Data integration for high–resolution, continental–scale estimation of air pollution concentrations

Matthew Thomas
Imperial College London

GEOMED 2019, Glasgow

28^{th} August 2019

Collaborators

- Gavin Shaddick (Exeter)
- Daniel Simpson (Toronto)
- Kees de Hoogh (STPH Institute)
- Jim Zidek (UBC)

INTRODUCTION

- Air pollution has been identified as a global health priority
- Fine particulate matter ($PM_{2.5}$) is associated with some adverse health outcomes
  - WHO estimate 4.2 million deaths globally are attributable to $PM_{2.5}$
- Estimation of health burden requires accurate estimates of exposures to air pollution
  - Localised levels
  - Associated measures of uncertainty

Ground Monitoring

- Information on exposures to air pollution traditionally comes from ground monitors
- Monitoring networks for $PM_{2.5}$ are growing worldwide
- Density of networks vary considerably
  - Urban and industrial areas
  - High and middle income countries

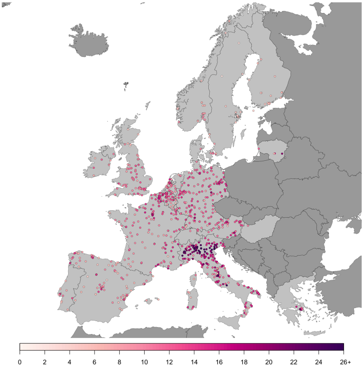
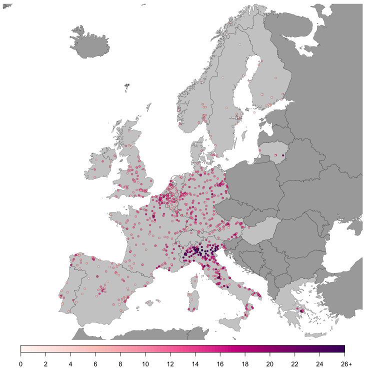


Figure: Locations of ground monitors measuring $PM_{2.5}$ in 2016. Colours denote the annual average concentrations (μgm^{-3}) of $PM_{2.5}$

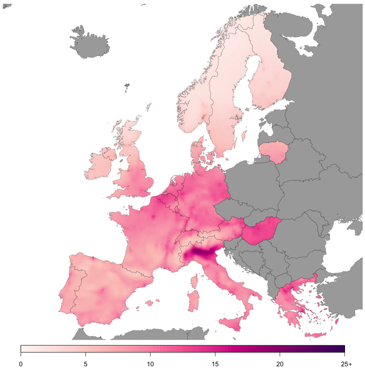
DATA FROM MULTIPLE SOURCES

- Ground monitoring
- Chemical transport models
- Land use regression
- Different resolutions
  - Ground monitors (points)
  - Chemical transport models (10km×10km)
  - Land use regression (1km×1km or 100m×100m)
- All will be subject to uncertainties and biases



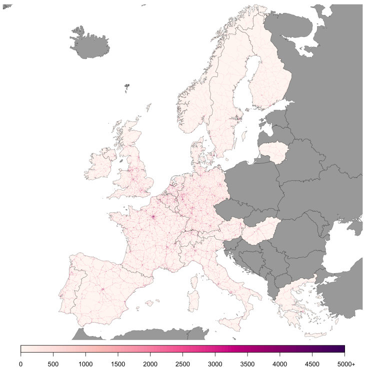
Data from Multiple Sources

- Ground monitoring
- Chemical transport models
- Land use regression
- Different resolutions
  - Ground monitors (points)
  - Chemical transport models (10km$\times$10km)
  - Land use regression (1km$\times$1km or 100m$\times$100m)
- All will be subject to uncertainties and biases



Data from Multiple Sources

- Ground monitoring
- Chemical transport models
- Land use regression
- Different resolutions
  - Ground monitors (points)
  - Chemical transport models (10km$\times$10km)
  - Land use regression (1km$\times$1km or 100m$\times$100m)
- All will be subject to uncertainties and biases



Statistical Calibration

- The aim is to calibrate estimates from chemical transport models, satellite remote sensing, land use regression and topography, X_{pl_st}, against measurements from ground monitors, Y_{st},

$$Y_{st} = \beta_0 + \beta_1 X_{1l_st} + \sum_{i=2}^{N} \beta_i X_{il_st} + \epsilon_{st}$$

- This will allow us to predict surface $PM_{2.5}$ where there is no ground monitoring information
- However, the relationship between ground monitors and other variables may vary over space and time

Statistical Downscaling

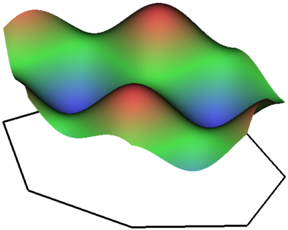
- Need to allow for the variation in the coefficients
- Coefficients can vary spatio-temporally

