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 nitrogen dioxide (NO_2) are obtained at a resolution of 1km $\times$ 1km for 20 member states of the European Union.

Data

Annual average concentrations of NO_2 (μgm^{-3}) in 2010 for 2433 ground monitoring sites were extracted from European Environment Agency's (EEA) Air Qualty e-Reporting database.

The study area consists of 20 countries within the European Union (EU): 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 at a 1km $\times$ 1km resolution.

Estimates of NO_2 were also obtained from the MACC-II ENSEMBLE CTM, at a 10km $\times$ 10km resolution.

Statistical Modelling

Ground measurements are calibrated against information from land use and the CTM. Let Y_s denote the square root of the annual NO_2 measurement from ground monitors, available at a discrete set of N_S locations $s = \{s_1, s_2, \ldots, s_{N_S}\}$,

$$Y_s = \tilde{\beta}_{0s} + \tilde{\beta}_{1s} X_{l_s} + \sum_{p=1}^{P} \gamma_p W_{pm_s} + \epsilon_s$$

where X_l denote gridded estimates from the MACC-II ENSEMBLE CTM, on grid of N_L cells $l \in \{l_1, l_2, \ldots, l_{N_L}\}$ with l_s denotes the grid cell containing ground monitor(s) at location s, W_{pm}, $p = 1, \ldots, P$ denote gridded estimates of roads, land use and altitude, for N_M cells $m \in \{m_1, m_2, \ldots, m_{N_M}\}$, with m_s denote the grid cells containing locations s and $\epsilon_s \sim N(0, \sigma_\epsilon^2)$ is a i.i.d random error term.

The coefficients, $\tilde{\beta}_{0s}$, and $\tilde{\beta}_{1s}$ denote random effects that allow the intercept and coefficient associated with the CTM to vary over space

$$\tilde{\beta}_{0s} = \beta_0 + \beta_{0s}$$
$$\tilde{\beta}_{1s} = \beta_1 + \beta_{1s}.$$

Here, β_0 and β_1 are fixed effects representing the mean value of the intercept and coefficient respectively, with β_{0s} and β_{0s} representing the local adjustment to these parameter values to allow the calibration functions to vary over space.

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.

Results

Figure 1 shows the spatial variation in predicted annual average concentrations and Figure 2 the uncertainty associated with those predictions. In the case of the latter, as might be expected, higher levels of uncertainty can be seen in areas where there is little, or no, ground monitoring.

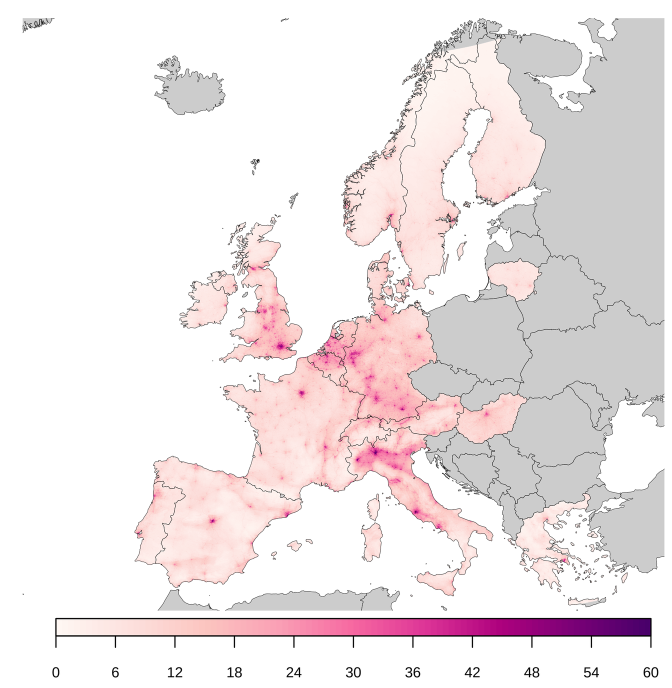

Figure 1: Median estimates of annual averages of NO_2 (μgm^{-3}) for 2010 for each grid cell (1km $\times$ 1km resolution).

