Global modelling of air pollution using multiple data sources

Matthew Thomas
SAMBa, University of Bath

Email: M.L.Thomas@bath.ac.uk

November 11, 2015

Outline

Motivation

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 East Asia
  - 6th in South Asia and
  - 7th in Africa and the Middle East
- There is convincing evidence for the need to model air pollution effectively.

MOTIVATION

- WHO and other partners plan to strengthen air pollution monitoring globally.
- Aim is to 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,
  - Atmospheric Modelling

Data Sources

Ground Monitoring

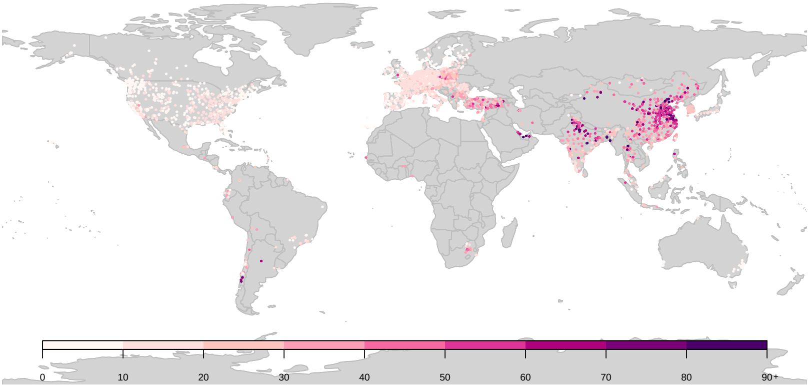


Figure: World map with ground monitor locations, coloured by the estimated level of $PM_{2.5}$ in $\mu g m^{-3}$.

Satellite Remote Sensing

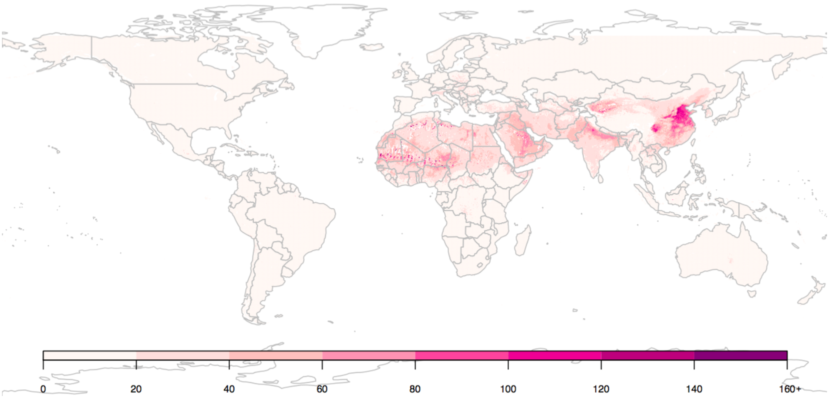


Figure: Global satellite remote sensing estimates of $PM_{2.5}$ in $\mu g m^{-3}$ for 2010 used in GBD2013

Atmospheric Modelling

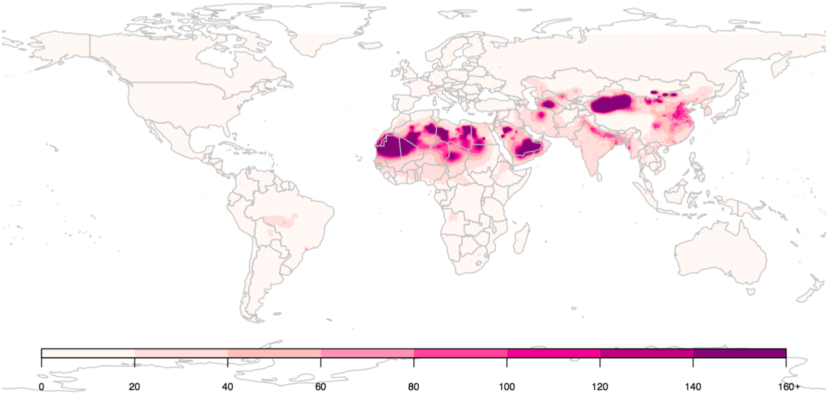


Figure: Global chemical transport model estimates of $PM_{2.5}$ in μgm^{-3} for 2010 used in GBD2013

