The percentage of missing data is reasonably small and is equal to 2.53% for CO, 3.02% for NO_2, 2.94% for PM_{10} and 2.61% for O_3. We assume that the missing-data mechanism is completely at random and we let the model estimate missing data by borrowing information across different variables.

Our model exploits the correlation between pollutants, temporally and spatially, due to the processes by which they are formed. For example, PM_{10} as well as the other pollutants considered here are driven by combustion processes and, in particular, diesel combustion. CO is a toxic gas emitted as a result of combustion processes which, in urban areas, are almost entirely from road traffic emissions, as is the nitrogen dioxide NO_2. Associations between the levels of pollutants are also due to their relations with meteorological conditions. The top panel of Figure 1a (see the Supplementary Material, Section A) shows the heat maps of correlation coefficients computed between pairs of pollutant time series, while the bottom panel shows the average of correlation coefficients within and between pollutants. The figure shows that, for many pairs of time series, the correlations are larger than 0.8. On average, the within-group correlations are around 0.7 for CO, NO_2, PM_{10} and 0.83 for O_3. The between-group correlations also tend to be moderately high and show that O_3 is negatively correlated with the other pollutants. A deeper inspection of the correlations suggests that the values decrease with distance between the monitors and that the rate of decay of the correlations depends on the type of pollutant (see Figure 2a of the Supplementary Material, Section A).

As regards health data, we work with population level summaries collected over time in a fixed study area, $\mathcal{D}_y$. Henceforth, the complete set of information for the health data will be denoted

---

[1]The monitoring networks are managed by the environmental agencies, ARPA, of Piemonte and Lombardia.
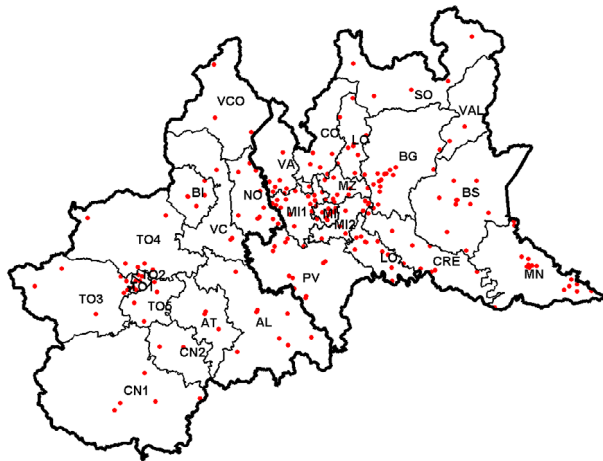
FIGURE 1 Map of the 28 districts and the complete monitoring network (dots). The thicker border in the middle separates Piemonte from Lombardia. The label within each district represents the acronymous of the Aziende Sanitarie Locali [Colour figure can be viewed at wileyonlinelibrary.com]

as $\left\{Y_k(\mathbf{s}_i, t), \mathbf{s}_i \in \mathcal{D}_y\right\}$, $k = 1, \ldots, n_y$; $i = 1, \ldots, N_y$; $t = 1, \ldots, T$, where n_y is the number of considered diseases, and N_y is the number of areal units, $\mathbf{s}_i$, within $\mathcal{D}_y$. For our case study, health data consist of counts of daily hospital admissions for cardiovascular disease (ICD-9, 390-429) and all respiratory diseases (ICD-9:460-519) in 2011. The two health outcomes (i.e., $n_y = 2$) are available at $N_y = 28$ districts (i.e. Aziende Sanitarie Locali, ASL) and refer to citizens with age 65 or older. Note that the classification of the diseases is consistent with the International Classification of Disease proposed in the APHENA (European and North American Approach) project (Katsouyanni et al., 2009).

It is worth noticing in this study that, while all the series are aligned in time, pollution and disease data result spatially misaligned because the geographical scales at which the data are measured are different (i.e. $\mathcal{D}_x$ and $\mathcal{D}_y$ have different spatial structures). In particular, pollution concentrations appear spatially sparse and are not available in all the areal units at which the disease data are recorded. For our case study, this feature is clearly shown in Figure 1 where both the monitoring sites and the 28 ASL of Piemonte and Lombardia are displayed.

Numerous studies have found positive associations between air pollution and hospital admissions for cardiovascular and respiratory diseases. The effect of air pollution to health in a population is usually represented by a concentration-response function, which is typically based on Relative Risk (RR) estimates. In practice, the differences in the population sizes and demographics between areal units are accounted for by computing the expected number of disease $E_k(\mathbf{s}_i, t)$ which is computed by external standardisation, using age and gender specific respiratory admissions rates for Italy. Based on the pair $(Y_k(\mathbf{s}_i, t), E_k(\mathbf{s}_i, t))$, a simple estimate of disease risks in areal unit $\mathbf{s}_i$, and time t, is represented by the *standardized morbidity ratio* (SMR), which is defined as the ratio $Y_k(\mathbf{s}_i, t)/E_k(\mathbf{s}_i, t)$, $k = 1, 2$.

Figure 3a (see the Supplementary Material, Section A) shows the map of correlation coefficients (ordered according to the longitude of the districts) computed between pairs of SMR time series, both for cardiovascular and respiratory diseases. About 60% of the correlations within the cardiovascular disease is larger than 0.45. The correlations for the respiratory disease are smaller;
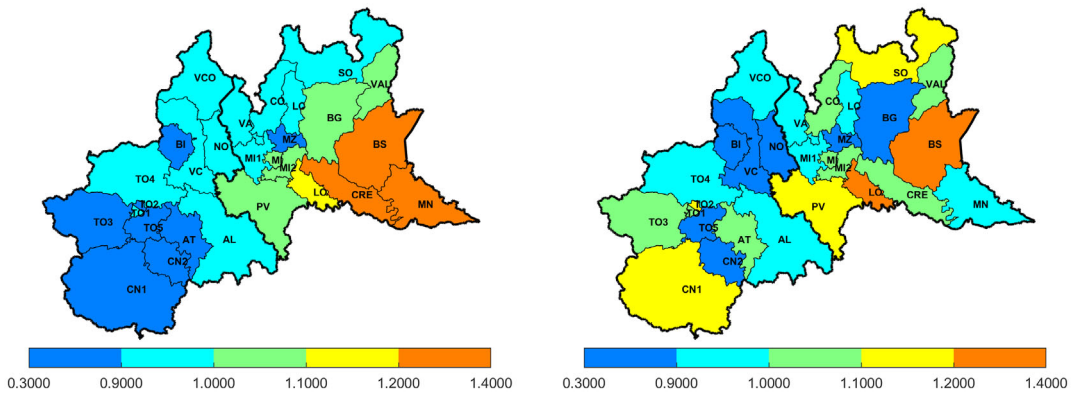
FIGURE 2   Map of the standardised morbidity ratio (SMR) for hospital admissions due to cardiovascular (left) and respiratory (right) diseases in 2011. The lable within each district represents the acronymous of the Aziende Sanitarie Locali [Colour figure can be viewed at wileyonlinelibrary.com]

only about 10% is larger than 0.45 and this is especially true for pairs of districts within Lombardia. Considering the between-disease correlations, the percentage of values above 0.45 is equal to 25% and, again, it refers to pairs of ASL within Lombardia region.

To provide more insight on the way in which the disease risks spread out to surrounding districts, Figure 2 below shows maps of raw standardized morbidity ratios obtained by averaging the SMR values across time. In general, the SMR maps show evidence of localized spatial clusters and suggest that, on average, Lombardia appears as the region most at risk. However, the strength of cross-sectional dependence is difficult to predict visually as the maps show evidence of some pairs of neighbouring areal units exhibiting similar risks and others having quite different values. Further analysis also gives an indication of overdispersion in some districts, especially in Lombardy region. In fact, as shown in Figure 4a (see the Supplementary Material, Section A), both for cardiovascular and respiratory diseases, the means and the variances of hospital admissions computed (across time) at each district are quite different.

Providing a spatio-temporal model useful to describe the spatio-temporal variability of the disease risks and its association with multiple pollutants, while accounting for the spatial misalignment and the presence of potential confounding factors that could bias the results of the analysis, is the objective of the next sections.

3 │ THE MODEL

This section introduces the GST-SEM model as a tool for modelling the spatio-temporal association existing between air pollution and hospital admissions. The model is developed within the framework of Structural Equation Modelling (Bollen, 1989) which provides a very general and convenient approach for multivariate statistical analysis.

There has been a growing literature developing different kinds of linear and nonlinear structural equation models and estimation methods for them. Our modelling approach outlines a *measuremt (response) model* for a vector of observed variables and a *structural model* for a corresponding set of latent variables. In particular, conditional on these latent variables, the

measurement model is a generalized linear model specified by a linear predictor, a link, and a distribution from the exponential family. The structural model, instead, reflects the dynamics of the latent variables through a set of regression and auto-regression equations. With the modelling of the (multivariate) spatial and temporal dependence shown by the data, the GST-SEM model can be viewed as a spatio-temporal generalization of the static structural equation model of Joreskog (1981), though many other dynamic spatial models could be specified by imposing zero restrictions on some model parameter matrices.

Assume that Y and X are two multivariate spatio-temporal processes observed at temporal instants $t = 1, 2, \ldots, T$ and generic locations, $\mathbf{s} \in \mathcal{D}_y$ and $\mathbf{u} \in \mathcal{D}_x$ respectively. Assume also that X is a predictor of Y, which thus represents the process of interest. For the two different processes, the spatial sites $\mathbf{s}$ and $\mathbf{u}$ can denote the same location but, in general, they need not be the same. Furthermore, $\mathcal{D}_y$ and $\mathcal{D}_x$ usually refer to different spatial domains.

Let n_y be the number of observed variables for Y and n_x the number of observed variables for X. The most informative case is represented by the isotopic configuration where, for each multivariate process, Y or X, all variables are 'collocated' and measured at all their respective sites. In this case, let $\mathbf{Y}(\mathbf{s}, t) = [Y_1(\mathbf{s}, t), \ldots, Y_{n_y}(\mathbf{s}, t)]'$ be the vector of the n_y values of Y at site $\mathbf{s}$ and time t. Equivalently, we write $\mathbf{X}(\mathbf{u}, t) = [X_1(\mathbf{u}, t), \ldots, X_{n_x}(\mathbf{u}, t)]'$ for the vector of the n_x values of X at site $\mathbf{u}$ and time t. The opposite case is the completely heterotopic case (Wackernagel, 2003) where not all the variables can be observed at the same site − this is especially true for X in our study. For the sake of simplicity, we use here the notation for the isotopic case, though our modelling approach allows to cope with the heterotopic case. Accordingly, the n_y variables of Y are observed at the same sites $\mathbf{s}_i$, $i = 1, \ldots, N_y$ and the n_x variables of X are observed at sites $\mathbf{u}_j$, $j = 1, \ldots, N_x$.

Let $\tilde{n}_y = n_y \times N_y$ and $\tilde{n}_x = n_x \times N_x$. At a specific time t, by using a site ordering, the $(\tilde{n}_y \times 1)$ and $(\tilde{n}_x \times 1)$ dimensional spatial processes are denoted as $\mathbf{Y}(t) = [\mathbf{Y}(\mathbf{s}_1, t)', \ldots, \mathbf{Y}(\mathbf{s}_{N_y}, t)']'$ and $\mathbf{X}(t) = [\mathbf{X}(\mathbf{u}_1, t)', \ldots, \mathbf{X}(\mathbf{u}_{N_x}, t)']'$. However, the data may also be ordered by variable. In this case, we write $\mathbf{Y}(t) = [\mathbf{Y}_1(t)', \ldots, \mathbf{Y}_{n_y}(t)']'$ and $\mathbf{X}(t) = [\mathbf{X}_1(t)', \ldots, \mathbf{X}_{n_x}(t)']'$, where $\mathbf{Y}_k(t)$ is the vector of N_y observations for variable Y_k, and $\mathbf{X}_l(t)$ is the vector of N_x observations for variable X_l. The two orderings are connected by a suitable permutation matrix. The variable ordering is often more convenient as it clearly shows the relevant contribution of each variable and the joint contributions of pairs of variables. Most of the models for regional data have been given in this form and, henceforth, unless differently stated, a variable ordering is assumed.

