Global modelling of air pollution using multiple data sources

Matthew Thomas – M.L.Thomas@bath.ac.uk

Supervised by Dr. Gavin Shaddick
In collaboration with IHME and WHO

July 17, 2018

Motivation

- Air pollution is an important determinant of health and poses a significant threat globally.
- It is known to trigger cardiovascular and respiratory diseases in addition to some cancers.
- Particulate Matter ($PM_{2.5}$) is estimated to be
  - 4th highest health risk factor in the world
  - attributable to 5.5 million premature deaths
- There is convincing evidence for the need to model air pollution effectively.

REQUIREMENTS

- WHO and other partners plan to strengthen air pollution monitoring globally.
- This will produce accurate and convincing evidence of risks posed.

Ground Monitoring

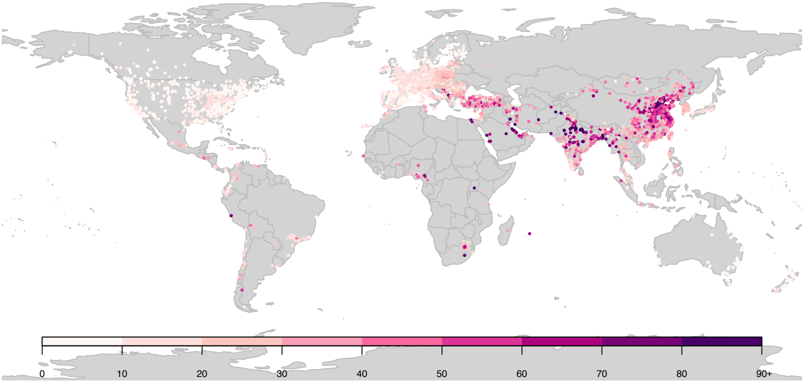

Figure: World map with ground monitor locations, coloured by the estimated level of $PM_{2.5}$ in $\mu g m^{-3}$.

Requirements

- WHO and other partners plan to strengthen air pollution monitoring globally.
- This will produce accurate and convincing evidence of risks posed.
- Allow data integration from different sources.
- This will allow borrowing from each methods respective strengths.
- Currently, three methods are considered:
  - Ground Monitoring,
  - Satellite Remote Sensing and
  - Atmospheric Modelling

Satellite Remote Sensing

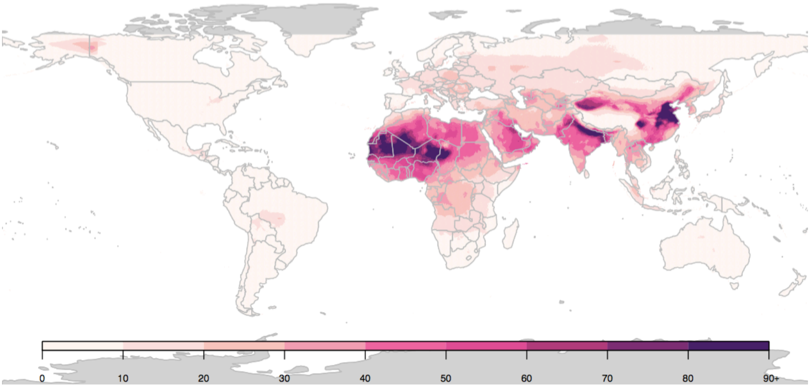

Figure: Global satellite remote sensing estimates of $PM_{2.5}$ in $\mu g m^{-3}$ for 2014 used in GBD2015

Atmospheric Modelling

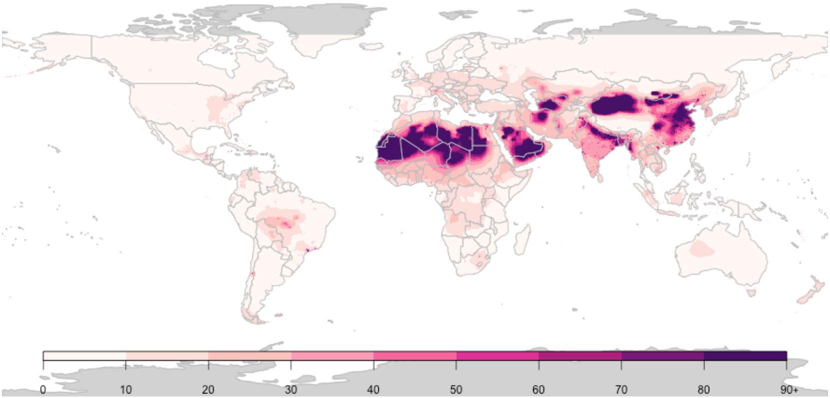

Figure: Global chemical transport model estimates of $PM_{2.5}$ in $\mu g m^{-3}$ for 2014 used in GBD2015

