Data Integration Model for Air Quality: A Hierarchical Approach to the Global Estimation of Exposures to Ambient Air Pollution

Matthew L. Thomas [1] Gavin Shaddick [1] Amelia Green [1] Michael Brauer [2] Aaron van Donkelaar [3] Rick Burnett [4] Howard H. Chang [5] Aaron Cohen [6] Rita Van Dingenen [11] Carlos Dora [7] Sophie Gumy [7] Yang Liu [8] Randall Martin [3] Lance A. Waller [5] Jason West [9] James V. Zidek [10] Annette Prüss-Ustün [10]

[1]Department of Mathematical Sciences, University of Bath, Bath, UK [2]School of Population and Public Health, The University of British Columbia, Vancouver, British Colombia, Canada [3]Department of Physics and Atmospheric Science, Dalhousie University, Halifax, Nova Scotia, Canada [4]Health Canada, Ottawa, Ontario, Canada [5]Department of Biostatistics, Rollins School of Public Health, Emory University, Atlanta, Georgia, USA [6]Health Effects Institute, Boston, Massachusetts, USA [7]World Health Organisation, Geneva, Switzerland [8]Department of Environmental Health,Rollins School of Public Health, Emory University, Atlanta, Georgia, USA [9]Department of Environmental Sciences and Engineering, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA [10]Department of Statistics, University of British Columbia, Vancouver, British Columbia, Canada [11]Institute for Environment and Sustainability, Joint Research Centre, European Commission, Italy

Overview

Air pollution is a major risk factor for global health, with an estimated 3 million deaths annually being attributed to ambient fine particulate matter ($PM_{2.5}$). The primary source of information for estimating exposures to $PM_{2.5}$ has been measurements from ground monitoring networks but, although coverage is increasing, there remain regions in which monitoring is limited. Ground monitoring data therefore needs to be supplemented with information from other sources. A hierarchical modelling approach for integrating data from multiple sources is proposed allowing spatially-varying relationships between ground measurements and other factors that estimate air quality. Set within a Bayesian framework, the resulting Data Integration Model for Air Quality (DIMAQ) is used to estimate exposures, together with associated measures of uncertainty, on a high resolution grid covering the entire world.

Introduction

It is vital that the risks, trends and consequences of air pollution are monitored and modelled to develop effective environmental and public health policy. Accurate measurements of exposure in any given area are required but this is a demanding task as, in practice, the processes involved are extremely complex and because of the scarcity of ground monitoring data in some regions. The locations of ground monitoring sites within the WHO Air Pollution in Cities database are shown in Figure 1 where it can be seen that the density of monitoring sites varies considerably, with extensive measurements available in North America, Europe, China and India but with little or no measurement data available for large areas of Africa, South America and the Middle East. For this reason, there is a need to use information from other sources in order to obtain estimates of exposures for all areas of the world.

Here, a model is presented for integrating data from multiple sources, that enables accurate estimation of global exposures to fine particulate matter. Set within a Bayesian hierarchical framework, this Data Integration Model for Air Quality (DIMAQ) estimates exposures, together with associated measures of uncertainty, at high geographical resolution by utilising information from multiple sources and addresses many of the issues encountered with previous approaches.

Data

The sources of data can be categorised into one of three groups:

- Ground monitoring data;
- Estimates of $PM_{2.5}$ from remote sensing satellites and chemical transport models;
- Other sources including population, land–use and topography.

Ground monitoring is available at a distinct number of locations, whereas the latter two groups provide near complete global coverage (and have previously been shown to have strong associations with concentrations of $PM_{2.5}$). Utilising such data will allow estimates of exposures to be made for all areas, including those for which ground monitoring is sparse or non-existent.