The GST-SEM is a hierarchical model with first level measurement equations for the conditionally independent variables,

$$Y_k(\mathbf{s}, t) | \eta_{y_k}(\mathbf{s}, t), \sigma^2_{y_k} \overset{ind}{\sim} F_y\left(\eta_{y_k}(\mathbf{s}, t), \sigma^2_{y_k}\right), \qquad k = 1, \ldots, n_y$$

$$X_l(\mathbf{u}, t) | \eta_{x_l}(\mathbf{u}, t), \sigma^2_{x_l} \overset{ind}{\sim} F_x\left(\eta_{x_l}(\mathbf{u}, t), \sigma^2_{x_l}\right), \qquad l = 1, \ldots, n_x,$$

where $\sigma^2_{y_k}$ and $\sigma^2_{x_k}$ are dispersion parameters. In general, the distributions F_y and F_x are allowed to be from any exponential family distribution. By choosing appropriate canonical link functions, the specification of the first level is completed with the specification of the following linear predictors

$$g_y[\eta_{y_k}(\mathbf{s}, t)] = \mu_{y_k}(\mathbf{s}, t) + \phi_{y_k}(\mathbf{s}, t)$$

$$= \mu_{y_k}(\mathbf{s}, t) + \sum_{i=1}^{m} h_{y_k, i}(\mathbf{s}) \, f_{y, i}(t) \qquad (1)$$

$$g_x[\eta_{x_l}(\mathbf{u}, t)] = \mu_{x_l}(\mathbf{u}, t) + \phi_{x_l}(\mathbf{u}, t)$$

$$= \mu_{x_l}(\mathbf{u}, t) + \sum_{i=1}^{r} h_{x_l,i}(\mathbf{u}) \, f_{x,i}(t), \quad (2)$$

where $\mu_{y_k}(\mathbf{s}, t)$ and $\mu_{x_l}(\mathbf{u}, t)$ are fixed effect terms representing the large-scale spatio-temporal variability of the processes, and $\phi_{y_k}(\mathbf{s}, t)$ and $\phi_{x_l}(\mathbf{u}, t)$, are random effects introduced to capture any residual spatio-temporal autocorrelation.

Equations (1) and (2) show that the random effect terms can be rewritten as truncated expansions in which, $h_{y_k,i}(\mathbf{s})$ and $h_{x_l,i}(\mathbf{u})$ are factor loadings of variables Y_k and X_l, and $f_{y,i}(t)$ and $f_{x,i}(t)$ are corresponding common factors. In general, we expect that $m << \tilde{n}_y$ and $r << \tilde{n}_x$. These equations also state that while the factor loadings are functions of space, the common factors are temporally dependent. Note also that, in matrix form, Equations (1) and (2) can be rewritten as

$$g_y \left[\boldsymbol{\eta}_y(t) \right] = \boldsymbol{\mu}_y(t) + \mathbf{H}_y \, \mathbf{f}_y(t), \quad g_x \left[\boldsymbol{\eta}_x(t) \right] = \boldsymbol{\mu}_x(t) + \mathbf{H}_x \, \mathbf{f}_x(t)$$

where $\mathbf{H}_y$ and $\mathbf{H}_x$ are $(\tilde{n}_y \times m)$ and $(\tilde{n}_x \times r)$ matrices of factor loadings.

In the second level of the hierarchy we model the dynamics of the common factors through the following equations

$$\mathbf{f}_x(t) = \sum_{i=1}^{s} \mathbf{D}_i \mathbf{f}_x(t - i) + \mathbf{v}_x(t) \quad (3)$$

$$\mathbf{f}_y(t) = \sum_{i=1}^{p} \mathbf{B}_i \mathbf{f}_y(t - i) + \sum_{i=0}^{q} \mathbf{C}_i \mathbf{f}_x(t - i) + \mathbf{v}_y(t) \quad (4)$$

where $\mathbf{B}_i$, $\mathbf{C}_i$ and $\mathbf{D}_i$ are coefficient matrices of dimension, $(m \times m)$, $(m \times r)$ and $(r \times r)$, respectively. Both $\mathbf{v}_x(t)$ and $\mathbf{v}_y(t)$ are zero mean normal error processes, i.e. $\mathbf{v}_x(t) \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{v_x})$ and $\mathbf{v}_y(t) \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{v_y})$.

Model completion requires specific forms for $\boldsymbol{\mu}_y(t)$ and $\boldsymbol{\mu}_x(t)$. The simplest specification of the mean components assumes the form of a linear regression function to take care of the effects of confounders, that is $\mu_{x_l}(\mathbf{u}, t) = \sum_{i=1}^{c} \sum_{j=0}^{g} \beta_{x_l,ij}(\mathbf{u}) \, z_i(\mathbf{u}, t - j)$, and $\mu_{y_k}(\mathbf{s}, t) = \sum_{i=1}^{c} \sum_{j=0}^{g} \beta_{y_k,ij}(\mathbf{s}) \, z_i(\mathbf{s}, t - j)$, where $z_i(\cdot, t)$, $i = 1, \dots, c$ are observed covariates or components representing seasonal and long-term trends introduced to take care of the effects of unmeasured confounders (Peng et al., 2006; Shaddick & Zidek, 2015). Note that the $z_i(\cdot, t)$ could also be smoothed versions of measured confounders represented by natural cubic splines with specified degree of freedom (see, for example, Bob et al., 2013).

Equations (1)–(4) give the complete formulation of the GST-SEM. More insight on the relationship exiting between the two sets of variables can be obtained from the following Figures 3 and 4 which show a simple example of path diagrams associated with the model.

Figure 3 shows the functional relationships between two latent factors (i.e. $r = 2$) and the corresponding measured air pollutants. The path also assumes that $\mathbf{f}_x(t)$ follows a first order VAR process — that is, $s = 1$ in Equation (3). However, the left side of Figure 4 shows the measurement part for the health variables which, here, are assumed to be represented by three factors — that is $m = 3$. The model completes with the structural part which provides a picture of the functional relationship among the latent factors, $\mathbf{f}_y(t)$ and $\mathbf{f}_x(t)$, represented by Equation (4), with $p = 1$ and $q = 0$. A feature of both path diagrams is that they clearly show that the factor loadings are written as function of space. As a result, one might expect they should be able to provide a picture of
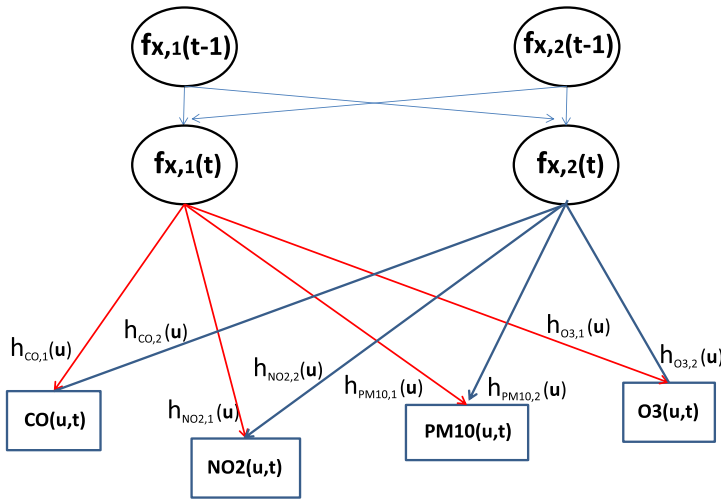
FIGURE 3   This path diagram is for a 2-factor model ($r = 2$) with $n_x = 4$ measured pollutants, where each variable loads on all the factors. The dynamics of the factors in $\mathbf{f}_x(t)$ is represented by a first order VAR process ($s = 1$) [Colour figure can be viewed at wileyonlinelibrary.com]
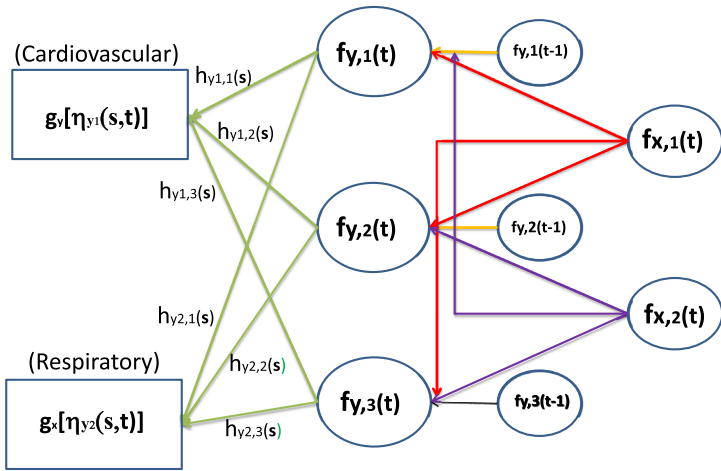

FIGURE 4   This path diagram is for a 3-factor model ($m = 3$) with $n_y = 2$ measured health data, where each variable loads on all the factors. The dynamics of the factors in $\mathbf{f}_y(t)$ is represented by a VARX process with $p = 1$ (i.e. only one lag for the autoregressive component) and $q = 0$ (i.e. instantaneous effect of $\mathbf{f}_x(t)$ onto $\mathbf{f}_y(t)$). Some of the arrows in the structural part have been omitted to facilitate the representation of the impacts [Colour figure can be viewed at wileyonlinelibrary.com]

the spatial distribution of the common factors $\mathbf{f}_y(t)$ and $\mathbf{f}_x(t)$. In this way, the model appears to be spatially descriptive in that it can be used to identify possible clusters of locations whose temporal behavior is primarily described by a potentially small set of common dynamic latent factors. Finally, we notice that the path diagrams are specified here according to an exploratory factor analytic approach, where each variable loads on all the factors. However, further specific patterns of the factor loadings can be favoured with the specification of flexible and spatially structured prior information through the columns of the factor loading matrices $\mathbf{H}_y$ and $\mathbf{H}_x$.

One clear advantage of the model is that, in the context of increasingly high-dimensional time series, the temporal relationship between dependent and regressor variables can be modeled in a reduced latent space. Equation (4), which represents the structural equation of the model, is in the form of a Vector Autoregressive model with exogenous variables (VARX, Lutkepohl, 2005), where the dynamics of the components in $\mathbf{f}_y(t)$ are also controlled for the effects of $\mathbf{f}_x(t)$. While Equations (1) and (2) allow to model independently Y and X at their own spatial scales, Equation (4) models their association in a latent space, which is independent of the spatial structure of the two processes. By marginalizing the effect of space, Equation (4) offers a strategy to overcome the spatial misalignment problem of regressing health and air pollution data. Also, with the help of Equation (3), if required, temporal forecasts of the variables of interest, Y, can be obtained by modelling the dynamics of a few common factors in $\mathbf{f}_y(t)$. We finally note that, differently from Rushworth et al. (2014), our modelling approach allows to take into account sources of uncertainty related to the modelled air pollutant concentrations and the factor loadings.

4 | BAYESIAN INFERENCE AND COMPUTATIONS

4.1 | Prior specification

In this section, we discuss the choice of model priors with particular reference to issues that arise in the estimation of the factor loading matrices and common factors. In practice, assumptions are needed to reduce the total number of possible parameters and appropriate choices of these priors allows us to regularize our inference.

PRIORS FOR FACTOR LOADINGS

The literature contains a variety of methods for specifying the common factors and the factor loadings. One approach places structure only on the estimated factors, leaving loadings free. For example, the factors could be the first few principal components, which are restricted to be mutually orthogonal, while the loadings are left unrestricted. Empirical orthogonal functions (Wikle & Cressie, 1999) and principal kriging splines (Fontanella et al., 2019; Sahu & Mardia, 2005), represent typical examples of this approach in the framework of spatio-temporal models.

A second approach, conversely, places structure only on the loadings, leaving factors free. To ensure identifiability, one can take independent normal priors for each element of the factor loading matrix and fix certain loadings to constant values. A widely used method is to restrict the factor loading matrix to be an upper or lower triangular matrix with strictly positive diagonal elements (Lopes & West, 2004). Further restrictions may also be imposed on the factor loading matrix to allow for dedicated measurements. These restrictions impose patterns of zeros in the factor loading matrix and postulates a priori the relationship between the observed variables and their underlying latent factors. This structure is often assumed in practical problems because it allows for simple interpretation of the underlying factors (Liu et al., 2005; Tsay, 2014). In practice, one can think of such *hard constraints* as the outcome of a Bayesian analysis with spiked priors. However, one may want to impose *soft constraints*, shrinking estimates in certain directions without forcing them. That is, one may want to perform Bayesian analysis incorporating qualitative prior information, so that likelihood information is blended with prior information rather than simply discarded. By following this approach, which is in between posing no restrictions and forcing restrictions, some structures may be favoured probabilistically instead of being imposed.

Let $\mathbf{h}_{xi} = \left(\mathbf{h}'_{x_1 i}, \mathbf{h}'_{x_2 i}, \ldots, \mathbf{h}'_{x_{n_x} i}\right)'$ denote the i-th column of $\mathbf{H}_x$ in variable order, so that $\mathbf{h}_{x_j i} = \left(h_{x_j i}(\mathbf{u}_1), h_{x_j i}(\mathbf{u}_2), \ldots, h_{x_j i}(\mathbf{u}_{N_x})\right)'$. In the context of spatial analysis, useful restrictions may refer to functional forms that impose smoothness of loadings across the sites. If $\mathbf{X}(\mathbf{u}, t)$ is an n_x-dimensional continuous spatial process observed on $\mathcal{D}_x$, we model the elements of each column of the measurement matrix $\mathbf{H}_x$ as realizations from a multivariate Gaussian spatial process with correlation structure defined through the linear model of coregionalization (LMC, Wackernagel, 2003).

