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

Matthew Thomas

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

20^{th} June 2017

Outline

- Introduction
- DIMAQ
- Results
- Conclusions

Introduction

- Air pollution has been identified as a global health priority.

Introduction

- Air pollution has been identified as a global health priority.
- In 2016, the World Health Organisation (WHO) estimated that over 3 million deaths can be attributed to ambient air pollution.

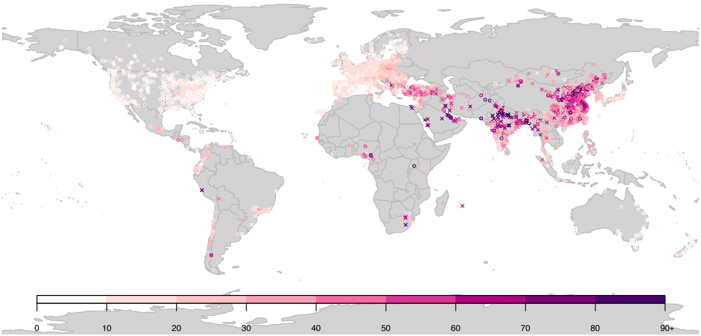
Introduction

- Air pollution has been identified as a global health priority.
- In 2016, the World Health Organisation (WHO) estimated that over 3 million deaths can be attributed to ambient air pollution.
- The Global Burden of Disease (GBD) project estimate that in 2015 ambient air pollution was in the top ten leading risks to global health.

Introduction

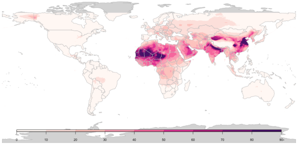
- Air pollution has been identified as a global health priority.
- In 2016, the World Health Organisation (WHO) estimated that over 3 million deaths can be attributed to ambient air pollution.
- The Global Burden of Disease (GBD) project estimate that in 2015 ambient air pollution was in the top ten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

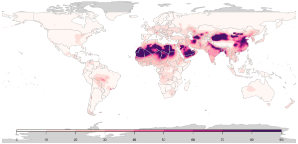
Estimating $PM_{2.5}$

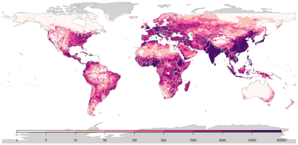


ESTIMATING $PM_{2.5}$

- 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.







Data Integration Model for Air Quality

- Developed to the Data Integration Model for Air Quality (DIMAQ).

Data Integration Model for Air Quality

- Developed to the Data Integration Model for Air Quality (DIMAQ).
- DIMAQ calibrates ground measurements to estimates from
  - satellite remote sensing,

Data Integration Model for Air Quality

- Developed to the Data Integration Model for Air Quality (DIMAQ).
- DIMAQ calibrates ground measurements to estimates from
  - satellite remote sensing,
  - specific components of chemical transport models

Data Integration Model for Air Quality

- Developed to the Data Integration Model for Air Quality (DIMAQ).
- DIMAQ calibrates ground measurements to estimates from
  - satellite remote sensing,
  - specific components of chemical transport models
  - land use

Data Integration Model for Air Quality

- Developed to the Data Integration Model for Air Quality (DIMAQ).
- DIMAQ calibrates ground measurements to estimates from
  - satellite remote sensing,
  - specific components of chemical transport models
  - land use
  - population.

Data Integration Model for Air Quality

- Developed to the Data Integration Model for Air Quality (DIMAQ).
- DIMAQ calibrates ground measurements to estimates from
  - satellite remote sensing,
  - specific components of chemical transport models
  - land use
  - population.
- The coefficients in the calibration model are estimated by country.

Data Integration Model for Air Quality

- Developed to the Data Integration Model for Air Quality (DIMAQ).
- DIMAQ calibrates ground measurements to estimates from
  - 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.

Data Integration Model for Air Quality

- Developed to the Data Integration Model for Air Quality (DIMAQ).
- DIMAQ calibrates ground measurements to estimates from
  - 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.
- Exploits a geographical nested hierarchy.
- Achieved using hierarchical random effects.