ATMOSPHERIC MODELLING

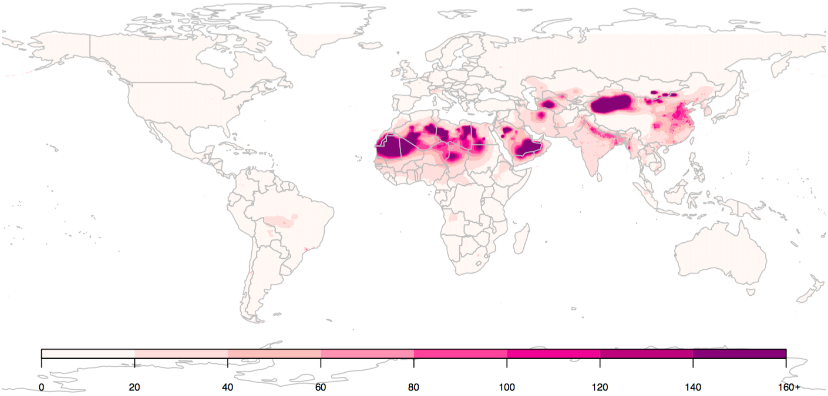


Figure: Global chemical transport model estimates of $PM_{2.5}$ in $\mu g m^{-3}$ for 2010 used in GBD2013

Existing Approaches

LINEAR MODELLING

- 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 (Least-Squares estimation).

PREDICTIONS

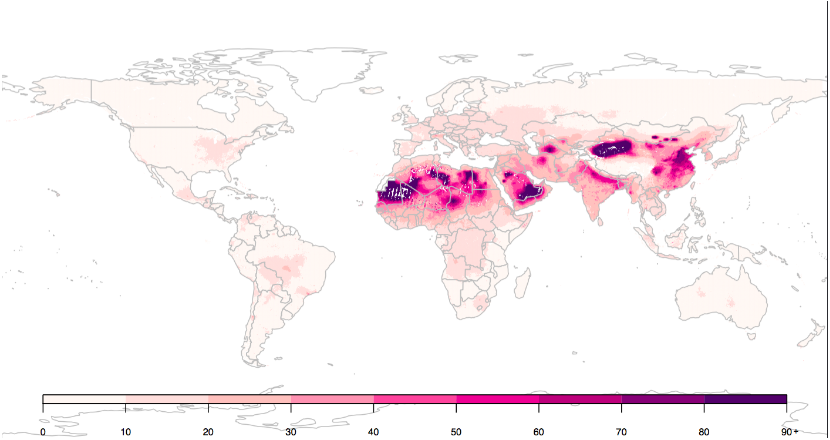


Figure: Predictions of $PM_{2.5}$ in $\mu g m^{-3}$ for 2010, from existing WHO/GBD model.

Predictions, By Region

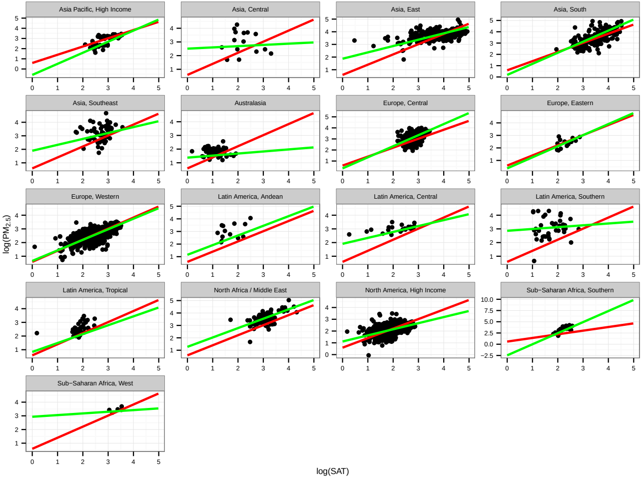


Figure: $PM_{2.5}$ measurements against satellite estimates on the log-scale, for 2010, split by region. The red and green lines denote the single 'global' and a region specific model respectively, estimated using all of the data.