Define $\mathbf{h}_{xi}(\mathbf{u}) = \mathbf{A}_i \boldsymbol{\omega}(\mathbf{u})$, $i = 1, 2, \ldots, r$, where $\mathbf{A}_i$ is a $(n_x \times n_x)$ full-rank lower-triangular matrix containing scale and regression parameters for LMC (see, Schmidt & Gelfand, 2003), and $\boldsymbol{\omega}(\mathbf{u})$ is a spatial process with n_x independent components, $\omega_k(\mathbf{u})$, $k = 1, \ldots, n_x$, each with mean 0, unit variances and stationary correlation functions, $\rho(|\mathbf{u} - \mathbf{u}'|, \varphi_{ki})$. Then, the covariance matrix of $\mathbf{h}_{xi}$ can be written as

$$\boldsymbol{\Sigma}_{h_{xi}} = \mathbf{P}' \left(\sum_{k=1}^{n_x} \mathbf{R}_k(\varphi_{ki}) \otimes \mathbf{a}_{ki} \mathbf{a}'_{ki} \right) \mathbf{P}, \quad i = 1, 2, \ldots, r$$

where $\mathbf{a}_{ki}$ represents the k-th column vector of $\mathbf{A}_i$ and $\mathbf{P}$ is a suitable permutation matrix used to respect the variable ordering of the data. Following Schmidt and Gelfand (2003), for the elements of $\mathbf{A}_i$, we assign inverse gamma prior distributions to the scale parameters and normal priors, with mean zero and a large variance for the regression parameters. One also needs to assign priors on the set, say φ_x, of the range parameters of the spatial processes $\boldsymbol{\omega}(\mathbf{u})$. The prior for the ranges depends upon the choice of correlation function. Following Ren and Banerjee (2013), we use an exponential correlation function and set prior distributions for the range parameters relative to the size of their domains. If $\mathbf{X}(\mathbf{u}, t)$ is univariate (i.e. $n_x = 1$), then the elements of each column of the measurement matrix $\mathbf{H}_x$ can be modeled as realizations from r univariate Gaussian processes, as described in Lopes et al. (2008).

Note that, through the use of spatially varying factor loadings, we allow the model to be spatially descriptive. For simplicity, assume only NO_2 is observed. We may have areas in which some factor loadings may be large, indicating that, for example, the traffic pollution is particularly high in NO_2, while, in other areas, they may be small indicating that traffic pollution is relatively low in NO_2. Differently, in the case of spatially constant factor loadings, their map would suggest that traffic produces a fixed amount of NO_2 relative to its abundance regardless of location.

The modelling of the spatial correlation for the factor loadings $\mathbf{h}_{yi}$, $i = 1, \ldots, m$, corresponding to multivariate spatial processes observed on a lattice, is also developed within the LMC. In this case, however, the univariate latent spatial processes $\omega_j(\mathbf{s})$, $j = 1, \ldots, n_y$, defined on $\mathcal{D}_y$, are in the form of conditional autoregressive CAR processes, also known as Gaussian Markov random fields, where the spatial dependence is based on a 0/1 contiguity adjacency matrix $\mathbf{W}$ − with elements equal to 1 if two districts share a common border and 0 otherwise − and the spatial autocorrelation parameters ρ_{ji}, $j = 1, \ldots, n_y$; $i = 1, \ldots, m$. Within the LMC, we thus set $\mathbf{R}_j(\rho_{ji}) = (\mathbf{W}_0 - \rho_{y_{ji}} \mathbf{W})^{-1}$, $j = 1, \ldots, n_y$, where $\mathbf{W}_0 = diag(\mathbf{W} \mathbf{1}_{N_y})$. As for the prior elicitation of the model parameters, we consider a uniform distribution in $(0, 1)$ for the spatial autocorrelation parameters and inverse gamma prior distributions for the scale parameters in matrix $\mathbf{A}_i$.

PRIORS FOR COMMON LATENT FACTORS

Assume, without loss of generality, that $p \geq \max(s, q)$, $\mathbf{C}_i = \mathbf{0}$ for $i > q$ and $\mathbf{D}_i = \mathbf{0}$ for $i > s$. By means of an appropriate concatenation of the common factors, i.e. $\mathbf{f}(t) = [\mathbf{f}_y(t)' \ \mathbf{f}_x(t)']'$, it is useful

to specify the joint generation process of Equations (3) and (4) as a VAR(p) process (Lutkepohl, 2005)

$$\mathbf{f}(t) = \mathbf{\Phi}_1 \mathbf{f}(t-1) + \cdots + \mathbf{\Phi}_p \mathbf{f}(t-p) + \mathbf{v}(t)$$

where

$$\mathbf{\Phi}_i = \begin{bmatrix} \mathbf{B}_i & \mathbf{C}_i \\ \mathbf{0} & \mathbf{D}_i \end{bmatrix}, \quad \mathbf{v}(t) = \begin{bmatrix} \mathbf{v}_y(t) \\ \mathbf{v}_x(t) \end{bmatrix}.$$

Let $\boldsymbol{\alpha}(t) = [\mathbf{f}(t)' \; \mathbf{f}(t-1)' \; \dots \; \mathbf{f}(t-p+1)']'$. Then, the evolution of the joint common factors can be represented by the following transition equation

$$\boldsymbol{\alpha}(t) = \mathbf{\Gamma}\boldsymbol{\alpha}(t-1) + \boldsymbol{\zeta}, \quad \boldsymbol{\zeta} \sim N(0, \mathbf{\Lambda}) \tag{5}$$

where $\mathbf{\Gamma}$ is a ($k \times k$) block coefficient matrix, with $k = (m+r)p$, characterizing the dynamic evolution of the joint common factors and $\mathbf{\Lambda}$ is a covariance matrix with elements $\lambda_{ij}, \; i, j = 1, \dots, k$. The prior for the latent process $\boldsymbol{\alpha}(t)$ is completed by $\boldsymbol{\alpha}(0) \sim N(\mathbf{a}_0, \mathbf{\Sigma}_0)$, with known hyperparameters $\mathbf{a}_0$ and $\mathbf{\Sigma}_0$.

Equation (5) represents exactly the dynamic evolution of a standard vector autoregressive process for which several priors can be elicited for the autoregressive parameters. In general, these differ in relation to three issues. First, VARs are not parsimonious models. Without prior information, it is hard to obtain precise estimates of many coefficients and thus, features such as impulse responses and forecasts will tend to be imprecisely estimated (i.e. posterior or predictive standard deviations can be large). For this reason, it can be desirable to 'shrink' forecasts and prior information offers a sensible way of doing this shrinkage. Second, the priors used with VARs differ in whether they lead to analytical results for the posterior and predictive densities or whether MCMC methods are required to carry out Bayesian inference. Third, the priors differ in how easily they can handle departures from unrestricted specifications (e.g. those allowing for different equations to have different explanatory variables or allowing for VAR coefficients to change over time).

Priors which are enjoying increasing popularity, which can be thought of as automatically selecting a restricted VAR (Fernandez et al., 2001), that are flexible enough in prior elicitation and simple in computation, are known as Stochastic Search Variable Selection (SSVS) priors (George & McCulloch, 1993). These priors favour shrinkage and lead to restricted VARs in an automatic fashion that requires only minimal prior input from the researcher. Posterior computation in the VAR with SSVS prior is also simple as it can be carried out using a Gibbs sampling algorithm. Let $\boldsymbol{\gamma}$ be a parameter vector collecting all the coefficients in $\mathbf{\Phi}_i, i = 1, \dots, p$, and let γ_j denote its generic element. Then, on each coefficient, SSVS specifies a hierarchical prior which is a scale (variance) mixture of two Normal distributions

$$\gamma_j | \delta_j \sim (1 - \delta_j) \; \mathcal{N}(0, \kappa_{0j}^2) + \delta_j \; \mathcal{N}(0, \kappa_{1j}^2)$$

where δ_j is a dummy variable. If δ_j equals one then, the autoregressive parameter, γ_j, is drawn from the second Normal and if it equals zero then γ_j is drawn from the first Normal. The prior is hierarchical since δ_j is treated as an unknown parameter and estimated in a data-based fashion. The SSVS aspect of this prior arises by choosing the first prior variance, κ_{0j}^2, to be small (so that the coefficient is constrained to be virtually zero) and the second prior variance, κ_{1j}^2, to be large

(implying a relatively noninformative prior for the corresponding coefficient). George et al. (2008) discuss a *default semi-automatic approach* to choosing the two variances which requires minimal subjective prior information from the researcher. Hereafter, to take care about model sparsity (in the structural part) and model selection, we shall focus on SSVS.

OTHER PRIORS

We place a non-informative multivariate normal prior with mean $\boldsymbol{\mu}_{\beta}$ and variance $\boldsymbol{\Sigma}_{\beta}$ on the slope parameters of the linear regression functions which parametrize $\boldsymbol{\mu}_y(t)$ and $\boldsymbol{\mu}_x(t)$. For all the model variances we have assigned independent Inverse Gamma prior distributions, with infinite variance and means equal to 1. Exceptions refer to $\boldsymbol{\Sigma}_{v_x}$ and $\boldsymbol{\Sigma}_{v_y}$ in $\boldsymbol{\Lambda}$ which are identity matrices for identifiability conditions.

4.2 | Posterior inference and computational issues

Let $\boldsymbol{\Theta}$ be the vector containing all unknown parameters in the model. Under the Bayesian framework, we make use of Markov chain Monte Carlo (MCMC) methods to obtain samples from the posterior distributions (see, for example, Gamerman & Lopes, 2006). Computationally, we only need to calculate the full conditionals of each parameter given all other parameters, which is usually not hard. A Metropolis-Hastings algorithm (Gamerman & Lopes, 2006), however, is used when the full conditionals are not of a known form, and therefore there is not a straightforward way to draw samples from them.

An important part of the estimation refers to the sampling of the common factors. For Gaussian data (i.e. identity link function), posterior inference for the proposed class of spatial dynamic factor models is facilitated by the fact that standard MCMC algorithms for dynamic linear models can be easily adapted to our model specification such that posterior and predictive analysis are readily available (Ippoliti et al., 2012; Valentini et al., 2013). Conditional on r and m, the number of common factors, the MCMC scheme described in Lopes and West (2004) can be easily adapted where the common factors are jointly sampled via the well-known *forward filtering backward sampling* (FFBS) scheme (Carter & Kohn, 1994; Frühwirth-Schnatter, 1994). However, Bayesian computation of the GST-SEM for non-normal data is more challenging. An efficient proposal is difficult to obtain. Especially for time series with large T, sampling of the common factors is computationally demanding. One possibility is to consider a block sampling scheme that combines techniques such as extended Kalman filter and block sampling, as proposed in Lopes et al. (2011). Another possibility is to consider a dynamic extension of the *IWLS algorithm* for maximum likelihood calculation in generalized linear models, as proposed in Gamerman (1998). A dynamic extension of block updating algorithms described in Knorr-Held and Rue (2002) may also be a possibility. However, these solutions are computationally expensive in our context and are not feasible in practice. A more computationally efficient solution will be discussed in Section 6.

4.3 | Model selection

Selecting an appropriate GST-SEM fit to a data set has some similarities to the problems in choosing a dynamic or spatial factor model for temporal or spatial series. For simplicity, in this paper we first rely on a particular form of the Deviance Information Criterion (DIC) which is particularly suited for latent variable models (Celeux et al., 2006). In this case, the likelihood function $f([YX]|\boldsymbol{\Theta})$ is evaluated conditional on a fixed number of common factors and the effective number

of parameters, which corrects the posterior mean deviance, $E_{\Theta}[-2\log f([YX]|\Theta)]$, is given by the posterior variance of the deviance, $2\text{Var}_{\Theta}[\log f([YX]|\Theta)]$, see Celeux et al. (2006) for more details.

Finally, we also check the adequacy of the model fits by evaluating the Bayesian *p*-values computed on two different statistics. The first, examines how well the model fits the observed data based on the quantity $Pr(y_k(\boldsymbol{s}, t) \geq \hat{y}_k(\boldsymbol{s}, t))$, where $\hat{y}_k(\boldsymbol{s}, t)$ is from the predictive distribution. The second, computes the Chi-square statistic

$$\chi^2_y = \sum_{k,s,t} \frac{(y_k(\boldsymbol{s}, t) - \eta_k(\boldsymbol{s}, t))^2}{\eta_k(\boldsymbol{s}, t)},$$

and then evaluates the Bayesian *p*-value, $Pr(\chi^2_{\hat{y}} \geq \chi^2_y)$, where $\chi^2_{\hat{y}}$ is the same as χ^2_y, but based on $\hat{y}_k(\boldsymbol{s}, t)$ - see for example Marshall and Spiegelhalter (2003).

Bayesian *p*-values close to 0.5 indicate that the generated data are compatible with the model while, as suggested by Tzala and Best (2008), values close to the boundary of the unit interval are an indication of poor fit.

4.4 | Uses of the model

FORECASTING AND INTERPOLATION

Since the GST-SEM is highly structured and flexible, it can be used to solve most of the statistical problems commonly encountered in the analysis of spatio-temporal data.

Time series is primarily concerned with forecasting into the future the variable of interest. The state space model proposed here enables easy calculation of the predictive distribution $p(\mathbf{Y}(T + k)|\mathbf{Y}, \mathbf{X})$ and details can be found in Lopes et al. (2011).