Regions

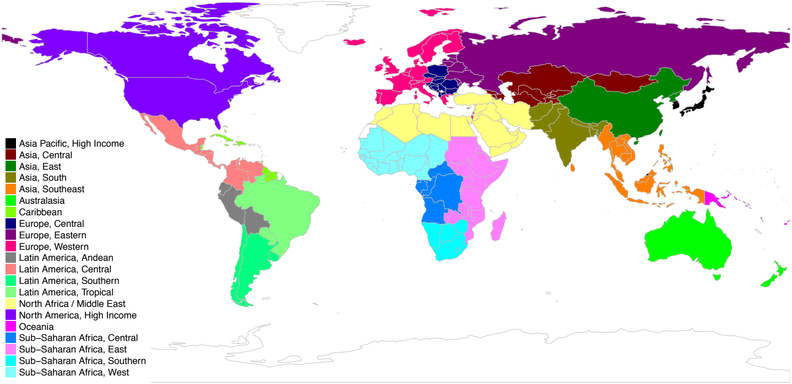


Figure: Map of regions.

SUPER-REGIONS

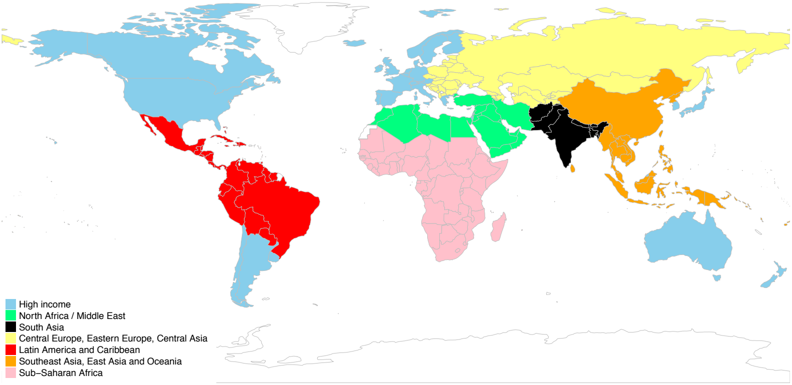


Figure: Map of super-regions.

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}.

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.

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

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.

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

Random Effects Structure

- The coefficients for super-regions are distributed with mean equal to the overall mean (β_0, the fixed effect) and variance representing the between super–region variation,

$$\beta_k^{SR} \sim N(\beta, \sigma_{SR}^2)$$

Random Effects Structure

- The coefficients for super-regions are distributed with mean equal to the overall mean (β_0, the fixed effect) and variance representing the between super–region variation,

$$\beta_k^{SR} \sim N(\beta, \sigma_{SR}^2)$$

- The coefficients for regions are distributed with mean equal to the coefficient for the super–region with variance representing the between region variation,

$$\beta_{jk}^{R} \sim N(\beta_k^{SR}, \sigma_{R,k}^2)$$

Random Effects Structure

- The coefficients for super-regions are distributed with mean equal to the overall mean (β_0, the fixed effect) and variance representing the between super–region variation,

$$\beta_k^{SR} \sim N(\beta, \sigma_{SR}^2)$$

- The coefficients for regions are distributed with mean equal to the coefficient for the super–region with variance representing the between region variation,

$$\beta_{jk}^{R} \sim N(\beta_k^{SR}, \sigma_{R,k}^2)$$

- The coefficients for a country is distributed with mean equal to the coefficient for the region with variance representing the between country variation,

$$\beta_{ijk}^{C} \sim N(\beta_{jk}^{R}, \sigma_{C,jk}^2)$$

INFERENCE

- Approximate Bayesian inference, such as Integrated Nested Laplace Approximations (INLA), provide fast and efficient methods for modelling with latent Gaussian models.

INFERENCE

- Approximate Bayesian inference, such as Integrated Nested Laplace Approximations (INLA), provide fast and efficient methods for modelling with latent Gaussian models.
- INLA performs numerical calculations of posterior densities using Laplace Approximations hierarchical latent Gaussian models:

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

INFERENCE

- Approximate Bayesian inference, such as Integrated Nested Laplace Approximations (INLA), provide fast and efficient methods for modelling with latent Gaussian models.
- INLA performs numerical calculations of posterior densities using Laplace Approximations hierarchical latent Gaussian models:

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

- Latent Gaussian models allows for sparse matrices, and therefore efficient computation.

Computation

- R-INLA was used to implement DIMAQ.

