Data Integration for high-resolution estimation of air pollution concentrations

Matthew Thomas

Joint work with Gavin Shaddick, Daniel Simpson,
Kees de Hoogh and Jim Zidek

14^{th} July 2018

Outline

- Introduction
- Statistical Calibration
- Air Quality in Europe
- Summary

Introduction

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 guidelines
  - Annual averages should not exceed 10 μgm^{-3}
- 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

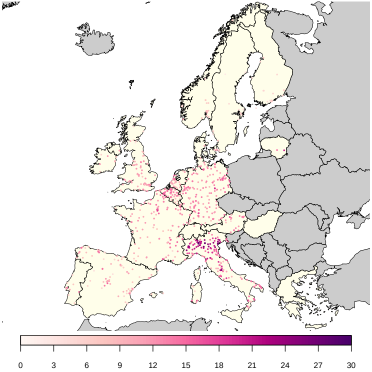


Figure: Locations of ground monitors measuring $PM_{2.5}$ in 2016. Colours denote the annual average concentrations (μgm^{-3}) of $PM_{2.5}$

DATA FROM MULTIPLE SOURCES

- Measurements from 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)
- All will be subject to uncertainties and biases

DATA FROM MULTIPLE SOURCES

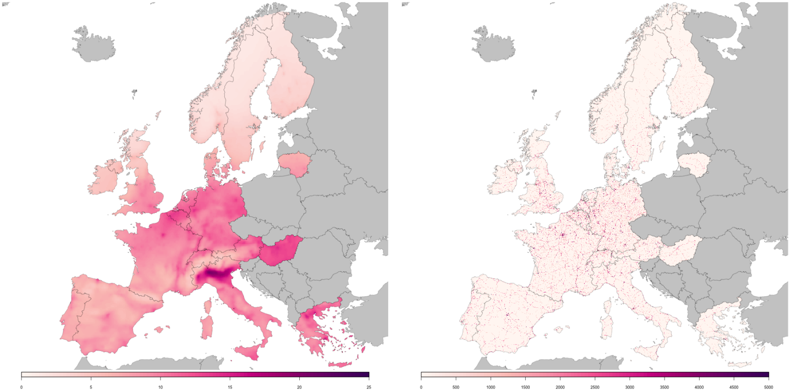


Figure: (Left) Estimates of annual average $PM_{2.5}$ (μgm^{-3}) from the MACC-II ENSEMBLE chemical transport model in 2016 for each grid cell (10km $\times$ 10km resolution) and (Right) Length of major roads within a 21km buffer of each grid cell (1km $\times$ 1km resolution)

Statistical Calibration

STATISTICAL CALIBRATION

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

$$Y_{st} = \beta_0 + \sum_{i=1}^{N} \beta_i X_{il_s t} + \epsilon_{st}$$

- 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

$$
\begin{aligned}
Y_{st} &= \beta_{0st} + \sum_{i=1}^{N} \beta_{ist} X_{il_s t} + \epsilon_{st} \\
\beta_{pst} &= \alpha_p + \theta_{pt} + m_{ps} + \kappa_{pst}
\end{aligned}
$$

- Generic coefficient $\beta_{st} \equiv \beta_{pst}$ comprises of
  - Fixed effect α
  - Temporal random effect θ_t
  - Spatial random effect m_s
  - Spatio-temporal random effect κ_{st}

PRIORS

- Fixed effects: $\alpha \sim N(0, 1000)$
- Temporal random effects: $\theta_t \sim N(\theta_{pt-1}, \sigma^2_\theta)$
- Example: Global air quality (more later)
  - Spatial random effects: $\boldsymbol{m} \sim N(\mathbf{0}, \sigma^2_m \Sigma_m)$
  - Matérn covariance function
- Example: Air quality in Europe
  - Spatio-temporal random effects:
  - AR1 in time

$$\begin{aligned} \kappa_{st} &= \rho \kappa_{st-1} + \omega_{st} \\ \boldsymbol{\omega}_t &\sim N(\mathbf{0}, \sigma^2_\omega \Sigma_\omega) \end{aligned}$$

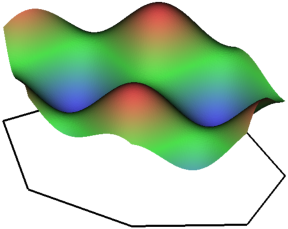
  - Matérn covariance function
  - Separable in space and time

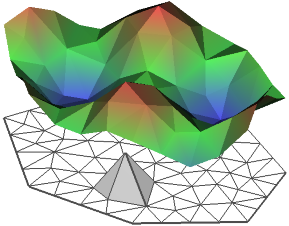
Approximation to the spatio-temporal fields

- Computationally challenging to fit multiple spatio-temporal processes
- Approximation to the spatial field using a triangulation
- Approximate using

$$\omega_s = \sum_{k=1}^{n} \phi_{ks} w_k$$

where n is the number of vertices (or nodes) of the triangulation, $\{\phi_{ks}\}$ are a set of bases functions and $\{w_k\}$ are a set of weights





APPROXIMATION TO THE SPATIO-TEMPORAL FIELDS

- If ϕ_{ks} are piecewise linear then ω_s is a Gaussian Markov Random Field
  - Conditional independence
  - Sparse precision matrices
- The approximation to the spatial field is the solution to Stochastic Partial Differential Equation (SPDE)
- Latent Gaussian model
- Inference based on Integrated Nested Laplace Approximations (INLA)
- Penalised complexity priors for model hyperparameters

Prediction

- High resolution estimates of air pollution concentrations are required over space and time
- Computationally expensive
- Monte Carlo Simulation
  - Draw M samples from the joint posterior of the model parameters
  - Produce M joint samples using the linear predictor
  - Aggregation is fairly straightforward
  - Summaries of the marginal posterior distributions can then be made