A spatial analysis of the pollutant variables may also require the interpolation of X_j at ungauged sites. From Equation (4), we know that this is not a necessary step to model the relationship existing between X and Y; however, in practice, spatial interpolation may be useful both for descriptive purposes and for multiplier analysis − see below. Interpolating a pollutant over the districts of interest, for example, is relatively easy and requires first the interpolation of the conditionally independent column random vectors, $\mathbf{h}_{xi}$, of $\mathbf{H}_x$. A full explanation of the procedure for the univariate case can be found in Ippoliti et al. (2012). A generalization of the procedure to the multivariate case can be found in Schmidt and Gelfand (2003) and we refer to them for detailed results.

MULTIPLIER ANALYSIS

This type of analysis (Lutkepohl, 2005) allows to describe how the disease risks observed at specific areal units change over time to an increase of the pollutant levels, thus providing an informative tool for characterizing the time course of risk of hospitalization. This is a question that may be raised by policymakers who, in general, are often interested in knowing how the occurrence of a certain future event would affect the future values of the variable of interest.

While simple in composition, effectively using a multiplier analysis in a spatio-temporal design can be challenging for at least two reasons. For example, obtaining precise estimates of the response function is difficult because the covariates are often highly collinear resulting in inflated standard errors. Furthermore, one may find it difficult to interpret the shape of the impulse response function as an effect of a pollution control policy since, in practice, policymakers can only control for general factors (e.g. traffic, combustion, etc), and not for single pollutants.

The procedures listed below provide details on how to perform multiplier analysis within our modelling framework and describe possible strategies to overcome the drawbacks listed above. In particular, we show that the model allows to investigate the relation between the risk of hospitalization and air pollution by tracing out the effect of an exogenous shock in all or some of the common factors of X. This approach, allows to estimate impulse response functions at different levels of the hierarchy.

1. **Second-level Impulse response analysis**. In this case, multiplier analysis is performed at the structural level of the model. The impulse responses are thus the coefficients in the vector moving averaging (VMA) representation of Equation (5). With VAR models, the parameters themselves, as opposed to functions of them such as impulse responses, are rarely of direct interest. In addition, the fact that there are so many of them makes it hard for the reader to interpret tables of VAR coefficients. Applying an impulse response analysis at the structural level of the model may thus be useful to understand which of the causal factors, $f_{x,i}(t)$, have more impact on the $f_{y,k}(t)$.

2. **Impulse response analysis for the Y variables**. An important part of the analysis is to quantify the day-to-day changes in the relative risk of hospitalization as an effect of the impulse on the causal factors $f_{x,i}(t)$. This can be done by estimating the impulse response, say $IR_{y_k}(\mathbf{s}, t)$, through the VMA representation of the linear predictor in Equation (1) - see also Equation (6) below — which, in turn, depends on the VMA form of the state Equation (5).

3. **Impulse response analysis for the X variables**. The multiplier analysis concludes by studying the response of the pollutant variables as a result of a unit shock in the causal factors, $f_{x,i}(t)$. In this context, the following steps describe how to estimate the impulse response function of $X_l(\mathbf{u}, t)$ within a generic area unit $\mathbf{s}$:

   a. write Equation (3) in its VMA form and substitute it into Equation (2) to estimate the impulse response of $X_l(\mathbf{u}, t)$, denoted $IR_{x_l}(\mathbf{u}, t)$, as an effect of exogenous impulses on the uncorrelated causal factors $f_{x,i}(t)$;

   b. estimate the day-to-day changes in air pollution levels within the district of interest $\mathbf{s}$. The desired impulse response, $IR_{x_l}(\mathbf{s}, t)$, of X_l, for area unit $\mathbf{s}$ is defined as

$$IR_{x_l}(\mathbf{s}, t) = \int_{\mathbf{u} \in \mathbf{s}} IR_{x_l}(\mathbf{u}, t) \quad d\mathbf{u},$$

   which is the average of the responses across a fine grid within the areal unit $\mathbf{s}$, computed by using Monte Carlo integration.

5 | SIMULATION

An initial assessment of the proposed methodology is performed in a simulation framework in which the health data and pollutants are generated from Poisson and Gaussian distributions respectively. The simulation is conducted to ensure appropriate identification of the features of the generating model and, in particular, of the second-level impulse response functions. Recovering the true path of these functions is of particular interest as the IR coefficients provide an estimate of the short-term effects of environmental exposures on daily hospitalization.

To keep the complexity of the simulation at a relatively simple level, the data are generated from a factor model with $n_y = n_x = 2$ and measurement equations

$$\ln\left[\frac{\boldsymbol{\eta}_y(t)}{E(t)}\right] = \boldsymbol{\mu}_y(t) + \mathbf{H}_y\ \mathbf{f}_y(t) + \boldsymbol{\epsilon}(t),$$

$$\mathbf{x}(t) \sim N\left(\boldsymbol{\mu}_x(t) + \mathbf{H}_x\ \mathbf{f}_x(t);\ \sigma_x^2 \mathbf{I}_r\right)$$

where $N_y = 28$, $N_x = 100$, $T = 200$, $r = 3$, $m = 2$, $\sigma_x^2 = 4$, $\epsilon \sim N(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}_2)$ and $\sigma_\epsilon^2 = 0.10$. The expected numbers of disease are the same as those used in the application section. For simplicity, the mean terms are assumed constant across the variables, that is $\mu_{x_k} = 30$ and $\mu_{x_k} = 2$, $k = 1, 2$. The factor loadings are considered as realizations of zero mean Gaussian processes with full-rank lower-triangular matrices $\mathbf{A}_x = \mathbf{A}_y = \begin{bmatrix} 1 & 0 \\ 0.5 & 1 \end{bmatrix}$. For the X variables, an exponential correlation function with spatial dependence parameter $\varphi = 10$ is considered for the latent spatial processes $\omega_k(\mathbf{u})$, $k = 1, 2$. For the Y variables, instead, we assume that the two latent spatial CAR processes both have a spatial correlation parameter equal to $\rho_y = 0.5$. The variances of the factor loadings are also considered constant with $\sigma_{hy_i}^2 = 0.1$, $i = 1, 2$ and $\sigma_{hx_j}^2 = 0.5$, $j = 1, 2, 3$. Finally, the structural Equations (3) and (4) for the common factors are based on the following matrices:

$$\mathbf{C}_0 = \begin{bmatrix} 0.75 & 0 & 0 \\ 1.50 & 0 & 0 \end{bmatrix}, \quad \mathbf{C}_1 = \begin{bmatrix} 0.75 & 1.50 & 0.50 \\ 0.50 & 0 & 0 \end{bmatrix},$$

$$\mathbf{C}_2 = \begin{bmatrix} 0 & -0.50 & -1.50 \\ 0 & 0.50 & 0 \end{bmatrix}, \quad \mathbf{C}_2 = \begin{bmatrix} 0 & 0 & 0 \\ -0.50 & 0 & 0 \end{bmatrix},$$

$\mathbf{B}_1 = 0.5 \mathbf{I}_2$, $\mathbf{B}_i = \mathbf{0}$, $i = 2, 3, 4$; $\mathbf{D}_1 = 0.5 \mathbf{I}_3$, $\mathbf{D}_2 = 0.2 \mathbf{I}_3$ and $\mathbf{D}_i = \mathbf{0}$, $i = 3, 4$.

Using this parametrization, the simulated IR functions represented (red colour) in Figures 5 and 6 show the short-term impacts of the common factors $\mathbf{f}_{x_i}(t)$, $i = 1, 2, 3$, of the pollutants on the common factors $\mathbf{f}_{y_j}(t)$, $j = 1, 2$, of hospitalizations. As it can be noticed, the proposed parametrization underlays different latencies and oscillations of the impulse response functions, including a scenario of mortality displacement in which the IR function is first positive at early lags and then negative at later lags (Welty et al., 2009) — see the bottom-left panel of Figure 5. Furthermore, based on 200 replicated datasets and 50,000 draws of the MCMC algorithm for each of the simulated datasets, all the subplots show the posterior mean IRFs (black) averaged across the 200 simulations with the 95% confidence bands (grey area). As discussed in Section 4.1, the model parameters of the state equation have been estimated by using the spike and slab procedure with s, p and q fixed a priori at the value of 4. A visual inspection of the results indicates that the MCMC procedure performs consistently well and that we are able to retrieve the different patterns and the different delayed responses of hospitalization. Although not shown here for conciseness, both the other model parameters and the spatial patterns of the factor loadings are also well estimated.

6 | SPATIO-TEMPORAL ANALYSIS OF HOSPITAL ADMISSIONS

In this section, we continue the statistical analysis of the data described in Section 2 and consider whether the GST-SEM is useful to understand how the pollutants impact on hospital admissions for respiratory and cardiovascular diseases.
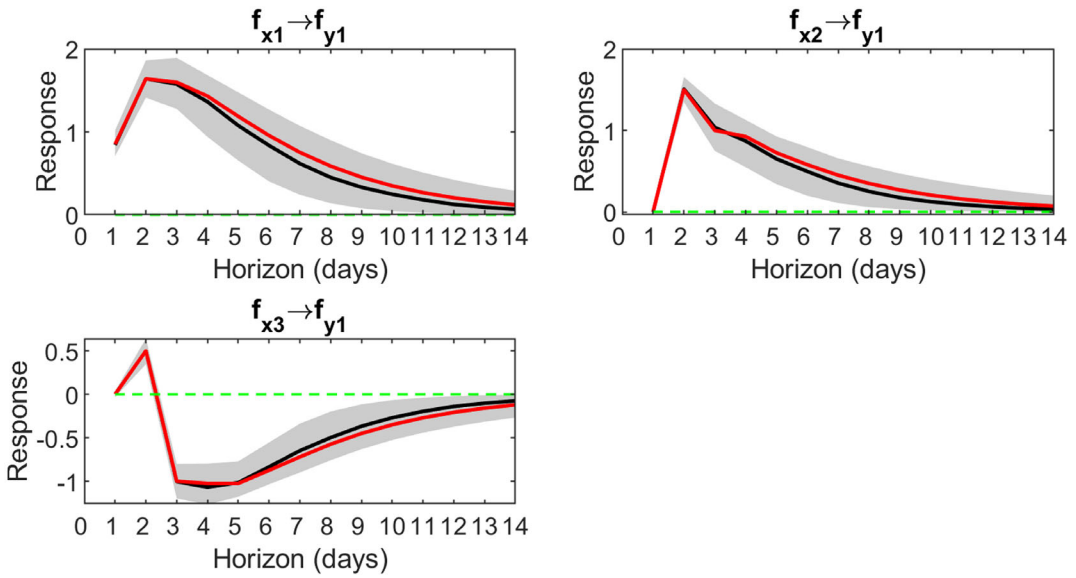
FIGURE 5 Posterior mean IR functions (black) and 95% confidence bands (grey) showing the short-term impacts of the common factors $\mathbf{f}_{x_i}(t)$, $i = 1, 2, 3$ on the first common factor $\mathbf{f}_{y_1}(t)$. The true IR functions are represented in red [Colour figure can be viewed at wileyonlinelibrary.com]

Poisson data are very common in public health research. Hence, we assume $Y_k(\mathbf{s}, t) \sim$ Poisson($\eta_{y_k}(\mathbf{s}, t)$) and consider a regression with a log link function, such that

$$\ln\left[\frac{\eta_{y_k}(\mathbf{s}, t)}{E_k(\mathbf{s}, t)}\right] = \mu_{y_k}(\mathbf{s}, t) + \sum_{i=1}^{m} h_{y_k,i}(\mathbf{s}) \, f_{y,i}(t) + \epsilon_k(\mathbf{s}, t), \ \ t = 1, \dots, T \tag{6}$$

where $E_k(\mathbf{s}, t)$, assumed fixed and known, serves as offsets in the model. The final component, $\epsilon_k(\mathbf{s}, t)$, is a zero mean normal pure error term with variance $\sigma^2_{\epsilon_k}$, introduced to accommodate for the overdispersion usually encountered in epidemiological studies — see the exploratory analysis of Section 2. As it is known (Banerjee et al., 2008), in the case of non-Gaussian data models, as long as the structure on the covariance associated with $\epsilon_k(\mathbf{s}, t)$ is simple, by including this residual error process one can also implement an efficient MCMC estimation. In fact, reparametrizing the process $\phi_{y_k}(\mathbf{s}, t)$ as in Equation (6), both the trend parameters in $\mu_{y_k}(\mathbf{s}, t)$ (see below) and the spatio-temporal process $\phi_{y_k}(\mathbf{s}, t)$ can be updated jointly by an efficient conjugate Gibbs step within the overall MCMC.

