UNIVERSITY OF BATH

SAMBa

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

Matthew Thomas

9th January 2017

# Outline

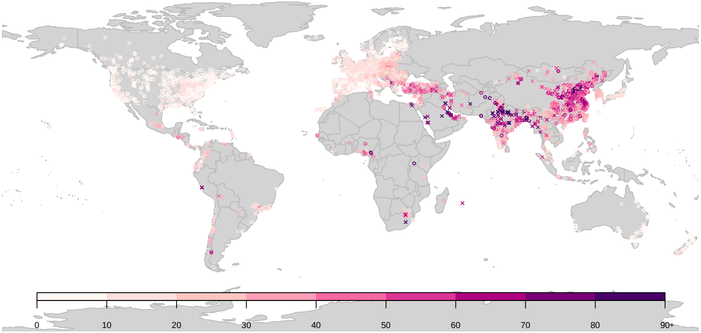
- Introduction
- Previous methods for integrating information from multiple data sources
- DIMAQ
- Results
- Conclusions

# Introduction

- Air pollution has been identified as a global health priority.
- In 2016, the WHO estimated that over 3 million deaths can be attributed to ambient air pollution.
- The Global Burden of Disease project estimate that in 2015 ambient air pollution was in the top ten leading risks to global health.
- Burden of disease calculations require accurate estimates of population exposure for each country.

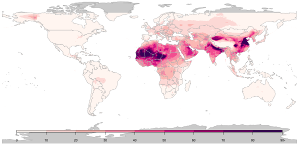
# Estimating $PM_{2.5}$

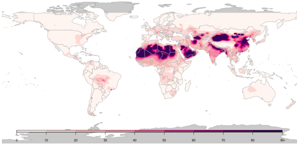
- Accurate estimates of exposure to air pollution are required
  - at global, national and local levels
  - with associated measures of uncertainty.
- While networks are expanding, ground monitoring is limited in many areas of the world.

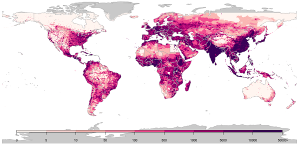


# Estimating $PM_{2.5}$

- Can utilise information from other sources
  - satellite remote sensing
  - atmospheric models
  - population estimates
  - land use
  - local network characteristics.
- Result of modelling and will be subject to uncertainties and biases.
- Provide a global coverage (1.4 million grid cells at a $0.1^{o}$ resolution).







# REQUIREMENTS

- Full global coverage with consistent methods.
- Reduced biases due to location or development status.
- High spatial resolution allowing better link to population data.
- Estimates of uncertainty.

# Linear Regression

- Exposures to $PM_{2.5}$ for the GBD study in 2010 and 2013 were obtained using simple linear modelling techniques.
- Combined estimates from satellites with those from a chemical transport model.
- Calibrated with available surface monitoring data using linear regression.
- Provides a single global relationship between ground measurements, satellite remote sensing and chemical transport model.
- This is unlikely to hold as relationships between the geographical regions are likely to be different.

# A More local Calibration

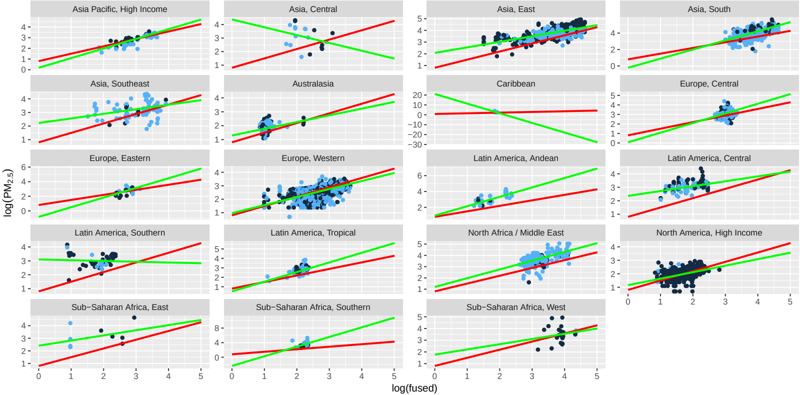


Figure: PM$_{2.5}$ measurements against a fused estimate from satellite remote sensing and chemical transport models, by region. In each panel, the red line shows the single, global, calibration model used by the GBD study. The green line represents a region specific calibration model. Dark blue dots denote measurements of PM$_{2.5}$ and light blue dots those that were converted from PM$_{10}$.

# Data Integration Model for Air Quality

- Developed to the Data Integration Model for Air Quality (DIMAQ).
- DIMAQ calibrates ground measurements to estimates
  - satellite remote sensing,
  - specific components of chemical transport models
  - land use
  - population.
- The coefficients in the calibration model are estimated by country.
- Model allows borrowing from higher aggregations and if information is not available on a country level.
- Achieved using hierarchical random effects.
- Exploits a geographical nested hierarchy.