Hierarchical Modelling

HIERARCHICAL MODELLING

- **Observation Level:** We assume the ground monitor data y_{ijkl}^{gm} comes from a measurement error model, on the log-scale:

$$\log\left(y_{ijkl}^{(gm)}\right) = z_{ijkl}^{(gm)} + \epsilon_{ijkl} \quad \epsilon_{ijkl} \sim N(0, \sigma_\epsilon^2)$$

- **Process Level:** Let x_{ijkl}^{sat}, x_{ijkl}^{am} and x_{ijkl}^{pop} denote the satellite, atmospheric model and population estimates respectively. The underlying process is modelled as follows:

$$z_{ijkl} = \tilde{\beta}_{0jkl} + \tilde{\beta}_{1jkl}\log(x_{ijkl}^{sat}) + \tilde{\beta}_{2jkl}\log(x_{ijkl}^{ratio}) + \tilde{\beta}_{3jkl}\log(x_{ijkl}^{pop})$$

$$\tilde{\beta}_{mjkl} = \beta_m^G + \beta_{mj}^{SR} + \beta_{mjk}^{R} + \beta_{mjkl}^{C} + \beta_m^P P_i + \beta_m^A A_i + \beta_m^U U_i, \quad (m = 0, 1, 2, 3)$$

- **Prior Level:** Vague priors were used, default in R-INLA to exploit conjugacy and therefore allow efficient computation.

Bayesian Hierarchical Modelling

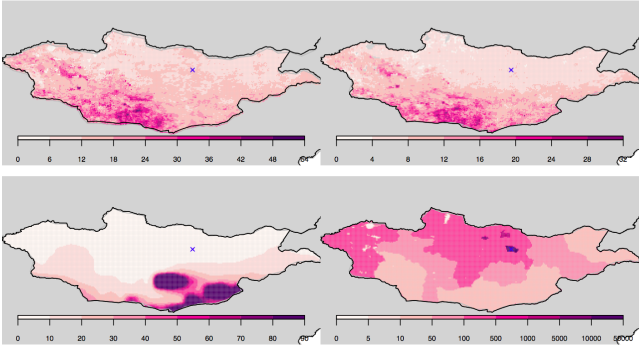
- Many spatial or spatio-temporal models that involve data inherently have a hierarchical structure.
- Hierarchical models extremely useful and flexible framework in which to model complex relationships and dependencies in data.
- Bayesian hierarchical models are commonly written:
  1. The observation level $\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{\theta}$ - Data $\boldsymbol{y}$, are assumed to arise from an underlying latent process $\boldsymbol{z}$, which is unobservable but measurements with error can be taken.
  2. The underlying process level $\boldsymbol{z}|\boldsymbol{\theta}$ - The latent process $\boldsymbol{z}$ assumed to drive the observable data and is the true underlying quantity of interest.
  3. The prior level $\boldsymbol{\theta}$ - This level describes known prior information about the model parameters $\boldsymbol{\theta}$
- Bayesian techniques to statistical modelling allow us to interpret levels in the model that weren't measured such as the underlying latent process.

Approach to Data Integration

- Data integration in the current framework uses a fused estimate.
- Atmospheric model estimates are numerically simulated data from a specified PDE
- Satellite estimates are modelled from images.
- Both estimation methods are very different; as they should provide different perspectives on the modelled system and have very different error structures.
- So, the terms were fitted separately within the model.