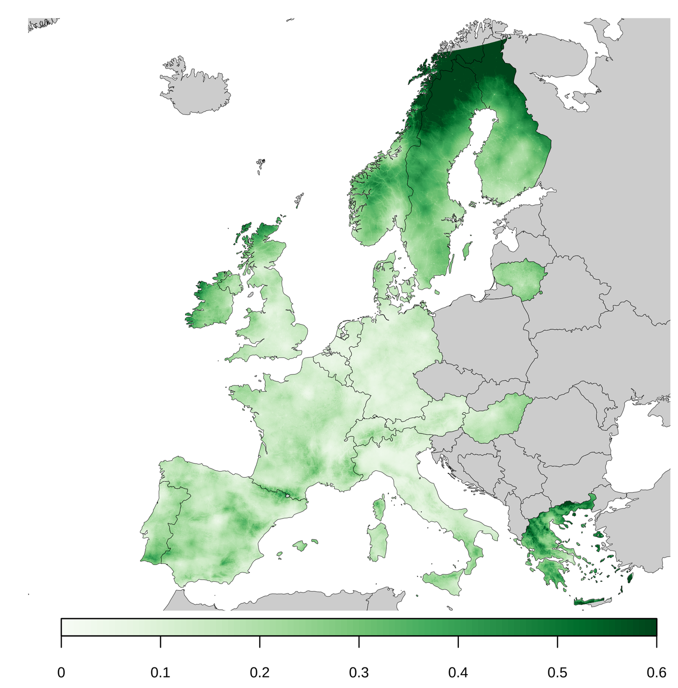

Figure 2: Coefficient of variation for 2010 for each grid cell (1km $\times$ 1km resolution)

Figures 3 and 4 show the spatial variation on in the coefficients associated with the downscaling component of the model. In each case, the deviation (of the random effects) from the fixed effects are shown. The intercept term indicates areas in which concentrations are below and above the overall mean with variations in the coefficient associated with the CTM reflecting the extent to which the CTM provides a biased estimate of surface level NO_2.

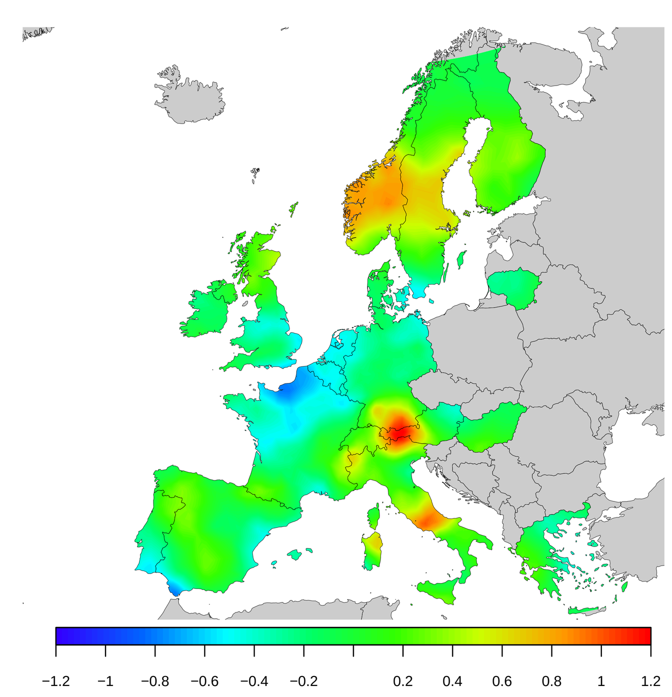

Figure 3: Deviation from the fixed effect associated with the intercept term for 2010.

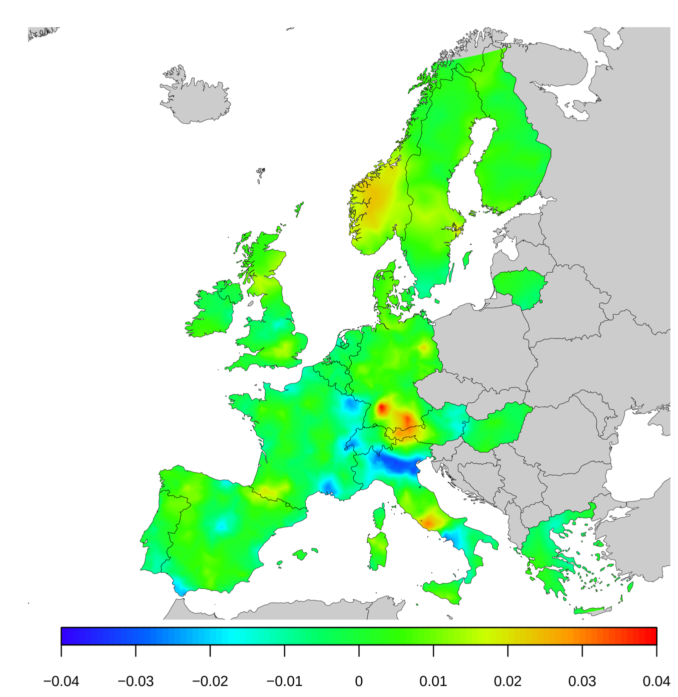

Figure 4: Deviation from the fixed effect associated with the CTM slope term for 2010.

Discussion

The ability to produce comprehensive, accurate, high-resolution, estimates of concentrations of pollutants 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.

Future research topics include extending the model to incorporate temporal data, allowing the calibration equations to vary over both space and time, and to allow multiple pollutants to be considered. In the case of the latter, the model would 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.