# REGIONS

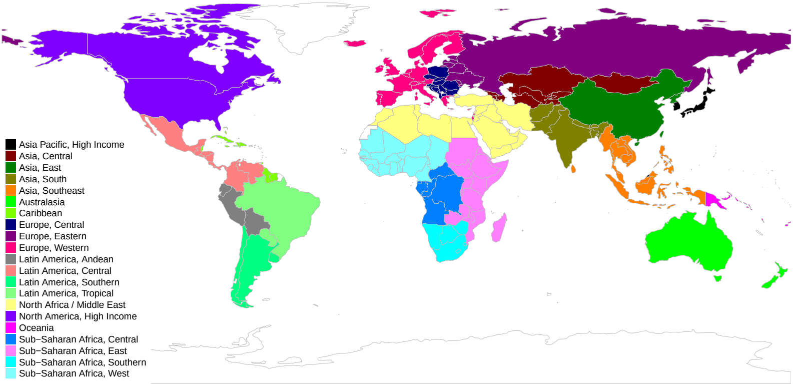


Figure: Map of regions.

# SUPER-REGIONS

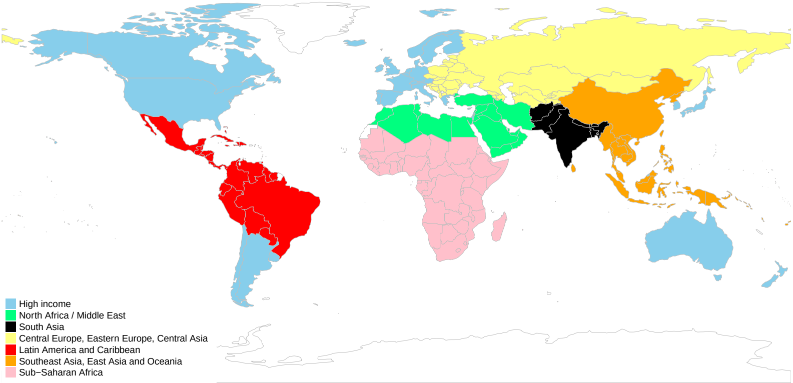

High income
North Africa / Middle East
South Asia
Central Europe, Eastern Europe, Central Asia
Latin America and Caribbean
Southeast Asia, East Asia and Oceania
Sub−Saharan Africa

Figure: Map of super-regions.

# 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 $Y_{slijk}$.
- The model consists of a set of fixed and random effects, for both intercepts and covariates, and is given as follows,

$$\begin{aligned}\log(Y_{slijk}) = \tilde{\beta}_{0,lijk} \quad &+ \quad \sum_{p \in P} \beta_p X_{p,slijk} \\ &+ \quad \sum_{q \in Q} \tilde{\beta}_{q,lijk} X_{slijk} \\ &+ \quad \epsilon_{slijk} \; .\end{aligned}$$

# Hierarchical Random Effects

- The random effect terms have contributions from the country, the region and the super–region.

$$\tilde{\beta}_{q,ijk} = \beta_q + \beta^{C}_{q,ijk} + \beta^{R}_{q,jk} + \beta^{SR}_{q,k}$$

- The intercept also having a random effect for the cell representing within-cell variation in ground measurements.

$$\tilde{\beta}_{0,lijk} = \beta_0 + \beta^{G}_{0,lijk} + \beta^{C}_{0,ijk} + \beta^{R}_{0,jk} + \beta^{SR}_{0,k}$$

# INFERENCE

- DIMAQ contains a large number of random effects and requires a large number of spatial predictions.
  - MCMC would have been impractical.
- Approximate Bayesian inference, such as Integrated Nested Laplace Approximations (INLA), provides fast and efficient methods for modelling with LGMs.
- INLA performs numerical calculations of marginal posterior densities using Laplace Approximations hierarchical latent Gaussian models:

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

- Implemented these techniques using R-INLA.

# Evaluation: Crossvalidation

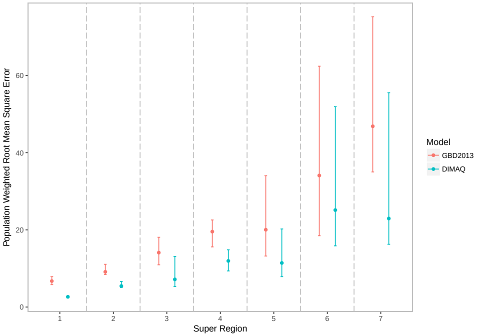


Figure: Summaries of predictive ability of the GBD2013 model and DIMAQ, for each of seven super–regions: 1, High income; 2, Central Europe, Eastern Europe, Central Asia; 3, Latin America and Caribbean; 4, Southeast Asia, East Asia and Oceania; 5, North Africa / Middle East; 6, Sub-Saharan Africa; 7, South Asia. For each model, population weighted root mean squared errors ($\mu$gm$^{-3}$) are given with dots denoting the median of the distribution from 25 training/evaluation sets and the vertical lines the range of values.