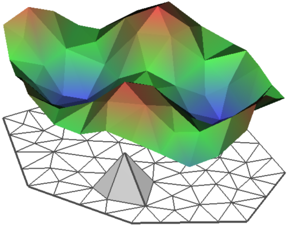
$$
\begin{aligned}
Y_{st} &= \tilde{\beta}_{0st} + \tilde{\beta}_{1st} X_{1l_s t} + \sum_{i=2}^{N} \beta_{ist} X_{il_s t} + \epsilon_{st} \\
\tilde{\beta}_{pst} &= \alpha_p + \beta_{pst}, \quad p = 0, 1
\end{aligned}
$$

- Coefficients $\tilde{\beta}_{pst}$, $p = 0, 1$, comprise of
  - Fixed effects α_p
  - Spatio-temporal random effect β_{pst}

Priors

- Spatio-temporal random effects
  - AR1 in time
  - Matérn covariance function
  - Separable in space and time
- Computationally challenging to fit multiple spatio-temporal processes
- The approximation to the spatial field is the solution to Stochastic Partial Differential Equation (SPDE)
  - Conditional independence
  - Sparse precision matrices





Approximation to the spatio-temporal fields

- Latent Gaussian model
- Inference based on Integrated Nested Laplace Approximations (INLA)
- Penalised complexity priors for model hyperparameters
- High resolution estimates of air pollution concentrations are required over space and time
- Computationally expensive
- Monte Carlo Simulation

Predictions

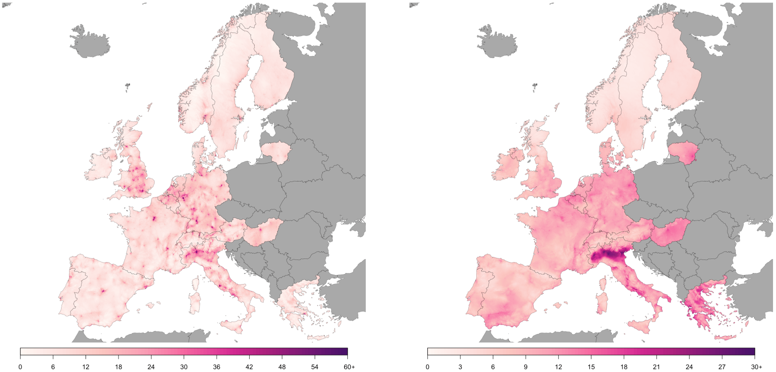


Figure: Predictive posterior median annual averages of (Left) NO_2 and (Right) $PM_{2.5}$ (in μgm^{-3}) for 2016 for each grid cell (1km $\times$ 1km resolution).

CTM VS PREDICTIONS

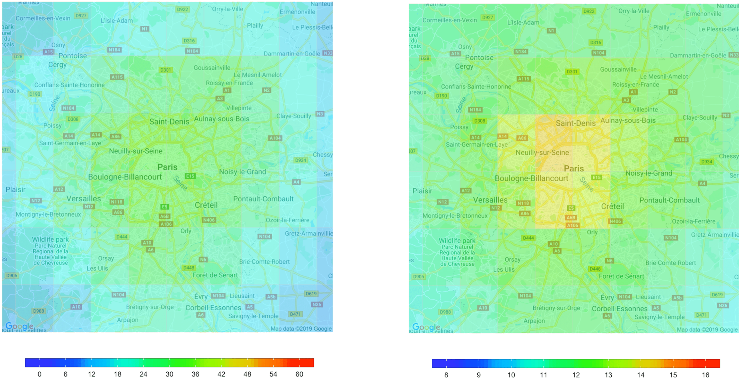


Figure: Estimated annual average concentrations of (Left) NO_2 and (Right) $PM_{2.5}$ (in μgm^{-3}) for 2016 from the MACC-II CTM for each grid cell in Paris (10km $\times$ 10km).

CTM VS PREDICTIONS

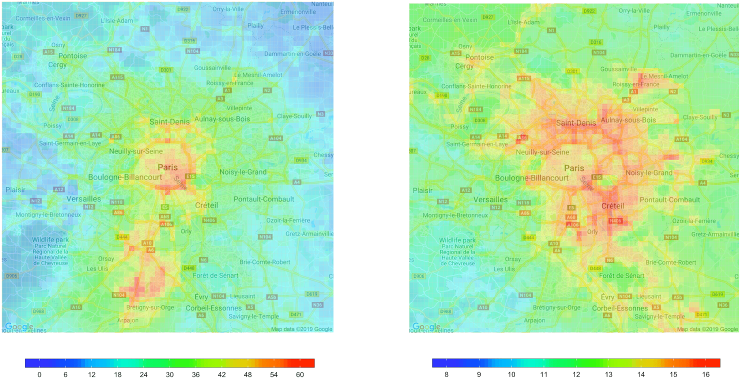


Figure: Predictive posterior median annual average concentrations of (Left) NO_2 and (Right) $PM_{2.5}$ (in μgm^{-3}) in 2016 for each grid cell in Paris (1km $\times$ 1km).

