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

Matthew Thomas

Supervised by Prof. Gavin Shaddick
Joint work with Kees de Hoogh, Daniel Simpson and Jim Zidek

9th August 2017

# Outline

- Introduction
- Statistical modelling
- Results
- Summary and Future Work

# Introduction

# Introduction

- Air pollution has been identified as a global health priority.
- Nitrogen dioxide ($NO_2$) is associated with some adverse health outcomes.
- Concentrations can vary considerably over short distances.
- WHO and EU guidelines:
  - annual averages should not exceed 40 $\mu$gm$^{-3}$.
- Estimation of health burden requires accurate estimates of exposures to air pollution
  - at local levels
  - with associated measures of uncertainty.

# AIR POLLUTION IN EUROPE

- Information on exposures to air pollution traditionally comes from ground monitoring
- Density of networks vary considerably
  - biased towards urban and industrial areas.

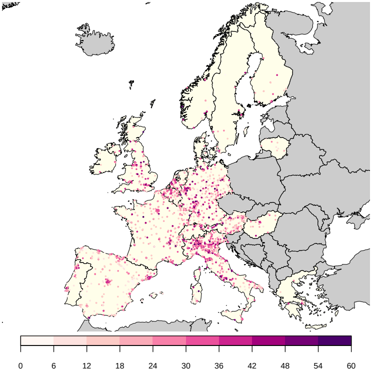



Figure: Locations of ground monitors measuring $NO_2$ in 2010. Colours denote the annual average concentrations ($\mu gm^{-3}$) of $NO_2$

# Data from Multiple Sources

- Ground monitoring data.
- Chemical transport models.
- Land use regression.
- Available on different resolutions
  - ground monitors (points)
  - chemical transport models (10km$\times$10km)
  - land use regression (1km$\times$1km).
- Estimates will be subject to uncertainties and biases.

# Data from Multiple Sources

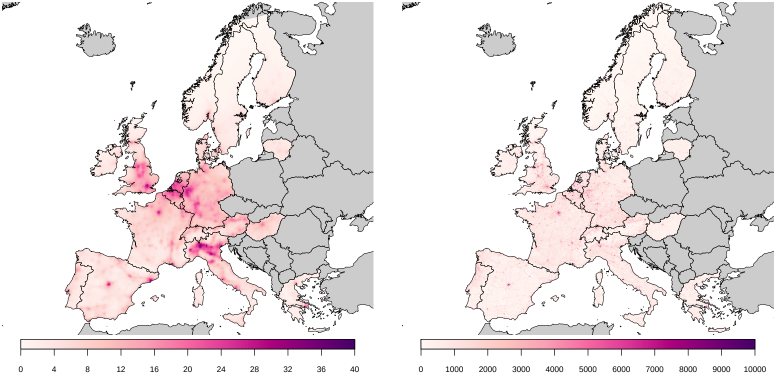



Figure: (Left) Estimates of $NO_2$ from the MACC-II ENSEMBLE chemical transport model for each grid cell (10km × 10km resolution) and (Right) Length of all roads with each grid cell from EuroStreets (1km × 1km resolution).

# Statistical Modelling

# Statistical Calibration

- The aim is to calibrate ground monitor information, $Y_s$, with estimates from a chemical transport model, $X_{l_s}$, and land use regression, $W_{qm_s}$.

$$Y_s = \beta_0 + \beta_1 X_{l_s} + \sum_{p=1}^{P} \gamma_q W_{qm_s} + \epsilon_s$$

- This will allow us to predict surface $NO_2$ where there is no ground monitoring information.
- However, the relationship between ground monitors and chemical transport models may vary spatially.

# Statistical Downscaling

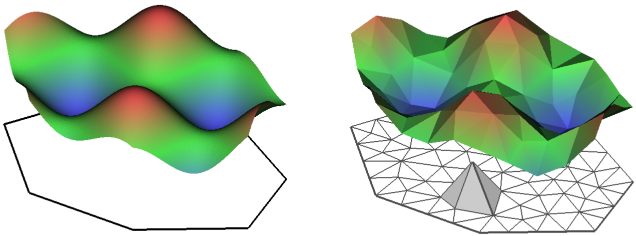
- We need to account for within grid-cell variability.
- We extend the linear regression model so that coefficients can vary spatially

$$Y_s = \tilde{\beta}_{0s} + \tilde{\beta}_{1s} X_{l_s} + \sum_{p=1}^{P} \gamma_q W_{qm_s} + \epsilon_s$$

- Coefficients, $\tilde{\beta}_{0s}$ and $\tilde{\beta}_{1s}$ are assigned spatial processes (Matérn class)

# COMPUTATION

- Bayesian hierarchical model.
- Integrated Nested Laplace Approximations (INLA), provide fast and efficient methods for latent Gaussian models.
- Use approximation to the continuous spatial field.
- Predictions at 1km $\times$ 1km = 3.5 million grid-cells.