# PREDICTIONS

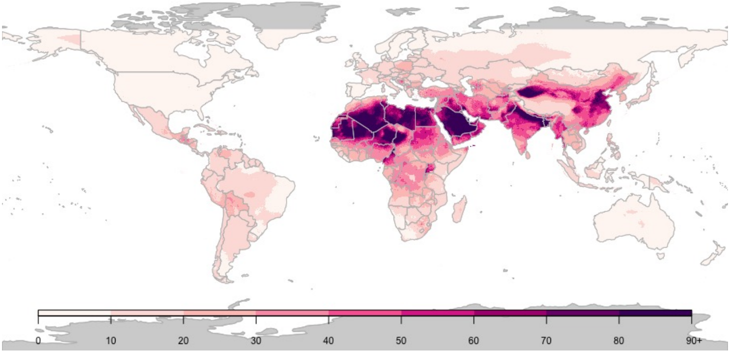


Figure: Median estimates of annual averages of $PM_{2.5}$ ($\mu gm^{-3}$) for 2014 for each grid cell ($0.1^o \times 0.1^o$ resolution) using DIMAQ.

# UNCERTAINTY

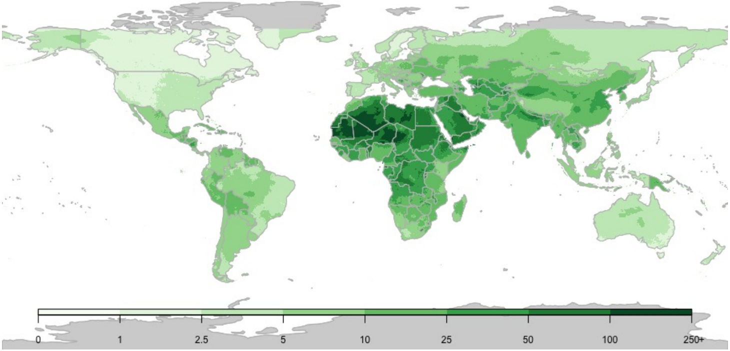


Figure: Half the width of 95% posterior credible intervals for 2014 for each grid cell ($0.1^o \times 0.1^o$ resolution) using DIMAQ.

# EXPOSURES TO $PM_{2.5}$

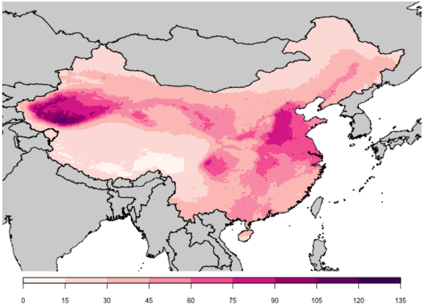


Figure: Medians of posterior distributions for estimates of annual mean $PM_{2.5}$ concentrations ($\mu gm^{-3}$) for 2014, in China.

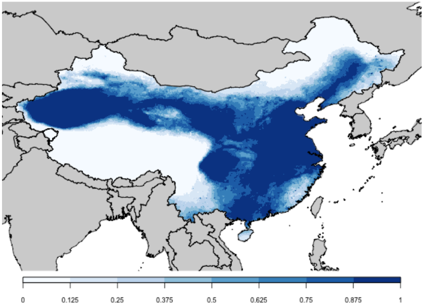


Figure: Probability of exceeding 35 $\mu gm^{-3}$ using a Bayesian hierarchical model for each grid cell ($0.1^o \times 0.1^o$ resolution) for 2014, in China.

# EXPOSURES TO $PM_{2.5}$

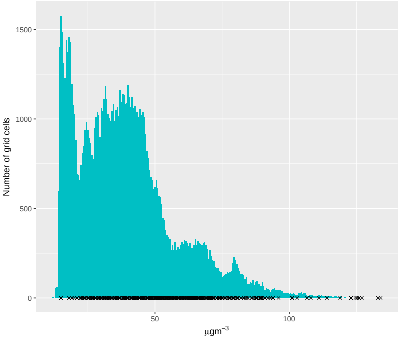


Figure: Estimated annual average concentrations of $PM_{2.5}$ by grid cell ($0.1^o \times 0.1^o$ resolution). Black crosses denote the annual averages recorded at ground monitors.

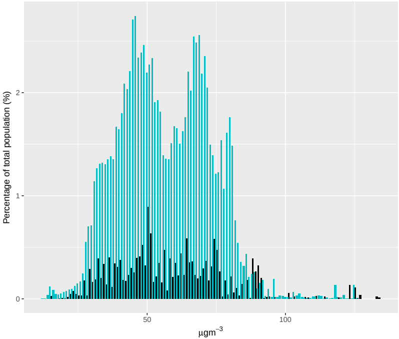


Figure: Estimated population level exposures (blue bars) and population weighted measurements from ground monitors (black bars).

# CONCLUSION

- Developed a model to integrate data from multiple sources with the aim of producing high-resolution estimates of population exposures to ambient particulate matter.
- DIMAQ
  - is currently based on a country-level spatial structure
  - uses covariates that have individual errors and biases
  - uses spatially-aligned gridded data.