Computation

- R-INLA was used to implement DIMAQ.
- 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).

COMPUTATION

- R-INLA was used to implement DIMAQ.
- 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.

Evaluation: Crossvalidation

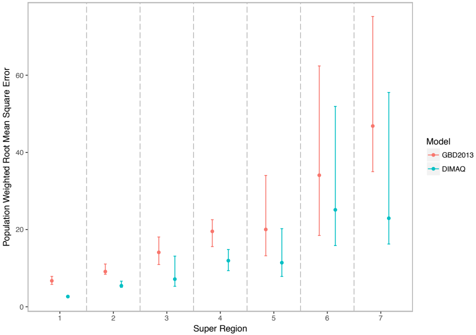


Figure: Summaries of predictive ability of the GBD2013 model and DIMAQ, for each of seven super–regions: 1, High income; 2, Central Europe, Eastern Europe, Central Asia; 3, Latin America and Caribbean; 4, Southeast Asia, East Asia and Oceania; 5, North Africa / Middle East; 6, Sub-Saharan Africa; 7, South Asia. For each model, population weighted root mean squared errors (μgm^{-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

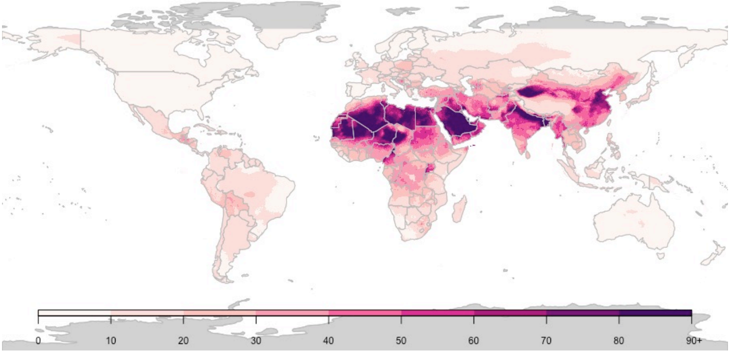


Figure: Median estimates of annual averages of $PM_{2.5}$ (μgm^{-3}) for 2014 for each grid cell ($0.1^o \times 0.1^o$ resolution) using DIMAQ.

UNCERTAINTY

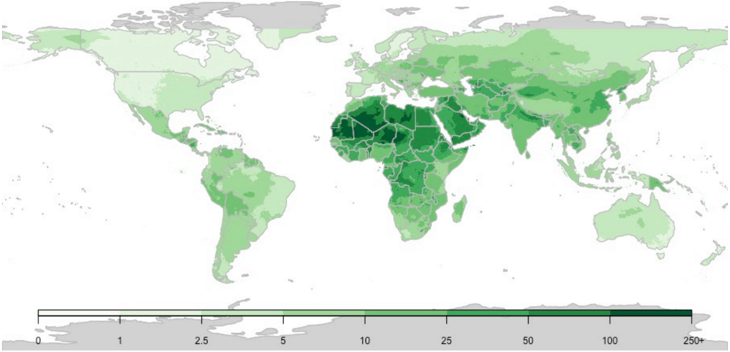


Figure: Half the width of 95% posterior credible intervals for 2014 for each grid cell ($0.1^o \times 0.1^o$ resolution) using DIMAQ.

Posterior Distributions

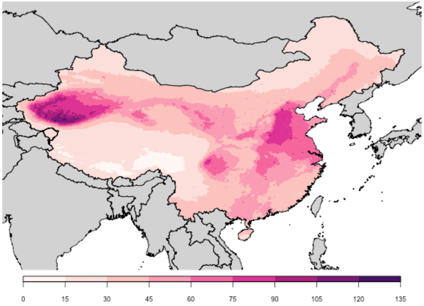


Figure: Medians of posterior distributions for estimates of annual mean $PM_{2.5}$ concentrations (μgm^{-3}) for 2014, in China.

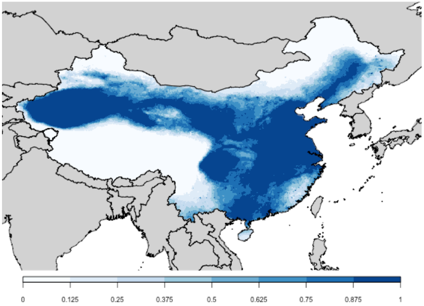


Figure: Probability of exceeding 35 μgm^{-3} using a Bayesian hierarchical model for each grid cell ($0.1^o \times 0.1^o$ resolution) for 2014, in China.