Population Estimates

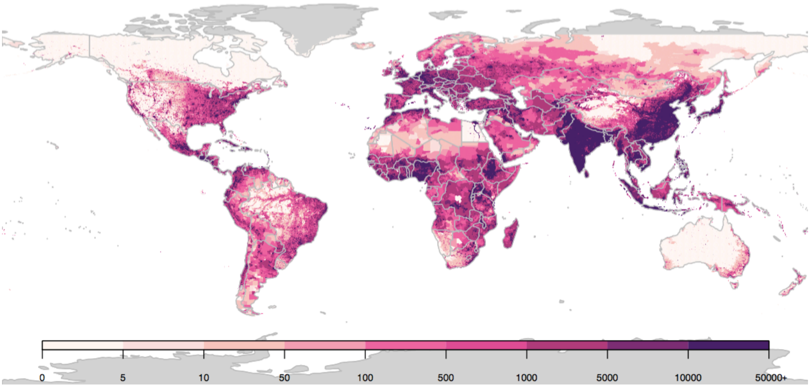

Figure: Estimate of population density per $0.1^o \times 0.1^o$ grid location for 2014 used in GBD2015

Previous Approach

- The current GBD approach to modelling combines estimates from atmospheric models and satellites into a 'fused' estimate.
- Let x_i^{am} and x_i^{sat} be atmospheric model and satellite estimates for grid cell i, then the fused estimate is defined as:

$$x_i^{fus} = \frac{x_i^{sat} + x_i^{am}}{2}.$$

- The ground monitors and grid data are calibrated, logged and fused data is used as an explanatory variable in a linear model to determine ground level $PM_{2.5}$:

$$\log\left(y_i^{gm}\right) = \beta_0 + \beta_1 \log\left(x_i^{fus}\right) + \epsilon_i \quad i = 1, \ldots, n.$$

- Ground level $PM_{2.5}$ is then estimated using tradition linear modelling techniques.

A Multilevel Random Effect Model

- Suppose that a ground monitor at location s is situated in grid cell B_j.
- To avoid non-negativity and skew we consider the estimates of $PM_{2.5}$ on the log-scale
- We then assume that the log estimates of $PM_{2.5}$ from ground monitors, y_s are normally distributed

$$y_s = z_{B_i} + \sum_{j=1}^{n} \gamma_j x_{s,j} + \epsilon_s$$

where

  - $x_{s,j}$ are covariate information for ground monitor at location s,
  - z_{B_i} is a mean trend for grid cell B_i
  - $\epsilon_s \sim N(0, \sigma_\epsilon^2)$ is measurement error.

A Multilevel Random Effect Model

- The mean trend z_{B_i} for grid cell B_i is modelled using the following,

$$z_{B_i} = \tilde{\beta}_0 + \sum_{j=1}^{k} \tilde{\beta}_j u_{B_i,j} + \sum_{j=k+1}^{m} \beta_j u_{B_i,j} + e_{B_i}$$

where

  - $u_{B_i,j}$ are covariate information for grid cell B_i,
  - $e_s \sim N(0, \sigma_e^2)$ is the within cell variability.

A Multilevel Random Effect Model

- To allow for more local variation we allow a series random effects

$$\tilde{\beta}_j = \beta_{j0} + \sum_{k=1}^{K} \beta_{jk}$$

- We propose these random effects to have a nested hierarchy to allow borrowing between levels.
- Here we aggregate countries into regions and regions into superregions
  - Using country level mortality levels and causes of death

