Data integration for high–resolution, continental–scale estimation of air pollution concentrations

Matthew L. Thomas [1] Gavin Shaddick [1,2] Kees de Hoogh [3,4] Daniel Simpson [5] James V. Zidek [6]

[1]Department of Mathematical Sciences, University of Bath, Bath, U.K. [2]Department of Mathematics, University of Exeter, Exeter, U.K.
[3]Swiss Tropical and Public Health Institute, Basel, Switzerland. [4]University of Basel, Basel, Switzerland. [5]Department of Statistical Sciences, University of Toronto, Toronto, Canada.
[6]Department of Statistics, University of British Columbia, Vancouver, British Columbia, Canada.

Introduction

Air pollution represents one of the most important environmental risk factors to human health globally. Traditionally, information on air pollution concentrations comes from ground monitoring networks however, there may be areas in which this information sparse or is not of sufficient quality. Here, ground monitoring information is supplemented with information from other sources, including road networks, land use and chemical transport models (CTM), within a Bayesian hierarchical modelling framework. Using statistical downscaling, this enables concentrations of pollutants to be predicted, together with measures of uncertainty, at a high-resolution. Here, concentrations of fine particulate matter ($PM_{2.5}$) are obtained at a resolution of 1km × 1km for 20 countries in Western Europe.

Data

Annual average concentrations of $PM_{2.5}$ (μgm^{-3}) between 2010 and 2016 for 1210 ground monitoring sites were extracted from European Environment Agency's (EEA) Air Qualty e-Reporting database.

The study area consists of 20 countries in Western Europe: Austria, Belgium, Denmark, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Liechtenstein, Lithuania, Luxembourg, Netherlands, Norway, Portugal, Spain, Sweden, Switzerland and the United Kingdom.

Information on roads (length of major roads and length of all roads), other land use (percentage of areas that are residential, industry, ports, urban green space, built up, natural land) and altitude were retrieved from the EEA Corine Land Cover 2006, EuroStreets digital road network and SRTM Digital Elevation Database respectively, all available at a 1km × 1km resolution.

Estimates of $PM_{2.5}$ were also obtained from the MACC-II ENSEMBLE CTM, at a 10km × 10km resolution.

Statistical Downscaling

Information from roads, land use, altitude, CTMs are calibrated against ground measurements. Let Y_{st} denote the log of the annual average $PM_{2.5}$ measurement from ground monitors, available at a discrete set of N_S locations $s = \{s_1, s_2, \ldots, s_{N_S}\}$ and N_T time points $s = \{t_1, t_2, \ldots, t_{N_T}\}$

$$Y_{st} = \beta_{0st} + \sum_{p=1}^{P} \beta_{pst} X_{pl_s t} + \epsilon_{st}$$

Here, X_{plt} denote gridded estimates of roads, land use, altitude and the MACC-II ENSEMBLE CTM, on grid of N_{pL} cells $l \in \{l_1, l_2, \ldots, l_{N_{pL}}\}$ with l_s denotes the grid cell containing ground monitor(s) at location s, and $\epsilon_s \sim N(0, \sigma_\epsilon^2)$ is a i.i.d random error term.

The coefficients, β_{pst}, $p = 0, 1, \ldots, P$ comprise of a series random effects that allow the intercept and coefficient associated with the CTM to vary over space and time

$$\beta_{pst} = \alpha_p + \theta_{pt} + \kappa_{pst}.$$

- α_p are fixed effects and are assigned $N(0, 1000)$ priors
- β_{pt} are temporal random effects and are assigned RW1 priors
- κ_{pst} are spatio-temporal random effects; a multivariate Normal distribution with covariance representing a separable space-time model using a Kronecker product of space and time.