Population Exposures to $PM_{2.5}$

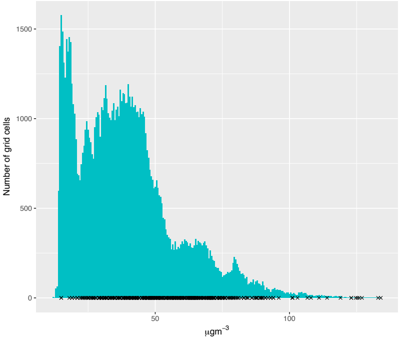


Figure: Estimated annual average concentrations of $PM_{2.5}$ by grid cell ($0.1^o \times 0.1^o$ resolution). Black crosses denote the annual averages recorded at ground monitors.

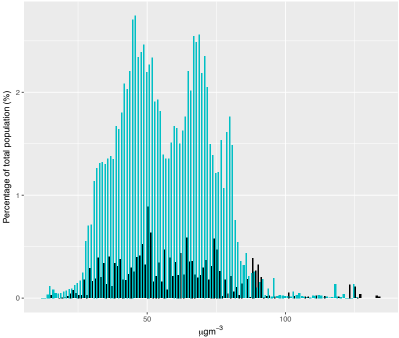


Figure: Estimated population level exposures (blue bars) and population weighted measurements from ground monitors (black bars).

CONCLUSION

- DIMAQ integrates data from multiple sources with producing high-resolution estimates of concentrations of ambient particulate matter.

Conclusion

- DIMAQ integrates data from multiple sources with producing high-resolution estimates of concentrations of ambient particulate matter.
- Estimates used by the WHO and GBD in burden of disease calculations.

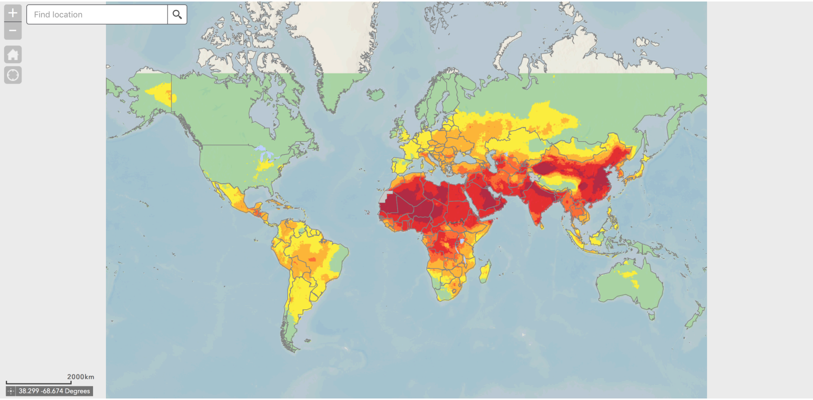
Conclusion

- DIMAQ integrates data from multiple sources with producing high-resolution estimates of concentrations of ambient particulate matter.
- Estimates used by the WHO and GBD in burden of disease calculations.
- Future Developments
    - Higher resolution estimates
    - Within country variability
    - Allowing for errors and biases in covariates
    - Use data at native resolutions

CONCLUSION

- DIMAQ integrates data from multiple sources with producing high-resolution estimates of concentrations of ambient particulate matter.
- Estimates used by the WHO and GBD in burden of disease calculations.
- Future Developments
  - Higher resolution estimates
  - Within country variability
  - Allowing for errors and biases in covariates
  - Use data at native resolutions
- Possible approaches to address these issues
  - Statistical downscaling
  - Bayesian melding.

Interactive Map



REFERENCES

- DIMAQ Paper:
  http://onlinelibrary.wiley.com/doi/10.1111/rssc.12227/full
- WHO Report:
  http://who.int/phe/publications/air-pollution-global-assessment/er
- GBD Paper:
  http://www.thelancet.com/journals/lancet/article/PIIS0140-6736(16)
- Interactive Map:
  http://maps.who.int/airpollution/

Any Questions?