DEFINED REGIONS

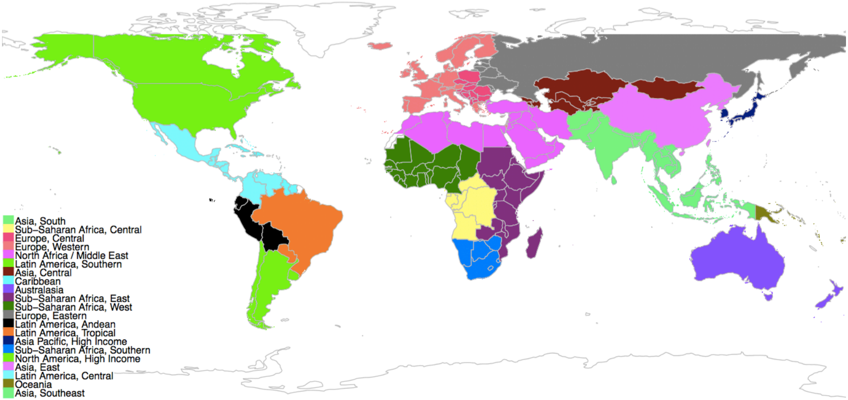

Figure: World map coloured by GBD defined Regions

Defined Super Regions

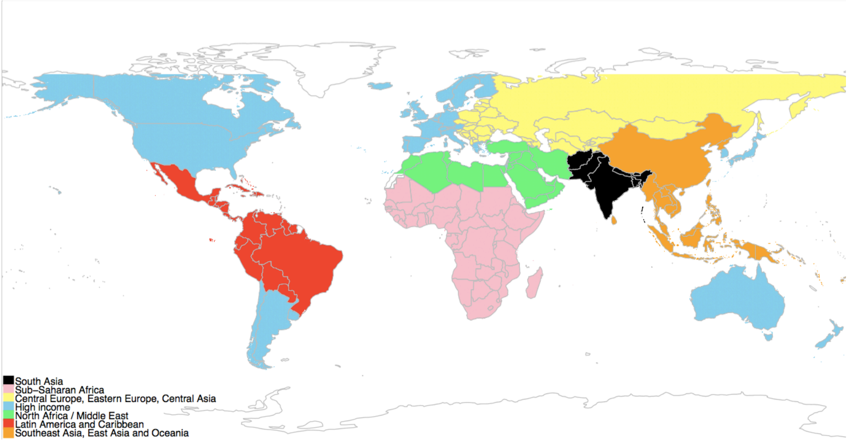

Figure: World map coloured by GBD defined Super Regions

Computation

- Bayesian models of this complexity do not have analytical solutions.
- 'Big' data means traditional MCMC techniques are impractical.
- Recent advances in approximate Bayesian inference provide fast and efficient methods for modelling, such as Integrated Nested Laplace Approximations (INLA).
- INLA performs numerical calculations of posterior densities using Laplace Approximations hierarchical latent Gaussian models:

$$p(\theta_k|\boldsymbol{y}) = \int p(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}_{-k} \quad p(z_j|\boldsymbol{y}) = \int p(z_j|\boldsymbol{\theta},\boldsymbol{y})p(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}$$

- A latent Gaussian process allows for sparse matrices, and therefore efficient computation.

Computation

- Already suite of programs to implement these (R-INLA).
- However, while INLA is computationally more attractive, R-INLA still requires huge computation and memory usage.
- Unable to run this model on standard computers (4-8GB RAM).
- Required the use of a High-Performance Computing (HPC) service.
  - Balena cluster at University of Bath.
  - $2 \times$ 512GB RAM nodes ($32 \times$ 32GB RAM cores).
- Took an iterative approach to prediction.

PREDICTIONS

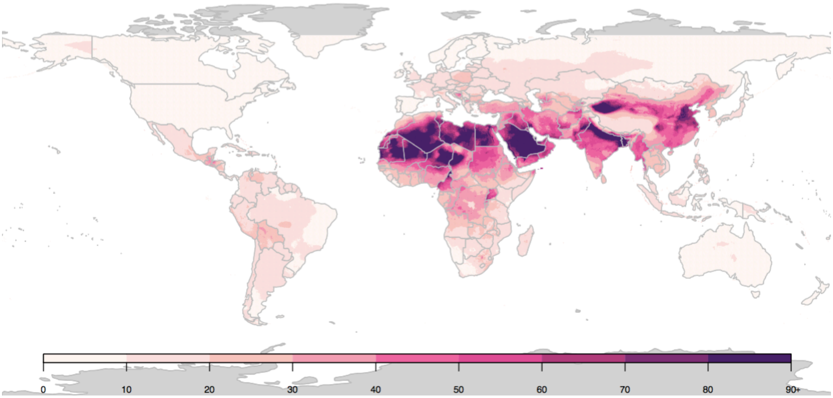

Figure: Predictions of $PM_{2.5}$ in $\mu g m^{-3}$, from hierarchical model for 2014.

