Multivariate spatio temporal models for large datasets and joint exposure to airborne multipollutants in Europe.

A. Fassó[1,*], F. Finazzi[1] and F. Ndongo[1]

[1] *University of Bergamo, Via Marconi 5, 24044 Dalmine BG, Italy; alessandro.fasso@unibg.it*
[*] *Corresponding author*

Abstract. *We consider the distribution of population by exposure to multiple airborne pollutuants at various spatial and temporal resolutions over Europe. The estimation of this distribution and its uncertainty are obtained via model based high resolution semiparametric estimates of daily average concentrations for seven pollutants in years 2009-2011. In order to exploit the spatial information content and allow the computation of daily multipollutant exposure distribution, uncertainty included, we use a multivariate spatio-temporal model capable to handle non Gaussian large datasets such as multivariate and multiyear daily air quality, land use and meteorological data over Europe.*

Keywords. *EM algorithm; D-STEM software; Air quality.*

1 Introduction

Following the development of modern fixed monitoring networks, human exposure to airborne pollution is often related to the so called "ambient exposure", namely the pollutant concentration at which people are exposed when outdoor, which is assessed by means of pollutant concentration data coming from the mentioned monitoring networks. Since using the ambient exposure as personal exposure may lead to an ecological fallacy, individual exposure is deserving an increasing attention in recent years. The exposure of an individual to airborne pollutants can be defined as the total or average amount of pollutants the person is exposed to (and can breathe) during a given period of time of his/her life. Clearly, this amount depends on a large number of factors, the main factors being the spatial location of the person and, conditionally on this, his/her life style and activity level. This is often called "personal exposure" as, in principle, it can be related to each specific person.

In the recent statistical literature, two main approaches to model personal exposure emerged. On the one side the probabilistic approach of Zidek et al. (2005) allows to simulate individual exposure temporal profiles for a finite set of individuals and, based on this, it allows to estimate a simulated population distribution. On the other side, an empirical approach based on personal monitors has been considered be few authors, see e.g. McBdride et al (2007) or Jahan et al (2013).

In principle, aggregating personal exposures bring us to population exposure of an entire city or country or continent. Since this *total* personal exposure is difficult or impossible to compute in practice,

Pollutant	C_6H_6	CO	NO_2	O_3	PM_{10}	$PM_{2.5}$	SO_2
Stations	344	591	1978	1800	1837	748	1340
Missing	47%	30%	22%	16%	23%	32%	26%

Table 1: European monitoring network details.

Finazziet al. (2011) and Shaddick et al (2013) consider a simplified concept of "population exposure", or exposure burden, which is intermediate between ambient and personal exposure and assumes that, for a large fraction of the population, ambient and personal exposure are roughly proportional. In other words exposure to background pollution is considered. Moreover the latter authors consider spatio temporal modelling to estimate the exposure burden at the country level while the former authors consider a purely spatial model at continental level.

In this paper we go on in this direction by extending the population exposure distribution introduced by Finazzi et al (2011) to the trans-Gaussian mutivariate case and by proposing a new method for computing its estimation uncertainty using methods that can handle large datasets. In particular the trans-Gaussian approach used is an evolution of Rister and Lahiri (2013) which proposes a bias correction based on Bootstrap. In particular, considering data for the European continent, we estimate multipollutant concentrations using a robust semiparamtetric extension of the model introduced by Fassó and Finazzi (2011) and Finazzi and Fassò (2014), which has been proven to successfully handle continental size airquality datasets. Model outputs are than used extensively to compute the exposure distribution as a tool for summarizing the risk related to air quality for European countries and some metropolitan areas.

2 Data

The European exposure data used here are daily data of airborne pollutants over Europe from background monitoring stations. In particular, we have seven pollutants for 3-years $(2009-2011)$, and we consider benzene (C_6H_6), carbon monoxide (CO), nitrogen dioxide (NO_2), ozone (O_3), particulate matters $(PM_{10}$ and $PM_{2.5})$ and sulphur dioxide (SO_2). The network is heterogeneous with extensive missing data, as shown in the in Figure 1 and Table 1, where we have more than 9×10^6 response variable observations with and more than 2×10^6 missing data. Covariates include Meteo (Wind speed, Pressure and Air Temperature), Land Use, Elevation (Alt) and Population, Weekend effects (WE).

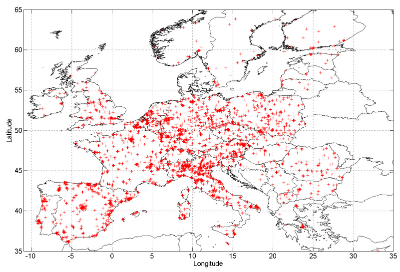


Figure 1: Monitoring network of NO_2

Subset name	Subset size	Subset use
D1	2/3	$STEM$ estimation $\rightarrow \hat{\Psi}$
D2	1/6	TG mixture estimation $\mid \hat{\Psi} \rightarrow \left(y_i \mid \hat{\xi}_i\right)$
D3	1/6	Model performance of $\hat{y}$ and $\check{y}$

Table 2: Data usage.