CTM VS PREDICTIONS

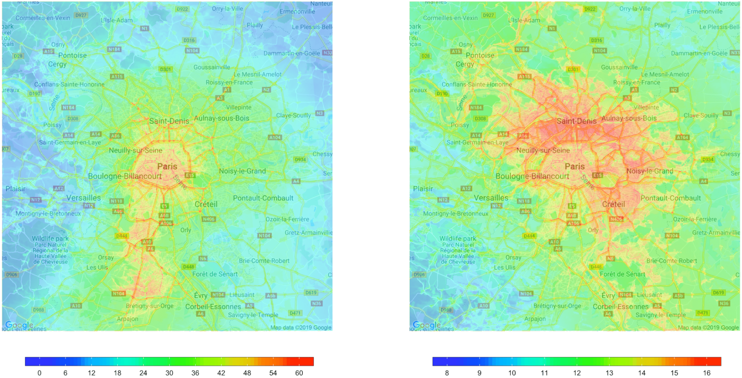


Figure: Predictive posterior median annual average concentrations of (Left) NO_2 and (Right) $PM_{2.5}$ in 2016 for each grid cell in Paris (100m $\times$ 100m).

EXCEEDANCES

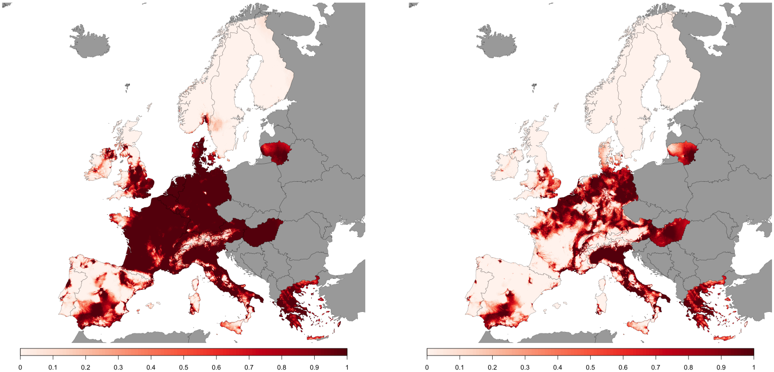


Figure: Probability that annual average $PM_{2.5}$ exceeds 10 μgm^{-3} (Left) for 2010 and (Right) for 2016 for each grid cell (1km $\times$ 1km resolution).

Changes over time

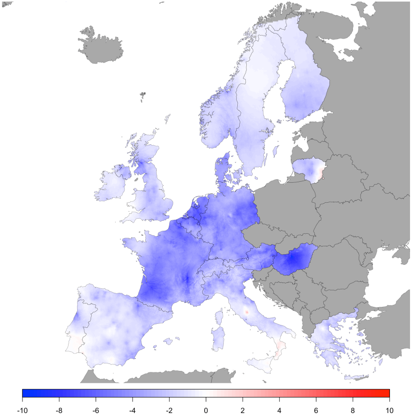


Figure: Median change in the annual average $PM_{2.5}$ between 2010 and 2016 for each grid cell (1km $\times$ 1km resolution).

SUMMARY

- Developed a model that
  - Integrates data from multiple sources
  - Integrates data with multiple resolutions
  - Produces high-resolution estimates of air pollution with associated measures of uncertainty.
- Future work
  - Multivariate ($PM_{2.5}$, NO_2, PM_{10} and O_3)
  - Higher temporal resolution
  - Burden of disease calculations

ArXiv Preprint

- Contact Details
  - matthew.thomas@imperial.ac.uk
  - `http://www.imperial.ac.uk/people/matthew.thomas`
- ArXiv Preprint (arXiv:1907.00093)

Statistics > Applications

Data integration for high-resolution, continental-scale estimation of air pollution concentrations

Matthew L. Thomas, Gavin Shaddick, Daniel Simpson, Kees de Hoogh, James V. Zidek

(Submitted on 28 Jun 2019)

Air pollution constitutes the highest environmental risk factor in relation to heath. In order to provide the evidence required for health impact analyses, to inform policy and to develop potential mitigation strategies comprehensive information is required on the state of air pollution. Traditionally, information on air pollution has come from ground monitoring (GM) networks but these may not be able to provide sufficient coverage and may need to be supplemented with information from other sources (e.g. chemical transport models; CTMs). However, these other sources may only be available on grids and may not capture micro-scale features that may be important in assessing air quality in areas of high population. We develop a model that allows calibration between data sources available at different levels of support, for example, GMs at point locations and estimates from CTMs on grid-cells by allowing the coefficients of calibration equations to vary over space and time. Using a Bayesian hierarchical framework, we address the computational issues that may arise when fitting varying coefficient models in larger scale problems, especially those using MCMC by using INLA. We assess the efficacy of the proposed model, through simulation and out-of-sample evaluation using data from Western Europe. The model is used to produce a comprehensive set of high-resolution (1km x 1km) estimates of NO_2 and $PM_{2.5}$ across Western Europe for 2010-2016. The estimates offer a wealth of information, including the ability to calculate exceedance probabilities and, by aligning to estimates of population, country-level population-weighted concentrations. We find that levels of both pollutants are decreasing in the majority of Western Europe, however there remain large populations exposed to levels that exceed the WHO Air Quality Guidelines and as such air pollution remains a serious threat to health.

ANY QUESTIONS?