Air Quality in Europe

PREDICTIONS

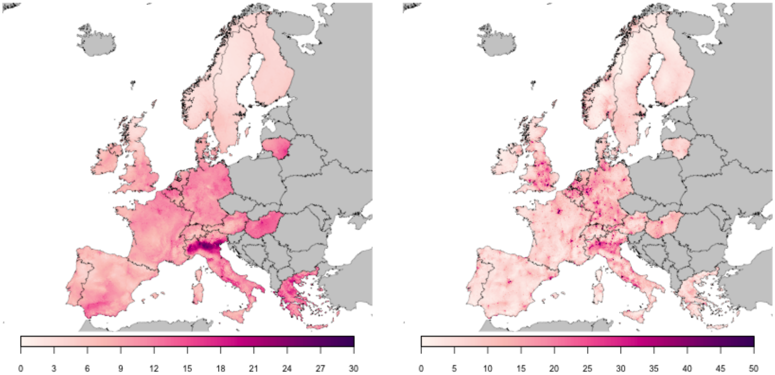


Figure: Median estimates of annual averages of (Left) $PM_{2.5}$ and (Right) NO2 (in μgm^{-3}) for 2010 for each grid cell (1km $\times$ 1km resolution).

CTM VS PREDICTIONS

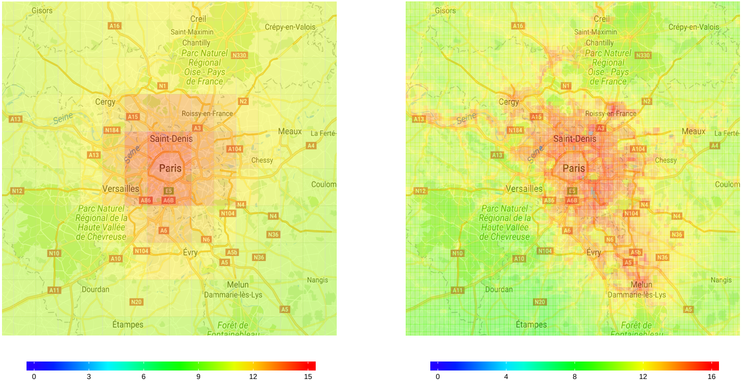


Figure: (Left) Estimated level of the annual average $PM_{2.5}$ in 2016 from the CTM for Paris and (Right) Predictive posterior median annual average of the annual average $PM_{2.5}$ in 2016

CTM VS PREDICTIONS

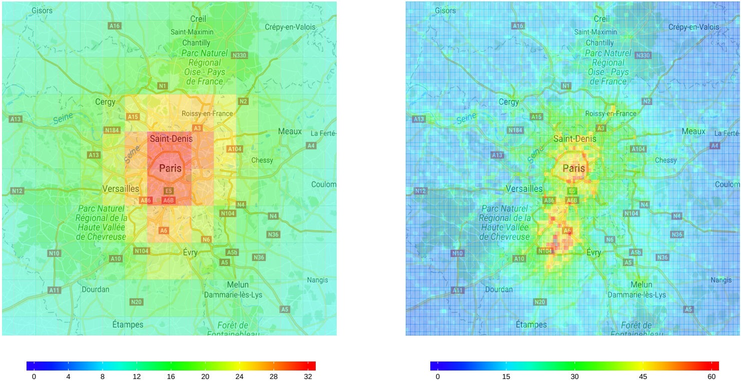


Figure: (Left) Estimated level of the annual average NO_2 in 2016 from the CTM for Paris and (Right) Predictive posterior median annual average of the annual average NO_2 in 2016

EXCEEDANCES

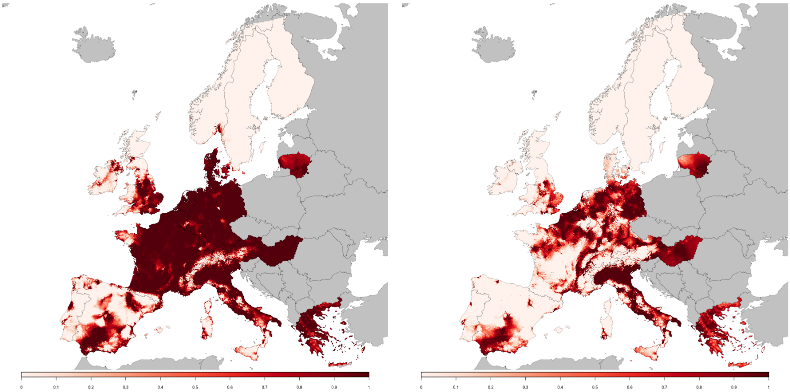


Figure: Probability that annual average $PM_{2.5}$ exceeds 10 μgm^{-3} (Left) for 2010 and (Right) for 2016 for each grid cell (1km $\times$ 1km resolution).

CHANGES OVER TIME

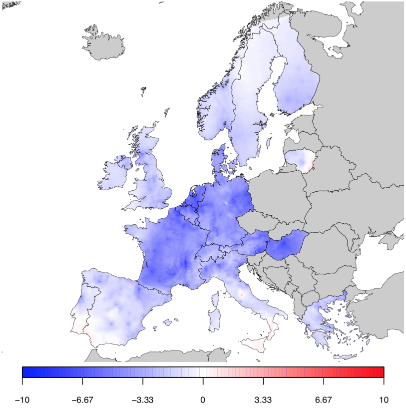


Figure: Median change in the annual average $PM_{2.5}$ between 2010 and 2016 for each grid cell (1km × 1km resolution).

Summary

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)
  - Burden of disease calculations

ANY QUESTIONS?