Inference

Downscaling models are often fit using Markov Chain Monte Carlo (MCMC) (Berrocal et al., 2010). However, with larger amounts of data, computation may be challenging. Here, inference is performed using integrated nested Laplace approximations (INLA), using an stochastic partial differential equation (SPDE) to provide a bridge between spatial data modelling on a continuous surface and Gaussian Markov Random Field (Lindgren et al., 2011). Fields that are utilised by R-INLA to provide efficient computation.

Prediction

Full posterior predictive distributions of air pollution concentrations are obtained using Monte Carlo simulation. This provides a computationally efficient method of performing prediction at high resolution over space and time and addresses the computational issues that may be associated with jointly fitting the model and performing prediction.

Results

Figure 1 shows the spatial variation in predicted annual average concentrations and Figure 2 the changes in annual average concentrations over the study period. In the case of the latter, it can be seen that $PM_{2.5}$ is decreasing in large parts of Western Europe.

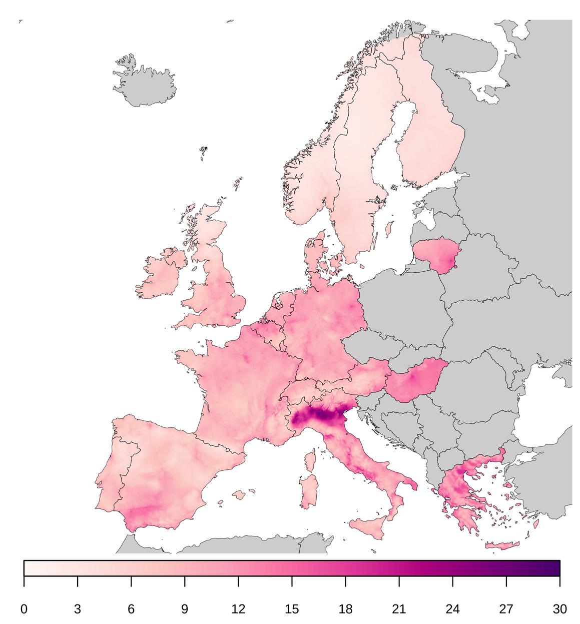

Figure 1: Median estimates of annual averages of $PM_{2.5}$ (μgm^{-3}) for 2016 for each grid cell (1km × 1km resolution).

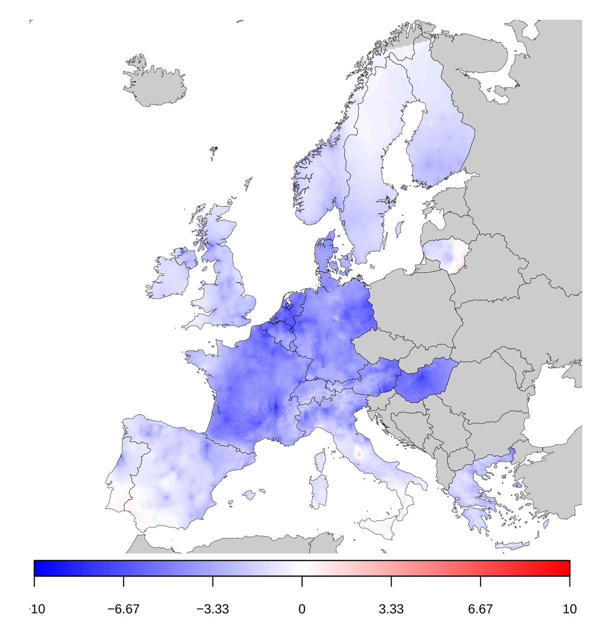

Figure 2: Median change in annual average $PM_{2.5}$ (μgm^{-3}) for each grid cell (1km × 1km resolution)

Figures 3 and 4 shows the estimated annual average $PM_{2.5}$ for Paris in 2016 from the CTM and the medians of the the posterior predictive distributions in each cell. Figure 4 shows a clear increase in granularity and provides substantial additional information on the spatial distribution of $PM_{2.5}$.