Regarding the measurement equation for X we note that is convenient to operate on a logarithmic scale to reduce the variability and to make data conform more closely to the Normal distribution. In principle, each pollutant could be included separately in Equation (6) in order to identify its relative impact on the health data. However, since air pollution measurements exhibit strong correlations (see Section 2), it seems inappropriate to include them directly in the model. Therefore, assuming $X_l(\mathbf{u}, t) \sim N(\eta_{x_l}(\mathbf{u}, t), \sigma^2_{xl}(\mathbf{u}))$, $l = 1, \dots, 4$, we model the mean as

$$\eta_{x_l}(\mathbf{u}, t) = \mu_{x_l}(\mathbf{u}, t) + \sum_{i=1}^{r} h_{x_l,i}(\mathbf{u}) \, f_{x,i}(t). \tag{7}$$
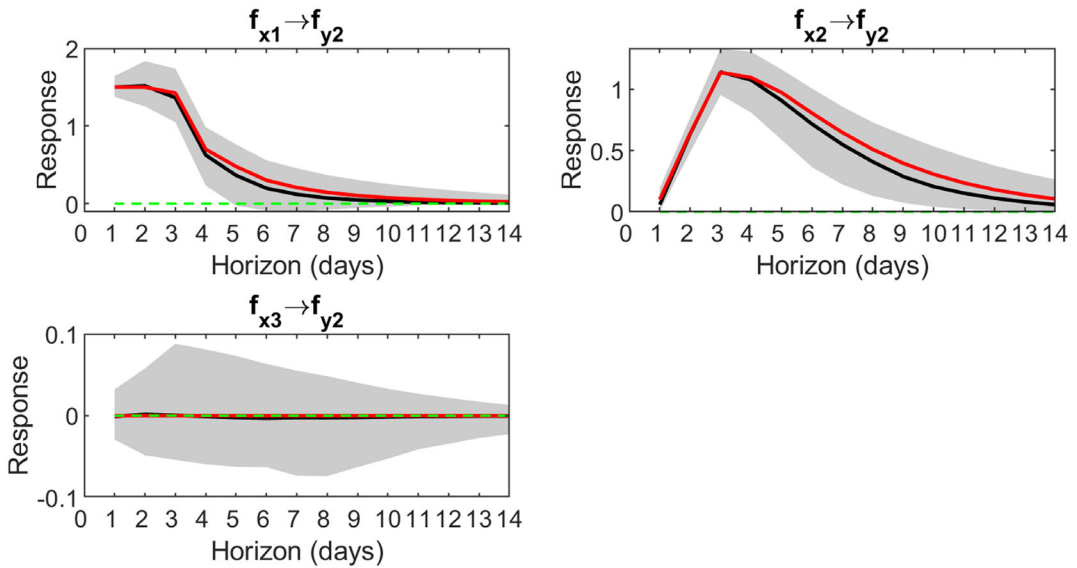
FIGURE 6 Posterior mean IR functions (black) and 95% confidence bands (grey) showing the short-term impacts of the common factors $\mathbf{f}_{x_i}(t)$, $i = 1, 2, 3$ on the second common factor $\mathbf{f}_{y_2}(t)$. The true IR functions are represented in red [Colour figure can be viewed at wileyonlinelibrary.com]

The dynamics of the common factors, $f_{x,i}(t)$ and $f_{y,i}(t)$, are modelled in the second level of the hierarchy through the structural Equations (3) and (4) - see below.

The structure of the large-scale variation follows typical specifications discussed in many time series studies and include time varying trends, the effects of temperature (*temp*) and humidity (*hum*) and dummy variables for calendar effects −, that is, weekend (WE) and festivities (FE). In particular, the mean term of Y takes the following form

$$\mu_{y_k}(\mathbf{s}, t) = \beta_{y_k,0}(\mathbf{s}) + \beta_{y_k,1}(\mathbf{s})\text{We}(t) + \beta_{y_k,2}(\mathbf{s})\text{Fe}(t) + ns_1(\text{temp}) + ns_2(\text{hum})$$
$$+ \beta_{y_k,5}(\mathbf{s}) \text{ temp}(t)_{0-2} + \beta_{y_k,6}(\mathbf{s})\text{hum}(t)_{0-2} + ns\left(t, \boldsymbol{\beta}_{x_j,7}(\mathbf{u})\right) \tag{8}$$

where $ns_1(\text{temp}) = ns\left(\text{temp}(t), \boldsymbol{\beta}_{y_k,3}(\mathbf{s})\right)$, $ns_2(\text{hum}) = ns\left(\text{hum}(t), \boldsymbol{\beta}_{y_k,4}(\mathbf{s})\right)$ and $ns\left(t, \boldsymbol{\beta}_{x_j,7}(\mathbf{u})\right)$, denote natural cubic splines with 3 degrees of freedom, and $\text{temp}(t)_{0-2} = 1/3\sum_{j=0}^{2}\text{temp}(t-j)$ and $\text{hum}(t)_{0-2} = 1/3\sum_{j=0}^{2}\text{hum}(t-j)$, are simply running means of the same variables for the specified lengths (Peng et al., 2009). We note that several daily time series of temperature and humidity are available over the two regions. However, because Lombardy and Piemonte tend to show very similar climate conditions, these time series appear highly correlated (in many cases the correlation is above 0.9) and, hence, only the averaged (across sites) time series of temperature and humidity were considered here.

For X the mean component takes a similar form and is given by

$$\mu_{x_l}(\mathbf{u}, t) = \beta_{x_l,0}(\mathbf{u}) + \beta_{x_l,1}(\mathbf{u})\text{We}(t) + \beta_{x_l,2}(\mathbf{u})\text{Fe}(t) + ns\left(\text{temp}(t), \boldsymbol{\beta}_{x_l,3}(\mathbf{u})\right)$$
$$+ ns\left(\text{hum}(t), \boldsymbol{\beta}_{x_l,4}(\mathbf{u})\right) + ns\left(t, \boldsymbol{\beta}_{x_l,5}(\mathbf{u})\right). \tag{9}$$

As it can be noticed, the mean functions allow for spatial varying coefficients. In general, estimation of these parameters can be performed by using structured priors as shown, for example,

in Gelfand et al. (2005). However, since the spatial correlation is captured through the factor loadings as well as through the effects of covariates, as specified in Section 4.1, we only use vague priors for these regression coefficients.

Furthermore, since the GST-SEM encompasses a large class of spatial-temporal models, we provide examples of the flexibility of the model by discussing two different parametrizations of the random effects, $\phi_{y_k}(\mathbf{s}, t)$ and $\phi_{x_i}(\mathbf{u}, t)$. The first, which represents a generalized common factor model, allows to perform separate analyses for each pollutant and disease. In this case, as in the simulation section, each variable is represented by one single common latent factor. The second model, instead, is much more general and favours the learning of complex interactions by borrowing strength across all the variables. Hence, the two parametrizations imply a trade-off between model complexity (and accuracy) and interpretability.

For all the fitted models, the MCMC algorithm was run for 500,000 iterations. Posterior inference was based on the last 40,000 draws using every 10th member of the chain to avoid autocorrelation within the sampled values. From the computational viewpoint, we first sample $\ln[\eta_y(\mathbf{s}, t)]$ from its marginal distribution using the adaptive rejection sampling (Gilks & Wild, 1992) and then we draw all the other parameters. Conditional on $\ln[\eta_y(\mathbf{s}, t)]$, the full conditional posterior distributions take convenient functional forms and can be easily sampled from. The only exceptions are for the conditional distributions of the spatial correlation parameters which are sampled through Metropolis-Hastings steps. Convergence of the chains of the model was monitored visually through trace plots as well as using the R-statistic of Gelman (1996) on two chains simultaneously started from different initial points.

6.1 | The common factor model

In this section, we briefly discuss the results from a generalized common factor model which is useful, for example, when initial separate analyses are performed for pairs of variables Y_k and X_j. Using this model parametrization the number of latent variables is fixed a priori as it is assumed that each variable is driven by its own single common factor. The model represents an extension to the space-time setting of the model introduced by Wang and Wall (2003). To provide an example of the potentialities of the model, consider the daily hospital admissions for respiratory diseases and PM_{10}. With the mean functions specified as in Equations (8) and (9), the measurement equations for this model are given by

$$\ln\left[\frac{\eta_y(\mathbf{s}, t)}{E(\mathbf{s}, t)}\right] = \mu_y + h_y(\mathbf{s})\, f_y(t) + \epsilon(\mathbf{s}, t),$$

$$x(\mathbf{u}, t) \sim N\left(\mu_x + h_x(\mathbf{u})\, f_x(t);\; \sigma_x^2\right)$$

and the structural equations, based on the single common factors $f_x(t)$ and $f_y(t)$, are

$$f_x(t) = \sum_{i=1}^{4} d_i f_x(t-i) + v_x(t)$$

$$f_y(t) = \sum_{i=1}^{4} b_i f_y(t-i) + \sum_{i=0}^{4} c_i f_x(t-i) + v_y(t).$$

To simplify model selection, as in the simulation, we set the values of the maximum lag-length of s, p and q at the relatively large value of 4 and use the spike and slab prior to favour a sparse solution
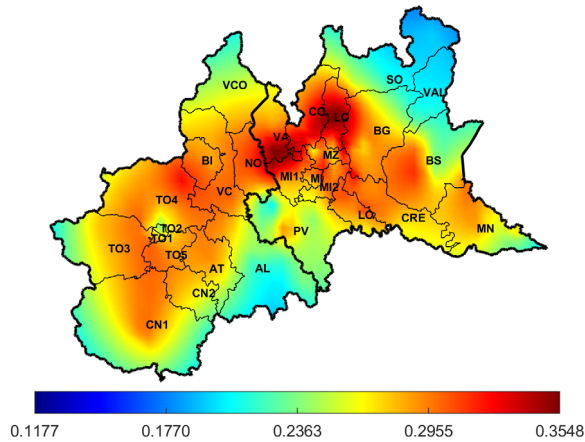
FIGURE 7   Posterior mean of the factor loadings associated with PM10 and interpolated over the two regions (left). Map of the posterior mean of the factor loadings associated with PM_{10} and averaged over the 28 districts (right). The labels within each district represents the acronyms of the Aziende Sanitarie Locali [Colour figure can be viewed at wileyonlinelibrary.com]

in the estimation of the parameters. To check possible differences with models commonly used in the time series literature, we compare results from our model with those obtained estimating a Bayesian distributed lag model (Welty et al., 2009) on the time series of Y and X averaged over all districts and monitoring sites.

Figure 7 shows the posterior mean of the factor loadings $h_x(\mathbf{u})$ interpolated, as described in Lopes et al. (2008), over the two regions. The map allows for a spatial descriptive analysis of the distribution of PM_{10} and clearly shows that the largest values are concentrated in the area of Milan (MI, MI1, and MI2), one of the most polluted cities of Europe, Varese (VA), Como (CO), Lecco (LC) and Monza (MZ). The area is bordered at the north by mountains which render very low air mass exchange. In general, the high degree of pollution is due to a greater contribution of emissions[2] as well as to adverse meteorological and thermodynamic conditions of the atmosphere. Also, the area counts many industrial facilities as well as small and medium enterprises which are major centres of activity and are therefore prime traffic generators. The average value of the factor loadings over these districts is about 0.35, which suggests that an increase of about $(\exp\{0.35\} - 1) \times 100 \approx 42\%$ of pollution is expected in the area as a consequence of a one-unit increase of $f_x(t)$. This means that, considering for example the district of Milan, where the averaged level of PM_{10} is about 50 $\mu g/m^3$, the expected effect of a unitary variation of $f_x(t)$ consists in an increase of 21 $\mu g/m^3$ in the same district.

Focusing on the temporal dynamics of the phenomenon, Figure 8 shows the result of an impulse response analysis at the structural level. The figure describes how the common factor $f_y(t)$, summarising the temporal dynamics of the hospitalization, reacts over two weeks to a one-unit shock of the factor $f_x(t)$ representing the dynamics of PM_{10}. Both the 90% and 95% credible intervals, represented by the shaded areas, suggest that the peak response of hospital admissions for respiratory diseases occurs after four days and then gradually decreases and dies out in about ten days.