Addition of Extra Covariates

- In many areas of the world air pollution estimates weren't very accurate.
- Example: Ulan Bator, Mongolia



- Other pollutant levels were not available
- Population was added into the model as a proxy (on the log-scale)

Random Effects

- Linear models used by WHO, assume a single global relationship.
- This is a massive assumption, that is unlikely to hold.
- Each country is assigned to a 'Region' and 'Super Region' (Nested Hierarchy).
- Could like earlier fit models by Super Region, Region or Country to look at more local relationships. However this comes with issues.
- Instead we added IID random effects for Super Region, Region and Country.
- This allows borrowing from hierarchy when there is limited data.

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).
- Unable to use INLA as parallelised code.
- Restricted to 1 $\times$ 32GB RAM node.
- Took an iterative approach to prediction.

PREDICTIONS

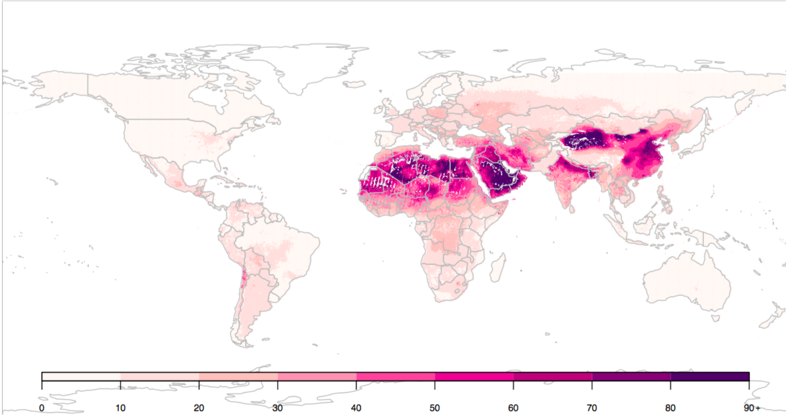


Figure: Predictions of $PM_{2.5}$ in $\mu g m^{-3}$, from hierarchical model for 2010.

UNCERATINTY

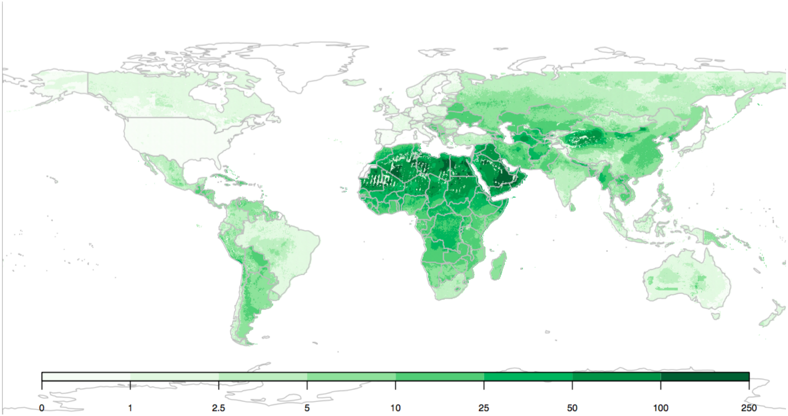


Figure: Uncertainty of PM2.5 predictions for 2010, for hierarchical model; half length of estimated 95% credible intervals