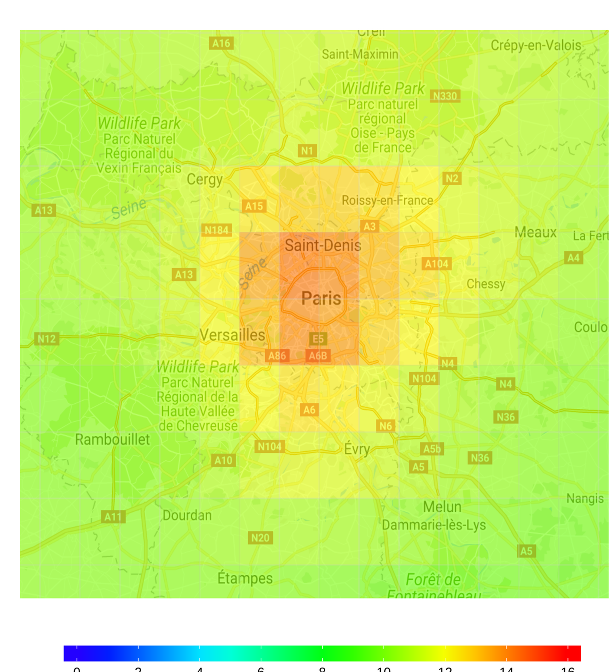

Figure 3: Estimated annual average $PM_{2.5}$ in 2016 from the CTM for Paris.

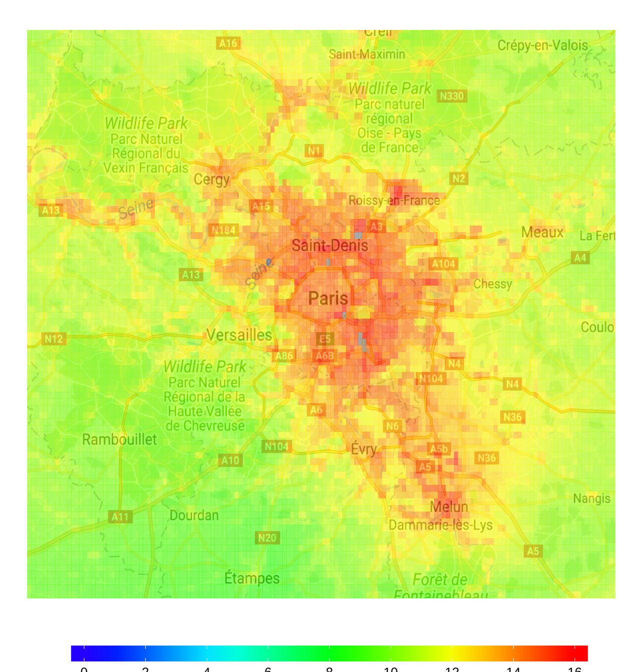

Figure 4: Predictive posterior median annual average $PM_{2.5}$ in 2016 for Paris.

Discussion

The ability to produce comprehensive, accurate, high-resolution, estimates of concentrations of pollutants over space and time is essential for policy support, epidemiological analyses and in estimating the burden of air pollution on human health.

When combining data from different sources, it is important to acknowledge possible changes in support, or differences in the spatial and temporal resolutions for which data is available. Here, ground measurements are available at point locations and the possibility of within grid-cell variability (the level at which CTM estimates are available) is incorporated in the calibration model by allowing the coefficient associated with CTM (and the intercept) to vary continuously over space and time.

Future research include extending the model to allow multiple pollutants to be considered to allow dependencies between pollutants, such as NO_2 and $PM_{2.5}$ to be exploited in order to reduce both bias and uncertainty.

References

Berrocal, V. J., A. E. Gelfand, and D. M. Holland (2010). A spatio-temporal downscaler for output from numerical models. *Journal of agricultural, biological, and environmental statistics 15*(2), 176–197.

Lindgren, F., H. Rue, and J. Lindström (2011). An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73*(4), 423–498.