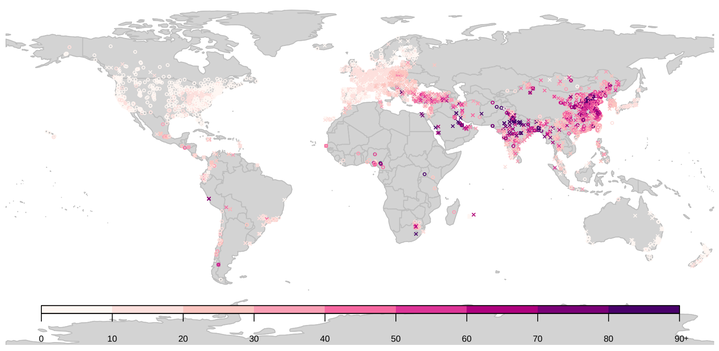

Figure 1: Locations of ground monitors measuring $PM_{2.5}$ (circles) and PM_{10} (crosses). Colours denote the annual average concentrations (μgm^{-3}) of $PM_{2.5}$ (or $PM_{2.5}$ converted from PM_{10}). Data are from 2014 (46%), 2013 (36%), 2012 (9%) and 2006-2011, 2015 (9%).

Statistical Modelling

The aim is to obtain estimates of concentrations of $PM_{2.5}$ for each of the 1.4 million grid cells, together with associated measures of uncertainty. This will be achieved by finding the posterior distributions for each cell, from which summary measures will be calculated.

The overall approach is statistical calibration: a regression model is used to express ground measurements, Y_s, available at a discrete set of N_S locations $S \in \mathcal{S}$ with labels $S = \{s_0, s_1, \ldots, s_{N_S}\}$, that are a function of covariates, X_{sr}: $r = 1, \ldots, R$, that reflect information from other sources. Covariate information may be available for point locations (as with the ground measurements) or on a grid of N_L cells, $l \in L$ where $L = l_1, ..., l_{N_L}$.
Considering a single covariate, X_{lr}, for ease of explanation,

$$Y_s = \tilde{\beta}_{0s} + \tilde{\beta}_{1s} X_{lr} + \epsilon_s \tag{1}$$

where X_{lr} is measured on a grid. Here, $\epsilon_s \sim N(0, \sigma^2_\epsilon)$ is a random error term. The terms $\tilde{\beta}_{0s}$ and $\tilde{\beta}_{1s}$ denote random effects that allow the intercept and coefficient 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 coefficients respectively, with β_{0s} and β_{1s} zero mean spatial random effects: providing (spatially driven) adjustments to these means, allowing the calibration functions to vary over space.

The structure of the random effects used here exploits a geographical nested hierarchy: each of the 187 countries considered are allocated to one of 21 regions and, further, to one of 7 super-regions. Each region must contain at least two countries and is broadly based on geographic regions/sub-continents and groupings based on country level development status and causes of death.

Data Integration Model for Air Quality

Ground measurements at point locations, s, within grid cell, l, country, i, region, j, and super–region, k are denoted by Y_{slijk}. There is a nested hierarchical structure with $s = 1, \ldots, N_l$ sites within grid cell, l: $l = 1, \ldots, N_i$, grid cells within country i; $i = 1, \ldots, N_j$, countries within region j: $j = 1, \ldots, N_k$, regions within super–region k: $k = 1, \ldots, N_k$. This structure can be seen in Figure 2.

Data Integration Model for Air Quality (ctd.)

In order to allow for the skew in the measurements and the constraint of non–negativity, the (natural) logarithm of the measurements are modelled as a function of the covariates, X_{lr}: $r = 1, \ldots, R$.

The model consists of a set of fixed and random effects, for both intercepts and covariates, and is given as follows,

$$\begin{aligned} \log(Y_{slijk}) = \beta_0 &+ \sum_{p=1}^{P} \beta_p X_{p,slijk} \\ &+ (\beta^G_{0l} + \beta^C_{0i} + \beta^R_{0j} + \beta^{SR}_{0k}) \\ &+ \sum_{q \in Q} (\beta^C_{qi} + \beta^R_{qj} + \beta^{SR}_{qk}) X_{slijk} \\ &+ \epsilon_{slijk} \; . \end{aligned} \tag{2}$$

Here, the set of R covariates consists of two groups, $R = (P, Q)$ where P are those with just fixed effects and Q those additionally assigned random effects. For the intercept terms: β^G_{0i}, β^C_{0i}, β^R_{0j} and β^{SR}_{0k} represent the coefficients for grid cell l, country i within region j and super–region k. For the covariate effects, the notation is the same but there is no grid cell effect, with β^C_{qi}, β^R_{qj} and β^{SR}_{qk} representing country, region and super–region effects respectively. The term ϵ_{slijk} denotes random error term with $\epsilon_{slijk} \sim N(0, \sigma^2_\epsilon)$.