However, by modulating the values of this impulse response function by the factor loadings $h_y(\mathbf{s})$, one can better understand the spatio-temporal variation of the risk of hospitalizations. For
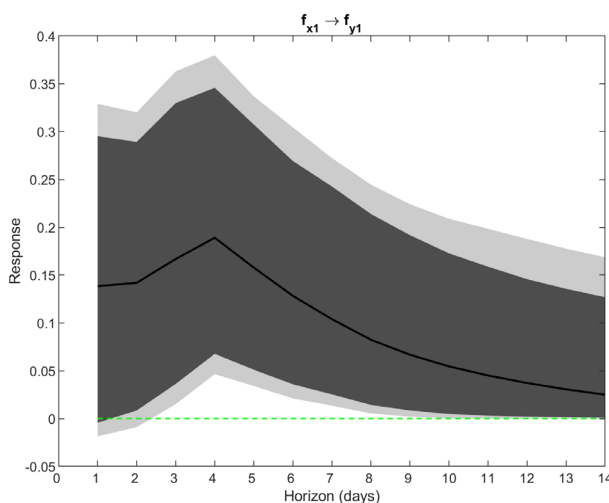
---

[2]http://www.inemar.eu/xwiki/bin/view/InemarDatiWeb/Emissioni+provinciali+2012

FIGURE 8 Posterior median impulse responses (solid line) of $f_y(t)$ to a one-unit shock of $f_x(t)$. The credibility intervals at 90% and 95% are represented by shaded areas [Colour figure can be viewed at wileyonlinelibrary.com]

example, Figure 9 shows the posterior mean of the impulse response functions, $IR_{y_k}(\mathbf{s}, t)$, for a selection of representative districts (located in different areas of the regions and with different levels of pollution). Controlling for confounding effects, the subplots show how the relative risk changes over districts and time as an effect of a one-unit shock of the common factor $f_x(t)$. As it can be noticed, the response functions appear similar in shape. However, the rate of decay, the persistence of the effect (i.e. the area which does not include the zero) and the levels of risk change with the districts. For example, considering the maximum peak, the risk varies between 7‰, for the third district of Torino (TO3), and 1.2% for the district of Milan. In general, the comparison of the impulse response functions shows that the effect is weaker at TO3 and worse at MI. The posterior mean of the function of Brescia (BS) is roughly comparable with that of Sondrio (SO); however, the effect estimated at SO is less persistent and the credible interval is larger than that of BS. It is also worth noticing that Sondrio (SO), the largest district in the Alpine administrative province, is the site for which the environmental agency Inemar-ARPA (Atmospheric Emissions Inventory) of Lombardy estimates the greater per capita consumption of wood. Hence, there is the hint that the effect of hospitalization in this district might be due to high values of Benzo(a)pyrene (BaP) which, formed by residential wood combustion, is one of the toxic components in the PM_{10} and promotes pulmonary injury.

In Figure 5b of the Supplementary Material, Section B, we also show the impulse response function obtained by estimating the Bayesian distributed lag (DL) model of Welty et al. (2009) with trend function similar to that specified in Equation (8). Based on the averaged (across space) time series, the IR function gives the percentage increase in daily hospitalization associated with a 1% increase in PM_{10} at lags 1, ..., 14. Results suggest that, for the DL model, the significant effects of the pollutant start after three days and that soon after an exponentially decaying pattern can be noticed. Furthermore, it also results that the expected short-term cumulative effect on daily hospitalization appears to be about 0.13% overall the region of interest. However, considering our common factor model, we have that the estimated cumulative effect on hospitalization for a one-unit shock of $f_x(t)$ is about 5.1%. Since over the whole region the expected increase of PM_{10}
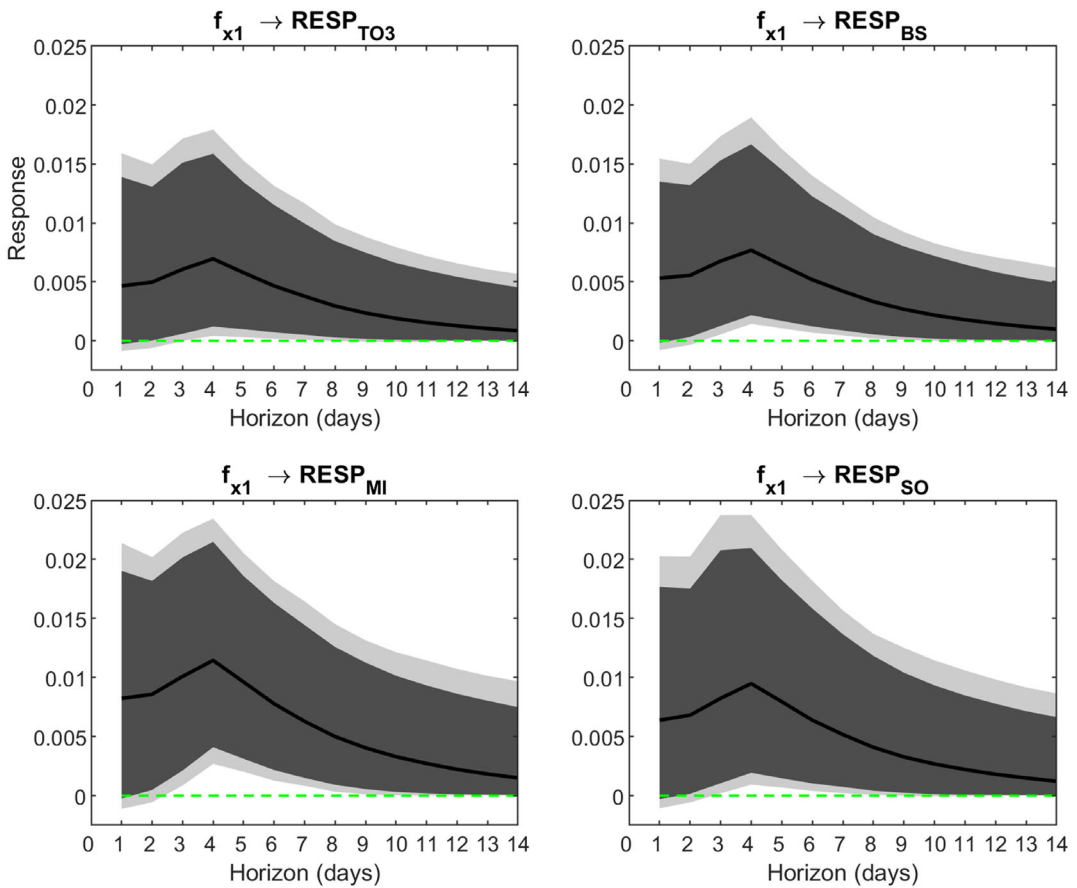
FIGURE 9 Posterior median impulse responses (solid line) of respiratory diseases to a unit shock of f_x. The credibility intervals at 90% and 95% are represented by shaded areas. As an example, the panels show the response functions at the third district of Torino (TO3), Brescia (BS), Milano (MI) and Sondrio (SO) [Colour figure can be viewed at wileyonlinelibrary.com]

is about 33% for a one-unit shock of $f_x(t)$, it follows that the expected cumulative effect on hospitalization for a 1% increase of PM_{10} is $5.1/33 = 0.15\%$. It thus appears that the two models tend to provide similar information and that small differences may be due to their different intrinsic characteristics. However, while the DL model only works with single time series and, as in our case, can only provide an estimate of a general effect over the whole region using spatially aggregated series, our model, is able to weight the effects of PM_{10} by means of the factor loadings, and can thus provide district-specific IR functions as in Figure 9.

6.2 │ The full GST-SEM model

In this section, we discuss results from the generalized spatio-temporal structural equation model which allows to examine relationships among all the variables. As an extension of the common factor model discussed above, it is assumed that there may be more than one underlying factor, and that relationships among the latent variables, modelled by means of Equations (3) and (4), are of primary interest.

For model selection, we again set the values of the maximum lag-length of the structural equations (i.e. s, p and q) at the large value of 4 and use the spike and slab prior to achieve sparsity in the coefficient matrices $\mathbf{B}_i$, $\mathbf{C}_i$ and $\mathbf{D}_i$. To facilitate the interpretation of the common factors $f_{x,i}(t)$, the AR parameter matrices $\mathbf{D}_i$ are also assumed diagonal. Then, by considering an increasing number of common factors, with r and m ranging from 1 to 6, the DIC criterion proposed for latent variable models (see, Section 4.3), suggests that the optimal choice is found for a model with four common factors for X and one common factor for Y. The DIC also suggests that the model fit is improved by accounting for the spatial distribution of the factor loadings, with a full model specification of the LMC performing much better than a separable one. Diagnostic results from the Bayesian p-values computed for the predicted checks (mean value, 0.44) and the Chi-square goodness-of-fit test (0.17), also suggest an adequate fit of the model as the estimated values lie within the interval [0.1, 0.9]. The posterior mean of the VARX model and the posterior inclusion probabilities of the associated parameters, are shown in the following Tables 1–3.

The 'median model', that is, the model containing the variables with posterior inclusion probabilities greater than 0.5 (Barbieri & Berger, 2004), shows that the SSVS prior excludes many of the estimated coefficients ensuring parsimony in the structural part of the model. The autoregressive parameters of both $\mathbf{f}_y(t)$ and $\mathbf{f}_x(t)$ are mainly concentrated at the first lag, with only one parameter found at the second lag for $\mathbf{f}_{x,2}(t)$ — see Tables 1 and 3. Regarding the regression parameters

TABLE 1 Posterior mean of $\mathbf{B}_i$ for lags 1 to 4 and its probability of inclusion (in brackets). In bold the parameters with inclusion probabilities greater than 0.5

$\mathbf{B_1}$	$\mathbf{B_2}$	$\mathbf{B_3}$	$\mathbf{B_4}$
0.10	0.01	0.06	−0.01
(**0.66**)	(0.32)	(0.49)	(0.29)

TABLE 2 Posterior mean of $\mathbf{C}_i$ for lags 0 to 4 and associated inclusion probabilities (in brackets). In bold the parameters with inclusion probabilities greater than 0.5

lag 0	0.01	0.01	0.03	0.00	lag 1	0.06	**−0.17**	0.05	0.04
	(0.29)	(0.32)	(0.37)	(0.28)		(0.49)	(**0.84**)	(0.48)	(0.41)
lag 2	0.00	**0.16**	−0.01	0.05	lag 3	**0.14**	−0.01	**−0.13**	0.00
	(0.33)	(**0.83**)	(0.34)	(0.47)		(**0.82**)	(0.33)	(**0.76**)	(0.30)
lag 4	0.00	0.00	−0.02	0.00					
	(0.32)	(0.28)	(0.37)	(0.26)					

TABLE 3 Posterior mean of diag($\mathbf{D}_i$) for lags 1 to 4 and associated inclusion probabilities (in brackets). In bold the parameters along the diagonal of $\mathbf{D}_i$ with inclusion probabilities greater than 0.5

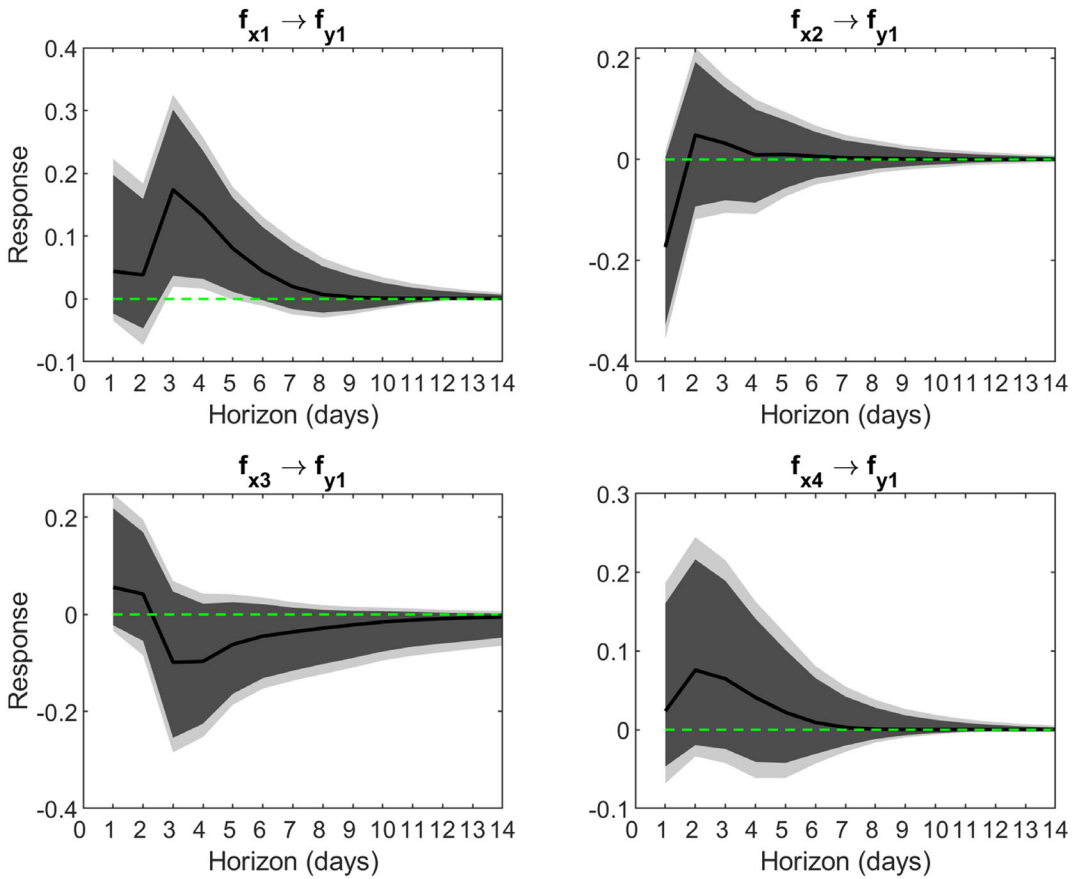| lag 1	**0.68**	**0.59**	**0.67**	**0.86**	lag 2	−0.06	−0.01	−0.02	**0.25**
	(**1.00**)	(**1.00**)	(**1.00**)	(**1.00**)		(0.21)	(0.95)	(0.08)	(**0.99**)
lag 3	−0.05	0.01	0.01	0.00	lag 4	0.00	0.02	0.06	0.01
	(0.36)	(0.10)	(0.11)	(0.07)		(0.07)	(0.15)	(0.46)	(0.07)

FIGURE 10   Posterior median impulse responses (solid line) of f_{y1} to a f_{xi} shock. The credibility intervals at 90% and 95% are represented by shaded areas [Colour figure can be viewed at wileyonlinelibrary.com]

in $\mathbf{C}_i$, we note that none of the coefficients on the fourth lag seem to be significant and only two coefficients on third lags are found to be important — see Table 2.