Predictions: Regional

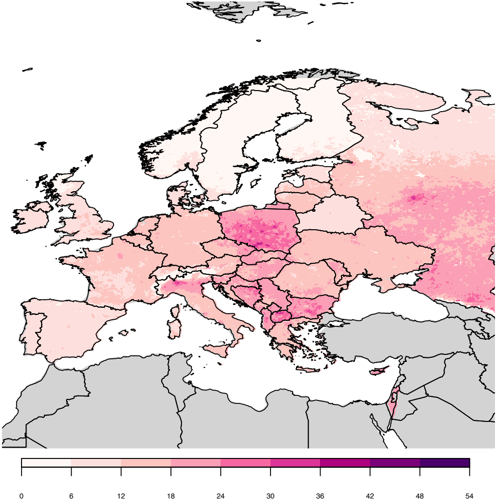

Figure: Predictions of $PM_{2.5}$ in $\mu g m^{-3}$, from hierarchical model for 2014 in Europe

Predictions: Local

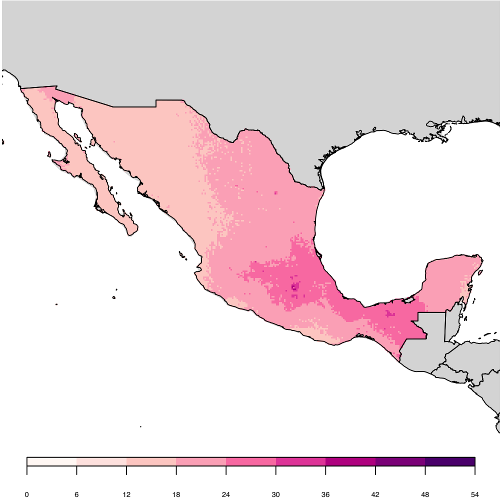

Figure: Predictions of $PM_{2.5}$ in $\mu g m^{-3}$, from hierarchical model for 2014 in Mexico

Evaluation: Crossvalidation

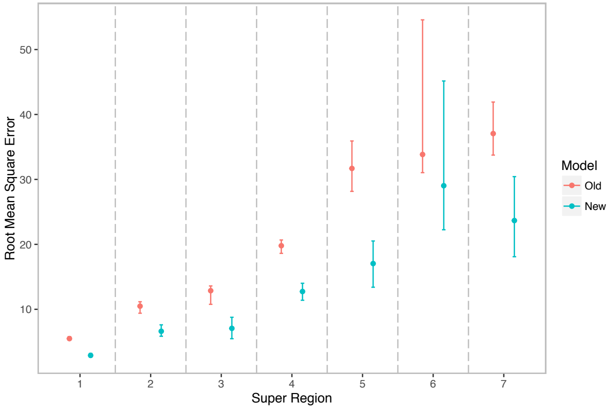

Figure: Comparison of RMSE between approaches. Dots denote the median of the distribution from 25 training/evaluation sets and the vertical lines the range of values. Super-regions are 1: High income, 2: Central Europe, Eastern Europe and Central Asia, 3: Latin America and Caribbean, 4: Southeast Asia, East Asia and Oceania, 5: North Africa / Middle East, 6: Sub-Saharan Africa and 7: South Asia.

Bayesian Melding

- Bayesian melding assumes there is one latent process z_s that drives all sources of data.
- **Data Level:** Ground monitor data is assumed to be a measurement error model i.e.

$$y_s^{gm} = z_s + \epsilon_s \qquad \epsilon_s \sim N(0, \sigma_\epsilon^2)$$

- The grid data is then modelled at point locations as a function of the true underlying process

$$y_s^{grid} = f(z_s) + \delta_s \qquad \delta_s \sim N(0, \sigma_\delta^2).$$

- As we cannot model grid data with a point process, we integrate and get the following integral:

$$y_{B_j}^{grid} = \int_{B_j} f(z_s) + \delta_s d\mathbf{s}, j = 1, 2, \ldots, m$$

Bayesian Melding

- **Latent Process Level:** In the second stage of the model, the true underlying process z_s is assumed to follow the model

$$z_s = \mu_s + m_s$$

where μ_s is a spatial trend and the m_s is a spatial process for location $\boldsymbol{s}$.

- **Inference:** It will be quantify the true levels of $PM_{2.5}$

$$p(z_s|\boldsymbol{y}^{gm}, \boldsymbol{y}^{grid}) = \int p(z_s|\boldsymbol{y}^{gm}, \boldsymbol{y}^{grid}, \boldsymbol{\theta})p(\boldsymbol{\theta}|z_s)d\boldsymbol{\theta}$$

Bayesian Melding

- Makes use of a flexible and coherent framework
- Allows user to assume one underlying process driving the
- Treats estimation methods as different quantities but are intrinsically linked
- To implement this framework on large-scale problems!
- Look at approximate Bayesian inference (INLA) for more efficient computation
- Allow for time effects.

Thank you for listening!

Any Questions?