# Results

# PREDICTIONS

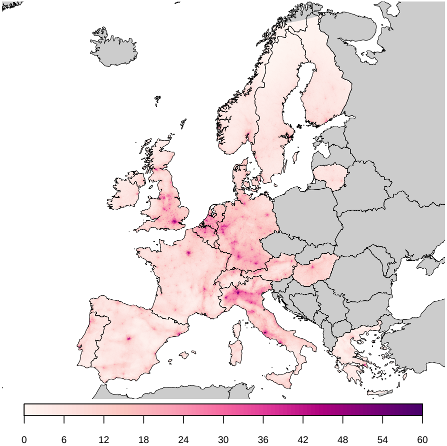



Figure: Median estimates of annual averages of $NO_2$ ($\mu gm^{-3}$) for 2010 for each grid cell (1km $\times$ 1km resolution).

# UNCERTAINTY

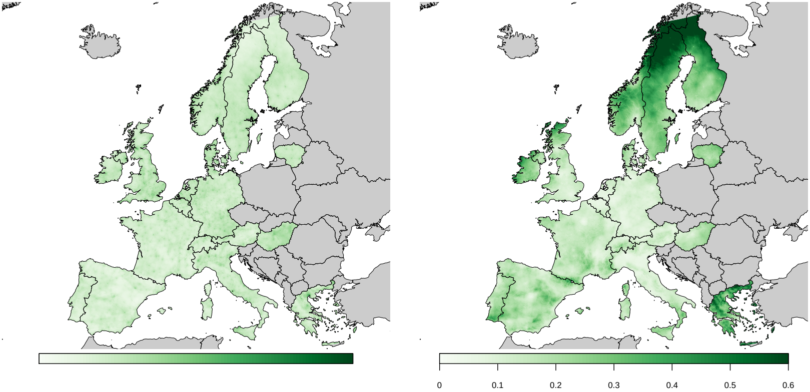



Figure: (Left) Half the width of 95% posterior credible intervals and (Right) Coefficient of variation for 2010 for each grid cell (1km $\times$ 1km resolution).

# EXCEEDANCES

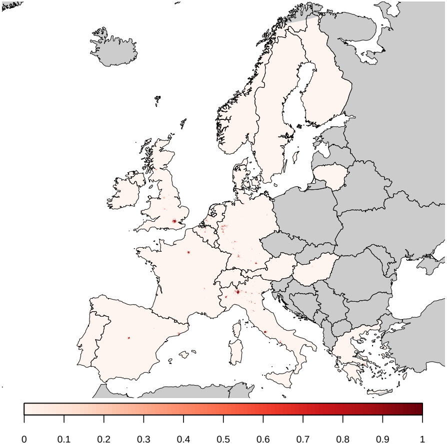



Figure: Probability of exceeding 40 $\mu$gm$^{-3}$ for 2010 for each grid cell (1km $\times$ 1km resolution).

# Model Coefficients

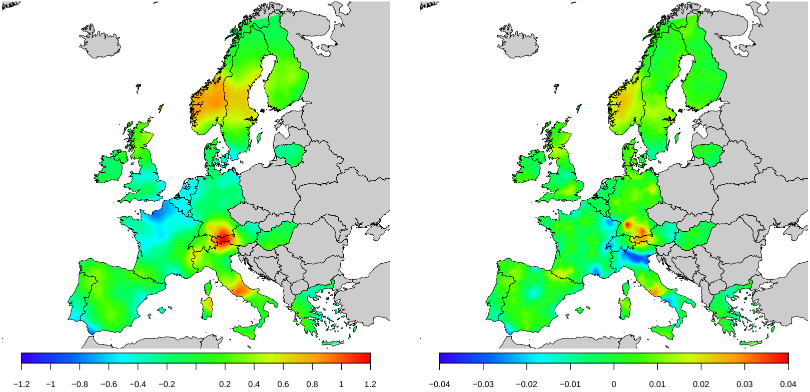



Figure: Deviations from the mean (Left) intercept and (Right) slope associated with the CTM for 2010.

# Summary

# Summary

- Developed a model that
  - integrates data from multiple sources
  - integrates data from multiple resolutions
  - produces high-resolution estimates of surface $NO_2$ with associated measures of uncertainty.
- Future work
  - Other pollutants such as $PM_{2.5}$.
  - Sensitivity of choice of mesh.
  - Extension to spatio-temporal models.
  - Burden of disease calculations.

# Any Questions?