The interpretation of the results is facilitated by exploring the spatial patterns of the interpolated surfaces of the factor loadings for CO, NO_2, PM_{10} and O_3, and the impulse response functions obtained from Equation (4). Figure 10 suggests that only the impact of $f_{x,1}(t)$ on $f_{y,1}(t)$ is of real interest as both the 90% and 95% credible intervals do not include the zero. As shown in the top-left panel, the associated impulse response function is hump-shaped with the peak response of hospital admissions for cardio-respiratory diseases occurring at the third day; the response then tends to gradually decrease dying out in about five/six days.

Considering the first common factor $f_{x,1}(t)$, which represents a summary of the amount of air pollution over the entail region, Figure 11 shows the posterior mean of the associated factor loadings. In particular, the maps on the left show the factor loadings interpolated over the two regions and those on the right show the areas in which the 95% credible intervals of the factor loadings do not include the zero. The posterior mean estimates of the other three factors are shown in Figures 6c–8c of the Supplementary Material, Section C, and an interesting feature is that some of these factors clearly separate Piemonte from Lombardia.
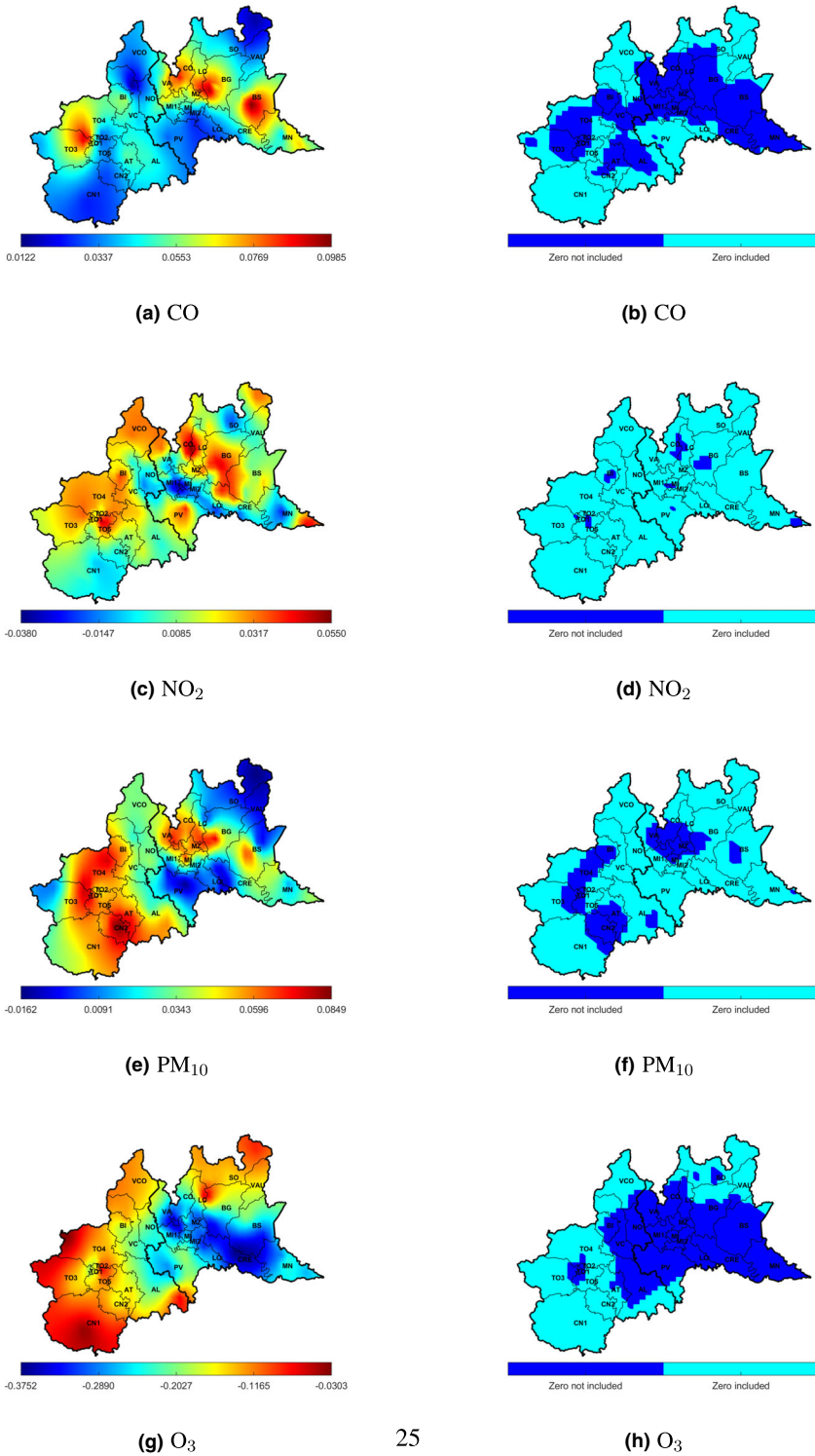
FIGURE 11 Posterior mean surfaces of the first factor loading $\mathbf{h}_{x1}$. Left: CO, NO$_2$, PM$_{10}$ and O$_3$. Right: in blue are shown the areas in which the 95% credible intervals do not include the zero [Colour figure can be viewed at wileyonlinelibrary.com]

As noticed in the previous analysis, the spatial patterns shown in Figure 11 suggest that, in general, the levels of pollution in Piemonte are lower than those observed in Lombardy and that, again, the higher values can be observed in the most industrialized and busy areas at the North-East part of Milan. It also appears that the patterns of NO_2 and PM_{10} are much more localized than other pollutants and it is interesting to note that these patterns do agree with those shown in the report published by Regione Emilia-Romagna[3] on the *Northern Italy Network to Forecast Aerosol pollution* (NINFA), which is a system based on the Chimere chemical transport model driven by the meteorological model COSMO-I7.

By focusing on the NO_2 and PM_{10} variables, we have that the average values of the factor loadings over the regions are about 0.023 and 0.033 respectively. Following the arguments given in Section 6.1, this suggests that an increase of about 2.3% and 3.3% of these pollutants is expected as an effect of a one-unit increase of $f_{x1}(t)$. Then, if we were interested in evaluating the impact over the whole area, we might consider, for example, the average levels of NO_2 (38.65 $\mu g/m^3$) and PM_{10} (38.05 $\mu g/m^3$), and find that the expected increase for these variables is about 0.86 $\mu g/m^3$ for NO_2 and 1.26 $\mu g/m^3$ for PM_{10}. Accordingly, considering that the EU limit values are fixed at 40 $\mu g/m^3$ for both variables, it follows that, in general, the risk of exceeding the threshold, mainly for traffic and industrial activities, is quite high.

For CO and O_3, the spatial maps suggest that these variable are widespread overall in the region and that a variation of about 4.5% and -15.0% is expected as a consequence of a one-unit increase of $f_{x1}(t)$. Their average levels, equal to 0.79 mg/m^3 for CO and 83.72 $\mu g/m^3$ for O_3, appear far from their EU limits of 10 mg/m^3 (CO) and 120 $\mu g/m^3$ (O_3) and they make us think of a quite good control policy for these variables.

Following the procedure described in Section 4.4, Figures 9d–12e (see the Supplementary Material, Section D) also show the estimated impulse response functions for the pollutants. For a specific district, each panel gives information about the day-today variation in air pollution levels as a result of a unit shock to the factor $f_{x,1}(t)$. For example, the first panel on the top-left corner of Figure 9d, suggests that giving a unit shock in f_{x1} at time t, we expect to observe about 15% increase in CO at time $t + 1$ in the district of Torino 1 (TO1). In general, these also show that, in any of the variables, a unit shock dies away quite rapidly due to the stability of the processes.

Extending the analysis at the level of hospitalization, Figure 12 shows the differences in the impulse response functions $IR_{y_k}(\mathbf{s}, t)$ of a selection of districts. The panels suggest that, controlling for confounding effects, the response to a unit shock in $f_{x,1}(t)$ starts after two days and the effect tend to die away in about six days. The peaks of the relative risk appear at the third lag and range between 0.08% (Torino 3) and 1.2% (Milano), with the higher values found for the cardiovascular diseases. Furthermore, wherever the factor loadings of a district are small, and close to zero, the impulse response functions do not show any relevant effect of the pollutants in that area. An example is shown in the last column of Figure 12 for the districts of Torino3 (TO3) and Milano1 (MI1).

In general, the response of the respiratory disease is weaker and the way it spreads over the area does not support the idea of a neighbourhood effect. This is probably due to the fact that the spatial correlation is not as strong as in the cardiovascular case, a feature already anticipated in Section 2. The impulse response functions, estimated for all the districts, can be examined in Figures 13e–16e of the Supplementary Material, Section E.

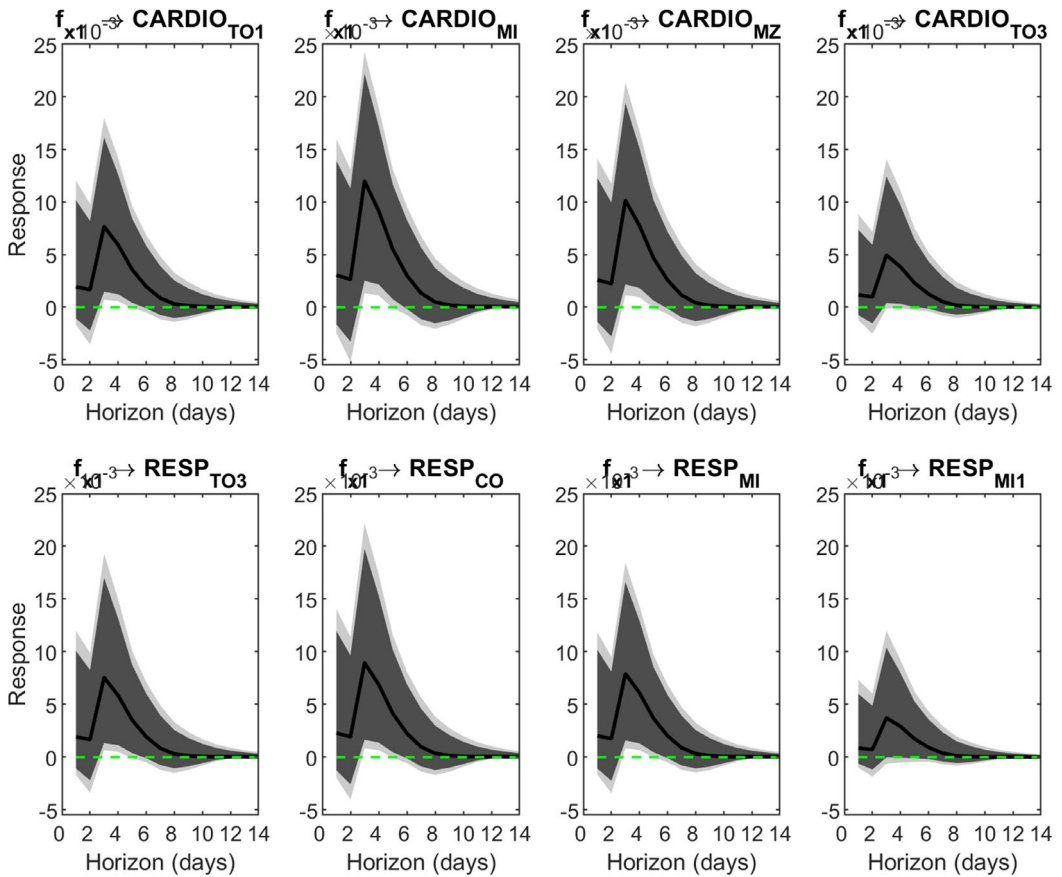[3]https://www.arpae.it/cms3/documenti/_cerca_doc/aria/rapporto_ninfa.pdf

FIGURE 12 Posterior median impulse responses (solid line) of respiratory and cardiovascular diseases to a unit shock of f_{x1}. The credibility intervals at 90% and 95% are represented by shaded areas. As an example, the panels show the response functions at TO1, TO3, MI, MZ and CO [Colour figure can be viewed at wileyonlinelibrary.com]

7 | DISCUSSION

It is common in public health research to have multivariate regional and geostatistical data. Often, it is desirable to examine relationships among these variables both in time and within and across regions. When the objective is to study the short-term effect of air pollution on health, it is common practice to consider one pollutant at a time, due to their high correlation. Multi pollutant methods have been recently proposed, mainly consisting of collapsing the different pollutants into air quality indexes. Other spatio-temporal approaches have proposed a multi-step analysis and have been found difficult to be extend to more than two pollutants (see, for example, Huang et al., 2018).