3 TG-STEM Model

We use a seven-variate trans-Gaussian model (TG-STEM) given by

$$\xi_i(u) = \alpha_i \omega_i(u) + \beta_{i,0}(t) + \beta_{i,1}(t)' Meteo(u) + \beta_{i,2}(t) Alt(s) + \beta_{i,3}(t) Pop(s) + \beta_{i,4}' WE(t) + \varepsilon_i(u)$$

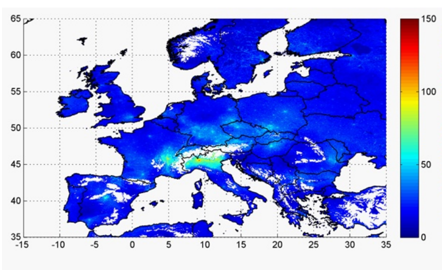
where $u = (s,t)$, $s \in D$, for some spatial domain D, $t = 1,2,...,n$; ξ_i is the standardized log-transformation of data y_i, $i = C_6H_6, CO, NO_2, O_3, PM_{10}, PM_{2.5}$ and SO_2 in $\mu g/m^3$; $\beta_{i,j}(t) = \beta_{i,j} + Z_{ij}(t)$ are time varying coefficients; $Z = [Z_{i,j}]$ is a Vector Markovian process; $W = (\omega_1, ..., \omega_7)$ is a linear coregionalization model ε_i are independent Gaussian errors.

Model selection and fitting are performed considering the original large European dataset D discussed in Section 2 and randomly splitting in three parts as described in Table 2. In particular in D1 estimation is performed using the EM algorithm in Finazzi and Fassó (2014). While, in D2 a nonparametric back transform is developed, while D3 is used for crossvalidation.

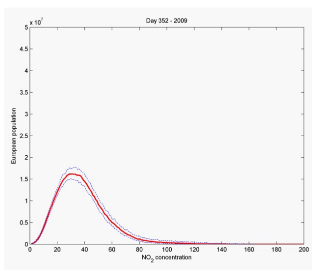
4 Exposure distribution

In Figure 2, we compute high resolution dynamic maps using TG-STEM $\check{y}(s,t) = E\left(y(s,t) \mid \hat{\xi}(s,t)\right)$ where $\hat{\xi}(s,t) = E(\log y(s,t) \mid Y)$ and, using plug-in approach, we compute the daily exposure distribution for each pixel, uncertainty included, namely

$$\hat{F}_i(y,t) = \frac{n\left(\hat{y}_i(\mathcal{B}_r, t) \leq y,\ r = 1,...,R\right)}{n(\mathcal{R})}.$$



High Resolution Dynamic Mapping



Exposure distribution and uncertainty

Figure 2: Mapping (left panel) and exposure distribution (right panel) for NO_2, day 352/2009.

We compute these distributions at various aggregation levels, in space from province to country, in time from day to year, and make various comparisons of European countries. For example in Figure 3, we compare three important metropolitan areas. The multivariate approach has two advantages. On the one side allows to improve spatial information for poorly monitored pollutants such as benzene. Moreover it allows to asses joint exposure distribution which is important for health protection.

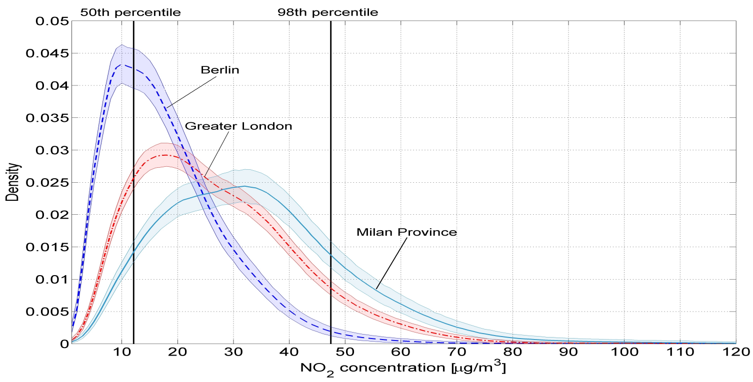


Figure 3: Exposure distribution for NO_2 in 2009, with 95% confidence bands. Vertical bars: $50th$ and $98th$ EU-percentiles.

References

[1] Fassò A., Finazzi F., (2011). Maximum likelihood estimation of the dynamic coregionalization model with heterotopic data. *Environmetrics*. Vol. 22:6, 735-748.

[2] Finazzi F., Fassò A. (2014). D-STEM: A Software for the Analysis and Mapping of Environmental Space-Time Variables. *Journal of Statistical Software*. **62**(6), 1–25.

[3] Finazzi F, Scott M.E, Fassó A. (2013). A model based framework for air quality indices and population risk evaluation. With an application to the analysis of Scottish air quality data. *Journal of the Royal Statistical Society, series C*. **62**(2), 287-308.

[4] Jahnan H.K., Kraemer A., Chen X.C., Chan C.Y., Engling G., Ward T.J. (2013). Ambient and personal PM2.5 exposure assessment in the Chinese megacity of Guangzhou. *Atmospheric Environment*, 74, 402-411.

[5] McBride S.J., Williams R.W., Creason J (2007). Bayesian hierarchical modeling of personal exposure to particulate matter. *Atmospheric Environment*, 41:29, 6143-6155.

[6] Rister K and Lahiri SN (2013). Bootstrap based Trans-Gaussian Kriging. *Statistical Modelling* 13(5&6): 509–539.

[7] Shaddick G., Yan H., Salway R., Vienneau D., Kounali D. & Briggs D. (2013). Large-scale Bayesian spatial modelling of air pollution for policy support, *Journal of Applied Statistics*, **40**(4), 777–794,

[8] Zidek, J. V., Shaddick, G., White, R., Meloche, J. and Chatfield, C. (2005). Using a probabilistic model (pCNEM) to estimate personal exposure to air pollution. *Environmetrics*, **16**, 481–493.