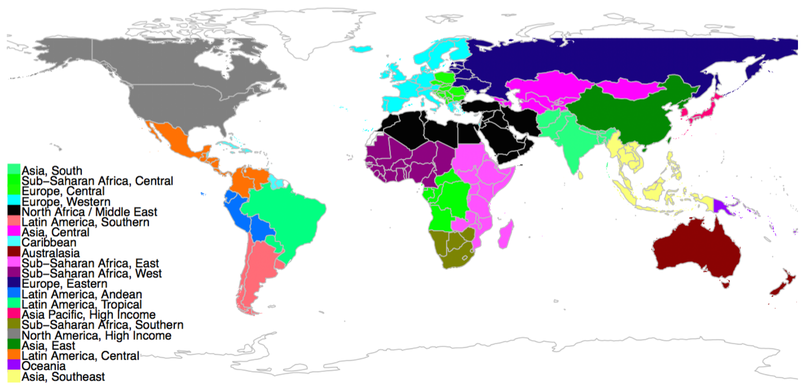

Figure 2: Schematic showing the nested geographical structure of countries within regions.

Inference

The model presented here is a Latent Gaussian Model which means that advantage can be taken of methods offering efficient computation when performing Bayesian inference. It was implemented using approximate Bayesian inference using integrated nested Laplace Approximations using the R-INLA software.

Results

A series of models based on the structure described above were applied with the aim of assessing the predictive ability of potential explanatory factors. The choice of which variables were included in the final model was made based on their contribution to within-sample model fit and out-of-sample predictive ability.

Model	R^2	DIC	RMSE†	PwRMSE†
(i)	0.54 (0.53, 0.54)	7828 (7685, 8657)	17.1 (16.5, 18.1)	23.1 (20.5, 29.3)
(ii)	0.90 (0.90, 0.91)	1105 (849, 1239)	11.2 (10.1, 12.9)	13.0 (11.5, 23.5)
(iii)	0.90 (0.90, 0.91)	986 (704, 1115)	11.1 (10.0, 13.3)	12.8 (11.2, 23.0)
(iv)	0.91 (0.90, 0.91)	877 (640, 1015)	10.7 (9.5, 12.3)	12.1 (10.7, 21.4)
(v)	0.91 (0.90, 0.92)	777 (508, 919)	10.7 (9.5, 12.5)	12.0 (10.7, 20.7)

† μgm^{-3}

Table 1: Summary of results from fitting five candidate models. Model (i) is the model used in the GBD2013 study. Model (ii) is a hierarchical model containing satellite based estimates of $PM_{2.5}$, population and local network characteristics. Models (iii-v) contain additional variables: model (iii), estimates of $PM_{2.5}$ from the TM5-FASST chemical transport model; model (iv), estimates of specific chemical components and dust from the GEOS-Chem chemical transport model and information on differences in elevation between a ground monitor and its surrounding grid cell (as defined by the GEOS-Chem chemical transport model); (v) both the estimates from the TM5-FASST and GEOS-Chem models. Results are for both in-sample model fit and out-of-sample predictive ability and are the median (minimum, maximum) values from 25 training-validation set combinations. For within sample model fit, R^2 and Deviance Information Criteria (DIC). For out-of-sample predictive ability, root mean squared error (RMSE) and population weighted root mean squared error (PwRMSE).

Discussion

We have developed a model to produce a comprehensive set of high-resolution estimates of exposures to fine particulate matter. This work presents an important step forward in large-scale data integration in this setting, allowing information on air quality to be drawn from a wide variety of sources, each potentially measured at different resolutions, with different error structures and with different levels of uncertainty. Ultimately, this will lead to more accurate estimates of air quality together with measures of uncertainty that acknowledge the uncertainty associated with the individual data sources.

This information can also be incorporated within a health effects model leading to improved characterisation of uncertainty when estimating disease burden. This in turn will lead to increased understanding of the effects of air pollution on health and the potential effects of mitigation strategies.