In this paper, we have proposed a factor analysis approach and introduced a multivariate spatio-temporal structural equation model to study the relations among multiple variables in a latent space. The model deals with both multiple diseases and multiple pollutants in a spatio-temporal design and also allows the observed variables to be from any exponential family distribution. By controlling for the effects of covariates which, if ignored can bias the estimated health effects of air pollution, the GST-SEM provides a flexible modelling of the residual

spatio-temporal variability by rewriting the random effect terms as a finite combination of factor loadings and dynamic common factors. Also, by allowing for spatially structured factor loadings, we have shown that their maps can reveal spatial patterns of interest overall the whole region. A further principal benefit of the GST-SEM is its ability to estimate the shape of the impulse response function relating increases in air pollution to health outcomes in short periods of time after an air pollution episode. This is a subject that has only been investigated in a time series context to date. In this paper, we have suggested possible strategies to produce impulse response functions at different levels of the hierarchy and this has allowed to estimate the response of the system both at global (whole region) and local (district) levels. All the models estimated here suggest that significant effects on hospitalizations can be noticed after two or three days from a given shock on pollution. However, effectively using a multiplier analysis in a multivariate spatio-temporal design remains challenging. According to our modelling strategy, we have studied how the effect of a shock (at the structural level) on the common factors propagates through the measurement equations on both health data and pollutants. The principle underlying this approach is that the air we breathe is a complex mixture of different pollutants which, because of the strong linear correlations, are generated by common processes or driven by similar factors, such as traffic or combustion. Hence, it is difficult in practice to disentangle the effect of the single pollutants and policymakers can only control for these general factors. However, although the results of our analysis suggest some possible hypotheses, more focused studies have to be conducted to obtain more precise information about the biological mechanisms. As suggested by one of the referees, this may be favoured by also controlling for further covariates related to the spatial topography of the territory and air mass exchange. This will be an issue for future work.

DATA AVAILABILITY STATEMENT

The Health data that support the findings of this study are available from Italian National Health Service (INHS). Restrictions apply to the availability of these data, which were used under license for this study. Data are available from the authors with the permission of INHS. The environmental data have no restrictions and are available from the authors. However, matlab codes to reproduce results reported in the simulation section are available. The complete set of functions can be easily adapted to analyze data similar to ours.

ORCID

Dani Gamerman ⬤ https://orcid.org/0000-0003-4778-0254
Luigi Ippoliti ⬤ https://orcid.org/0000-0003-2335-746X

REFERENCES

Banerjee, S. & Gelfand, A.E. (2002) Prediction, interpolation and regression for spatial misaligned data points. *Sankhya*, 64, 227–245.

Banerjee, S., Gelfand, A.E., Finley, A.O. & Sang, H. (2008) Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society, Series B*, 70, 825–848.

Barbieri, M.M. & Berger, J.O. (2004) Optimal predictive model selection. *The Annals of Statistics*, 32, 870–897.

Blangiardo, M., Pirani, M., Kanapka, L., Hansell, A. & Fuller, G. (2019) A hierarchical modelling approach to assess multi pollutant effects in time-series studies. *Plos One*, 14, 1–16.

Bob, J.F., Dominici, F. & Peng, R.D. (2013) Reduced hierarchical models with application to estimating health effects of simultaneous exposure to multiple pollutants. *Journal of the Royal Statistical Society, Series C*, 62, 451–472.

Bollen, K.A. (1989) *Structural equations with latent variables*. New York, NY, USA: John Wiley & Sons.

Bruno, F., Cameletti, M., FrancoVilloriac, M., Greco, F., Ignaccolo, R., Ippoliti, L. et al. (2016) A survey on ecological regression for health hazard associated with air pollution. *Spatial Statistics*, pp. 1–15.

Carter, C.K. & Kohn, R. (1994) On gibbs sampling for state space models. *Biometrika*, 81, 541–553.

Celeux, G., Forbes, F., Robert, C. & Titterington, D. (2006) Deviance information criteria for missing data models. *Bayesian Analysis*, 1, 651–674.

Choi, J., Fuentes, M. & Reich, B.J. (2009) Spatial-temporal association between fine particulate matter and daily mortality. *Computational Statistics & Data Analysis*, 53, 2989–3000.

Dockery, D.W., Pope, C.A., Xu, X., Spengler, J.D., Ware, J.H., Fay, M.E. et al. (1993) An association between air pollution and mortality in six US cities. *New England Journal of Medicine*, 329, 1753–1759.

Dominici, F., Daniels, M., Zeger, S.L. & Samet, J.M. (2002) Air pollution and mortality: estimating regional and national dose-response relationships. *Journal of the American Statistical Association*, 97, 100–111.

Dominici, F., Peng, R.D., Barr, C.D. & Bell, M.L. (2010) Protecting human health from air pollution: shifting from a single-pollutant to a multipollutant approach. *Epidemiology*, 21, 187–194.

Fernandez, C., Ley, E. & Steel, M.F.J. (2001) Benchmark priors for bayesian model averaging. *Journal of Econometrics*, 100, 381–427.

Fontanella, L., Ippoliti, L., Sarra, A., Nissi, E. & Palermi, S. (2019) Investigating the association between indoor radon concentrations and some potential influencing factors through a profile regression approach. *Environmental and Ecological Statistics*, 26, 185–216.

Frühwirth-Schnatter, S. (1994) Data augmentation and dynamic linear models. *Journal of Time Series Analysis*, 15, 183–202.

Gamerman, D. (1998) Markov chain monte carlo for dynamic generalised linear models. *Biometrika*, 85, 215–227.

Gamerman, D. & Lopes, H.F. (2006) *Markov Chain Monte Carlo*, 2nd edn. New York: Chapman & Hall/CRC.

Gelfand, A.E., Zhu, L. & Carlin, B.P. (2001) On the change of support problem for spatiotemporal data. *Biostatistics*, 2, 31–45.

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

Gelman, A. (1996) Inference and monitoring convergence. In: Gilks, W.R., Richardson, S. & Speigelhalter, D. (Eds.) *Markov Chain Monte Carlo in practice*, Boca Raton: Chapman & Hall, pp. 131–143.

George, E.I. & McCulloch, R.E. (1993) Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 85, 398–409.

George, E., Sun, D. & Ni, S. (2008) Bayesian stochastic search for var model restrictions. *Journal of Econometrics*, 142, 553–580.

Gilks, W.R. & Wild, P. (1992) Adaptive rejection sampling for gibbs sampling. *Journal of the Royal Statistical Society, Series C*, 41, 337–348.

Gotway, C.A. & Young, L.J. (2002) Combining incompatible spatial data. *Journal of the American Statistical Association*, 97, 632–648.

Greven, S., Dominici, F. & Zeger, S. (2011) An approach to the estimation of chronic air pollution effects using spatio-temporal information. *Journal of the American Statistical Association*, 106, 396–406.

Huang, G., Lee, D. & Scott, M. (2018) Multivariate space-time modelling of multiple air pollutants and their health effects accounting for exposure uncertainty. *Statistics in Medicine*, 37, 1134–1148.

Ippoliti, L., Valentini, P. & Gamerman, D. (2012) Space-time modelling of coupled spatiotemporal environmental variables. *Journal of the Royal Statistical Society, Series C*, 61, 175–200.

Joreskog, K.G. (1981) Analysis of covariance structures. *Scandinavian Journal of Statistics*, 8, 65–92.

Katsouyanni, K., Samet, J.M., Anderson, H.R., Atkinson, R., Le, T.A., Medina, S. et al. (2009) Air pollution and health: a European and north American approach (aphena). *Research Reports Health Effects Institute*, 142, 5–90.

Knorr-Held, L. (2000) Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in Medicine*, 19, 2555–2567.

Knorr-Held, L. & Rue, H. (2002) On block updating in markov random field models for disease mapping. *Scandinavian Journal of Statistics*, 29, 597–614.

Lawson, A., Choi, J., Cai, B., Hossain, M., Kirby, R. & Liu, J. (2012) Bayesian 2-stage space-time mixture modeling with spatial misalignment of the exposure in small area health data. *Journal of Agricultural, Biological, and Environmental Statistics*, 17, 417–441.

Lee, D. & Sahu, S.K. (2016) Estimating the health impact of environmental pollution fields. In: Haining, R., Lawson, A., Banerjee, S. & Ugarte, M.D. (Eds.) *Handbook of spatial epidemiology*, Boca Raton: Chapman and Hall/CRC, pp. 271–288.

Lee, D. & Shaddick, G. (2010) Spatial modeling of air pollution in studies of its short-term health effects. *Biometrics*, 66, 1238–1246.

Lee, D., Rushworth, A. & Sahu, S.K. (2014) A Bayesian localized conditional autoregressive model for estimating the health effects of air pollution. *Biometrics*, 70, 419–429.

Lepeule, J., Laden, F., Dockery, D. & Schwartz, J. (2012) Chronic exposure to fine particles and mortality: an extended follow-up of the harvard six cities study from 1974 to 2009. *Environmental Health Perspectives*, 120, 965–970.

Liu, X., Wall, M.M. & Hodges, J.S. (2005) Generalized spatial structural equation modeling. *Biostatistics*, 6, 539–557.

Liu, Y., Sun, J., Gou, Y., Sun, X., Zhang, D. & Xue, F. (2020) Analysis of short-term effects of air pollution on cardiovascular disease using bayesian spatio-temporal models. *International Journal of Environmental Research and Public Health*, 17, 879–891.

Lopes, H. & West, M. (2004) Bayesian model assessment in factor analysis. *Statistica Sinica*, 14, 41–67.

Lopes, H.F., Salazar, E. & Gamerman, D. (2008) Spatial dynamic factor analysis. *Bayesian Analysis*, 3, 759–792.

Lopes, H.F., Gamerman, D. & Salazar, E. (2011) Generalized spatial dynamic factor models. *Computational Statistics and Data Analysis*, 55, 1319–1330.

Lutkepohl, H. (2005) *New introduction to multiple time series analysis*. Berlin: Springer.

Marshall, E.C. & Spiegelhalter, D.J. (2003) Approximate cross-validatory predictive checks in disease mapping models. *Statistics in Medicine*, 22, 1649–1660.

Peng, R.D., Dominici, F. & Louis, T.A. (2006) Model choice in time series studies of air pollution and mortality. *Journal of the Royal Statistical Society, Series A*, 169, 179–203.

Peng, R.D., Dominici, F. & Welty, L.J. (2009) A bayesian hierarchical distributed lag model for estimating the time course of risk of hospitalization associated with particulate matter air pollution. *Journal of the Royal Statistical Society: Series C*, 58, 3–24.

Ren, Q. & Banerjee, S. (2013) Hierarchical factor models for large spatially misaligned datasets: a low-rank predictive process approach. *Biometrics*, 69, 19–30.

Rushworth, A., Lee, D. & Mitchell, R. (2014) A spatio-temporal model for estimating the longterm effects of air pollution on respiratory hospital admissions in greater London. *Spatial and Spatio-Temporal Epidemiology*, 10, 29–38.

Sahu, S. & Mardia, K. (2005) A Bayesian Kriged-Kalman model for short-term forecasting of air pollution level. *Journal of the Royal Statistical Society, Series C*, 54, 223–244.

Schmidt, A. & Gelfand, A.E. (2003) A bayesian coregionalization model for multivariate pollutant data. *Journal of Geophysics Research*, 108, 8783.

Shaddick, G. & Zidek, J. (2015) *Spatio-temporal methods in environmental epidemiology*. Boca Raton: Chapman & Hall.

Tsay, R. (2014) *Multivariate time series analysis: with R and financial applications*. Princeton: Wiley.

Tzala, E. & Best, N. (2008) Bayesian latent variable modelling of multivariate spatio-temporal variation in cancer mortality. *Statistical Methods in Medical Research*, 17, 97–118.

Valentini, P., Ippoliti, L. & Fontanella, L. (2013) Modeling us housing prices by spatial dynamic structural equation models. *The Annals of Applied Statistics*, 7, 763–798.

Wackernagel, H. (2003) *Multivariate geostatistics*. Heidelberg: Springer-Verlag.

Wang, F. & Wall, M.M. (2003) Generalized common spatial factor model. *Biostatistics*, 4, 569–582.

Welty, L.J., Peng, R.D., Zeger, S.L. & Dominici, F. (2009) Bayesian distributed lag models: estimating effects of particulate matter air pollution on daily mortality. *Biometrics*, 65, 282–291.

Wikle, C.K. & Cressie, N. (1999) A dimension-reduced approach to space-time kalman filtering. *Biometrika*, 86, 815–829.

Wilson, A., Zigler, C.M., Patel, C.J. & Dominici, F. (2018) Model-averaged confounder adjustment for estimating multivariate exposure effects. *Biometrics*, 74, 1034–1044.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Gamerman, D., Ippoliti, L. & Valentini, P. (2022) A dynamic structural equation approach to estimate the short-term effects of air pollution on human health. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 71(3), 739–769. Available from: https://doi.org/10.1111/rssc.12554