Exceedance Probabilities

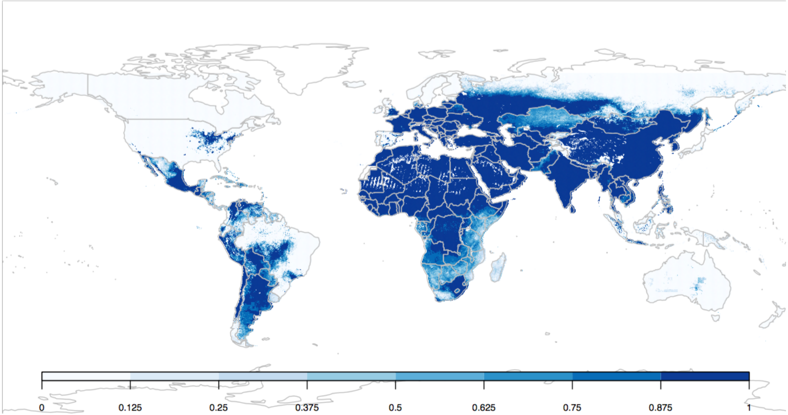


Figure: Probability that level of $PM_{2.5}$ in each cell exceeds $10\mu g m^{-3}$ in 2010, for hierarchical model

Next Steps

BAYESIAN MELDING

- Bayesian melding makes use of a Bayesian hierarchical model.
- Assumes a latent process $z(\mathbf{s})$ that represents the true level $PM_{2.5}$.
- **Data Level:** Ground monitor data is assumed to be a measurement error model i.e.

$$y^{gm}(\mathbf{s}) = z(\mathbf{s}) + \epsilon(\mathbf{s}) \qquad \epsilon(\mathbf{s}) \sim N(0, \sigma_{\epsilon}^2)$$

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

$$y^{grid}(\mathbf{s}) = f(z(\mathbf{s})) + \delta(\mathbf{s}) \qquad \delta(\mathbf{s}) \sim N(0, \sigma_{\delta}^2).$$

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

$$y^{grid}(B_j) = \int_{B_j} f(z(\mathbf{s})) + \delta(\mathbf{s}) d\mathbf{s}, j = 1, 2, \ldots, m$$

Bayesian Melding

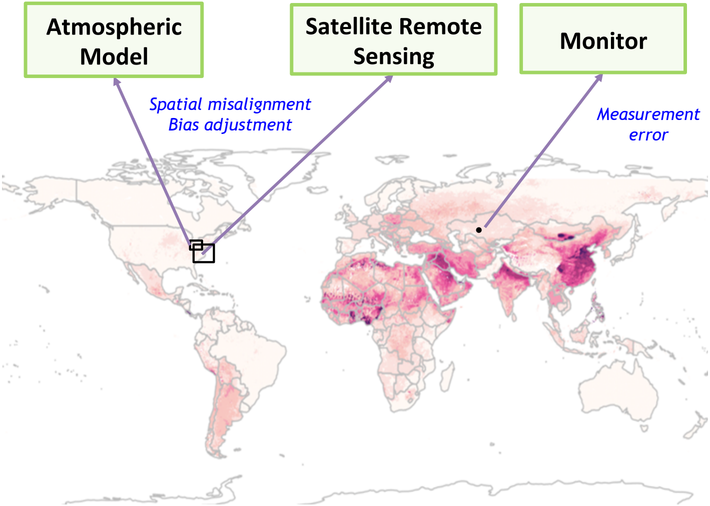
- **Latent Process Level:** In the second stage of the model, the true underlying process $z(\boldsymbol{s})$ is assumed to follow the model

$$z(\boldsymbol{s}) = \mu(\boldsymbol{s}) + m(\boldsymbol{s})$$

  where $\mu(\boldsymbol{s})$ is a spatial trend and the $m(\boldsymbol{s})$ is a spatial process for location $\boldsymbol{s}$.
- **Prior Level:** It will also be necessary to specify relevant priors for model parameters.
- **Inference:** It will be quantify the true levels of $PM_{2.5}$

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

Bayesian Melding



Bayesian Melding

Advantages:

- 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

Disadvantages:

- Very computationally demanding (particularly with MCMC)
- Only implemented in small-scale problems ($\tilde{2}$0 Monitors)

Aims:

- To implement this framework on large-scale problems!
- Look at approximate Bayesian inference for more efficient computation

Any Questions?

Thank you for listening!